UNIVERSITY OF CALIFORNIA
SANTA CRUZ

PRINTED TEXTS AND DIGITAL DOPPELGANGERS: READING LITERATURE IN THE 21ST CENTURY

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

LITERATURE

By

Jeremy Throne

December 2018

The Dissertation of Jeremy Throne is approved:

Professor Susan Gillman, chair

Professor Kirsten Silva Gruesz

Professor Johanna Drucker

Lori Kletzer
Vice Provost and Dean of Graduate Studies

Table of Contents

Abstract

Printed Texts and Digital Doppelgangers: Reading Literature in the 21st Century

Jeremy Throne

Much ink has been spilt worrying over the death of the book. It may be, however, that we find ourselves facing a situation where, as Whitman himself prophesies, “To die is different from what any one supposed, and luckier.” My dissertation tests the truth of this prediction by exploring the potential for digitization to create spaces where unexpected relationships between texts, authors, and readers may appear. The dissertation begins with an overview of several options for conducting literary history in a digital environment. I look at a number of sources for gathering information about Twain and discusses their potential. In the second chapter I proceed to an extended reading of data contained in a single digital source, the Chronicling America project. I discuss ways in which the data archived in Chronicling America may be used to suggest popular topics of conversation in the news at the time Mark Twain was working on his autobiography. The third chapter looks at how computer simulation can be used to stage interaction between Twain’s text and media accounts of the past. Together the chapters of the dissertation test the extent to which digitization invites new forms of literary inquiry, as well as the extent to which digital technologies may be united with printed texts to form the terrain for reading literature in the 21st century.

Dedications

To Nathan, Ella, and Ethan; so they know it can be done.

To Susan, Kirsten, and Johanna; with gratitude and admiration for their dedication to helping others.

To Rob and to Nancy; who believed in me.

To the friends, family, and colleagues who accompanied me along the way.

Introduction

On the main street in my hometown, next door to the bank my family frequented when I was a child, sits a used bookstore whose name escapes me now, but that was once a central feature of my world. Somewhere around the age of 7 or 8 I developed an insatiable love of the Hardy Boys. The teenage mystery solving duo of Frank and Joe went everywhere with me, first by means of a small collection housed on the shelves of my elementary school library; then, when that supply was exhausted, often courtesy of my grandfather--a plumbing contractor, son of a plumbing supply shopkeeper, and father of a plumber, all of whom share my middle name. Most gift-giving occasions that presented themselves around that time in my life saw me receive another installment in the ghostwritten series credited to Franklin W. Dixon. *The Tower Treasure*, *The Sting of the Scorpion*, even the detective's handbook, I read them all, and when I think about what I read growing up, these are some of the first books that come to mind. In fact, I still have them. For many years they sat together, occupying two of the six shelves in the overflowing bookcase that stood in our living/family room, right above the set of World Book encyclopedia my mother purchased for us for Christmas one year (that collection still occupies the bottom of the same bookcase, now ensconced in a different living room, in a different house about a half hour from the old). When I moved away for college, the Boys were carefully taken from their home and packed into a plastic milk crate that sat in a corner in the house I grew up in awaiting my return. When my family lost that house, the crate moved to my grandparents until I reclaimed it and moved it into an unused closet in my then-girlfriend, now wife's parents' home. Somewhere along the line I lost the grandfather who originally presented me these books: sometimes in ones and twos, sometimes packed, a dozen or so, in an old cardboard box that had been repurposed for the occasion. I lost the giver, but retained the gifts; and somehow or another those gifts became the books I would give to my children to read...except that they aren't. Dixon's tales, riddled with unflattering stereotypes and xenophobic innuendo, not to mention wooden characters and questionable plots, cannot, in my opinion, compete with much of the material aimed at the budding young readers of today.[1] Yet the objects themselves retain a certain talismatic quality for me that serves as a continual reminder not only of the passing of time, but of the permanence of memory, and the ways in which the two compete with each other for the attention of the present and the future.

Many of the significant details in my memories of the Hardy Boys—the library, the bookshop, the books, cardboard boxes, milk crates, bound encyclopedias—are, like my grandfather, fast becoming part of another era; and yet, also like my grandfather, they continue to live on in the traces they have left behind. All around us, bold new futures are quickly taking root and spreading rhizomatically through soils enriched by the passing of prior generations of both texts and readers. Private corporations, non-profit entities, colleges, universities, and government agencies of all kinds are feverishly farming this terrain in hopes of cultivating its

[1] To be fair, my assessment deals only with texts from the first version of the Hardy Boys series, published between 1927 and 1958. Wikipedia's entry for "The Hardy Boys" states: "beginning in 1959, the books were extensively revised, largely to eliminate racial stereotypes." Perhaps the revised texts read differently, I cannot say; however, at least one critic cited in the entry, Meredith Wood, argues that the revisions were not an improvement. See "Hardy Boys."

products to their advantage. The Fourth Great Information Age is upon us, Cathy Davidson is fond of saying, and indeed, many signs seem to point toward the radical transformation she credits Robert Darnton with naming.[2] Yet in the rush to embrace our digital future, much would be lost if our analogue pasts were overwritten. Arguments in favor of the digital utopias awaiting us just around the corner—when all the texts are scanned, when all the books are counted, when all the words are tagged—have little appeal for me if it means that the physicality of these entities is effaced. Prophecies of digital nirvana share much with visions of "the End of History" proclaimed by Francis Fukuyama, and a "Science Neither by the People nor for the People" identified by Paul Humphreys; in all three cases human potential, and with it the possible worlds this potential may bring forth, is confined to an increasingly small sphere of activity. Derrida responds to Fukuyama by invoking the ghost of Marx: hauntology, he argues, will forever offer those who seek it an opportunity to transform the world. Donna Haraway offers the cyborg as a technology-infused version of the will to resist ideological ossification.[3] While I am sympathetic to, and inspired by these responses, I worry that they cede too much ground to the forces they seek to oppose: the turn toward technological and/or supernatural kinship may reinvigorate humanity (as Derrida and Haraway no doubt intend) but it may also render it catatonic: a world populated by revenants and replicants--even if they are victorious in their battles with the agents of global capital-

[2] See Darnton for his assessment. Davidson has drawn attention to Darnton's work in a number of settings, ranging from blog posts and interviews to more formal pieces of academic writing. See Davidson and Goldberg, Future 19; Davidson, "What's It Like" and "My Commencement"; as well as many of the top hits returned by searching the Web for the phrase "fourth great information age."

[3] See Fukuyama; Humphreys, especially 6-9; Derrida; and Haraway.

-does not, to my mind, inspire much hope for the future. The inability to conceive of the regenerative processes these figures mark in terms that do not forsake humanity stands as one of the great dangers of the present moment, a time when the once dissonant voices of the spectre and the cyborg now seem to harmonize more with the digital dreams of multinational corporations than with their opposition.

The movement from Man to man, a movement which has enlisted much of the cultural criticism of the past five decades, has made important strides toward the creation of a more compassionate world; however, this movement risks a potentially devastating misstep if in embracing the post-human it leaves behind, pushes aside, or otherwise obscures our human pasts. Such is a significant risk in the push toward large-scale, data-driven approaches to cultural criticism typified by the Digging into Data Challenge, an international competition driven by the belief that "as the world becomes increasingly digital, new techniques will be needed to search, analyze, and understand these everyday materials" ("About"). Although they draw upon technological developments in fields as varied as natural language processing and geographic information systems; text-mining and computer simulation, many of the "new, computationally-based research methods" sought by the Challenge may be united by their belief in a methodological distance that separates their efforts from prior human endeavors ("About"). In literary studies, this belief has made some headway in the guise of "distant reading."

Amid the talk of epochal shifts, apocalyptic prophecies, and visions of birth and rebirth that accompanied the arrival of the year 2000, Franco Moretti introduced in the pages of the New Left Review "distant reading" as a method for resurrecting the vision of global literary studies articulated by Goethe and Marx in the nineteenth century (“Conjectures”). Goethe proclaimed “national literature is now a rather unmeaning term; the epoch of world literature is at hand, and everyone must strive to hasten its approach” (qtd. in Damrosch, What 1). Marx later echoes that prediction with the observation that “national one-sidedness and narrow-mindedness become more and more impossible, and from the numerous national and local literatures, there arises a world literature” (qtd. in Damrosch, What 4). Both claims describe shifts from local to global concerns, but they characterize these shifts with varying degrees of precision: Goethe’s proclamation leaves the relationship between the local present and the global future unclear; and, a similar vagueness permeates his criticism of “national literature,” particularly evident in the use of the qualifier “rather.” In comparison, Marx is more exact in his criticism: world literature grows out of regional antecedents that are problematic not because they lack meaning, but because they are provincial. Despite their different diagnoses, they share a remedy: increase the scale on which literature operates by turning attention from national to global concerns—mankind’s “universal possessions” (per Goethe) or “common property” (Marx). In reviving this approach to literary studies, Moretti has more in common with recent theorists of world literature than with these foundational moments in the field. Rather than focusing on what is owned by all, distant reading looks to uncover what moves through all. The lack of an emphasis on possession is one of the ways in which distant reading avoids simply replaying the naive universalism shared by

Goethe and Marx. David Damrosch, for example, suggests world literature names a concern with circulation, rather than with the content of a work: “I take world literature to encompass all literary works that circulate beyond their culture of origin, either in translation or in their original language” (What 4). A similar emphasis on circulation can be found in Moretti’s approach, which concerns itself with literary systems, rather than literary texts, and argues that “if we want to understand the system in its entirety, we must accept loosing something” (“Conjectures” 57). That “something,” he argues, is the intimate experience that comes with reading any particular work. Like Goethe and Marx, however, Moretti is committed to the idea that changing the scale upon which literary works are studied is necessary for combating stagnation. Distant reading, he claims, reading in which distance from a text is an essential feature of one’s relation to it, can revive interest in supranational approaches to literary studies following the inability of comparative literature to move beyond an infatuation with a small number of Western European literatures. Moretti is not alone in his criticism of comparative literature or his desire to revive it. Gayatri Chakravorty Spivak, for example, delivered her Wellek Library Lectures calling for “a new comparative literature” the same year that Moretti’s text appeared (xii). What sets Moretti’s work apart is a belief that world literature, with the aid of digital technology, can be a study of literature in the aggregate and need not concern itself with the reading individual of texts.

As Moretti himself predicted, distant reading has provoked the ire of many literary critics. When contacted by the NY Times for comment, Harold Bloom simply labeled distant reading an absurdity (Eakin); and he is not alone in his condemnation. The approach has made some inroads, however, in the digital humanities, where scholars working on questions of authorship attribution and stylometrics have traced the idea of studying literature without reading texts back at least as far as the 1850s (Hoover). Martin Muller sees this history as more than a prelude to contemporary attempts to provide dates and authors for unidentified texts using Principal Components Analysis; Cluster Analysis; Delta, Zeta, and Iota scores; T-tests; or any number of other measures of textual difference constructed by transposing literary texts into lists of recurrent features. Reducing texts to lists, he argues, may allow us to distinguish textual characteristics that would otherwise escape our attention: “the impossibly impoverishing reduction of a text into lists of its constituent parts may let you see some salient differences and resemblances across many texts that you could not as readily see by reading” (Mueller 294). Distant reading transforms this method into a tool for reading on larger scales as well; it seeks to track across vast literary collections the presence of “units that are much smaller or larger than the text: devices, themes, tropes—or genres and systems” (Moretti, “Conjectures” 57), while at the same time fostering an awareness that “units” are not neutral categories but strategically deployed interpretive acts.

The understanding of “unit” upon which distant reading depends is perhaps most clear if we temper the scientific rhetoric favored by Moretti (who at times seems to revel in seeking out literature’s “universal laws”) with the advice Damrosch offers to readers of world literature: "to be effective, a comparison of disparate works needs to be grounded in some third term or set of concerns that can provide a common basis for analysis. Without some meaningful ground of comparison, we would be left with a scattershot assortment of unrelated works" (How 46). Damrosch’s emphasis on “effective” comparison and his examples of grounds for meaningful comparison—which include the treatment of plot, character, setting; use of imagery; social, political, and economic issues—outline an alternative to the scientific terminology embraced by Moretti that approaches the construction of the unit of comparison that enables a distant reading as an interpretive act whose value is indicated by its ability to build conversational communities out of what may otherwise appear to be unrelated texts; and by its ability to show how the construction of what may appear to be self-evident communities depends simultaneously on the construction of, and deployment of, the units of comparison that bring them into being.

The bifocal perspective championed by distant reading brings into focus a fertile middle ground between Muller’s “Literary Informatics” and work in the newly proposed field of “Culturomics” (“Culturomics”; Michel) where literary criticism may mingle with new and established approaches in the digital humanities to produce hybrid forms of literary analysis that treat texts as, simultaneously, large and small objects of study, and that blur distinctions between quantitative measurement and theoretical attention. In exploring this ground, one point of departure may be the edition of the Autobiography of Mark Twain released by UC Press in 2010. Praised as a model of digital scholarship and as a scholarly tome, the Autobiography, as digitization project, scholarly critical edition, and New York Times bestseller, offers multiple points of entry into debates over the future of literary studies that have attracted the attention of literary critics and digital humanists alike. Above all, the acclaim the text has found as both a digital and a print edition points to the potential for print and digital media to coexist. This means a decisive departure from the homogenizing prospect of an all-digital future and toward a view of how printed texts and their digital doppelgangers work together to constitute the terrain for literary criticism in the 21st century. The goal here would be to spur both literary critics and digital humanists to forgo drawing lines in the sand and begin the collaborative development of approaches to the study of literature that embrace the technological developments of the last 60 years as warmly as the theoretical developments of the last 60 years.

Twain, through both his life and his writing, provides ample opportunity to explore the convergence of literature and technology. Not only was he fascinated by inventors and inventions (as were many of his contemporaries) he was active as both an inventor and an investor in the push for technological innovation that drove the mechanization of industry and growth of capitalism throughout the United States in the nineteenth century. “No major American author of the nineteenth century,” argues Bruce Michelson, “participated more actively and imaginatively in that revolution than Samuel Clemens” (591). For Twain, who followed the success of *The Innocents Abroad* (1869) with a spate of publications between 1871 and 1873 that includes *Mark Twain’s (Burlesque) Autobiography*, *Roughing It*, and *The Gilded Age*, as well as patents for Mark Twain’s Elastic Strap and Mark Twain’s Self-Pasting Scrapbook, technological and literary production were never very far apart.[4] “An inventor,” he wrote in the midst of this period, “is a poet—a true poet—and nothing in any degree less than a high order poet” (qtd. in Oxford 306; “SLC and OLC”). Twain’s interest in technology, on display in business ventures (the Paige Typesetter, the Kaolatype engraving process) and in texts like Connecticut Yankee and Pudd’nhead Wilson—forms a central preoccupation in the Autobiography, where he struggles to combine a conception of the self as plural and the practice of life writing in order to produce a polyvocal account of his life capable of supporting his characteristic attention to the intricate details of the languages, dialects, and linguistic nuances of the events—public and private, formal and informal, historical and contemporary, ceremonial and personal—that constitute the text.

Twain worked for the majority of his career to produce an autobiographical account of his life that satisfied him. It was not until he hit upon the idea of dictating his autobiography that he felt he had found the appropriate tools with which to construct his story. Carving several hours out of his daily schedule, he employed stenographer Josephine Hobby and dictated reflections on his life and the events of the day several times a week between January 1906 and October 1909. Dictation, Twain thought, provided the tools needed to balance the inner feelings of a would be autobiographer and the external events of his or her life. “It is the first time in history,” he wrote, “that the right plan has been hit upon” (“Second Preface”).

[4] For a chronology of Twain’s output during this period, see Oxford 792. For Twain’s account of the development of Mark Twain’s Elastic Strap, see “SLC to Mortimer.” For his plans for the scrapbook, see “SLC to Orion.”

My dissertation is inspired by Twain's autobiographical experiment and the ways that it freely combines and recombines texts and contexts to produce an archive that calls all archives, and especially his own, into question. This is a kind of archival impulse that combines technical and theoretical ambition, a willingness to risk failure, a desire to trouble established literary conventions, and the curatorial projects of recent theorists (Stephen Ramsay, Johanna Drucker). "Once you have programmatic access to the content of the library, screwing around suddenly becomes a far more illuminating and useful activity," argues Stephen Ramsay (6). But archival exploration is not all fun and games, reminds Johanna Drucker: "formal logic, with its grounding in *mathesis* and claims to objectivity, can be challenged only by an equally authoritative tradition of aesthetic works and their basis in subjective forms of knowledge production" (SpecLab xiii). Curation from this point of view can provide a way both into and out of the whole issue of what and how to read that plagues literary studies and the digital humanities.

Why turn to curation now? Above all, it brings into view common ground between literary criticism and the digital humanities, ground that has been inadvertently and unfortunately covered over as a result of a scuffle between advocates of close reading and their opponents. On the site of this battle curation stages a reconciliation, bringing to the fore issues (from questions of access and audience, to debates over scope and scale) that trouble both fields as they struggle to navigate the changes wrought by the emergence of a widespread digital culture. Chief among these issues, for literary critics and digital humanists of every stripe and specialty (to say nothing of their colleagues in libraries, schools of information science, and elsewhere): what are we going to do with all these texts? In response to this question, curation provides the opportunity to craft a dynamic participatory model of literary studies that canonization, with its emphasis on institutional stricture (in the Academy, in the Press, in the Library), has a tendency to obscure.

The key implications of carrying the concept of curation beyond the museum and library communities are, first, that in principle, every text becomes part of a single archive and exists on the same plane within it, and second, that processes of selection (rather than the qualities of a text) are responsible for disturbing this arrangement. We are used to thinking of curation in terms of hierarchal arrangements that proceed from the top down. We are also used to thinking of curation in terms of counter-cultural movements that proceed from the bottom up. We are less familiar with forms of curation that leave us (or appear to leave us) to determine our own relationships to the materials available for display. These, however, are some of the forms of curation that most often confront us during our interactions with texts in the digital world. Showing how these forms of curation work and why they are used is one goal of my project; determining how their presence separates the experience of reading a text on-line from reading it in print is another; and imagining alternative forms of curation is a third.

Thinking through the lens of curation could account for a spectrum of ways in which digital editions reference their printed predecessors, from those that simulate the reading of a printed text to those that seek to transcend it. The PageTurner interface developed by the HathiTrust and the standard interface Archive.org uses to make texts available for reading on-line are two examples of digital environments that simulate the experience of reading a printed text. In both cases, the on-line environments emphasize a print-digital continuum by inviting users to “flip” through scans as if they were turning the pages of a physical book. In contrast the Online Reader at Project Gutenberg presents digital texts that reproduce the linguistic content of printed texts, but make only limited reference to their non-linguistic features. Texts in the eText Archive at the University of Virginia frequently reference their printed counterparts, but do not replicate them. The on-line reading environment created by the Mark Twain Project attempts to transcend the experience of reading a book by juxtaposing text, table of contents, and editorial commentary; and allowing users to navigate between the three using hyperlinks. Each of these presentation methods provides a different experience of the text. Archive.org and the HathiTrust use methods of digitization that draw attention to the physical object consulted during the digitization process. Project Gutenberg jettisons the physical object while drawing attention to the text it contains. At Virginia, the incorporation of images into the text via hyperlinks casts those materials in a decidedly secondary role. The editions of the Mark Twain Project place the editorial work that went into the preparation of an edition on equal footing with the text itself.

The concepts of duo, double, and doppelganger are useful for drawing distinctions between these various acts of representation. Double and doppelganger in particular, conventional targets for literary critics, encourage thinking of the ways in which digitization projects may (con)fuse issues of literary interpretation (what is a text, how should it be read) and concerns over preservation and access (what needs to be digitized, how should it be stored). Twain himself provides something of a model for the exploration of these issues through what Lawrence Howe describes as “the characteristic doubleness operating at every level of his literary conception” (1). Twain’s fascination with twinning appears in his treatment of characters (Huck Finn and Tom Sawyer, Tom Canty and Edward Tudor, Tom Driscoll and Valet de Chambre), settings (in his depictions of life on the Mississippi, his travel writings, and more fancifully in the time traveling jaunts of Hank Morgan and the “Mysterious Stranger”), and themes (freedom and slavery, progress and decay, technology and tradition, civilization and savagery). In each of these cases it is the complex relationships between parts, rather than the parts themselves, that are responsible in large part for elevating his texts above typical examples of the genres—adolescent fiction, travelogue, mystery, comedic sketch—in which he often works. Authors of, and audiences for, digitization projects may have much to gain by following Twain’s lead and focusing not only on the creation of increasingly feature-laden digital texts, but also on the relationships these texts form with their printed predecessors.

A reading practice attuned to the ways in which reading environments vary—among digital texts, between digital and printed texts, and amid printed texts themselves, irrespective of the digital world—could address the limits and possibilities for reading literature in an environment where digital objects and printed volumes coexist by inviting us to see both digital and printed texts as forms of curation, or what Lawrence Lessig calls “remixed media”: media that “succeed when they show others something new; they fail when they are trite or derivative. Like a great essay or a funny joke, a remix draws upon the work of others in order to do new work. It is great writing without words. It is creativity supported by a new technology.” Digitization projects, from this perspective, are not second order forms of scholarship, nor are they ideologically neutral. Like the “Read Write” or “RW culture” Lessig describes, digitization projects have the potential to move literary studies away from the respect/reject logic of an authoritarian culture and toward a view of culture “as a draft” that invites revision (82-85).

This approach is a significant departure from the focus on developing comprehensive coverage and content that haunts many current digitization projects. The global aspirations of the Google Books project are perhaps the best-known example of this drive, but they are far from alone on the quest to create a complete archive of our literary heritage. Microsoft’s now defunct Live Search Books, which was absorbed by the Internet Archive, and the HathiTrust have similar goals. These projects share a desire to create comprehensive archives that contain a digital double of every available text in the world, even as their interpretations of “availability” differ—for example, available when talking about Google Books means that the work is held by an entity that is a member of either Google’s Partnership Program or Library Project; and, that the text has not been deemed “too fragile” to scan. Available, in the context of the HathiTrust, means a text that meets the non-profit organization’s guidelines for deposit and is held by an eligible institution that is willing to share in the costs of digitizing material.[5] Another group of digitization projects proceeds along more thematic lines. The Alex Catalogue of Electronic Texts seeks to catalogue "great works" from the Western Tradition. The Perseus Digital Library Project offers an extensive collection of texts in Greek and Latin. Early English Books On-line brings together a range of literary work from the English Tradition. The Transcribing Bentham project is devoted to bringing the unpublished writings of Jeremy Bentham before the public. Numerous governments have expressed interest in creating digital repositories as a means of preserving cultural identity. Looking beyond these efforts to digitize printed texts, a number of second order digitization projects may also be distinguished. The Internet Archive's Open Library project, for example, is devoted to cataloguing bibliographic data for texts without regard for whether they are available to be digitized or not, while OCLC's WorldCat seeks to create a global catalogue that will allow readers to search the holdings of every library in the world, but has chosen not to digitize the texts themselves. However, as is the case with efforts to build complete collections of texts, the global catalogues promised by these organizations come in various sizes. OCLC, for example, boasts of gathering bibliographic data from a network of partners in 124 countries, with 491 languages and dialects represented to build its WorldCat catalogue (by comparison the U.N. puts the number of countries in the world at 193, the US Department of State recognizes 195) but how much of this database one has

[5] See "About Google Books" for a general overview of Google Books; and, "How library book" for a brief policy statement concerning scanning and fragile materials. For general information about the HathiTrust, see "Our Partnership," eligibility requirements may be found at "Eligibility and Agreements" and "Getting Content Into HathiTrust."

access to depends on how it is accessed: users that access WorldCat's global catalogue via WorldCat.org have more limited options for locating texts than do users who access the same database directly through the catalogue of a library that has partnered with the organization.[6] Another group of projects is also interested in providing tools for navigating archives that go beyond bibliographic data. The Orlando Project offers its users an avowedly feminist literary history of women writers in Britain. The NINES group brings together digital archives of Nineteenth Century US literature compiled by other digital projects. The MONK project provides users with several collections of digital texts and a set of tools for generating statistical summaries of their content. Google's N-Grams Viewer gives users the ability to explore changes in word frequency across a subset of texts drawn from the Google Books project, and Voyeur allows one to perform a statistical analysis on any text available in digital form. In contrast to the highly organized efforts of many of the projects just mentioned, the efforts of Project Gutenberg and the Oxford Text Archive shed the desire for comprehensive coverage and control in favor of archives that are decentralized and avowedly piecemeal. Anyone can add a text, any text can be added, and their efforts proceed without any aspirations to a master plan for total coverage of the world's literary resources. These archives avoid the authoritarian pose

[6] WorldCat.org searches the holdings of more than 10,000 libraries that have created the WorldCat Registry profile necessary to link a library with the site; while the WorldCat database searches the holdings of all of the libraries that have partnered with OCLC. For a more detailed explanation of how WorldCat.org differs from other ways of accessing WorldCat, see "WorldCat.org frequently asked questions." Information on the current number of languages represented on WorldCat is available at "Inside WorldCat" and the size of the WorldCat network may be found at "About." The U.N. count of countries in the world may be found at "UN Member States | On the Record" and the count maintained by the U.S. Department of State count may be found at "Independent States in the World."

of their more closely controlled curatorial cousins by leaving it to their audiences to determine the content of the archive. At Project Gutenberg, anyone can volunteer to help build the archive; and, any text free of copyright restrictions can be uploaded. A similar spirit of openness can be found at the Oxford Text Archive, which has an open invitation seeking "literary or linguistic primary source research material of interest to UK Higher Education."[7] Both archives offer a user-driven approach to assembling archives to build collections that are collaboratively shaped by individual contributors, rather than being shaped by the desires of collectors acting from above. The risk of failure is great, but so is the reward: a model of literary study powered by both the elite company it has kept (rightly or wrongly) for so long and the populist energy generated by multiple actively engaged communities of readers working (sometimes together, sometimes at cross-purposes) to steward the construction of their own literary archives.

The dissertation is motivated by a similar belief in the importance of promoting literary stewardship as an alternative to literary consumption. It proceeds from one act of stewardship, the 2010 publication of Twain's autobiography, part of a self-described effort at "giving Mark Twain the texts he always wanted, but never got," into a wide ranging exploration of the how digitization alters our perceptions of Twain and his work that moves through three themes over the course of three chapters (Hirst, "Textual"). I argue that digitization doesn't simply give us the same

[7] See "Actions" for the submission system currently in place at Project Gutenberg. The egalitarian ethos of Project Gutenberg is evident in much of the advice it provides volunteers; see, for example, the FAQs about volunteering ("Volunteers' FAQ") particularly "How do I get started as a Project Gutenberg volunteer?" For information on contributions to the Oxford Text Archive, see "Depositing with the University of Oxford Text Archive."

old Twain in a new wrapper, it gives us what we might call, following Ed Folsom and Kenneth Price's ongoing work to create a digital (re)presentation of the writings of Walt Whitman, a "re-scripted" Twain: a Twain whose existence depends upon replacing the printed texts we have held in our hands for so long with digital texts that we may never hold. Throughout their work on the Walt Whitman Archive, Folsom, Price, and their collaborators portray digitization as a dialogue with, rather than a departure from, our print cultural pasts.[8] My argument mines a similar vein: in the move from print to digital texts, much ink has been spilt worrying over the death of the book. It may be, however, that we find ourselves facing a situation where, as Whitman himself prophesies, "To die is different from what any one supposed, and luckier." My argument tests the truth of this prediction by exploring the potential for digitization to create spaces where unexpected relationships between texts, authors, and readers may appear. The dissertation begins with an overview of several options for conducting literary history in a digital environment. I look at a number of sources for gathering information about Twain and discusses their potential. In the second chapter I proceed to an extended reading of data contained in a single digital source, the Chronicling America project. I discuss ways in which the data archived in Chronicling America may be used to suggest popular topics of conversation in the news at the time Mark Twain was working on his autobiography. The third chapter looks at how computer simulation can be used to stage interaction between Twain's

[8] For an extended discussion of the ways in which the Walt Whitman Archive compliments prior work on Whitman, see Folsom and Price, esp. the appendix "What Whitman Left Us." (The text may also be found as part of the Walt Whitman Archive at "Re-Scripting.") For a more general discussion of the relationship between digital and print editions, see Price.

text and media accounts of the past. Together the chapters of the dissertation test the extent to which digitization invites new forms of literary inquiry, as well as the extent to which digital technologies may be united with printed texts to form the terrain for reading literature in the 21st century.

Chapter One draws upon a digital archive of information about the publication and reception of Twain's works culled from the holdings of multiple on-line sources in order to chart a trajectory for Twain's career. These sources call into question narratives of decline commonly associated with the end of Twain's literary career and draw attention to the ways in which he remains a prominent figure until times much closer to our own. Beginning with audience responses to Twain, rather than his authorial acts also draws attention to the ability contemporary audiences have to shape the reputations of authors and their works for future readers. Acknowledging this role is a step toward understanding reading (and other forms of cultural engagement) as acts of stewardship, rather than consumption. A logic of consumption lends itself to understanding relationships between authors and audiences in terms of market dynamics (exchanges of capital; market trends; self-promotion; fads). The prevalence of this logic contributes to the marginalization of the humanities by foregrounding questions about the utility, profit, and loss of the humanities; and, by obscuring questions about the longevities, kinships, connectivities, possibilities, and alterities the humanities both record and enable. I conclude by arguing that digital approaches to our literary records offer opportunities to explore established regional, national, temporal, and linguistic concerns in the humanities on a global, multi-lingual scale.

Chapter Two uses a variety of digital tools to compare the subjects Twain takes up in his autobiographical writings with those circulating through the contemporary media of his day. Comparing Twain's interests with those preserved in the nineteenth-century newspapers collected by Chronicling America draws attention to where the topics he chooses to address overlap with and depart from popular concerns shared by his contemporaries. Identifying where other voices reverberate through Twain's text helps to distinguish his trailblazing moments from more pedestrian conversations. It is also a step toward understanding the inner-workings of what Twain calls the "apparently systemless system" ("26 March: Paragraph 27") of the text that leaves us better positioned to evaluate three facets of the future engendered by the work: the future Twain envisions for his autobiography; the future detailed in our historical records; and our own visions of what may be in store for Twain and his text.

Chapter Three stages the interaction between Twain's text and the conversations of his contemporaries using agent-based models. Modeling discussion of Twain's Autobiography is one way of evaluating the impact his text may have had on his contemporaries if it had been published during his lifetime instead of being released after his death. Developing a sense of how Twain's contemporaries may have responded to the text provides benchmarks against which Twain's claims about the popularity of the text may be evaluated. Understanding the reception Twain predicted for the text provides a window into his relationships with contemporary audiences and future generations that directs our attention toward Twain's worldview and our own place within it.

New media, Meredith McGill and Andrew Parker argue, brings forth new versions of literary history that are "different from what we had anticipated" (966). My dissertation tests this idea. The versions of Twain I have assembled from the archives are neither totally new nor totally known. They provide alternating glimpses of Twain as producer (setting forth a new autobiographical tradition) and product (recycled into new editions on a regular basis), pilot (seeking control over his image even from beyond the grave) and passenger (like his audience, taken along for the ride as the future unfolds). Digital media offer new frames for bringing together these competing and complementary facets of authorship together. These frames are no replacement for the experience of sitting down to turn the pages of a printed text, but they offer new avenues to that experience. Approaching familiar settings via unexpected paths allows us the experience of seeing text that has grown familiar with fresh eyes, holding books that have passed through our hands countless times as if for the first time. Digitization projects, be they duos, doubles, or doppelgangers, grant us the opportunity to recognize printed texts anew, in ways that return to us a sense of wonder—at the vividness of an illustration, the strength of a binding, the texture of a page, the crispness of a font—that once commonly characterized the relationship between readers and the written word, and that becomes increasingly hard to find as printed pages give way to their digital descendants.

Chapter One: Mark Twain's Literary Legacies: A Digital Perspective

Introduction

> "He was essentially an actor—that is, a child—that is, a poet—with no taint of mere histrionism, but always suffering the emotions he expressed. He suffered them rather than expressed them in his later years, when his literature grew less and less and his life more and more." (Howells 312)

Offered in the pages of Harper's Magazine three years after Twain's passing, William Dean Howells's portrait of a Mark Twain rendered mute in his final years by the weight of emotions he could not express may be as influential as any review the critic, recognized by his contemporaries as "The Dean of American Letters," penned while Twain was alive.[1] Howells's text is an early example of what Michael Shelden argues has become a tradition of viewing Twain's final years as "a time of bitterness and retreat" (xxxix). So deep is the shadow cast by Twain's allegedly sad figure that it even colors the work of sympathetic contemporary critics like Susan Gillman, who presents Twain's career arc in terms of evolution, rather than desiccation. In her hands the transition away from writing identified by Howells becomes an extension of Twain's artistic voice. Pointing to a visit made to Congress to discuss copyright legislation, Gillman argues: "the kind of performance typified by the 1906 copyright episode always was and came even more to be a mode of artistic expression for Mark Twain, not a repression of the self but a means of self-expression" (Dark Twins 188). Building on Gillman's work, Shelden names this same event a turning point in Twain's career, both "a dramatic break from the past" and the start of "Mark Twain's

[1] For a book length exploration of the period during which Howell's reputation as "The Dean" was cemented, see Crowley.

last great adventure" (xxiv, xxxviii). Other critics have been less kind. Karen Lystra, for example, sees Twain's later career as a period in which he loses control of his life to the point "that he came very close to being taken over" by Isabella Lyon and John Ashcroft (232). "They thought they owned him body and soul," she asserts in her, at times, tawdry expose of Twain's final years (Lystra 223). Sidestepping the issue of the nature of the change Twain undergoes, John Tuckey, drawing upon evidence from Twain's correspondence, suggests that near the end of his career Twain simply tired of writing (27-28).

Despite their differences, each of these critics places Twain at the center of their analysis. My dissertation explores the trajectory of Twain's career from a different perspective. I focus on how audiences respond to Twain, rather than on the author himself. Using the publication and circulation of Twain's texts as measure of audience response, my approach places the decline in Twain's literary career much closer to the present. Pushing the decline closer to our own time makes it easier to see that Twain ends his life more popular than at any point in his career. It also allows one to see that his popularity does not fade in his time, but in the time of his heirs, and even more so in our own. These points come together to question the appropriateness of the narratives of decline that are often associated with Twain and his texts. Justin Kaplan, for example, argues the writing of *Connecticut Yankee* marks "a stage in his own disintegration," the beginning of the end for an author who "stood at the peak of his life and powers at the end of 1885" (11, 9). Beginning with audience responses, rather than authorial acts also emphasizes the ability contemporary audiences have to shape the reputations of authors and their works for future readers. Acknowledging this role is a step toward understanding reading (and other forms of cultural engagement) as acts of stewardship, rather than consumption. A logic of consumption lends itself to understanding relationships between authors and audiences in terms of market dynamics (exchanges of capital; market trends; self-promotion; fads). The prevalence of this logic contributes to the marginalization of the humanities by foregrounding questions about the utility, profit, and loss of the humanities; and, by obscuring questions about the longevities, kinships, connectivities, possibilities, and alterities the humanities both record and enable.

The trajectory I chart for Twain's career draws upon a digital archive of information about the publication and reception of Twain's works culled from the holdings of multiple on-line sources, including: the Reading Experience Database, Google Books, Google Trends, Open Library, and Wikipedia. I investigate this archive by coupling digital tools with more traditional tools of literary criticism—bibliography, criticism, and primary research—in order to demonstrate how large scale bibliographic databases and other digital technologies can be used to unlock the latent potential of the alternative histories that are subsumed by and contained within any one particular view of the past. The macroscopic lenses enabled as part of a digitally infused literary criticism provide a vision of Twain that resembles the portrait drawn by Albert Bigelow Paine. In place of the gloom Howells wraps around Twain's final years, Paine, Twain's official biographer, paints the end of Twain's career in a more cheerful light: "Advancing years did little toward destroying Mark

Twain's interest in human affairs. At no time in his life was he more variously concerned and employed than in his sixty-seventh year--matters social, literary, political, religious, financial, scientific. He was always alive, young, actively cultivating or devising interests--valuable and otherwise, though never less than important to him" (1150). Paine has been accused of glorifying Twain (Trombley 92, 248). The emergence of digital tools for literary history provides an opportunity to revisit assessments of Twain and his legacy. These tools also invite us to investigate and aggregate the discrete portrayals offered by Paine, Howells, and other scholars interested in Twain by drawing upon maps, models, and other kinds of visualizations that cast an extremely wide net around the bibliographic footprints Twain and his activities have left behind. I will not argue that this net is unique because of its size; I grant that Paine, Howells, or any other biographer could have undertaken a study of equal scope given enough time and resources to construct it by roaming the literary archives of the world. I will argue, however, that the net digital tools allow us to cast around Twain is unique because it is mutable: growing, shrinking, changing as the archive of information it depends expands (and contracts) in conjunction with the evolution of our understanding of the publication and reception histories of Twain's works. The flexibility and extensibility digital tools provide allows literary histories to be both archives and sites of revision: their ability to function as both snapshots of the past and platforms for revision is one of their great strengths. As snapshots, digital tools make possible the collation of massive amounts of information, a process that raises the level of detail in and definition of our portraits of the past. As platforms, digital tools make it possible to explore how these portraits may be redrawn, a process that unlocks the latent potential of the alternative histories that are subsumed by and contained within any one particular view of the past. The Reading Experience Database is one such tool.

The Reading Experience Database

Begun in 1995 as an effort to create a digital record of the ways texts were experienced by "readers born or resident in the British Isles reading in any language whatsoever" between 1450-1914, the Reading Experience Database has expanded its focus over the past two decades to include national projects of similar scope in Australia, Canada, New Zealand, and the Netherlands.[2] Touted by founder Simon Eliot as a "new and prestigious academic project" at the time of its launch, the growth of the database is notable given a significant difficulty Eliot makes no attempt to downplay: "the truth is that, although not exclusively so, the evidence for reading is obscure, hidden, scattered and fragmentary. Its discovery is often a matter of serendipity." Further troubling the situation, he suggests, is the fact that available information about reading practices is often far from typical: "any reading recorded in an historically recoverable way is, almost by definition, an exceptional recording of an uncharacteristic event by an untypical person." If the Reading Experience Database is limited as an archive of mainstream reading habits, the project still excels

[2] See Eliot for his remarks and "Welcome to RED" for the current roster of associated projects.

as a collection of oddities, specimens in a modern day cabinet of curiosities that provide inspiration for alternative readings of literary history even as they offer their own versions of the past.

Records in the database may be browsed by author, reader, or reading group; and searched according to more detailed criteria that further describe an audience or text. Browsing or querying the database returns a number of pieces of "evidence" volunteers have contributed to the project. These contributions range from first-person accounts of reading habits recorded in autobiographical texts to secondary accounts about the reading practices of others. Taken as a set, these individual sketches of reading habits provide a multifaceted portrait of the audience reached by a given text or author.

Searching for Twain shows a divide between people who describe turning to his "masterpieces" for "intellectual manna" and readers who present Twain's work as a precursor to more demanding literature.[3] However divides like this are harder to spot as the pool of data grows to include more ambiguous accounts such as that attributed to the "family of Rose Gamble":

> On the wall at the side of the chimney Dad put up the bookshelves which Dodie began to fill with secondhand penny books. Over the years we had Conrad and Wodehouse, Eric Linklater and Geoffrey Farnol, Edgar Wallace, Jane Austen, Thomas Hardy, Mark Twain, Arnold Bennett, Robert Louis Stevenson, John Buchan, and a host of others, good, bad and awful, and we read the lot, some of them over and over. ("Record Number: 11428")

[3] Twain's champions on the site include Charlie Chaplin ("Record Number: 5314") and Joseph Conrad ("Record Number: 29001"). Portraits of Twain's works as juvenile entertainment are attributed to Neville Cardus ("Record Number: 5279") and James Williams ("Record Number: 5044") among others.

Digital tools can help to account for this larger pool of ambiguous, fragmentary, evidence. The increase in the scope, scale, and availability of evidence they provide opens a new chapter in the history of reading.

Reading 2.0

The army of readers enlisted by the Reading Experience Database in a quest to find evidence of texts being read has a digital double in information collected in real time by the purveyors of access to the massive databases of digitized texts that have become prominent features of our digital world. These databases, closer to sci-fi fantasies than boardroom realities at the time the Reading Experience Database was conceived, offer varying levels of access to evidence of reading around the globe. The public and private datasets they contain rival printed archives as significant records of reading activity and help to create an environment where pageviews, edit histories, and other access statistics jostle with more traditional forms of literary research for the attention of scholars interested in the history of reading.

In this environment astute works of literary criticism such as Raymond Williams's *Keywords* compete with the likes of Google's Ngram Viewer to illuminate how the use of a text, word, or phrase has developed over time. Williams investigates the cultural significance of a term at several specific moments in time; the Ngram Viewer offers an opportunity to wrap his investigations in general portraits of the frequency with which specific texts, terms, and phrases are used in a given context.

These portraits help to contextualize and complicate Williams's observations by wrapping specific examples in larger patterns of use. For example, Williams calls culture "one of the two or three most complicated words in the English language" (87). Ngrams show that between 1800 and 2000 the term "culture" appears with increasing frequency in books published in English.[4] For much of this period, texts published in English in Great Britain use the term with less frequency than it appears in English language texts published elsewhere. Beginning in the early 1990s, however, the term appears more frequently in texts published in Great Britain than it does in texts published in other parts of the world. English language texts all use "culture" with less frequency than it appears in a sampling of other languages: French texts consistently show higher levels of usage of the term; Spanish texts begin to match these levels by 1940; and Italian texts surpass them beginning about 1960. Details like these allow us to place the specific texts Williams dwells upon—often noting the date as part of his citation—within larger patterns that show the ebb and flow of interest in the topic under discussion. This functionality is particularly useful for charting how literary reputations—of real and fictional events, as well as of the people and literary characters taking part in those events—have captured the public consciousness.

Mark Twain at Whittier's 70th Birthday

[4] Ngrams were constructed using the 2012 corpuses of English language texts; texts published in the United States; texts published in Great Britain. Additional comparisons searched for "culture" in the 2012 French language corpus and "cultura" in the 2012 Spanish language and Italian language corpuses. See Fig. 1 and Fig. 2 in the appendix.

The quality of Twain's performance at the Whittier birthday dinner may be beyond our ken, but the attention commanded by this performance and the reputations it engendered are not. Whether good or bad, Twain's performance—and other events like it--have left trails through time that become increasingly legible as the mass digitization of our cultural heritage moves forward. Tools like the Ngram Viewer can help us to explore these trails. They assist us not by presenting the past as it really was, but by increasing our ability to explore the ways in which the past may have been.

Ngrams for several key figures involved with the dinner provide a sense of context for Twain's performance and a window into its aftermath. A search of the 2012 corpus of books published in English collected by Google Books for "Mark Twain," "Oliver Wendell Holmes," "Ralph Waldo Emerson," "John Greenleaf Whittier," and "Henry Wadsworth Longfellow" shows that around the time Twain arose to address his audience at the Hotel Brunswick in Boston on December 17, 1877, his name was appearing in print more frequently than any of the prominent literary reputations he takes as texts for his speech. Having equaled the notoriety of Ralph Waldo Emerson and Oliver Wendell Holmes in 1870, by 1877 Twain is nearly twice as likely to have his name mentioned in a text. In 1879, however, his popularity appears to reach a plateau: references to his name maintain a near constant level for the next fifteen years. During this time Holmes and Emerson gradually rise to equal

Twain's popularity.[5] This pattern of development helps to explain Twain's ambivalence about the event.

30 years later, while dictating his autobiography, Twain is still trying to come to terms with his performance.[6] Richard Lowry has suggested the dinner marks the emergence of Twain's interest in authorship and authority. Ngrams suggest why Twain may have been drawn to these themes: the dinner marks a shift in the trajectory of his career. Perhaps it is a leveling off that marks his establishment as a star; perhaps it is a slowdown or detour in his development. The inability to interpret this visual change in his trajectory with any certainty mirrors Twain's own inability to interpret the dinner and helps to explain why the event has become, as Lowry notes, "a kind of locus classicus of Twain scholarship" (14).

Lystra's *Dangerous Intimacy*, Shelden's *Mark Twain: Man in White*, and the publication by UC Press of a new edition of autobiographical writings by Twain are signs of enduring interest in the arc of his career. These texts are united in a determination to revisit Twain's final years by focusing on his personal life, rather than his writings. The focus on biographical detail these texts display, while seeking to inspire new visions of Twain, continues the practice of constructing portraits that place his actions and activities at their center. Ngram Viewer offers an opportunity to approach Twain's legacy from a different vantage point, one that begins by taking conversations about Twain, rather than the man himself, as a starting point.

[5] Fig. 3 in the appendix

[6] See "11 January 1906" for Twain dictating a glowing review of the speech; and, "23 January 1906" for an equally negative response.

In several of the collections available through the Ngrams Viewer the year of Twain's 1895 world tour stands out as a point near which his notoriety rapidly begins to distinguish itself—both in terms of volume and volatility—from the contemporaries he was called upon to honor as part of Whittier's birthday celebration. Collections of books published in English, Spanish, or Italian; works of fiction published in English; and books published in England or in the United States all show a strong upturn in references to Twain beginning near this point. These observations seem to confirm what Shelden has called "The Grand Adventure" of Twain's final years. Other collections, however, leave this claim in doubt. In collections of books published in French and books published in the UK, consistently higher levels of interest in Twain date from near the appearance of his first books. In the German corpus distinctive attention to Twain is present early in his career, then drops off before appearing again as his career ends. Twain's My point here is not to make a definitive statement about the arc of Twain's career, but to suggest that expansive collections of texts contain multiple narratives about authors and their works that can both confirm and complicate our understanding of literary history.[7]

Twain Today

Google Trends provides the ability to explore Twain's notoriety in a more contemporary setting; it tracks the relative popularity of searches conducted using

[7] Fig. 4 – 11 in the appendix

Google. Terms are charted by region, over time, and within multiple additional categories. Videos on the help page—on creating advertising messages, evaluating brands, and measuring campaign impact--imply the tool is aimed at the advertising industry. This targeted marketing masks the potential of Trends to be a general purpose tool for investigating the circulation of words and phrases since 2004 (the earliest year for which data is supplied). Many scholars, particularly those interested in the newly coined field of "Culturomics" ("Culturomics"; Michel), have turned to Google's Ngram Viewer in order to discover patterns in the use of terms within very large collections of digitized texts. Trends offers an opportunity to explore how contemporary audiences treat similar objects of study: timelines created by Trends show Twain consistently attracting high levels of attention, followed by Emerson, and a third group constituted by Holmes, Whittier, and Longfellow. Emerson's spot is different from his position in Ngrams Viewer.[8] Ngrams often group Emerson and Holmes together, followed by a third group constituted by Whittier and Longfellow.[9]

Pairing the data available through Trends with the results of Ngram Viewer creates a window into the notoriety of a subject over an extended period of time. The combined result from the two tools suggests that Emerson's current position behind Twain has roots that extend back to the nineteenth century, where he and Holmes frequently form a group in language corpuses. It also shows that at some point Emerson began attracting enough attention to separate himself from Holmes, but not

[8] Fig. 12 in the appendix

[9] These groups appear clearly in the English language corpus (Fig. 3 and 4); Spanish language corpus (Fig. 5); corpus of books published in England (Fig. 8); books published in the United States (Fig. 9); and the German language corpus (Fig. 10). The separation is less clear in the Italian language corpus (Fig. 6); the corpus of works of fiction published in English (Fig. 7) and the French language (Fig. 11).

enough to rival Twain's notoriety. Trends shows interest in all five authors tends to decrease as we move forward from 2004. The decline is cyclical, with lows frequently occurring in July, June, and August before climbing again, presumably as the school year resumes. The presence of a pattern in our worldwide search results that follows the school year in the Northern hemisphere, specifically the school year in North America, suggests that interest in these authors reflects regional preferences. These preferences can be further investigated using heat maps: created by Trends to show how interest in a topic is distributed across a region, these maps show Twain and Emerson attract the most widespread following of the five authors; Holmes and Longfellow form the middle; and interest in Whittier is the most localized.[10]

Exporting data from Trends to csv files allows for the construction of more detailed visualizations.[11] A stacked bar graph shows the US is the only country to register interest in all five authors.[12] This result further confirms the link to North America suggested by the cyclical pattern observed in timeline. The US registers to highest number of queries for every author but Longfellow; the Philippines registers the most interest in Longfellow. Canada and the United Kingdom register interest in everyone but Whittier. Both countries divide their queries in a similar fashion: Twain and Holmes capture slightly more interest than Longfellow and Emerson, but queries are fairly well-balanced among the four authors. India and the Philippines show interest in Twain, Emerson, and Longfellow. In both countries interest in Longfellow

[10] Fig. 13-17 in the appendix

[11] For data exported from Trends for figures at the Whittier birthday dinner, see WhittierDinner.csv in Supporting Files for Chapter 1

[12] Fig. 18 in the appendix

runs high; and, India favors Twain over Emerson; the reverse holds for the Philippines. Germany and Australia show interest in Whittier, Emerson, and Twain. Germany shows over 60% of its queries are for Twain, while Australia shows more balanced interest in all three men. Singapore, South Africa, Mexico, Poland, Italy, Spain, the Netherlands, Turkey, Brazil, China, France, and Japan register interest in Twain and Emerson. Singapore and South Africa stand out in this group for favoring Emerson with 60% of its queries to Twain's 40%. Japan, China, and Brazil nearly split their interest. The remaining countries favor Twain more than 60% of the time. Out of 69 countries that register an interest in Twain, 50 countries register an interest in Twain alone. These results don't tell us whether authors are the subject of queries because they are less known than their companions or more known. However, the percentage of queries provides a gauge of how active the reputation of an author is within a particular country. A donut chart shows how well balanced the spread of interest is around the globe.[13] Interest in Twain is fairly evenly distributed across a number of countries. Emerson has also reached a number of countries, but over 50% of his reputation is concentrated in three: the US, the Philippines, and Canada. Almost half of the interest shown in Longfellow originates in the Philippines. Whittier has not made much of an impression outside of the US.

Linking the maps created via Trends to data available through Wikipedia creates several opportunities to explore how readers interact with the materials they take an interest in. According to a 2012 study, the article that appears in the most

[13] Fig. 19 in the appendix

languages on Wikipedia is about the “True Jesus Church,” which appears in 254 of 283 Wikipedias (Warncke-Wang et al.). 97 versions of the article can be traced to users that appear to be from New Zealand and another 79 were started by a single user; the authors of the study argue that this is evidence that a dedicated group of users can affect content in Wikipedia on a global scale. The authors also note that if articles are ranked by size, “True Jesus Church” doesn’t appear in the Top 20. This observation suggests that it may be more difficult for organized groups to influence the development of an article, perhaps because language barriers and other factors may limit an article’s potential for growth. Using tools like machine translation, these same groups may create at least a minimal presence in a wide variety of languages. In addition to providing a tool for dampening the impact of organized groups of article writers, article size may be a metric for gauging interest in a topic: the larger the article, the more interest it can be said to attract. A prohibition on original research and the embrace of communal editing practices on Wikipedia make it unlikely that large articles will be produced without input from multiple users. Wikipedia’s New Pages Patrol is an example of one way articles begin drawing attention shortly after they are created (“Wikipedia:New pages”).

We can further evaluate interest in a Wikipedia article like “True Jesus Church” using Google Trends. According to Trends, “True Jesus Church” has consistently attracted attention since April 2005.[14] Interest among speakers of English in the True Jesus Church is concentrated in three regions of the globe: Malaysia, the

[14] Fig. 20 in the appendix

United States, and the United Kingdom.[15] Of these regions, Malaysia shows the highest level of interest and the United States shows the longest duration of interest.[16] In Malaysia, the majority of interest is located in and around the city of Kuala Lumpur.[17] In the US, interest is located in the states of California, Texas, and New York.[18] In the United Kingdom, interest is centered in and around England.[19] In Malaysia, interest registers for the period from April, May and June of 2012.[20] In England interest registers in August 2012, then drops out until appearing again in March 2013.[21] In the US interest registers consistently from April 2007 to the present.[22] These results suggest that despite the global presence on Wikipedia, the movement has not reached audiences in many places. They also suggest that the activity in New Zealand may be anomalous: judging from the data available on Trends, most New Zealanders are not actively seeking information on the True Jesus Church on-line, even though some New Zealanders seem to have made an attempt to establish an on-line presence for the organization. Trends data also shows that on-line interest was especially prevalent in Malaysia for a brief period and has been consistently been expressed in the United States, which may lead one to believe that the Wikipedia pages in Malay and English would reflect similar patterns of activity. However, data collected by Wikipedia suggests that Malay and English language

[15] Fig. 21 in the appendix
[16] Fig. 22 in the appendix
[17] Fig. 23 in the appendix
[18] Fig. 24 in the appendix
[19] Fig. 25 in the appendix
[20] Fig. 26 in the appendix
[21] Fig. 27 in the appendix
[22] Fig. 28 in the appendix

interest in “True Jesus Church” may not be as different as Trends suggests. The history page for “True Jesus Church” on Malaysian Wikipedia shows interest in the topic beginning in 2005 and continuing to the present. The page averages five views a day between December 10, 2007 and November 14, 2013, compared with fifty-five views averaged by the English article over the same period of time.[23] It also has attracted an average of 31 edits a year for the period between 2005 and 2013, compared with the 119 edits averaged by the English page.[24] While the scale of number of interactions is greater in English, visualizations of the interactions on both sites show both articles attracting a small, but stable following over the long term.[25] Much of this attention comes in the form of minor edits, which suggests both a high level of attention to detail and a sense of consensus concerning the content of the articles on both sites. These findings suggest that the spike in Malaysian interest

[23] See Fig. 29 in the appendix. Pageview statistics for individual articles were collected using the now deprecated Wikipedia article traffic statistics tool, Wikistats, developed by Wikipedian Emw (see Emw, “Wikistats”; “403: User Account Expired”; and, “7.7 Wikistats Tool”). The query used was http://toolserver.org/~emw/wikistats/?p1=Gereja_Jesus_Benar&project1=ms&project2=en&from=12/10/2007&to=11/14/2013&plot=1. The Wikistats tool added additional functionality to another article traffic tool developed by Wikipedian Henrik, who appears to have stopped development on his tool by 2014 (see Henrik and “User talk:Henrik”). Without Henrik’s tool to draw upon Wikistats is useless, perhaps explaining why its developer allowed it to become obsolete. (The dependent relationship between the two tools is documented at “User Talk:Emw,” see in particular the entries “3 2010 3.5 Statistics,” “4 2011 4.1 Year Wikipedia traffic,” and “4 2011 4.8 wikistats on toolserver.”) Both Emw’s Wikistats and Henrik’s tool provided users with an easy way to search an archive of article traffic between 2007 and 2016 that remains available online (see “Page view statistics” for the data; see Henrik for documentation attesting to this data as the source for the statistics provided in this chapter). Technically speaking, this data tracks search queries rather than page views (“Page view statistics”); which means that pageview statistics may exist even if pages do not exist, a useful feature for gaining a sense of interest in a topic before it becomes an article. In addition to the “Page view statistics” archive, readers interested in the statistics I provide in this chapter may also be interested in consulting the Legacy PageCounts API provided by Wikimedia (“Analytics/AQS/Legacy”) for an additional source of pagecount data between January 2008 and July 2016. Readers interested in pageview statistics from 2015 to the present should see “Analytics/AQS/Pageviews” for information about the current Pageview API; or try suite of tools available at “Pageviews Analysis.”

[24] Stats collected using Wikipedian Aka’s Wikipedia Page History Statistics tool. For the collected data, see TJC_Histories.xlsx in Supporting Files for Chapter 1.

[25] See Fig. 30 and 31 in the appendix.

captured by Trends is likely to be the product of some short term phenomenon driving traffic through Google, perhaps a class assignment or some other form of organized query likely to be repeated on a large scale, rather than a general spike in interest in the True Jesus Church.

This brief survey of the presence of True Jesus Church on Wikipedia suggests several different measures we can consult to gauge interest in a topic. Intra-Wikipedia measures--length of attention to article; size of Wikipedia hosting the article; distribution of Major and Minor Edits—are useful for evaluating interest in a topic within a specific Wikipedia audience. Inter-Wikipedia measures like number of languages and size of article are useful for evaluating interest in a topic across Wikipedia audiences. Word count; and number of visitors, editors, links, and sources have all been shown to indicate article quality (Nielsen). Using many of these variables, the developers of Wikibu have created a tool that allows users to evaluate the reliability of articles on German Wikipedia.

One inter-Wikipedia measure, inter-language links, has proven particularly popular with researchers seeking tools for investigating patterns in the creation and dissemination of knowledge on a global scale. Petzold et al. argue that inter-language links "can provide vital indications of both the knowledge relationship among languages and the currency of concepts among languages (some universal/cosmopolitan and some particular/regional)." A group of researchers at Northwestern University has developed Omnipedia (Bao et al.) in order to visualize the kinds of patterns that interest Petzold and many others. "Omnipedia highlights the similarities and differences that exist among the language editions, making salient information that is unique to each language as well as that which is shared more widely" ("About Omnipedia"). It uses inter-language links to group excerpts from articles written in different languages into "multilingual articles" that can then be read in a number of target languages via machine translation.

Despite their popularity, inter-language links are not without their limitations. Warncke-Wang et al. observe "unique articles are generally about people, places, organizations, historic events, and cultural artifacts like music, artists, and TV/radio" and they often lack inter-language links. The authors argue this omission is a sign of the "limited scope of interest" in these topics. To the creators of Omnipedia, however, "mounting evidence suggests that some of these differences are due to variation in world knowledge across language-defined communities, not simply missed opportunities for translation" (Bao et al. 1076-1077). Comparative study of wikis, they conclude, provides "a big picture view of how much topics were being discussed in different language editions," and also provides English-centric audiences with the ability "to comprehend the magnitude of information that was not available to them in the English Wikipedia" (Bao et al. 1083). Inspired by their effort to understand the reach of articles on Wikipedia, I substitute expressions of interest in texts for links in order to explore relationships between audiences around the world in order to explore a possible path of translation between the world of data analysis and the world of literary study. In addition to developing a map of the uneven spread of interest in particular works of literature, my approach is a step toward redrawing portrayals of “Big Data” scientists and “close reading” humanists as antagonistic combatants jousting for control over the future of the humanities.

Wikipedia: A Gateway to Big Data and Literary Studies

Taking Omnipedia as a potential model for a union of big data and literary studies, I mine Wikipedia in order to transform data about individual articles into maps of when, where, and from whom authors and texts are drawing attention. These maps sketch the terrain that our own readings travel; they provide a sense of where the paths we follow merge with others and where they begin to diverge. Knowledge of the travels of the readers that precede us—a sense of what texts have drawn attention, where that attention originated, and how it has been received—underscores the atypical nature of the expansive popularity of a figure like Mark Twain. Twain’s works clearly do not conform to the limited scope observed by Warncke-Wang et al. Whether as a result of his connections with an American Empire-building machine working full force to spread its reach around the globe, or perhaps because of his efforts to bring this machine to a halt, Twain’s reach is rivaled by few other authors.

A window into Twain’s global appeal is provided by the Wikimedia Foundation, which makes available a number of statistics that can be used to draw comparisons between its various projects (“Wikimedia Statistics”). In addition to providing an overview of the activities of the language communities interested in Wikipedia, these aggregate statistics contain narratives about Twain that a fine-grained approach to the data can bring to the surface. Developing these narratives invites a turn away from the prepackaged tools and datasets offered by the kinds of large institutional presences—both public and private—that have been the focus of my attention thus far and brings into view less well traveled territory where the interests of individual scholars may merge with the stores of "big data" that have become a hallmark of the landscape in which we now move. This terrain is host to a number of institution-driven entities—places like Stanford's Literary Lab, the Metalab at Harvard, and the Scholar's Laboratory at the University of Virginia—that provide in a kind of scholarly counter-point to the din of the data deluge. It is also fertile terrain for the development of what I will call parasites: small, experimental projects that depend upon more established data-stores for their existence and that provide an alternative model for the future of humanistic scholarship beyond their laboratory-like competitors.

Parasites

Akin to content aggregators and mashups, parasites feed off content provided by other sites. They may be distinguished by two of their more prominent features. First, parasites draw from a single source (although multiple parasites, working together, may come to resemble a more traditional aggregator). Second, they explicitly reformat and repurpose the information on which they feed, rather than striving for transparency: where content aggregators seek to convey information, parasites seek to construct arguments with and often against the sites they draw upon. Parasites are digital equivalents of paratexts, a version of re-mix culture grounded in code.[26] Like the paratexts from which they take inspiration, parasites position data for consumption; both are an unavoidable element of the process of presenting information to readers. Conscious efforts to exploit the possibilities offered by parasites, should they become widespread, may mark a new phase in the development of the Web.

Histories of the Web often note two phases in its development as a tool for presenting information: an early period where websites were static sites for distributing content—something like digital billboards that made whatever was put up available for all the world to see—and a second phase, commonly labeled Web 2.0, where users begin interacting with content and with each other, transforming these billboards into something more like bulletin boards where users could comment on what they saw, leave messages for others, and engage in numerous other forms of more and less social activity. Howard Rheingold calls this second period "the social web." Its growth has created conditions that may be conducive to the emergence of parasites on a grand scale. These conditions include: a large, well connected audience interested in sharing content and interacting with each other; content providers interested in making material available for others to use; and, the availability of tools and training for bringing these groups together.

[26] See Genette on paratexts and Lessig on remix culture.

Visualizing Chronicling America (VCA) is a parasite hosted by Chronicling America, a website maintained by the Library of Congress as part of an effort to digitize and distribute access to newspapers published in the United States. As of March 5, 2018, the project boasts of 12,918,917 available pages spanning the years 1789 to 1963; and a directory of the titles of newspapers published between 1690 and the present ("Home"). The standard interface for browsing this material is text based: users provide a combination of states, years, and terms to search; the response to the query is available as a list of links to images of the appropriate pages or as a gallery of thumbnail images. These responses are useful for finding particular documents within the collection and for mindless browsing, but frequently leave users that desire to explore the holdings of the collection with multiple pages of results to flip through. Recognizing that this situation may not be ideal for some researchers, the developers of Chronicling America have provided an application programming interface (API), a choice that allows users to access the data provided by Chronicling America on their own terms and creating an environment that invites parasites like VCA to take hold.

In its current form, VCA allows users to interact with Chronicling America using a timeline. It was inspired by Jer Thorp's visualizations of the New York Times (NYTimes 365/360 in particular) and was written in Processing, the simplified, yet powerful Java-based programming environment created by Ben Fry and Casey Reas.[27] VCA provides a sense of the frequency with which authors, texts, and other

[27] See Thorp for visualizations of the New York Times; and, "Processing" for more on Processing. For the Processing code I developed for Visualizing Chronicling America, see linegraphScale2chronicleAmerica in Supporting Files for Chapter 1.

ngrams appear in the LOC's extensive effort to digitize US newspapers. It extends the search interface for Chronicling America by offering a visual summary of search results in the form of a timeline. The timeline reveals trends in media coverage of a topic. Because it quantifies the data we put into it, the timeline can also be read as a window into our interest in making a topic part of our historical record. Accessing documents via the timeline serves as a reminder that documents are snapshots from a particular moment in time, not unbiased or ungrounded accounts of past events. The interface I have created stresses the temporal distribution of documents in Chronicling America. Alternative parasites may emphasize other aspects of the data. For example, spatial distribution of interest could be explored by mapping documents by place of publication, circulation of publication, etc. Enhancements like these would offer even more finely grained engagement with our historical records.

Another parasite, biblioGrapher, provides an example of what more finely grained attention may look like. biblioGrapher has yet to be built, but planning for the project was begun as part of a hackathon held by the DPLA in Chattanooga, TN prior to the launch of the website ("Appfest"). biblioGrapher allows users to visualize search results from multiple perspectives. If Visualizing Chronicling America invites us to view the past as a collection of blocks of time, biblioGrapher invites greater scrutiny of the activities taking place within those blocks. The target audience for the project includes literary scholars, librarians, and others interested in developing a sense of the temporal, linguistic, and spatial travels of a text, author, topic, or phrase. biblioGrapher offers a more finely grained view of the past by paying attention to more of the metadata associated with each of the records in the database being accessed. For example, the DPLA records more than a dozen different languages for items it holds. Counts of the holdings over time for each of these languages follow their own trajectories. Visualizing these trajectories and explicating the ways in which they harmonize with each other and with more traditional versions of literary history is one way in which the union between bibliography and literary criticism at the end of the 20th century under the guise of Book history may continue to assert its vitality via a digital flowering at the start of the 21st century. The construction of parasites is one way in which scholars may help to nurture this process by taking an active role in interacting with content providers to create the environment within which literary studies in the 21st century may grow.

Parasites are not simply a technical endeavor; nor are they defined by the creation of new tools and techniques. Construction is one form of what I will call parasitic thinking, a frame of mind which is as likely to involve repurposing existing tools as it is to call for the creation of new ones. For example, sonification is a technique for turning data into sound.[28] Its more well-known uses include the Geiger counter and the heart-rate monitor; it also allows us to explore what literary history sounds like.

Sonification allows us to hear publication histories as musical events. In the case of a collection of 686 bibliographic records in the HathiTrust database that list Mark Twain as an author, sonification allows us to hear Twain's history of

[28] For an extensive introduction to sonification, see Hermann et al.

publication as a series of musical “movements.” These movements convey a sense of dynamic range latent in otherwise static representations of the past; they are examples of what I call the texture of history, the grooves in our historical records. A feeling for these grooves--the peaks and valleys, plains and pockmarks that populate our attempts to record the past--deepens our sense of the attention Twain’s texts have attracted at various points since they first appeared.

My approach to sonification takes up the investigation of paratextual elements begun by Genette, who persuasively argues that the bits of information that accompany a text--but are not often considered parts of the text themselves--make framing arguments about how a work should be received. Scaling up his observation, I argue that looking for patterns in the distribution of paratextual elements is one way of gaining a sense of the attention texts have attracted attention around the world. Jerome McGann has likened pattern in publication and reception histories to the DNA of literary studies (Textual 16). Extending his comparison, we might say that paratexual elements are among the base pairs. Examining patterns of paratextual elements provides a sense of how texts have been put together; a window into the kinds of attention they have attracted; and an indication of audiences they hope to garner. It may also provide insight into regions and periods of time that could be otherwise inaccessible to us—paratextual elements may be legible where linguistic elements might not—and a baseline against which to gauge interest in other forms of cultural circulation.

One problem with studying paratextual elements is that they can be hard to spot. Genette asserts “a paratextual element may appear at any time, it may also disappear, definitively or not, by authorial decision or outside intervention or by virtue of the eroding effect of time” (6). Access to a large number of editions of a work provides one way of meeting this challenge. Access to the work of bibliographers, the under sung heroes of literary history, is another. Another challenge pertains to their longevity: “one may doubtless assert,” Genette claims, “that a text without a paratext does not exist and never has. Paradoxically, paratexts without texts do exist, if only by accident: there are certainly works—lost or aborted—about which we know nothing except their titles” (3-4). The checklists, handbooks, and pamphlets produced by bibliographers are a treasure trove of paratextual information and one of the domains where the longevity Genette observes in them asserts itself. The ability of paratexts to outlive the texts they may accompany provides a point of connection between projects like HathiTrust and projects like Open Library. Finally, as the amount of paratextual information we consider increases, it brings with it many of the challenges that users of large datasets commonly face. These challenges may give the impression that drawing conclusions about literary history from large bibliographic databases may be no more appealing than the more traditional methods of literary study that contemporary approaches like “distant reading” have set out to overturn.

The sonifications I have created from the HathiTrust’s collection of Twain’s works are shaped by, even as they attempt to overcome these challenges. They were produced by using Sonification Sandbox, a Java application overseen by Bruce Walker of the Sonification Lab at Georgia Institute of Technology. Walker created the software as part of an effort to popularize sonification by offering “broadly usable, platform-independent, and user-friendly tools” (Walker and Cothran 3). I was unable to get the program operating on a new Windows 8 PC; but was able to use it with some success on a Mac running an older operating system (OS 10.4).

I assembled my data using the HathiTrust Research Center portal to locate 686 records with “Mark Twain” in the author field. I downloaded the MarcXML files for these records and then used Python to extract information from them.[29] I focused my attention on the leader and control fields; the Python scripts create csv files populated with variables like place of publication, date of publication, and language of publication. The csv files are then imported into Sonification Sandbox.

As is the case with visualization, sonification is as much an art as a science. Where visualizations work with elements like color, shape, and pattern to arrange visual space; sonifications work with parameters like pitch, volume, tempo, placement, and type of instrument to arrange aural space. The line between aesthetic embellishment and clarifying insight is equally hard to draw in both fields. Both of the sonifications I created focus on control field 8 in the XML, which appears in every record in my dataset. The first is a sonification of the primary date of publication as defined by character positions 07-10 in control field 8, aka “Date 1.”[30]

[29] For the MarcXML files I collected and the information I extracted from them, see myInputfiles and HathiXMLDataforTwain.xlsx in Supporting Files for Chapter 1

[30] See MARC 21 for a detailed guide to MARC encoding. For a look at possible types of information encoded specifically in control field 8, see “008 – Fixed-Length.”

This sonification is an example of what one hears in the dataset with minimal processing.

One of the things that stands out to me in this example is the emergence of a structure that seems to separate the dataset into three “movements”: a period of rapid oscillation, but little dynamic movement at the beginning; a period of rapid oscillation and wide dynamic shifts in the middle; and a silent period punctuated by occasional bursts of pulsing, near rhythmic activity.[31] The pulse in the final section in this series is particularly interesting to me: perhaps it signals the stabilization of a publishing cycle, perhaps the death-rattle of a publishing cycle coming to an end. From our present position it may be impossible to tell; however, the assumption that we hear the sound of stability would seem to place the future of Twain’s legacy in greater risk.

A more elaborate sonification allows us to hear the interplay of language and place of publication in Twain’s works as an intricately structured relationship that gives way, over time, to a monotonous paring of English language texts published in New York.[32]

Recognizing the presence of more and less musical structures like these in our bibliographic records is an example of a method for creating multi-dimensional models of history that I call “textural histories.” Textural histories offer alternatives to the homogenized, sanitized histories cobbled together by nation-states in order to consolidate their power; they also complicate the narratives of resistance those who

[31] The sonification is available as a midi file in Fig. 32 of the appendix. See Fig. 33 for a visualization of the data.

[32] The sonification is available as a midi file in Fig. 34 of the appendix. See Fig. 35 for a visualization of the data.

answer Benjamin's call to brush history against the grain often produce. They deploy digital tools in order to represent the past as a mosaic of fragmentary bits of evidence, rather than a seamless narrative. These narratives are a step toward comparative, poly-vocal literary histories that make visible the connections and disconnections that give shape to our world. The more apparent these relationships become, the more textured our sense of the past may be, and the greater the opportunity we will have to experience our history in meaningful and transformative ways. Sonification collapses the distance between person and page. A form of what has been called "peceptualization," it is a technique that encourages us not simply to survey, but to experience the contents of our historical records (Hermann et al. 3; van der Heide and Lamers).

Wikipedia: A Host

The availability of data plays a key role in shaping the alternative engagements parasites enable. On a basic level, without a data source to feed upon parasites cannot exist. Beyond this baseline requirement, data provides specific information about past events and acts as a benchmark for evaluating the scope of our knowledge. Data for literary studies is available from a mix of private and public sources. The data available through Wikipedia stands out in this mix because it is produced through active and ongoing curation: Wikipedia articles are both a presentation and an evaluation of knowledge. To close this chapter I will look at some of the problems and possibilities for enlisting Wikipedia data in the service of literary studies. I begin by discussing pageview statistics; next I introduce a tool I have developed for analyzing these statistics; and, finally, I apply the tool to a three-part case study exploring interest in articles on American literature, Mark Twain, and Huckleberry Finn. The study demonstrates how pageview statistics may be combined with other metrics in order to develop a portrait of how people from different linguistic communities interact with topics of share interest on Wikipedia.

One approach to working with articles on Wikipedia that has gained some traction involves compiling article traffic statistics by tracking pageviews.[33] Pageviews are problematic because they obscure distinctions between the types of activity that may have led to a page being viewed: composing, revising, and viewing a page are lumped together as part of a pageview count; as are repeated edits by a single user. This practice inflates pageview counts and makes it difficult to determine the contexts in which a page is being viewed. One way to address these issues is to supplement pageview statistics with editing statistics; joining the two helps make multiple types of interaction with a page more visible. For example, pageview statistics for the American literature article on English Wikipedia record over 2.4 million views, an average of 1,296 per day, during a 1,872 day period between December 10, 2007 and February 4, 2013.[34] Editing stats show that editors will make

[33] See "Wikipedia: Web statistics tool" for a current list of tools for investigating pageviews. Deprecated Wikipedia article traffic statistics visualizers that used this approach include the tools developed by Emw and Henrik (see note 23 above for more about these tools).

[34] Stats collected using the now deprecated Wikipedia article traffic statistics tool, Wikistats, developed by Wikipedian Emw (see Emw, "Wikistats"; "403: User Account Expired"; and, "7.7 Wikistats Tool"). (See note 23 above for discussion of Wikistats.) The query used was

12.1 changes to the page in an average month and 146.8 edits in an average year. An edit occurs every 2.5 days; and, 22.5% of the edits are proclaimed "minor" edits.[35] "A minor edit is one that the editor believes requires no review and could never be the subject of a dispute." For example, correcting spelling and grammar, fixing broken links, and adjusting the layout of a page are all considered minor edits; while the addition and deletion of content and links to the page are considered major edits. Edits can only be flagged as minor by registered users. This is one of the ways Wikipedia displays a distrust of anonymous editors. While anyone can edit the site, there is a clear preference for editors that are willing to have their use of the site tracked. The stated justification for this preference is that it helps to prevent vandalism ("Help:Minor edit"); this choice also creates an environment where anonymous users are assumed to lack judgment.

One useful supplement to pageview statistics is the number of languages an article has been taken up in. This readily accessible metric provides a rough gauge of how much attention a topic has attracted across the language communities represented on Wikipedia. The Wikimedia Foundation lists official Wikipedias in 285 languages as of January 13, 2013 ("List of Wikipedias"). "American literature" articles appear in 35 of these languages (12% of the languages available on

http://toolserver.org/~emw/wikistats/?p1=American_literature&project1=en&from=12/10/2007&to=2/4/2013&plot=1

[35] For editing data for the American literature article on English Wikipedia, see ChartComparingAmLitOnWikis .xls in Supporting Files for Chapter 1.

Wikipedia).[36] Uniting pageview statistics and language links from an article can provide a sense of how vigorously a topic has been taken up by speakers of languages we may not be familiar with and an opportunity to explore how these patterns relate to those languages we may know more intimately. The Arabic Wikipedia article on American literature is one of many places where we can see the benefits of a more complex approach to the analysis of individual pages on Wikipedia.

Created on August 10, 2009 by Nayrouz Aly, the Arabic Wikipedia article on American literature is part of the list of interwiki language links included on the American literature article on English Wikipedia.[37] Pageview statistics compiled between December 10, 2007 and January 25, 2013 (1,859 days) record 27,357 views, an average of 15 views per day. A peak of 118 views is recorded for November 30, 2011. The data record a constant stream of visitors looking for the page beginning in February 2008, over a year before the creation of the American literature article.[38] A comment included by Nayrouz Aly in the revision history labels the page as a translation created using Google Translation Toolkit and provides evidence of the widely accepted practice among Wikipedians of using material on the English Wikipedia as source material for starting an article in another language.[39] At the time

[36] For a list of languages, see “American literature.” For the interwiki language links collected from the American literature page on English Wikipedia, see AmLitLangs.txt and AmLit01212013.xls in Supporting Files for Chapter 1

[37] This and other article creation dates were collected using the Contributors tool formerly located at https://toolserver.org/~daniel/WikiSense/Contributors.php and now available at “Page History.”

[38] Stats collected using the now deprecated Wikipedia article traffic statistics tool available at http://toolserver.org/~emw/wikistats/ (see note 23 above). See “Wikipedia:Pageview statistics” for further discussion of page view statistics on Wikipedia.

[39] See “Revision History” for Nayrouz Aly’s comment; see Open Translation Tools for a collaboratively authored chapter from 2009 promoting the use of the Google Translate Toolkit on

this article was created, Arabic Wikipedia was near 106k articles following its creation in 2003 (“Wikipedia Statistics Article Count”). Since its creation, the article has been edited 54 times—30 times for self-declared “minor” changes. The mean time between edits is 23 days, with an average of 15.9 edits per year. 27 unique editors--6 of whom chose to identify themselves as anonymous--have made alterations to the article.[40] In addition to the activities of the users that created the page and the users that have viewed it, a third tool may be used to open a window into the activities of Wikipedians who wish to be alerted to any changes made to the page. Fewer than 30 people are actively watching Arabic Wikipedia’s American literature article.[41] The article appears to function more as a point of reference and record of consensus than as a platform for discussion and debate. The limited number of watchers suggests the article does not change much or in ways that have attracted much scrutiny; and, at 15 views per day and 15.9 edits per year, editing activity on Arabic Wikipedia is below average. (The average number of edits per year across all of the Wikipedias with an American literature article is 20.7, with a standard deviation of 27.8.) However, if the frame of analysis is limited to the more restrictive

Wikipedia; and, see “Wikipedia:Translation” for a more current introduction to translation practices on Wikipedia.

[40] Page history attained using Wikipedian Aka’s Wikipedia Page History Statistics tool. For the collected data, see statistics.webarchive and HTML_conversion_of_statistics_webarchivefile in Supporting Files for Chapter 1

[41] Page watcher stats collected using the now deprecated Watcher tool developed by Wikipedian MZMcBride (see MZMcBride, “Watcher” and “Hello”). Instead of using his tool, MZMcBride now advises people interested in current watch statistics for a page to consult the Page Information that accompanies a Wikipedia article (see “UserTalk:MZMcBride,” particularly section “94 Number of watchers”; and, MZMcBride, “Hello”). As a security precaution, the exact number of watchers is not provided if less than 30 users are watching a page. (“UserTalk:MZMcBride” offers an account of the emergence of this decision particularly section “107 Watcher tool count.”) See “Help:Page Information” for more on the information that accompanies each Wikipedia article.

category of editing activity in the previous year, editing activity on the Arabic page is slightly above average (13 edits vs. an average of 11.3 across all Wikipedias). The average number of edits per user provides further evidence that levels of interest on Arabic Wikipedia in American literature may be more typical than they first appear: the page averages 2 edits per user, compared with 2.2 edits per user across all Wikipedias with a page for American literature. If we consider that the median number of edits per user across all Wikipedias with an American literature page is 1.7 edits, users of Arabic Wikipedia appear to rank among the more active groups with an interest in American literature (on par with users of English Wikipedia). Similar statistics may be compiled for the other 34 non-Arabic American literature articles; and, with this table in hand we begin to develop a multifaceted portrait of interest in American literature.[42]

"American literature" in Global Relief

The detailed records kept by Wikipedia are a rich source of information about interest in American literature across multiple language communities. One of the ways in which their richness may be articulated is to transform them into feeding grounds for parasites that compliment data-driven approaches to literary studies. For example, the parasites I have enlisted to examine Wikipedia show—perhaps unsurprisingly—that the American literature article on English Wikipedia receives

[42] For this table, see ChartComparingAmLitOnWikis.xls in Supporting Files for Chapter 1.

more pageviews, both on an average day and in aggregate, than any other project; they also allow us to see that on a few exceptional days this is not necessarily the case. The average number of English pageviews (1295) is approached or surpassed by the peak number of pageviews on the Chinese (1017), German (1267), and Spanish (1327) pages. These peaks outline surges of interest in American literature expressed by a heterogeneous, multi-lingual audience. While accounting for these surges is difficult, the ability to identify them opens for play a broad field of comparative studies into the relationships between texts, authors, and audiences along a number of spatial, temporal, and linguistic vectors. While certainty about the impetus for a surge may be beyond our reach, the effort to explain these movements reveals a field of possible events that together help to chart the contours of a variety of possible histories and potential futures for American literature that come into being through the union of digital tools and scholars trained in the humanities.

My argument, in other words, is that the data collected by Wikipedia are not only statistics about how users have interacted with the site; they are a shadow produced by past events and a glimpse into a possible future: they chart prior interest in American literature and gesture toward paths of inquiry yet to be followed. They show, for example, that on the day views of the American literature page on English Wikipedia reached 4243, more than 3 times the average daily count, an article on Ernest Hemmingway was featured on the English Wikipedia Main Page ("English Main Page"). The day before the American literature article on Spanish Wikipedia reached 1327 views, Mario Vargas Llosa won the Nobel Prize for Literature (a feat also noted on the Main Page for that project).[43] These two facts serve as a reminder that pageviews do not occur in a vacuum. They are indicators of activity, we may never know for sure whether they lead or follow. Narrating the relationships between pageviews and content is one of the avenues through which literary studies in the 21st century may continue develop.

German Wikipedia shows a 10-fold spike in interest in American literature on the day Google honored Mark Twain’s 176th birthday by depicting Tom Sawyer whitewashing Aunt Polly’s fence (“Mark Twain’s”). A spike in viewers to Chinese Wikipedia occurs on the 109th birthday of Chinese novelist Shen Congwen, a presence who “looms large in the history of Chinese literature not because he wrote an unusually monumental work but, on the contrary, because his contributions to literature were so diverse and pervasive” (Gargan). These findings gesture toward an evaluation of what the contours of American literature may be and what places we see it holding in our world: peaking interest in American literature connected with Congwen and Vargas Llosa, for example, suggests that American literature maintains a certain global appeal by virtue of the fact that it is one facet of a larger conversation about Literature; this claim down plays national origins in favor of emphasizing literary roots and offers an alternative to more nation-based approaches to the field. Meanwhile, more traditional representations of American literature are supported by the spikes in pageviews on the German and English language projects. With further

[43] See “Spanish Main Page.” For a sample of media coverage see Bosman and Romero; Kakutani; and Dunne, who notes Vargas Llosa is the first Latin American win in 12 years. See “Why is Mario Vargas” for a record of the event trending on Twitter.

attention—and a richer pool of data--we may come, at the very least, to a point where we can weigh national and global interests against each other in order to evaluate whether contemporary readers see American literature as a site of regional interest or an avenue of more transcendent connections.

Another perspective on the balance between local and global concern for literature opens when we turn our attention toward individual authors or groups of authors who have drawn enough attention over the years to be identified for their relationships with particular literary traditions. Twain, for example, is often taken as a prototypical, if not the prototypical, American author. English Wikipedia lists pages for Twain in 103 additional languages.[44] This stat speaks to the breath of his appeal: he has pages even on wikis where American literature does not. He gets more attention in many categories, including number of editors and number of edits. However Twain and American literature both average slightly more than 2 edits per user and have similar average percentages of anonymous users (13% for American literature; 14% for Twain).[45] These stats suggest that while Twain may attract a large audience, the level of engagement of that audience is similar to that of the audience for American literature.

[44] For a list of language codes for Wikipedias with a Mark Twain article, see TwainHomePages011713.xls in Supporting Files for Chapter 1

[45] For the Wikipedia data collected for the Mark Twain and American literature articles, see MarkTwain020313.xls and AmLit020313.xls in Supporting Files for Chapter 1; and, for a comparison of this data that also includes information about the Huckleberry Finn article, see StatsCompared.xls in Supporting Files for Chapter 1.

Statistics for Huckleberry Finn further underscore Twain's popularity.[46] Interest in Huck Finn measures below Twain, but above American literature in many categories—number of edits; number of minor edits; mean time between edits (where a lower number signals more activity on the page); average number of edits per month and year; number of edits within the previous week, month, and year (but not within the previous day). Measures of the number of editors and the average number of edits per user do not follow this trend: Huckleberry Finn has the most named editors, most anonymous editors, and smallest average number of edits per user of the three pages. Together these metrics suggest that Huck Finn draws attention from a greater number of editors, but that these editors typically make fewer alterations to the article. The history page for the Huck Finn article on English Wikipedia reveals that a large portion of the edits to the page are acts of vandalism and their subsequent correction. The frequency of these entries on the history page is a reminder of the controversies that have surrounded the text at various points in its history and an indication that these controversies are not only a part of the past. The large number of edits to the page (which averages slightly over an edit a day), the high number of editors, and the large percentage of anonymous editors are indications that the page is an attractive platform for attention seeking perpetrators of juvenile acts of rebellion. The fact that the page has 128 watchers also indicates that the page draws a lot of attention, both positive and negative, from its audience.[47]

[46] For the Wikipedia data collected for the Huckleberry Finn article and a comparison of this data to the Twain and American literature articles, see HuckFinn020313.xls in Supporting Files for Chapter 1

[47] For the current number of watchers, see "Information for 'Adventures of Huckleberry Finn.'"

History pages from other wikis bring more detail to this nascent portrait of user interest in the text. German Wikipedia has hosted a page on Huckleberry Finn for nearly as long as its English counterpart. It is the second highest scoring page in terms of many editing metrics. The page averages an edit every 13 days and has an average of 28 edits per year. Unlike the English Wikipedia page for Huck Finn, the German page has been relatively unscathed by vandalism. This fact suggests that the controversy surrounding this text may be contained to English speaking audiences. The construction of parasites like Visualizing Chronicling America, biblioGrapher, and others is a possible approach to testing this and other claims by marshalling the growing resources available through digital archives alongside the regional, national, temporal, and linguistic concerns of established forms of literary criticism in order to study works of literature on a global, multi-lingual scale.

Chapter Two: "Mark Twain" in Chronicling America: Contextualizing Mark Twain's Autobiography

Introduction

> …life does not consist mainly—or even largely—of facts and happenings. It consists mainly of the storm of thoughts that is forever blowing through one's head. Could you set them down stenographically? No. Could you set down any considerable fraction of them stenographically? No. [Fifteen] stenographers hard at work couldn't keep up. Therefore a full autobiography has never been written, and it never will be. It would consist of three hundred and sixty-five double-size volumes per year—and so if I had been doing my whole autobiographical duty ever since my youth [all the] library buildings on the earth [could not] contain the result. ("10 January 1906")

> I intend that this autobiography shall become a model for all future autobiographies when it is published, [after my death,] and I also intend that it shall be read and [admired] a good many centuries because of its form and method—a form and method whereby the past and the present are constantly brought face to face, resulting in contrasts which [newly] fire up the interest all [along] like contact of flint with steel. ("26 March 1906: Paragraph 26")

Twain's wholesale denial of the existence of a comprehensive autobiography functions as criticism of existing work in the genre and a preemptive strike against anyone seeking to point out gaps in the narrative he offers. The distance he places between his autobiography and a complete, linear narrative also enables the alternative approach he brings to the genre: framing life as impossible to contain—by a stenographer, a book, building, or even a planet—frees Twain to explore the possibility of an open-ended, curatorial approach to autobiography. Although Twain does not use the term curation (which I read as an amalgam of at least four subprocesses: selection, notation, annotation, and juxtaposition) it names the union of form and method he places at the heart of his autobiographical effort.

The zeal with which Twain advances his curatorial approach to autobiography—an approach he frames as both a break with the past and a model for the ages—calls into question his relationship to existing traditions of autobiographical writing and his familiarity with curatorial practice. Les Harrison argues that the development of the museum in the United States during the nineteenth century is marked by two conflicting paradigms, which he names "the temple" and "the forum." As temples, museums function as didactic spaces where narratives are presented for public consumption; as forums, museums function as exploratory spaces where the public is charged with constructing an understanding of the items on display (Harrison). Much like Harrison divides the history of the museum in the United States in two using an archetypal temple (the collection displayed by Charles Willson Peale) and an archetypal forum (Barnum's American Museum) Twain bifurcates the development of autobiographical writing into a didactic/consumptive period and an experiential/constructive period. In addition to creating a point of origin for his own text, the space Twain marks out at the rupture between the two/four periods brings the nineteenth century divide Harrison sees at work in the development of the museum in the United States into the realm of autobiography.

Twain's insistence that all autobiographies are incomplete and hope that all future autobiographies will descend from his own draws attention to existing autobiographical texts and establishes distance from them. These existing texts divide into at least two groups that help to frame Twain's approach to autobiographical writing: autobiographies he mentions in his text and those he does not acknowledge. The former category includes texts by Ulysses S. Grant and Benvenuto Cellini, while among the latter group are works by Helen Keller and P.T. Barnum. The content and circulation of texts in these two groups provides points we may use to triangulate Twain's conception (and reception) of the autobiographical traditions with which he is engaging. Identifying aspects of Twain's autobiography that overlap with and depart from existing autobiographies brings into view the ways in which his project attempts to distinguish itself and the points where his protestations of difference fall flat. Investigating the circulation of texts he references (and ignores) further sketches the contours of the autobiographical framework against (and within) which his text may be positioned. Locating Twain's text within this larger network of autobiographical efforts provides context for his intervention: it enables us to see, for example, which authors he chooses to explicitly place himself in conversation with and which voices are left out. Focusing on how other autobiographical voices reverberate through Twain's text helps to distinguish his trailblazing moments from more pedestrian conversations. It is also a step toward understanding the inner-workings of what Twain calls the "apparently systemless system" of the text that leaves us better positioned to evaluate three facets of the future engendered by the work: the future Twain envisions for his autobiography; the future detailed in our historical records; and our own visions of what may be in store for Twain and his text ("26 March 1906: Paragraph 27").

Contextualizing the Autobiography: Mapping an Intervention

In his work on the ancient novel Dan Selden has popularized the idea of reading via "text networks," "autopoetic bodies of related compositions" (7). Inspired by Selden's model, which seeks to show how an individual text "acquires its historical significance less singularly or diachronically…than associatively in relation both to earlier and later texts" (8-9), I read the Autobiography with an eye toward uncovering the ways in which Twain works to create a network of contrasts designed to spark interest "like contact of flint with steel." The Autobiography is particularly suited to this approach because it is a text about—and constituted in large part by—other texts: public speeches, personal anecdotes, and private correspondence share space with newspaper clippings and literary criticism; and, the juxtaposition of these texts is as much a part of the work as the contents of the texts themselves. Sketching the features of the Autobiography's network makes visible the meandering paths around which the text was constructed. Identifying these paths, the known landmarks they circle, and the unknown terrain they skirt is one way of expanding our reading of Twain's autobiographical effort to include what Jerome McGann encourages us to see as the "rhetorical dimension/dementian" of the text: an angle of approach from which the effort "to establish forms of readerly attention – to select and arrange textual materials of every kind in order to focus the interest of the reader (audience, user, listener) and establish a ground for response" comes into view ("Marking Texts"). Ignoring the ways in which the Autobiography is constructed to evoke a response from its audience hampers our ability to evaluate the text in terms of one of the explicit goals first set by Twain for the project: to produce a work that "shall be read and [admired] a good many centuries because of its form and method" ("26 March 1906: Paragraph 26").

Landmarks

Texts directly referenced in the autobiography, such as those by Grant and Cellini, are undoubtedly a small sample of the autobiographical texts with which Twain was familiar. Nevertheless, these texts provide a baseline for evaluating Twain's presentation of the existing body of autobiographical texts that precedes his effort in the genre. Grant and Cellini appear frequently in Twain's text. Grant's popularity and connection to Twain's publishing house provide logical reasons for his appearance in Twain's text. Cellini's connections with the American Renaissance unfolding around Twain may help to account for his presence.[1] In addition to their joint popularity, the two figures pair in a number of other ways: recognized for his autobiographical account of daily life in sixteenth century Italy, Cellini is steeped in

[1] Cellini held enough currency to become a nickname for at least one prominent American architect, Stanford White (Glazer and Field 31). A partner in the architectural firm of McKim, Mead & White (designer of the Boston Public Library, among other structures) White's murder in 1906 spawned what was dubbed by the press of the day "The Trial of the Century," ensuring that he was something of a household name when the bulk of Twain's autobiographical dictations were made (Linder).

the kind of classical associations Twain attempted to bestow upon Grant (most clearly when comparing Grant's autobiography to the writings of Julius Cesar). Dubbed by Twain "the simple soldier," Grant also functions as an exemplar of the common man for Twain; cast in this light, Grant offers as foil to the artistic heights scaled by Renaissance artists of Cellini's ilk. Juxtaposing references to the two men links high cultural and the common interests of Twain's day through a simple binary. Framed by Grant on one side and Cellini on the other, Twain's narrative resonates with both elite and the popular concerns.[2]

A list of autobiographies that escape comment includes texts by several notable figures, including Helen Keller. Twain's lack of comment on Keller's autobiography is particularly surprising given her strong presence in the first volume of the MTP edition of the work: references to Keller open and close the first volume of Twain's text, which also includes a photo of the two together.[3] P.T. Barnum's texts also avoid comment. Twain's neglect of autobiographies by these popular contemporary figures is notable because it points to other prominent autobiographies that he doesn't mention. For example, Terence Whalen places Barnum, Franklin, and Douglass at the center of an American autobiographical tradition (viii-ix): none of their work draws Twain's attention. The absence of these typical American figures

[2] For a depiction of Grant as "simple soldier," see Twain, Mark Twain's Civil War 86; for his comparison of the writings of Caesar and Grant, see the autobiographical dictation from "1 June 1906"; and, for evidence of the continued influence of Twain's comparison, see Perry 234; Wilson 132; Aaron 178-179; and Garrity. Peter Betjemann sees a central role for Cellini in nineteenth century discussions of the relationship between art and craft (6-7).

[3] As if to further underscore their connection, the caption to this photo reads "Helen Keller and Clemens, 1895. The inscription is in Clemens's hand" ("Photographs").

suggests a Twain in search of an alternative lineage for his autobiography that downplays his national origin. The presence of Cellini invites speculation that this alternative heritage is designed to emphasize a more international/multi-temporal autobiographical tradition constructed in part by avoiding close association or reference to iconic American texts, perhaps out of fear that these texts would taint his project. Grant, in this reading, is prized by not for his associations with America, but for his associations with more abstract, universal values: freedom, ingenuity, and integrity among them. Twain's choice of moniker reflects this: his Grant is frequently "General Grant," rather than U.S. Grant; "General" in this usage reads both as a noun identifying military rank and an adjective which brings to mind the shirking of distinction. In order to further explore the nature of Twain's usage of Grant, Cellini, and other prominent autobiographical voices, the ability to evaluate the spatio-temporal reach and range of figures circulating in Twain's time would be helpful. Visualizing Chronicling America (VCA) is one approach to meeting this goal.[4] I have applied VCA in a US-centric context, but it could be expanded to include other locations where data is available.

Boundary Lines

[4] An explanation of Visualizing Chronicling America and its development appears in Chapter 1. Data gathered by VCA is available in VCADATA in Supporting Files for Chapter 2 and screenshots of VCA may be found in the appendix for this chapter as indicated in the footnotes below.

VCA can be used to create an outline of the media coverage of figures Twain incorporates into and excludes from his autobiography. This outline allows us to explore the circulation of topics and figures in Twain's time and compare Twain's use of these figures with their presence in the media. Introducing VCA to Chronicling America suggests that many of the figures that have prominent positions in Twain's text were also popular in the media. Cellini appears as an increasingly popular feature of news articles for much of the period between 1836 and 1922.[5] References to Hellen Keller follow a similar trend, as do those to Twain himself.[6] Benjamin Franklin also increases in popularity during this period, but in a more stable way: references to Twain, Keller, and Cellini all show precipitous drops in circulation near the end of the period under consideration, while Franklin does not.[7] References to "U.S. Grant" follow a different pattern: they peak much closer to the middle of the period and appear roughly evenly distributed on either side of their high.[8] The deployments of "Ulysses S. Grant" and "President Grant" both peak slightly earlier and are less symmetrically distributed than results for "U.S. Grant."[9] The distribution of references to "General Grant" is also left-skewed and has peak levels of reference that follow more closely the distribution of mentions of "President Grant" than the pattern of references to "Ulysses S. Grant." [10] References to "President Grant" and "General Grant" show high levels of activity for multiple years, while "Ulysses S.

[5] See Appendix, Figure 1.
[6] See Appendix, Figs. 2 and 3
[7] See Appendix, Fig. 4
[8] See Appendix, Fig. 5
[9] See Appendix, Figs. 6 and 7.
[10] See Appendix, Figure 8.

Grant" and "U.S. Grant" show a narrower period of high activity. The dominance of references to Grant as "General" in all years except those during which he was the sitting President offers evidence that his military reputation exceeded his identity as a political figure in a way that parallels Cellini's status as an Artist: both men circulate via monikers that trumpet their individual prowess and mute the national contexts to which they may be linked. Barnum shares a similar pattern of circulation with "U.S. Grant": a high central peak buttressed by lower levels of distribution to either side. This trend holds true whether one looks at results for "P.T. Barnum" or "Phineas Taylor Barnum," with "P.T." being (then, as now) the much more popular form of reference.[11] The fact that the two men share similar patterns of circulation further supports the supposition that it was not Grant's fame that drew Twain's attention, but rather Grant's evolution into (and possibly Barnum's inability to become) a persona that transcends the limits of national identity.

In addition to charting individual popularity trends, VCA can be used to compare levels of attention paid to multiple figures. Importing VCA data into existing software packages allows for rapid prototyping of alternative versions of the perspective provided by the parasite. One promising substrain overlays trends for multiple searches in the same visual space. Using the line chart functionality offered by Excel to explore this possible mutation emphasizes the dominance of Grant's military persona over his political identity. Overlaying the results of the searches for "General Grant" and "President Grant" also illuminates the mobilization of Grant's

[11] See Appendix, Figs. 9 and 10.

military persona for political gain: three of the four highest peaks in the appearance of "General Grant" in the press occur co-occur with Grant's campaigns for the presidency; the remaining peak marks the year of his death.[12] Peaks marking the year of death also appear in the remaining search results for Grant, as well as "Mark Twain," "Frederick Douglass," "P.T. Barnum," and "Phineas Taylor Barnum." The consistency with which this observation appears suggests a relationship between death and media coverage that invites further study.

A version of VCA that extracts text from pages where a phrase is mentioned offers an opportunity to review news coverage in order to develop a sense of the contexts in which a phrase appears, the popularity of a topic, and the ways in which a work or phrase was deployed. In 1907 Twain was feared lost at sea and received an honorary doctorate from Oxford University. Charting the simple distribution of articles containing the phrase "Mark Twain" by month suggests the doctorate may have been a popular topic with the media, but his loss at sea was not. The highest number of appearances of "Mark Twain" in my corpus of news coverage occurs in July of 1907; the second highest in June. Twain traveled from the United States to England to accept the award and returned during this period. May of the same year saw Twain refuting reports of his death at sea.[13] The volume of appearances of "Mark Twain" during this period closely follows the results for most other months of the

[12] See Appendix, Figs. 11 and 12. For the data used when comparing four VCA searches involving "Grant" see All_Grant.xlsx in Supporting Files for Chapter 2.

[13] See Twain, Mark Twain: The Complete Interviews for press coverage of his trip (610-646) and the circumstances surrounding his rumored demise (583-587).

year, suggesting that Twain's reported brush with death did not receive an unusual amount of attention from the press.[14]

Term frequency counts for these months are dominated by stop words—90 of the first 100 terms deployed in news coverage for May 1907 are stopwords, 72 of the next hundred, and 53 of the third hundred. Results for June and July are similar.[15] Removing the stop words from the results still produces a set of highly overlapping results for each month.[16] Comparison of the top 10,000 results for each month in the trio shows the first unique result, "Howe," ranks 436 on the list; the first unique OCR artifact "thelr" appears a 2633; and, the first unique common noun, "rods," ranks 3171.[17] Overlap between these lists sketches a web of discourses that constitutes the linguistic contexts or terrain within which "Mark Twain" circulates; less common (and more unique) results on the lists highlight particular discourses at play within this landscape. Term frequency-inverse document frequency (tf-idf) scores can be used to make these particular discourses more visible.

Charting a Course: Transversing Linguistic Terrain with TF-IDF

[14] See Appendix, Figure 13. See TWAINCHRONICLINGAMERICA1907 in Supporting Files for Chapter 2 for the data related to this observation, as well as the entire corpus of newspaper coverage of upon which the following discussion is based.

[15] See Appendix, Figure 14. For the stopword list I used, see "SEASR Stopwords." For files of the term frequency counts and files identifying the number of stopwords in the term frequency counts, see 1GRAMS in Supporting Files for Chapter 2.

[16] See Appendix, Figure 15. For files of the term frequency counts with stopwords removed, see NOSTOPWORDS in Supporting Files for Chapter 2.

[17] For an Excel file of this comparison, see CompareMayJunJuly1907.xlsx in Supporting Files for Chapter 2.

Often deployed as a metric for identifying key terms in a text, tf-idf scores provide a window into the contexts in which references to Twain appear.[18] The calculation of tf-idf weights allows for studying all terms in a collection of documents: high frequency words that appear in a large number of documents (typically stop words) have low tf-idf weights, while more unique terms have higher values. Focusing on terms with high tf-idf values in a corpus effectively screens out stop words without altering the result set derived from a corpus; this is an effective advantage over examining a corpus using stop word lists because it avoids the problem of having the choice of stop word list color the analysis of a corpus.

Applying tf-idf weights to the corpus of Twain results produced by searching Chronicling America for “Mark Twain” suggest that in May of 1907 Julia Ward Howe, her Civil War anthem “The Battle Hymn of the Republic,” and the Rip Roarer mine were prominent topics of conversation in the news. Among 1-grams produced for the May 1907 corpus, “roarer” ranks second; and, the 2-gram “rip roarer” appears fifth in lists of tokens sorted by tf-idf weight from greatest to least. The 3-grams “is marching on” and “god is marching” (both parts of the refrain from Howe’s work) are first and fourth in similar lists; while the phrase “god is marching on” appears sixth in a list of 4-grams. In June 1907 attention turns to the “Revolte du Midi,” a series of confrontations between wine-makers and government forces in southern France and a murder trial in Virginia involving a former judge.[19] Results for July bring attention

[18] For a brief overview of tf-idf, see Dittenbach.

[19] For discussion of the Revolte du Midi, see Smith, Frader; and, for an account of the trial see Hamm.

the growth of the steel industry in Duluth, Minnesota and a Japanese admiral's lowly opinion of the United States Navy.[20]

While evaluating the popularity of terms within a single month is relatively easy to do with my corpus and the computing power available to me, I found evaluating the usage of a term over an extended period of time to be a very resource intensive process. In order to mitigate the limitations of the hardware on which my research in this regard was conducted, I trimmed my corpus of Twain results to include only the top 10000 1-grams in 1907 when ranked by tf-idf weight from greatest to least likely to appear. The number of terms chosen for inclusion in the corpus is arbitrary; I selected 10000 in an effort to balance coverage of the corpus with the limits of the processing power available to me: as the number of terms increases, the query takes longer and becomes more difficult to run.

To begin exploring the use of terms over time, I limited my query to only those 1-grams that appear in a maximum of ten percent of the new corpus (roughly a single month). Querying my modified corpus of Twain results with this limitation in place reveals that "roarer" is highly representative of the media coverage in May: not only does the term rank among the most heavily weighted terms in the corpus for May, "roarer" does not appear with enough frequency in other months to garner a rating. Increasing the percentage of the corpus in which a 1-gram may appear further

[20] For files containing ngrams and tf-idf weights for the 1907 corpus, see NGRAMSWITHTFIDF in Supporting Files for Chapter 2. Sort files by row from greatest to least to reproduce the results discussed here. Each row of the file holds data for a month in 1907: sort on row 6 to view results for May; row 7 for June, etc. For discussion of the development of the steel industry in Duluth, see Alanen; and, for context on US/Japan naval relations of the period, see Possner.

underscores the extent to which “roarer” is characteristic of media discussions in May. As the percentage of the corpus in which the 1-gram “roarer” may appear is increased in ten percent increments, the tf-idf weight assigned to the term decreases; but, “roarer” remains among the top 10000 1-grams in the corpus until seventy percent of the corpus is taken into consideration for a query. Somewhere between seventy and eighty percent the term falls out of the top 10000 terms. In other words, even when the queried term is allowed to appear in up to seventy present of the corpus, “roarer” remains highly descriptive of media coverage for May. While it is no longer the second most descriptive term for the month (having been equaled or surpassed in terms of tf-idf weight by the 1-grams “procter,” “kuroki,” “wenatchee,” “rosenfeld,” and “garlln”) the term “roarer” still does not appear in other months with enough frequency to garner a rating. In contrast seven months reference the 1-gram “kuroki” with enough frequency to garner a rating, a possible reflection of the scope of media interest in General Kuroki Tamemoto Tamesada, an honored guest at the founding of New York’s Japan Society on May 19, 1907.[21] A common noun, the 1-gram “fiesta” (which enters the top 10000 terms once fifty percent of the corpus is taken into consideration) appears with enough frequency to be noted in the six month period between April and September of 1907 when terms are allowed to appear in up to seventy percent of the corpus. Results for the 2-gram “rip roarer” are similarly concentrated in May. However, once 50 percent of documents in the corpus are taken

[21] For a brief account of Kuroki’s visit to the Japan Society see ”Japan Society Timeline” and for an example of the media coverage his visit inspired see “Kuroki in New York.” Mitziko Sawada notes “the New York Times gave him front-page coverage and faithfully reported his every movement and his impressions of New York City, West Point, horse racing, and American women” (17).

into consideration the term drops out of the top 10000 2-grams. The lower threshold for the 2-gram signifies that “rip roarer” is less prominent in the corpus of 2-grams; this suggests that some of the results for the 1-gram “roarer” are likely to refer to topics other than the mine, but the mine remains a likely topic for a substantial component of the media coverage for May. Expanding to the 3-gram “the rip roarer” further decreases the percentage of documents in which the term appears: the 3-gram drops from the top 10000 results once 30 percent of documents in the corpus are taken into consideration. The 4-gram “the rip roarer mine” holds a place in the top 10000 results only when tokens that appear in a maximum of 10 percent of the corpus (roughly a single month) are taken into consideration. Like its predecessors in this analysis, “the rip roarer mine” only registers as a significant component of the May document.

Using a keywords in context tool, such as the one provided as part of the Voyant suite of text analysis tools, provides additional perspective on these results. The KWIC tool in Voyant shows the 1-gram “roarer” appears 15 times in the May document; the 2-gram “rip roarer” appears 13 times; “the rip roarer” 6 times; and “the rip roarer mine” 5 times.[22] While the results provided by a KWIC tool are less abstract, they quickly become unwieldy as the number of results grows. Tf-idf weights are a useful tool for mitigating this unwieldiness. Coupled together, KWIC and tf-idf weight provide an expedient way to summarize topics present in large

[22] For a Voyant query for “roarer” in media coverage from May 1907, see voyantforRoarer.docx in Supporting Files for Chapter 2.

amounts of texts. When used in conjunction the two tools show that every occurrence of the 1-gram “roarer” in May refers to some aspect of the Rip Roarer Mine; that the same result should be expected for the 2-gram “rip roarer,” and that the reason the 2-gram appears to refer to the mine less frequently than the 1-gram is due to errors in the OCR, rather than differences in subject matter. In other words, both the 1-gram and the 2-gram function as adequate identifiers for determining the popularity of the Rip Roarer Mine as a topic of conversation in my corpus of media coverage for May 1907.

Lacuna: OCR Errors And Data Cleaning

Tools like VCA and its variant may work best when applied to error-free texts, but their use is not limited to contexts where digital texts have been scrubbed of imperfections in an effort to make them palatable to human readers. OCR errors are present throughout the texts provided by Chronicling America. These errors can make reading the texts difficult and/or unpleasant for human readers. However, they need not have an effect on the results attained by reading texts using the kinds of computer-aided reading strategies I have presented thus far. The random distribution of OCR errors throughout a corpus may mar a text, but these unsightly blemishes have little effect on reading methods that depend on sensors other than the human eye. To test the strength of this proposition I assembled a dataset of Mark Twain metadata obtained from the HathiTrust. My experiments with this data suggest that the time consuming process of data manipulation may not always lead to a significant change in the quality of the visualization produced from a dataset.

I built the data set by downloading 686 MarcXML records selected by searching for “Mark Twain” in the author field. Using Python, I extracted publication information about these records from several different fields and combined the results into a csv file. I then made a copy of the csv file and loaded that copy into OpenRefine, an open source software package for cleaning data. I used OpenRefine to standardize data related to date of publication, name of publisher, and location of publisher. When I was finished, I exported the data from OpenRefine into a new csv file. I then opened two workbooks in Excel, one based on the original csv and one based on the csv produced by OpenRefine. Using the Pivot Table and Pivot Chart functions in Excel, I produced three sets of line graphs in order to compare the impact of cleaning data on the resulting visualization.[23]

The results of my tests show a mixture of improvements and complications arising from the cleaning process. Charting the dataset by date of publication before and after processing suggests that the cleaning process introduces slightly more variation into the data: a comparison of trend lines plotted through the data using Excel reveals a reduction in the value of R^2 from 0.9447 to 0.9437.[24] Plots for distribution of publisher locations reveal a stronger trend that moves in the opposite

[23] The MarcXML dataset discussed below may also be found in the Supporting Files for Chapter 1. For files pertaining to the extraction and cleaning of this dataset, see HATHITRUST in Supporting Files for Chapter 2.

[24] See Appendix, Figs. 16a and 16b.

direction. The increase in the value of R^2 from 0.9454 to 0.9503 suggests that processing has brought about a slight reduction in the amount of variation in the dataset.[25] Finally, processing the names of publishers has the greatest effect on the shape of the dataset and produces the most interesting observation. Prior to processing, publisher names are accounted for with an R^2 value of 0.912. After processing, this value falls to 0.8695.[26]

After the dataset is cleaned, the dominance of the Harper Brothers publishing house stands out clearly. Among the changes to the data, the cleaning process united publishers identified as "Harper & Brothers,$" "Harper & brothers$" and "Harper and brothers,$" into a single group that accounts for nearly half of the volumes in the dataset. Combining these groups reduces variability among the names of publishers found in the dataset; it also increases the range separating more and less prolific publishing houses. The dominance of the Harper Brothers publishing house is interesting because it is unexpected in a way that the dominance of the high performing nodes in the publication date and place of publication data is not. The distribution curves for publication data and place of publication confirm to reasonable expectations for a dataset with a few high performing points and a large number of points that perform less well. In other words, we should not be surprised that New York dominates the publishing locations because the shape of the dataset suggests (as reflected in the trend line and the high R^2 value) that some location in the dataset

[25] See Appendix, Figs. 17a and 17b.
[26] See Appendix, Figs. 18a and 18b.

should dominate at about that level. The fact that it is New York that dominates is, from a statistical point of view, irrelevant: if it weren't New York at the top, the makeup of the dataset encourages us to expect some other location to perform equally well in comparison to the other locations in the dataset. The same point can made regarding the prevalence of the "1899-1918" date among publication dates. If that date was not the most commonly occurring point in the dataset, the shape of the dataset suggests that we should expect some other period of time to operate in a similar fashion. The performance of the Harper Brothers publishing house is different. The frequency with which the Harper Brothers appear in the dataset is not to be expected: the low R^2 value attests to this fact. The Harper Brothers are an outlier, a statistical anomaly, in comparison to the other publishers in the dataset. Their domination over other publishing houses in the dataset is so complete as to be unexplained by pure statistical probabilities.

Given the fame and longevity of the Harper Brothers publishing house, their outlier status and the praise attributed to them is perhaps not unexpected: knowing that this dataset is limited to works by Mark Twain, one might even be able to guess that Harpers (the press Twain selected to publish his complete works) would feature prominently. Quantifying the extent to which the Harpers outperformed other presses interested in Twain adds to our store of knowledge about the press and provides a metric for evaluating the performance of other publishing houses that may be at its most useful when if functions as a lodestar in the exploration of datasets where we know little or nothing about the press in question. Calculating the distance between a point representing the Harper Brothers performance and a trend line that accounts for some arbitrarily high portion of the dataset (a line with an $R^2 > 0.95$, for example) provides a numeric description of just how anomalous the performance of the house is in relation to the rest of the publishers in the dataset. This calculation is so dependent upon the value selected for R^2 as to be, in itself, ultimately meaningless. When placed in conversation with the voluminous amount of scholarship on the Harpers, however, this meaningless statistic becomes a measure of what at least one historian of the publishing industry has called "the shrewdness, skill, and vitality of the founding brothers" (Winger 61). In other words, the distance between a single point in the dataset and a trend line that attempts to account for all points in the dataset provides a quantitative compliment to qualitative assessments of the past.

Quantifying the range of interest displayed by presses is a possible way to evaluate the likely of a press within a dataset and provides a means of drawing comparisons across datasets. By calculating the frequency with which the name of a publishing house appears we can begin to develop comparative insight into how publishing industries operated in any number of environments. These comparisons may operate across cultures—for example, the performance of the Harpers may be used as a barometer for evaluating the performance of presses in other locals—or they may operate across periods of time, identifying, for example, the Harpers of today. These comparisons are likely to be most informative when they look in multiple directions: focusing attention on how a press fits into its own environment as well as identifying anomalous performances in distant locals.

A comparative approach also invites consideration of how different datasets describe a topic. In order to probe the strength of the observations made concerning the Harper Brothers press using the dataset assembled from HathiTrust, I conducted a similar study on a dataset of 2500 records collected from Open Library. One goal of this analysis is to see if the anomalous position of the Harpers within the dataset is a product of the dataset itself or if it can be confirmed via a second source. I was particularly interested in seeing if the amount of variation exhibited by the Harper Brothers publishing house in the HathiTrust data is the product of a problematic dataset or if it offers a statistical measure of the qualities of the publishing house itself.

Continuing with the procedure outlined above, I downloaded 2998 records with “Mark Twain” in the author field from Open Library. I created a Python script to extract the publisher, place of publication, and date published from these records and saved the result as a csv file. I then made a copy of this file and processed it using OpenRefine. The original data and the cleaned data were then charted using Excel to create Pivot Tables and Pivot Charts.[27]

As was the case with the HathiTrust dataset, cleaning the location fields removed variability from the dataset collected from Open Library. R^2 before cleaning was 0.9205; after processing R^2 climbed slightly to 0.9478.[28] A significant number of records in my Open Library dataset did not contain a place of publication. I removed

[27] For the Open Library dataset and Excel files produced after cleaning this dataset, see OPENLIBRARY in Supporting Files for Chapter 2.
[28] See Appendix, Figs. 19a and 19b.

these records and charted the location fields a second time. Before processing, the dataset showed an R^2 of 0.9349; the value of R^2 climbed slightly after processing to 0.9564.[29] While removing the blank fields further decreased the amount of variability in the dataset, it did not disrupt the trend observed in the prior visualization; this suggests that the presence of records with blank location data does not overly distort the shape of the curve that describes the distribution of publisher locations in the dataset and removing them may be unnecessary.

Charts of the dates of publication from the Open Library dataset showed a trend similar to the one found in the publication locations. Before processing, the dataset could be described with an R^2 of 0.924. After cleaning the data, the value of R^2 climbed to 0.9331.[30] This upward movement runs counter to the relationship found in the HathiTrust dataset, where R^2 fell slightly from 0.9447 before cleaning to 0.9437 after cleaning the publication date data. In both cases the difference in R^2 values before and after cleaning is small enough that it calls the value of the time spent cleaning the date data into question: the data may look nicer after cleaning, but the picture it paints is very similar in either form.

The results of processing the names of publishers in the Open Library dataset also call the value of cleaning the dataset into question. The original dataset can be described with an R^2 of 0.9444. After cleaning, the value of R^2 climbs to 0.96.[31] In contrast the value of R^2 fell from 0.912 to 0.8695 when the publisher names obtained

[29] See Appendix, Figs. 20a and 20b.
[30] See Appendix, Figs. 21a and 21b.
[31] See Appendix, Figs. 22a and 22b.

from HathiTrust were cleaned. While the change in R^2 does not confirm the decrease observed after cleaning the HathiTrust data, it does mirror the minimal change observable in the other five comparisons.

Further tests with additional datasets will be required to determine definitively whether or not processing has a significant impact on the distribution of values within a dataset, but for the two datasets I have explored it appears to be the case that processing the data does not have a significant impact on the trends observed in the data in most cases. My observation is consistent with other studies that set out to explore the impact of OCR error on computational analysis. One recent study on the impact of OCR error on word frequency counts obtained by digitizing historic newspapers (the same kind of material collected by Chronicling America) concludes clean data is "desirable but not essential" (Strange et al.). Another scholar cited in the same study, after noting that a clean corpus is preferable, acknowledges that the impact of OCR error depends on the type of information sought from a corpus (Eder). While OCR errors may present problems for sentiment analysis and other techniques focused on extracting meaning from a corpus, when applied to the description of the features of a corpus their presence may be overlooked. Document clustering and topic modeling, two techniques that fall into the latter category, have been shown to be effective in the presence of OCR error. Walker et al., for example, note that "clustering methods should perform almost as well on OCR data as they do on clean data" (249). In their attempt to topic model a corpus of historical newspapers Yang et al. observe "we found that although our corpus contains noise from OCR errors, it may not need expensive error correction processing to provide good results when using topic models" (103). My experiments with HathiTrust and Open Library data provide an example of how these techniques, which often are brought to bear on narrative objects like books and newspapers, may be applied to bibliographic records.

I am particularly interested in the potential for topic modeling as an approach for working with texts that contain OCR errors or other imperfections that may make them difficult for humans to read. Topic modeling is often used to determine, as the name suggests, what topics are present in a given document. However, if we forgo this frequently sought goal and seek instead to simply explore a document mathematically, rather than linguistically, the errors introduced by OCR into a text should not matter because they will constitute only a small portion of the topics found in a given document. This is not to argue that OCR errors are not troublesome, but rather to suggest that the trouble they cause has been overstated.

Topic Modeling and Messy Data

The information obtained by approaching media coverage of "Mark Twain" in terms of word frequencies and tf-idf weights provides a guide for determining when terms that are prominent in the corpus begin to appear in the output of a topic model. Definite signs of topics that address prominent terms in my corpus begin to appear in the collection of May 1907 results once 60 topics are sought in the corpus.[32] At this scale, which is slightly higher than the recommended number of topics suggested by some topic modelers for a corpus of 236 documents, topics relating to Julia Ward Howe and the Rip Roarer mine are evident in the results.[33] Ratcheting the number of topics up to 150 brings to prominence lesser known figures like “Miss Bland,” a central character in a short story attributed to Alexander Ely; and an installment of *The Lion and the Mouse*, a serialized legal thriller written by Charles Klein and Arthur Hornblow.[34] Below 60 topics, clear references to these topics recede from view: Julia Ward Howe remains simply as “Howe,” while the other topics are obscured completely. The token “Howe” remains present in topic models that seek as few as five topics. The resilience of this token is an indication of the position “Howe” holds in the corpus: as a frequently occurring token, “Howe” maintains a descriptive presence in nearly any reading of the corpus, from one that divides the corpus in two

[32] For results from using SEASR to produce topic models of 5, 10, 20, 30, 60, 90, and 150 topics found in news articles from May 1907, see MAY1907TOPICMODELS in Supporting Files for Chapter 2. I make this data available with the caveat that my interest here is not in topic modeling the corpus. The production of a topic model of the corpus is a necessary pretext to the discussion of agent-based modeling I develop in the remainder of this chapter.

[33] There is no standard for determining the appropriate number of topics and recommendations vary widely. For example, the developers of the topic-modeling-tool, a GUI interface for Andrew McCallum’s MALLET, suggest looking for 10-20 topics in a 1K document corpus; 20-60 in a 10K document corpus, and 50-200 in a 100K document corpus (“topic-modeling-tool”). Documentation for MALLET notes simply that “the number of topics should depend to some degree on the size of the collection, but 200 to 400 will produce reasonably fine-grained results” (“Topic modeling”).

[34] To read “Women’s Diplomacy,” a short story about meeting Ms. Bland, see Ely. An advertisement in Publishers’ Weekly claims Hornblow’s novels sold more than five hundred thousand copies by 1913 (see “Mask”) and the Harry Ranson Center at the University of Texas at Austin has an archive of Klein’s work as a playwright (see “Charles Klein”). To read an installment of The Lion and the Mouse, a Klein play adapted by Hornblow, see Klein and Hornblow. WorldCat Identities summarizes the plot as follows: “Judge Ross, on the Federal Bench, rules in favor of a large company in litigation before him, unaware that a smaller company in which he owns considerable stock has been subsumed by the larger firm, thus creating appearance of a conflict of interests. When one of the Judge's enemies plots to ruin the Judge over this apparent improper behavior, Judge Ross's daughter Shirley sets out to prove her father's innocence” (“Hornblow, Arthur 1865-1942”).

to a reading that seeks to an extremely fine-grained approach to the texts. The persistence of a token like "Howe" across multiple levels of granularity is useful because it illustrates how topic models also function as models of reading strategies tuned to different levels of engagement with a corpus. The reader who skims through a corpus quickly we may equate with a topic model that seeks fewer topics in a text; the more fastidious reader will identify more topics. Topic modeling shows us that both the fastidious reader and the skimmer are likely to notice references to "Howe" during their engagement with the corpus.

When approached as a mathematical model of how a text is likely to be experienced under a given set of conditions, topic modeling offers a tool for investigating relationships between readers and texts in addition to its widely recognized role as a tool for exploring what a corpus is about. Ted Underwood has suggested we approach topic models as indications of "discourses" present in a corpus.[35] Whether in terms of textual content or the more philosophical approach endorsed by Underwood, the numerical output produced by topic models is underutilized as a tool for documenting the way in which a corpus may have been read. Treating topic weight as an indication of the decisions that have lead up to the production of a particular vision of the corpus is of particular interest to me because it provides a tool for modeling reading practices in a manner that is reproducible, and

[35] This argument runs throughout most of Underwood's blog posts on topic modeling (tedunderwood, "Category Archives: Topic Modeling") and is most directly made in "What can topic models of PMLA teach."

therefore fertile ground for experimentation.[36] This consistency invites the willful manipulation of topic modeling I would like to advance in the remainder of this chapter.

Identifying topics within a corpus is extremely useful for tasks like information retrieval, but I find more interesting the way in which topics and their weights function as a numerical summary of the particular constellation of choices made in the construction of a given topic model—choices which include decisions concerning the construction and use of a stop-word list, the normalization of tokens, and the number of topics sought. Changing these parameters not only alters the textual findings a model produces, but should also be understood as an alteration to the way in which a corpus is read—quickly and sporadically in one scenario; in painstaking detail under another. I am particularly intrigued by the limits and possibilities for using a blend of agent-based models and sonification to explore the conversations these reading scenarios are likely to produce.

Model Models and Meta-Models: The Agent-Based Articulation of Topic Models

Agent-based modeling is a form of computer simulation Nigel Gilbert has likened to a Sims videogame with better theory and worse graphics (2). To explicate Gilbert's humorous, yet entirely accurate comparison, I will close this chapter by

[36] While the specific mix of tokens that make up each topic may change each time a model is produced, topic modelers are in general agreement that the material to which a topic points remains consistent: in other words, a given set of inputs will repeatedly produce recognizably similar outputs.

presenting an agent-based model seeded with information gleaned by topic modeling newspaper coverage of “Mark Twain” from May 1907. Staging this data in an agent-based model allows us to investigate how these topics may have circulated among Twain’s contemporaries. An understanding of how media coverage may have been taken up by Twain’s audience as he was writing his autobiography helps us to understand the intervention he makes with his text, a topic which I take up in the following chapter.

Modeling the media landscape embedded in an archive of nineteenth-century newspapers as an agent-based model invites us to think through the characteristics of Twain’s audience, the circulation of information, and the ways in which his text may have been produced by and exerted an influence upon their relationship. I begin this project by outlining the creation of an agent-based model that invites users to adopt the perspective of an agent navigating a digital environment in which “Mark Twain” is a central topic of interest. I then demonstrate that occupying space within the model provides users an opportunity to experience the conversation surrounding “Mark Twain” as it unfolds in real-time; and, I show how sonification may be employed to enhance this experience by providing as sense of the sonic environment created by conversations as they unfold. I conclude with a comparison of insights obtained by visualizing my dataset with those produced by modeling the same data and a discussion of what makes agent-based modeling a compelling avenue for literary research in the humanities.

An Agent-Based Approach to Modeling a Media Landscape

The model I have developed is purposely minimal. It is a sketch of an environment, rather than an attempt at faithful reproduction or historical reenactment: the blank canvas of my model world functions in at least two ways: it serves as a literal conversation space in that it marks the boundaries of a location within which conversations may be held. At the same time, this conversation space is itself the product of conversation: dependent for its form on conversation about what characteristics it should contain beyond the primitive grid of uniform squares that provides its essential structure. Agents are tiled throughout the conversation space, one per square, each seeded with a topic of conversation and a color that is associated with the topic. The use of color allows the distribution of topics throughout the environment to be quickly assessed, but currently has no other intended meaning.[37]

At each time step during a run of the model, every agent turns to a neighbor and attempts to hold a conversation. If the pair share a common topic of conversation, a count tracking the number of times that topic is discussed is increased. If the two agents don't share a topic of conversation, no discussion occurs; a counter tracking the number of times no discussion takes place is increased; and, the agent that initiated the interaction selects a new topic of conversation and changes to the color

[37] See Appendix, Figure 23. For the Netlogo and SuperCollider code to create and sonify the model, see MODELCODE in Supporting Files for Chapter 2.

of the newly selected topic. The code governing the behaviors described above is as follows:

```
ask turtles-on neighbor
  [ ifelse [topic-of-conversation] of self = [topic-of-conversation] of myself ;; if there is a topic match
    between neighbor and asker
        [ let counter table:get conversation-table topic-of-conversation ;; identify the topic in the table
          and get the count
          let increasedcounter counter + 1 ;; increase the count for the topic
          table:put conversation-table topic-of-conversation increasedcounter  ;; put the topic back into
          the table with the updated count
        ]
        [ set countnodiscussion countnodiscussion + 1 ;; if there isn't a match, increase count of no
          discussion
          ask myself [ set topic-of-conversation (random-weighted active-conversations list-of-
          probabilities) ];; if there isn't a match, asker picks a new topic
        ]
  ]
```

When agents select a new topic of conversation (and also when they are initially seeded with a topic to initialize the model) their choice is governed by an array of probabilities determined by topic modeling media coverage from a particular month. For example, the code below describes topics active in media coverage from January 1907 and the proportion of media coverage devoted to each topic during that month:

```
to load-jan-topics
  set active-conversations ["Topic0" "Topic2" "Topic4" "Topic6" "Topic9" "Topic17" "Topic25"]
  set list-of-probabilities [0.1331 0.0649 0.1878 0.056 0.1124 0.2657 0.1802]
end
```

The variable “active-conversations” names a list of topics identified in the media coverage and “list-of-probabilities” names a list containing the proportion of media coverage devoted to each topic. The specific combination of topics and proportions that describes the media coverage for each month is determined by topic modeling a collection of newspaper articles from 1907 in which the phrase “Mark Twain” appears. The results of the topic modeling algorithm are treated in my model as a proxy for the likelihood that a given topic will be discussed as agents navigate their environment. In other words, my model assumes that the more media coverage a topic receives in the media, the more likely it is to be a topic of conversation in the world I have designed. The work of complicating this assumption, for example by creating feedback loops between media coverage and agent interests, has been left for another stage in the development of the model.

A Model Experience

In its current form the model suggests what conversations are likely to evolve when agents in a controlled environment are made familiar with a set of discussion topics extracted from a digital archive of newspapers. Staging these discussions provides predictions about how media coverage may circulate (or may have circulated) through an audience under a given set of conditions. Knowledge of statistically likely patterns of circulation helps us to better understand the conversation dynamics that are part of the discussions preserved in our media archives and other sources of historical documentation. Modeling the circulation of historically attested discussions through a simulated environment also provides benchmarks that may be useful for predicting patterns of conversation in situations where documentation pertaining to the circulation of a topic may be lacking. For example, running the model ten times shows several different ways in which conversations seeded with the topics identified in media coverage from January 1907 may spread through the environment I have constructed: a minimally defined environment that is currently free of constraints on the ability of agents to acquire and discuss information, but which could be readily reconfigured to include such details. Both divergence and overlap are evident after just one time step in the development of the set of topics explored over this collection of ten runs, which were generated by running the model with random seeds initialized from 1 through 10 (inclusive).[38]

Topic 17, one of three topics that appear in every month of my dataset and the most prevalent topic within the dataset, appears in the upper most corner in each chart. Eight of ten times Topic 17 grows in popularity after one time step; it remains flat once; and decreases once. Similarly unpredictable development patterns may be observed for the other topics. The uncertainty captured by the model is a product of the behavior of the agents, which choose at random what they will discuss with whom in accordance with the constraints imposed on their behavior during the design of the model. As a result of this randomness, each run of the model is apt to be distinct in its details. However, as the number of runs of the model grows it becomes possible to distinguish a general range of expected outcomes for this topic data in this environment. My ten run sample indicates, among other things, that Topic 17 is likely to increase in popularity after one time step. Increasing the number of model runs to 10,000 and tracking number of agents interested in Topic 17 lends further support to this assumption: plotting the distribution of agents for each of the 10,000 runs shows

[38] See Appendix, Figure 24.

that both averages and outliers for Topic 17 tend to increase after one time step.[39] Increasing the number of runs of the model also brings into view likely and less likely upper and lower bounds on the level of attention agents in the model may be expected to devote to Topic 17. Identifying these limits provides a sense for when, where, and by how much any particular observation conforms to or deviates from the expectations suggested by a given dataset.

Juxtaposing observations taken from our present environment, or an environment preserved in our historical records, with patterns of behavior described by simulated norms provides additional perspective on exceptional and quotidian events by situating them within a range of possible outcomes. For example, in the eighth run of the model the observed level of engagement with Topic 17 decreases from an initial value of 305 agents down to 295 agents after one time step.[40] This decrease stands out within the context of the ten run dataset. Expanding the dataset to include ten thousand runs of the model shows a decrease is likely to occur in 593 runs (or roughly 6% of the time); these decreases may range between 1 and 47 agents; the average amount of decrease is 9 agents; and the distribution of runs with a decrease over these ten thousand runs closely follows an exponential curve: small decreases are common; and, decreases become more infrequent as the size of the decrease grows.[41] These observations suggest the decrease of interest in Topic 17 observed in run eight may be described as fairly typical: the level of decrease observed in run

[39] See Appendix, Figure 25.
[40] See Appendix, Figure 24, panel eight (second row, 3 from left).
[41] See Appendix, Figure 26.

eight (-10 agents) is very close to the average level of decrease (-9 agents) seen in the 593 runs where a decrease is observed. Viewed in the context of a large number of model runs the decreased number of agents discussing Topic 17 observed in run 8 of the initial ten run dataset looks uncommon (given that decreases in the dataset occur only 6% of the time) but not as unusual as it may initially appear to be. Extending the collection of data out over many time steps allows us to make similar observations about the ways a topic (or set of topics) may be expected to evolve over time. Distinguishing common and uncommon patters of development provides a greater appreciation of those exceptional moments in our recorded histories that stand apart from the expected ebbs and flows of everyday life. For example, running the January data out to 1000 time steps shows topics are not likely to shift away from their initial distribution in the environment through chance alone. In all ten runs (again with random seeds 1 to 10, inclusive) the arrangement of topics in the environment stays close to the initial distribution identified by the topic model with which the environment was seeded: three groups of two topics beneath a single topic.[42] Comparing the percentage of media coverage attributed to a topic by the topic modeling algorithm I employed with the minimum, maximum, and average number of turtles with the topic generated by the model indicates the presence of deviations from this observed tendency that can be, at times, quite large. In the case of Topic 17, which accounts for 26.57% of the media coverage according to the topic model, interest fluctuates from a minimum of 24.52% of agents in the environment to a

[42] See Appendix, Figure 27.

maximum of 41.97% of agents over the course of ten one thousand time step runs. As a whole, the data from this experiment suggests that a widely reported topic, such as Topic 17, may be likely to circulate through an audience with even greater frequency than it appears in the media.[43] This result suggests that topic modeling may indicate lower limits of circulation for prominent topics in an archive of media coverage, but may be approach the upper limit of circulation for topics that are less well covered.

In addition to tracking the spread of individual topics through an audience, the model may be used to compare the discussion of individual topics with the development of discussion in the environment as a whole. There are 1,089,000 opportunities for discussion during a 1000 time step run of my model (1089 agents x 1000 time steps = total discussion opportunities in the environment). The vast majority of these discussion opportunities fail: an agent turns to its neighbor and is unable to connect via a shared topic of interest nearly 73 percent of the time on average. Of those conversations that do succeed, discussion is dominated by Topic 17, which accounts for more than half of the conversations that take place in the environment after 1000 time steps.[44] Visualizing a summary of the conversation data shows that the percentage of active conversation for each topic is established in the model soon after a spike in the percent of failed discussions. This spike in failed discussions is observable early in the development of each run of the model. Once established, the percentage of discussion devoted to each topic remains largely

[43] See Appendix, Figure 28.
[44] Conversation data summarizing ten runs of the model to 1000 time steps (with random seeds 1 through 10, inclusive) is available in the appendix, Figure 29.

unchanged over the course of the remaining time steps. In every run the percentage of the total discussion devoted to each topic is far less than the percentage of media coverage devoted to each topic and the percentage of agents familiar with a topic.[45]

Summarizing comparisons between the percentage of the corpus attributed to each topic by topic modeling, the maximum percentage of the total discussion produced by modeling the circulation of each topic, and the maximum percentage of agents familiar with the topic at any point during ten runs of the model to 1000 time steps (with random seeds 1 through 10, inclusive) shows that discussion between agents in the model occurs infrequently.[46] The low levels of conversation observed suggest that while high levels of media coverage may spur high levels of familiarity with a topic among agents in the environment, conversation between agents is fairly uncommon in the egalitarian conversation space of the current model. (All agents are equally likely to converse with each other and simple proximity is the primary characteristic that enables conversation between agents.) While the maximum percentage of agents familiar with a topic consistently meets or exceeds the percentage of media coverage devoted to that topic, discussion between agents falls well below the level of coverage of topics evident in the media--even in the case of popular topics like Topic 17. These results suggest that the proportions determined by topic modeling indicate the hierarchal arrangement of topics in terms of their maximum levels of discussion and in terms of the maximum number of agents a topic

[45] See Appendix, Figure 30.
[46] See Appendix, Figure 31.

may reach. All three measurements order the topics from least to greatest as: Topic 6, Topic 2, Topic 9, Topic 0, Topic 25, Topic 4, Topic 17. The proportions of media coverage indicated by topic modeling also appear to describe a lower limit on the maximum % of agents a topic may reach. While the accuracy of this description wanes as the popularity of a topic increases, it appears to be a good predictor of audience size for uncommon topics. Altering the ways in which agents engage with each other is likely to produce substantially different observations. Modifying agent behavior is one way in which the model provides opportunities to create multiple versions of the past preserved in our historical records and exploring where these alternatives may lead.

Modeling “Mark Twain”: A Comparison of Modeling and Visualization

Using the combination of agent-based models and topic modeling discussed above I have drawn upon topic data from the remaining months to create a composite portrait of the conversations that may have been inspired by media coverage involving “Mark Twain” in 1907. This portrait is based on data generated by 100 runs of the model.[47] Each individual run provides information about the potential range of conversation dynamics that appear over the course of the year as media attention shifts from topic to topic and conversations between agents evolve.[48] With each

[47] See Appendix, Figs. 32a and 32b.

[48] See Appendix, Figs. 33a and 33b for a sample of the data from running the model with a random seed value of 44.

additional run the definition of the composite image increases as additional information about the possible behavior of agents and topics in the environment is added to the image. Analyzing these patterns of behavior allows us to observe how much attention a topic is likely to garner from an audience; and, how an audience may respond to changes in media coverage as it evolves over time. These observations provide a useful metric for investigating how the topics addressed by particular cultural commentators overlap with and depart from the interests of their audiences (a subject I take up in the final chapter of my dissertation with regards to Twain).

When plotted over time, the circulation of topics in the model appears as a series of peaks, valleys, convergences, and divergences.[49] Framing data produced by the model in this way encourage us to evaluate the movement of topics through the environment in terms of trends and to observe these trends from a distance. The desire to hold the media environment of Twain's time at arm's length and observe it from the outside is part of the legacy of the Cartesian divide between mind and body, observer and observed. Agent-based modeling offers a challenge to the Cartesian worldview where it encourages us to reevaluate the default tendency to remove ourselves from our data.

On Uri Wilensky's NetLogo modeling platform this challenge may be enacted by taking advantage of the option to view models in 3D. Depending on how it is

[49] See both the composite portrait (Appendix, Figs. 32a and 32b) and the visualization of an individual run (Appendix, Figs. 33a and 33b) for examples.

deployed, 3D in NetLogo provides the experience of traveling behind, among, or even as an agent navigating the model world. From this perspective we experience what an agent experiences: we see what an agent sees, we may hear what an agent hears, and even, with the addition of hardware designed to provide haptic feedback, feel what an agent feels. Joining agents as they navigate their world provides an opportunity to develop methods of data analysis that encourage us not only to observe, but to experience the environments preserved in our data.

Descending from the grand overlook which brings conversational arcs generated by the model into full view and attempting to survey the same data at close range provides a window into the ways in which the discussions circulating through the environment appear to the agents who experience them. For instance: when accompanying three randomly selected agents through the model, Topics 25, 9, and 4 (signified by light blue, aqua and brown in the model) all appear to be more prominent topics of conversation than Topic 17 (signified by peach) when the model is initially created.[50] What seems plausible from the perspectives of these individual agents runs counter to the distribution of topics that appears when we step outside of the model and survey the world from a distance. Running counts of all activity in the model show that at Tick 0 Topic 17 (peach) is present in 282 agents (26%); while Topic 25 (light blue) is of interest to 198 agents (18%), Topic 9 (aqua) is 129 (12%), and Topic 4 (brown) accounts for 202 agents (19%). After the first tick in the run of the model the presence of Topic 17 in the environment appears to increase, as it is

[50] See Appendix, Figure 34.

now clearly visible to all three agents. A numerical overview shows that, in fact, Topic 17 is now an object of interest for 308 agents (28%). Even at this point, however, Topic 17 is once again overshadowed by other topics when approached through the eyes of these three randomly selected agents. Agents 1038 and 182 are presented the light blue and brown topics (respectively), both of which have risen from the background to become potential subjects of conversation for agents 1038 and 182; while agent 554 has the opportunity to discuss Topic 2 (red in the model), a topic that was present within the environment but not within the field of vision of any of the three agents at the prior time step. By the final time step in this brief series, Topic 17 has captured the attention of two of the three agents and appears in the field of vision of the third. At this moment the prominence observable in the raw numerical data on topic circulation in the environment appears align with the level of visibility Topic 17 now has for these three agents. Even at this point, however, Topic 17 does not appear to be an overwhelmingly popular point of conversation: against the field of green, light blue, and brown agents visible in the background, one would be hard pressed to determine that at this point Topic 17 has captured the interest of 325 (or 30%) of the agents in the environment while the green, light blue, and brown topics account for 45, 203, and 207 agents (4%, 19%, and 19% respectively).

Fluctuations in the presence and absence of topics within an agent’s field of vision casts the circulation of topics through the environment in more uncertain light than numerical overviews and visual surveys made from a distance may lead us to believe. Embracing these fluctuations and inhabiting our models is one way of overcoming the one dimensional, hierarchal, homogenous perspective distance-based approaches bring to data analysis. Embracing proximity, rather than distance, allows us to capture individual moments from trips through the model from the perspectives of agents themselves. These static snapshots provide access to a host of subjective experiences of another world and another time: a past (or future) not as it was (or may be) but as it has been preserved in and predicted by our historical records.

Viewing the data produced by the model through the eyes of agents undermines the desire to construct a hierarchal arrangement of the circulation of topics in the model. From many of the agent-based vantage points the numerical domination of Topic 17 within the environment is not as clear as it appears to be when viewed from a distance; neither is the context within which this topic was discussed. What may appear to the outside observer as clear and distinct features of the environment—more and less prominent topics; cycles of boom and bust; developmental progress and/or decay; points of origin and endpoints—blend together to provide an embedded observer with an emergent sense of the environment that is much closer to the way in which we experience our own surroundings; and, I suggest, is much closer to the ways in which Twain and his contemporaries may have experienced theirs.

Listening to conversations in the model unfold also alters the way we perceive the circulation of conversations in the environment; this time by drawing our attention to patterns within the environment that may be difficult to see from the outside.

Extending the model by coupling NetLogo with SuperCollider is one way in which the frequency with which a topic circulates in the environment can be made audible. After sonifying the data produced by the model by creating a real-time link to SuperCollider, the circulation of topics plotted above in Figure 12 appears to divide into four groups. Conversation between agents in June and July sounds subdued: discussion among agents in these months is characterized by low frequency conversations. In March, November, and December conversations in the environment are slightly more animated, but dominated by low frequency discussions. May, August, and September see more elevated levels of discussion characterized by the presence of high frequency topics. January, February, April, and October are the site of very active conversations dominated by high frequency discussions.[51] Listening to discussion in the model evolve over the course of a run of the model also provides a feeling for how audiences may experience topics as they circulate through the environment. Topic 17, for example, fades in and out of prominence as agents converse: at times dominating the discussion and at times being hardly noticeable. The oscillation of this topic in the environment is much more difficult to detect in my visualizations, which encourage us to see Topic 17 as a steady presence in the environment throughout the year.

The raw conversation data produce by the model, subjected to minimal processing, encourages two very different readings of Topic 17: one in which it is a notable topic of conversation throughout the year; and, one in which it fades in and

[51] See Appendix, Figs. 35 and 36 for a visual representation of the sound of conversation each month.

out of prominence. My purpose here is not to suggest that one or the other is the more "correct" reading, but rather to suggest that both are equally defensible: Topic 17 is and is not a prominent topic of conversation among agents in the model. Grasping this fact requires both the visualization and the sonification of the data produced by the model. Ignoring one or the other hampers our ability to appreciate the richness of the perspectives preserved in the historical archive upon which the model is based, an archive in which Topic 17 figures simultaneously as a prominent point of conversation that dominates discussion throughout the year; and, as one interest among many—at times distinct and at times overshadowed by other threads of conversation.

The experience of exploring digital models of the worlds preserved in our historical archives, which we may choose to call "embodied data analysis," is difficult to capture in the abstract. Each model is distinct and offers its own perspectives on the data around which it is build. Each agent within a model provides a distinct temporally and spatially bound subjectivity through which we may experience the model world. In the current model there are 1089 possible subjectivities at play in each run of the model, one for each agent in the environment. While exhaustively iterating through them all may be impossible, sampling the experiences they provide of the past remains instructive. These agent-lead explorations are likely to bring us as close as we may ever come (outside of a dream) to the experience of time travel. In likening embodied data analysis to time travel, however, we must grant that it is a particular type of time travel at play: navigating agent-based models from the perspectives of the agents themselves transports us into an environment that is not as it was (or may be) but as we remember, predict, and imagine it to be.

Discussing the alternative worlds agents bring into view without viewing them for ourselves is akin to reading a review for a film, performance, or a work of art. Like these other events, the perspectives agents provide invite us into another world. Individual moments taken from a trip through the model function, at best, like postcards from that world: each an invitation to inhabit another space, rather than an attempt to make known to the viewer the experience of being there. The combination of NetLogo and SuperCollider is one possible way of accepting this invitation. As the images above from a trip through the model world may suggest, embodied data analysis does not feel natural; it feels strange, disorienting, uncanny. That it feels abnormal or disconcerting to experience our data and the worlds it conjures using not just vision, but all of the senses we use to experience our surroundings every day, is a testament to the strength of the Cartesian drive to separate observer from observation: it feels somehow more appropriate to posit some artificial distance across which we may stand and survey our findings, rather than embrace them for ourselves as our own. The English language itself supports this distinction between subject and object, lacking as it does the ability to articulate the possibility of an entwinement between the two (a grammatical construction familiar to scholars of Ancient Greece as the middle voice). If agent-based models are able to bridge the Cartesian divide, it is where they offer an experience akin to time travel; not in the sense of an effort to transport us from our present world into another, but through their ability to call elements from other worlds--past worlds and future worlds--into view from where they linger around us in the present moment. Agent-based models are one of many technologies with the power to expand our sense of the present: David Eagleman, for instance, has demonstrated through his V.E.S.T. project another potential avenue for enhancing through technology our ability to perceive the world in which we live (“Sensory Substitution”; “VEST”). In a popular survey of current work in neuroscience by Eagleman and others, Michio Kaku heralds a future in which digitization may one day allow us to plumb the furthest reaches of the Universe without ever leaving our home planet.[52] Agent-based models provide a readily available environment through which this kind of breath-taking exploration can be brought to bear on our historical records today.[53]

[52] Kaku discusses possibilities for digitization and space travel over two chapters in The Future of the Mind: chapter 13 (“The Mind as Pure Energy”) and chapter 14 (“The Alien Mind”). Of particular relevance to the model of literary study I am advancing here is Kaku’s warning—set forth within the context of a speculative discussion with Paul Davies about possible reasons why the Earth has not been visited by aliens—that the temptation to “play out imaginary lives in different virtual worlds” may come to supplant our desire to interact with our surroundings as virtual reality technologies advance (312). This age-old fear of confusing shadow for substance seems apropos in situations where somatic experience and digital simulation entwine.

[53] For video from a trip through the model world, see Appendix, Figure 37.

Chapter Three: From Mining to Modeling: An Agent-Based Approach to Mark Twain's Autobiography

"Data mining" is a popular subject of conversation among many digital humanities scholars who work with literary texts. Scott Weingarten, for example, notes in his analysis of submission data to Digital Humanities 2015 that "a full 21% of submissions include some form of Text Analysis, and a similar number claim Text or Data Mining as a topic."[1] The use of the phrase is mildly troubling given that it is a misnomer, as Wikipedia points out, that ignores the fact that "the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (*mining*) of data itself" ("Data mining"). More troubling are the connections mining references suggest between digital humanities scholars and efforts to liberate ready-made materials from their surroundings. Such implications mistakenly give the impression that computer-assisted readings of literary texts have more in common with information retrieval than with the critical and interpretive traditions that are a defining feature of humanities research; and, provide unnecessary ammunition to critics who see digital humanities as a mechanistic approach to scholarship that lends support (willingly or not) to a dehumanizing neoliberal agenda.[2]

Moving beyond mining rhetoric may leave digital literary scholars on unstable ground, but does not leave them without touchstones. Anne Burdick and her collaborators, for example, place instability at the heart of digital humanities research

[1] Scott Weingarten has made discussion of submissions to the Digital Humanities conference an annual feature on his blog ("dhconf"), see "Submissions" for his discussion of DH 2015.

[2] See Allington et al. for a sample of the critique of digital humanities as neoliberal and Clements for a selection of responses.

by deploying the construction “Digital_Humanities” to mark the known margins and unknown center of the field through the visual play of positive and negative space. Read from the outside in, Digital_Humanities renders the field as a bounded space created by the meeting of other approaches to research; reading from the inside out, the underscore at the heart of their construction draws attention to the presence of recognizably undefined and limitless potential that gives way to other areas of research as one moves toward the margins. Mary Louise Pratt’s notion of the “contact zone” echoes the space Burdick and her colleagues mark visually. Imported from linguistics to describe life along the western edge of the U.S. during the nineteenth century, Pratt’s phrase captures both the bounded uncertainty Burdick and her colleagues underscore; and the frontier spirit evident in many digital humanities projects. When considering the dynamics at work within the meeting space underscored by Burdick and named by Pratt, James Clifford’s understanding of culture as “rooted/routed” provides a potential model for explaining how devotees of the digital and the humanities may come together to produce distinctive, yet fluid forms of scholarly research. Each of these scholarly frameworks draws attention to active processes of translation between existing constructs—the machined and the manned; the West and the rest; the bound and the unbound—that offer alternatives to the rigid, hierarchal ordering of space suggested by connecting digital humanities research and mining. Jerome McGann gestures toward these processes by giving translation a central role in his understanding of digital humanities. In our contemporary moment, he argues, “we lay foundations for translating our inherited archive of cultural materials, including vast corpora of paper-based materials, into digital depositories and forms" ("Marking"). McGann's insight encourages us to see digital humanities as an act of building futures, rather than mining the past. His vision, which moves from one cultural archive to multiple foundations, offers a prophetic (rather than parasitical) version of digital humanities that casts the field as more than—and even primarily not—a technological undertaking, but as a translation project.

In Walter Benjamin's work translation is associated with the production of echoes. "The task of the translator," he argues, "consists in finding the particular intention toward the target language which produces in that language the echo of the original" ("Task" 258). Benjamin describes that echo as a translator's personal recreation of his or her source: "it is the task of the translator to release in his own language that pure language which is exiled among alien tongues, to liberate the language imprisoned in a work in his re-creation of that work" (261). The movement from source text to distinct reproduction described by Benjamin has a contemporary analogue in the production of digital texts, which Matthew Kirschenbaum has shown to be distinct representations of their source materials with histories that may be unwound on the level of the nanoscale just as surely as book historians can articulate the stories of printed and handwritten texts. Casting these digital texts as translations invites consideration of the ways in which they echo their sources while building connections between past, present, and future worlds.

Digital Humanities and/as Translation Studies

In what has become a foundational text in the field of Translation Studies, Roman Jakobson identifies three forms of translation: interlingual, intralingual, and intersemiotic (139). Within the digital humanities, evidence of engagement with the first two modes of translation ranges from longstanding projects like The Perseus Project ("Perseus") to more recent undertakings such as Folger Digital Texts ("Folger"). However it is "intersemiotic" translation that provides the most adequate characterization of the kind of cross-medial/trans-medial translation projects that McGann encourages us to associate with digital humanities research. Intersemiotic translation, which Jakobson defines as a movement "from one system of signs into another, e.g., from verbal art into music, dance, cinema, or painting" (143), not only names the transition McGann describes, but situates digital projects within a longstanding tradition of humanistic knowledge-making projects. While Jakobson's examples draw from the existing art forms of his day, his definition may be generalized to denote not only a theory of movement from one form of media to the next, but also the negotiation of relationships between source and target texts. Reframed as translation projects, digital literary research becomes a process of rewriting the work of prior ages in order address the present and prepare the future. Defined specifically as intersemiotic translation project, the production of digital approaches to literary texts is also a project of articulating—through selection from existing forms and creation of new ones—the context appropriate constraints within which this rewriting process may take place.

Susan Bassnet and André Lefevere remind us that “all rewritings, whatever their intention, reflect a certain ideology and a poetics and as such manipulate literature to function in a given society in a given way. Rewriting is manipulation, undertaken in the service of power, and in its positive aspect can help in the evolution of a literature and a society” (vii). “But rewriting,” they warn, “can also repress innovation, distort and contain” (vii). Our ability to come to terms with the rewriting projects in which digital literary scholars are engaged is greatly enhanced by unmasking what Lawrence Venuti has termed the translator’s “invisibility,” particularly with regards to the translator’s interest in producing “foreignizing” and/or “domesticating” versions of their source texts (1-42). Venuti’s project, which seeks “to elaborate the theoretical, critical, and textual means by which translation can be studied and practiced as a locus of difference, instead of the homogeneity that widely characterizes it today” (42) offers scholars interested in using technology to investigate literary texts an opportunity to avoid duplicating digitally the kind of reductive, uniform, homogenizing experiences Venuti sees as the outcome of many projects in literary translation, particularly those seeking to make texts written in languages other than English available to English speaking audiences; such translation projects, he argues, are united by “the translator’s effort to insure easy readability by adhering to current usage, maintaining continuous syntax, fixing a precise meaning” (1). These efforts, he claims, mask the translator’s role in producing a text in order to create the illusion that audiences have an unmediated relationship with the translator's source materials. To ascertain that a parallel project to Venuti's critique of literary translation is necessary in the digital humanities, one need look no further than the static visualizations that characterize much digital scholarship carried out in the wake of Franco Moretti's call to remake literary history in the style of "graphs, maps, and trees."

Graphs, Maps, and Trees: Beginning Anew Literary History

Derided as "an absurdity" and praised for bringing a novel approach to the study of literature since its appearance in 2003 and 2004 as a series of articles published in the New Left Review, practitioners of "distant reading," as the approach to literary criticism Moretti has helped to popularize is known, share a desire to assemble large collections of text and extract insights from them, often dealing with their content (if the texts are digitized) or investigating their publication and reception histories.[3] For example, Moretti's line graphs of "The Rise of the Novel, 18th to 20th Century" draws out similarities in the production of the form over five regions and three centuries by compiling data spread out over a number of regionally and temporally specific bibliographies with titles like *A Check List of English Prose Fiction, 1700-39* and *The Novels of the 1740s*, both of which chronicle the

[3] Moretti published his argument over three articles in the journal. For early criticism, see Harold Bloom's reaction as reported by Eakin; and, see Koktsidis for praise from Homi Bhabha.

appearance of the novel in Britain.[4] Ted Underwood and Andrew Goldstone give us dueling network graphs of trends in literary criticism as reflected in 5940 articles published over the past century in the Modern Language Association's flagship journal, PMLA, and available through JSTOR.[5] Matthew Jockers and David Mimno use word clouds investigate discussions of what they label "female fashion" in a corpus of 3,346 works of fiction published in the United States and Great Britain between 1750-1899.[6] The preference for clean, crisp, and clarifying visualizations evident in the work of these scholars and many other digital humanists interested in working with literary texts has strong roots in a Cartesian divide between subject and object; and unashamedly promotes a vision of literary history in which critics stand at a distance, surveying from heretofore unknown heights—with the aid of technologically-assisted vision—their objects of interest. The scene is eerily reminiscent of the description Benjamin provides of the Angel of History, "who looks as though he were about to distance himself from something which he is staring at. His eyes are opened wide, his mouth stands open and his wings are outstretched" ("On the Concept"). Before the Angel, Benjamin continues, "one single catastrophe, which unceasingly piles rubble on top of rubble." Unlike the Angel, however, who is blown back by a storm ("that which we call progress") and prevented from acting on the objects of its attention, the data-driven distant reader is deeply embedded in a mode of scholarship that seeks to accomplish what Benjamin's Angel could not: "to

[4] See Appendix, Figure 1, reprinted from Moretti, "Graphs, Maps, Trees: Abstract Models for Literary History--1."

[5] See Appendix, Figure 2, reprinted from tedunderwood [Ted Underwood], "What Can Topic Models of PMLA Teach."

[6] See Appendix, Figure 3, reprinted from Jockers and Mimno.

pause for a moment so fair, to awaken the dead and to piece together what has been smashed." There is a strong tendency among distant readers to frame this pause and the angelic perspective it depends upon as objective knowledge (evident, for example, in the drive to establish Culturomics as a "scientific" pursuit) ("Culturomics"; Michel). Embracing the language of objectivity undercuts the revolutionary fervor that is among the most appealing aspects of distant reading and ignores another of Benjamin's powerful insights: "to articulate what is past does not mean to recognize 'how it really was.' It means to take control of a memory, as it flashes in a moment of danger."

The challenge set by Benjamin, to take control of a memory, to make it speak to us, invites more flexible accounts of literary history than those Moretti's work has brought into play. The appearance of distant reading, framed from the outset in Moretti's work as a moment of danger heralding the transformation of literary studies into a quest for natural laws and objective truths, has inspired an astounding array of perspectives on our literary pasts. These perspectives remain, however, static portraits of our literary past constructed from the top down and the outside in, products of observation and surveillance that are complicit in the same kind of authoritarian silence that undergirds the official histories they seek to overturn. Here I strive for a less permanent version of literary history, one that gives voice to the perspectives contained in our records of our past by employing agent-based models to work from the bottom up and inside out to investigate perspectives on literary history that are not entirely under our control.

An Agent-Based Approach: Literary History through the Eyes of Agents

Nigel Gilbert likens agent-based simulation to a SIMS game with better theory and worse graphics (2). Agent-based simulations work by giving highly autonomous agents the freedom to explore, interact with, and essentially create their environment. The approach has been used to tackle everything from problems of racial segregation to disaster response scenarios to insect behavior. Its use in the humanities has included the exploration of historic events and cultural evolution, both real and imagined (Dean et al.; Epstein and Axtell). Among literary scholars it remains a niche field in need of greater recognition. Michael Gavin has begun the project of drawing the attention of literary scholars in the digital humanities to this area of research; and this work is an effort to continue that project while continuing my own research into how digital tools may be used to conduct literary history (Gavin; Throne).

The agents in the model I have developed navigate a given landscape and converse about given topics of interest under their own power. Simulating these interactions is an attempt to bring to life the “mediascape” surrounding Mark Twain as he worked to compose an autobiography near the end of his life (Appadurai 9). Exploring this environment provides a window into how topics in the media may have circulated among Twain’s contemporaries and a forum for examining how the introduction of topics discussed by Twain in his autobiography may have influenced these patterns of circulation. Undertaking this exploration through agent-based modeling environment like NetLogo allows scholars to join the overarching surveys familiar to us as visualizations with the limited understanding agents have of their environment in order to produce surrogate perspectives through which the worlds preserved in our historical records may be experienced anew. Immersing ourselves in these simulations by engaging the past vicariously through the eyes of agents navigating virtual worlds provides a wide range of subjective accounts of historical events that, when brought together, outline the contours of the past worlds and "possible futures" (Gillman, "Humbolt" 526) contained in and produced by our historical records. Knowledge of these contours helps to distinguish between and articulate relationships among expected and exceptional pasts; and the futures to which these pasts may lead.

The conversations I simulate in my model are based upon media coverage preserved in the Chronicling America project at the Library of Congress, which contains over 9.4 million pages of digitized newsprint published in the United States between 1836 and 1922 as of May 28, 2015. I take the archive of media coverage Chronicling America provides as a record of the topics of conversations people are likely to have discussed after encountering content published in the press.[7] Topic modeling is a popular tool among digital humanities scholars interested in working with the holdings of large digital archives like Chronicling America; and, network visualization is a common approach for organizing the output of topic models. Using

[7] For information on the current holdings of the project, see "Home."

Complex Adaptive Systems (CAS) to work with the output of topic models provides an alternative to the failure of the combination of topic models and network visualization to transcend the limits of what Merleau-Ponty and many other theorists working in the field of visual studies refer to as a "Cartesian" understanding of vision, a "realist paradigm, which turned vision into a view *on* the world, rather than *in* it" (Levin 163). A CAS approach mitigates the rigidity of the Cartesian divide between subject and object by providing opportunities to observe the development of interactions between agents in an environment as they unfold from multiple perspectives: the observer perspective offers the kind of omniscient view favored by many practitioners of network visualization, agent-level perspectives provide a sense of what it may be like to occupy a position within a network. Agent based models also provide an opportunity to generate multiple experiences of the same historical information. These experiences, when bundled together, provide access to what I call the texture of history: the range of potential pasts that Susan Gillman encourages us to see as possible, "unfinished" futures, "futures yet to come" ("Humboldt" 525). Agent-based modeling provides a powerful tool for exploring the trajectories these unrecognized pasts put in place and the pasts upon which these alternative futures may depend.

The historical textures that interest me here outline the contours of the media conversation surrounding Mark Twain as he worked on the composition of his autobiography. A complicated and highly self-conscious work, the bulk of the Autobiography consists of a series of daily dictations Twain carried out between 1906 and 1909. Twain himself explains the construction of the text as follows: “the thing uppermost in a person’s mind is the thing to talk about or write about. The thing of new and immediate interest is the pleasantest text he can have—and you can’t come here at eleven o’clock, or any other hour, and catch me without a new interest—a perfectly fresh interest—because I have either been reading the infernal newspapers and got it there, or I have been talking with somebody; and in either case the new interest is present—the interest which I most wish to dictate about” (“16 January 1906”). Simulating the circulation of topics Twain is likely to have encountered as part of his daily regimen of newspaper reading and conversation allows us to see how the text of the autobiography adheres to and deviates from the construction of a possible representation of “the thing uppermost” in Twain’s mind during the composition of the text. Assuming that, at a minimum, he paid attention to newspapers he himself was mentioned in, I have limited my exploration to contexts in which a reference to Twain appears.

Using software of my own design, I extracted from Chronicling America a corpus of 3174 pages of newsprint (nearly 95MB of text) that contain a reference to “Mark Twain” during the year 1907, a year in which Twain could be expected to figure prominently in the media. In May of 1907 Twain was feared lost at sea after a ship on which he was traveling failed to reach its appointed destination on time; in July he traveled to Oxford to receive an honorary doctorate—an honor so significant to Twain that he wore his ceremonial robes on a variety of occasions from that point on, including his daughter’s wedding. I have further limited my analysis of the 1907 corpus to the three month period during which these events occurred in order to compensate for the limitations of the computer on which I constructed topic models of the corpus. Topic models suggest that neither event figured prominently in the pages of newsprint in which “Mark Twain” appears. Readers in May, for example, were more likely to be exposed to reportage on Julia Ward Howe or the Rip Roarer mine.

Armed with a set of topic weights produced by means of a topic modeling flow developed in the Software Environment for the Advancement of Scholarly Research (SEASR) to provide a general overview of the contents of the corpus of media coverage within which the phrase “Mark Twain” operates, I turn to agent-based modeling to recreate the media conversations surrounding Twain as they are recorded in our historic newspapers. Actively exploring various incarnations of the model, each one populated by agents programmed to discuss with each other a set of topics derived from the holdings of Chronicling America, allows me to examine how levels of interest in particular topics of conversation change over time and how different variables within the environment, and within the agents themselves, may influence these observations. Staging this environment as an agent-based model also provides an opportunity to investigation how Twain’s Autobiography could have influenced the development of the conversations that inspired him if he had made the text available to his contemporaries, rather than withholding publication of most of the work until after his death.

The media of the day, or more precisely what Appadurai calls the "mediascape," an amalgam of physical, social, and technological processes which give shape to and are in turn shaped by currents of information circulating throughout our world, is positioned by Twain in the quote above as a fundamental inspiration for his text: it is the source of his "perfectly fresh interest," and provides a sounding board against which he develops his autobiographical account. Just as a sounding board amplifies a speaker's voice so that it may be heard in a crowd, a feeling for Twain's sounding board enables us to better hear (and more clearly hear) what he was talking about. In the absence of this board, we are only receiving a portion of Twain's text. The combination of topic modeling and the millions of pages of digitized newsprint contained in Chronicling America is a potentially powerful way of accessing this sounding board and developing a sense of what Twain and his contemporaries may have been reading. However, topic modeling only gives us half of the story: it provides access to what may have been read, but it has little to say about how this material may have been discussed. Agent-based modeling helps to address this issue by providing a possible method for reconstructing the missing half of Twain's sounding board.

In the prior chapter I discussed the circulation of topics in the media surrounding "Mark Twain" during 1907. A similar combination of the techniques used to model those patterns of circulation shows the phrase "Mark Twain" is a constant feature of newspaper coverage throughout 1906 as well. Coverage peaks in April and December and reaches its lowest point in July. The pattern of peaks and valleys in the 1906 coverage draws further attention to the amount of media interest in Twain’s trip to Oxford in June and July 1907, which garners more attention than either of the peak months for Twain coverage in 1906.[8] Comparison of the two datasets is complicated, however, by the fact that the data was collected from a constantly changing database at two different points in time: the 1907 data dates to August 16, 2014; while the 1906 data was collected on October 11, 2016. The holdings of Chronicling America that reference Twain are likely to have changed in the interim, but more research is needed on this point. Although the data offer imperfect points of comparison across the two years, each data set makes a useful contribution toward underlining the constant presence maintained by “Mark Twain” in the media in its respective year. Ascertaining Twain’s level of participation in the media is important because it provides a context for evaluating his engagement with other topics circulating in the media at the same time. Assessing Twain’s level of engagement with the media allows readers of the Autobiography to determine, for example, whether his interest in commenting on stories in the press is tied to the levels at which he himself appears in the news. Answering this question allows us to see where the text may be read as a highly filtered account shaped by Twain’s own presence in and personal relationship to the media; and, where the autobiography functions more as a record of the times, devoted to and developed out of stories with Twain is no more or less personally involved than that of any other reader of his day.

[8] See Appendix, Figure 4. For an Excel file of data comparing my 1906 and 1907 Chronicling America search results for “Mark Twain” see Mark_Twain_1906_1907_comparedv2.xlsx in Supporting Files for Chapter 3.

Term frequency counts for the first three months of 1906 consistently show a general mix of formality, masculinity, and temporality dominating the top of high frequency term lists for each month of press coverage and an array of more idiosyncratic words occupying the bottom of the frequency count associated with each month.[9] 2grams made with the same data highlight the inward looking nature of the press and draw attention to some of the more common topics in the media with which Twain shares the spotlight.[10] Surpassed by the 2gram "United States" and rivaled by "Mr Mrs" (the phrase "Mr and Mrs" with the stopword "and" removed) "Mark Twain" appears more often in the news than "President Roosevelt," "Standard Oil," the "White House," and multiple U.S. cities. Corrected for variations such as "Mark Twains" and "Mr Clemens," both of which appear in lists of the top 25 2grams that appear in each month, Twain may even surpass the "United States" in terms of coverage. Even without these corrections Twain's outsized popularity is clear: "Mark Twain" has no rival among the other phrases unearthed by the search for 2grams, a list that includes: simple phrases ("years ago," "young man," "short time"); public figures that remain recognizable today ("Standard Oil," "President Roosevelt," "White House," "supreme court"); and figures whose appeal may be more limited, such as "Mr Ir" and "Mrs Irs," subjects of a multimillion dollar divorce proceeding that was popular enough in the press to appear at the tenth and twenty-first most frequent 2grams in the February 1906 press coverage.

[9] See Appendix, Figure 5 and 6. Figure 5 shows the top of the term frequency list and Figure 6 shows the bottom. For ngram data for the first three months of 1906, see NGRAMS in Supporting Files for Chapter 3.
[10] See Appendix, Figure 7 for common 2-grams.

Moving the search for term frequencies out to 3grams and 4grams draws attention to a host of additional topics, including newspaper coverage in January of “Cook St Bank,” “Oklahoma Indian Territory,” a “Philippine tariff bill,” the “Womans Kansas Day club,” “Miss Alice Roosevelt,” the trial of “Albert Patrick York,” Mark Twain’s involvement with Robert Ogden, and Grover Cleveland’s connection to John Carlisle; coverage of the land claims of Manuel Otero are highlighted in February; and a “railroad rate bill,” the “Congo Free State,” and the “United Mine Workers (of) America” are among the stories in March.[11] Each of these topics functions as a feature on the media landscape Twain transverses as part of the daily process of working on his autobiography. Identifying these features even as briefly as I have done here provides an opportunity to investigate how closely Twain weaves contemporary media coverage into the fabric of his autobiography.

Tf-idf weights provide an additional perspective on the popularity of topics in the contemporary media of Twain’s day. A list of 4grams ranked by tf-idf weight from greatest to least reinforces the characterization of Grover Cleveland and John Carlisle; and Mark Twain and Robert Ogden as popular pairs in the press during January 1906. Tf-idf scores also draw attention to new topics that received a lot of press, such as the trial of Albert Patrick and the Woman’s Kansas Day Club in January; the poem “Good-By, Sweet Day” by Celia Thaxter (set to music by Kate Vannah) in February; and activity involving the “United States Supreme Court” in

[11] See Appendix, Figure 8 for common 3-grams and Figure 9 for common 4-grams.

March.[12] 3grams continue to register the presence of "standard oil company" and "kansas day club" in January; Thaxter's poem and Manuel Otero place near the top of the tf-idf list for February; and "railroad rate bill" and "standard oil company" rank highly in March. 2grams weighted by tf-idf consistently place "united states," "mark twain," and "mr mrs" at the top of each month with even less variation in the rankings that is shown by simple term frequency counts.

Topic models of the 1906 corpus offer a description of how popular terms identified using tf-idf are deployed throughout the corpus. For example, constructing 30 topics out of the corpus of coverage assembled for January 1906 (roughly one topic for each day of the month) shows references to "Ogden" as concentrated in two locations: Topic 11 and Topic 24. These topics, in turn, are connected with three days of media coverage: January 8, 9, and 23. Searching this coverage for "Ogden" reveals that he appears in connection with Twain because the two are listed among the featured speakers at a Carnegie Hall fundraising event for the Tuskegee Institute, which was held January 22 and announced in multiple papers on January 8th and 9th.[13] In addition to serving as a tool for drawing our attention to particular aspects of a corpus, topic modeling may also be used to develop an understanding of how press coverage may be likely to divide under a reader's gaze. This second use of topic modeling highlights the ways in which topic modeling may be used not only as a tool

[12] See Appendix, Figure 10 for 4-grams from January; Figure 11 for 4-grams from February; and Figure 12 for 4-grams from March 1906 ranked by tf-idf weight. For Excel files of ngrams with tf-idf weights, see TFIDF in Supporting Files for Chapter 3.

[13] For the 30 topic model that produced these observations, see Twain_1906_Jan_30T_A1_67_B_1 in Supporting Files for Chapter 3.

for document discovery and data mining, but also as a reading strategy and a tool for knowledge creation.

From Tf-Idf to Topic Models and Beyond: A Computation-Based Approach to Reading

To understand the ways in which topic modelling may be used to model the attention of a reader, it may be worth dwelling for a moment on how topic modeling works. Topic models are, in their most basic form, collections of probabilities: word A appears frequently with word B, so those two words are likely to be referring to the same thing. The number of things—aka topics—a topic model seeks in a collection of texts is predefined at the outset. This initial condition determines not only the construction of the topics and the usefulness of topic modeling as a whole, it also functions as a mathematical equivalent of the level of attentiveness a hypothetical reader may bring to a text or body of texts. An attentive reader, one who pays painstaking attention to each and every word that falls under her or his gaze, is likely to identify more topics/discourses/points of conversation at play in any particular work or collection of works. Such a reader may be represented mathematically as a topic modeling algorithm with the capacity to recognize a large number of topics. In contrast, the hurried, distracted, or disinterested reader we may liken to a topic model that seeks a smaller number of topics in the corpus. The makeup of the topics themselves is, for my purposes here, less important that the proportions with which they constitute the corpus.

Mapping the 1906 corpus over a variety of levels of attention shows a tradeoff between the number of topics found and the efficiency with which the search is conducted: as the number of topics sought in the corpus increases, the average number of topics associated with each day of coverage increases from an average of 2.5 topics per day when 30 topics are sought; up to an average of 8 topics per day when 900 topics are sought. At the same time, increasing the number of topics sought in the corpus decreases the likelihood that a topic will be found during the search. Seeking 30 topics in the model produces a 100 percent success rate in the corpus, while searching for 900 topics leads to a success rate of 10 percent. This data suggests that readers who read the press expecting to encounter a large number of topics will be not be rewarded for their efforts, while those readers who invest less effort in engaging with the media are likely to be satisfied by the encounter when the amount of information found in the press is weighed against the amount of effort invested in seeking information. In other words, the highly attentive reader who picks up the newspaper and reads motivated by the desire to identify 900 distinct topics over the course of a month (roughly 30 topics a day); and assumed to have the mental capability to track these topics accurately over the course of a month; is likely to identify only 91 distinct topics covered in the press for January 1906. A less motivated reader, one who seeks only one or two topics a day in the news, is likely to find that information a very high proportion of the time.[14] Using this information as a guide we can reconstruct the reading experience of a reader with particular characteristics and evaluate how satisfied and/or frustrated that reader is likely to be as she or he navigates the presentation of information in the press.

In addition to describing how a reader with a particular level of attention may be likely to interact with the press, the same measure of attention may be used to model a reader's interactions with another corpus of texts. Carrying these levels of attention and reading satisfaction metrics over to other texts allows us to investigate how audiences were likely to experience additional texts. For example, assuming that a reader is equally interested in reading the newspaper and Mark Twain's autobiographical writings, we see that a reader interested in the twelve January dictations in the autobiography and reading with enough intensity to track 30 topics would find 19 topics in the text, a success rate of 63.3%. The twelve January dictations would appear to this reader to contain one topic 8 times; 3 topics 3 times; and 5 topics one time, giving each dictation the appearance of containing an average of 1.8 topics. Of the 30 topics sought, no topic would be found 11 times. Of the 19 topics found: a topic would appear a single time in the corpus 16 times; twice three times; and would never appear three or more times. This hypothetical reader, in other words, would experience Twain addressing most topics of interest once or not at all. A small percentage of the time Twain would address topics twice; but no topic or set of topics would appear to be a focus of Twain's attention in the month of January.

[14] For data comparing topic models of 1906 media coverage, see ApproachesToTopicModeling1906.xlsx in Supporting Files for Chapter 3.

Twain’s focus in the dictations would appear to be limited to a single topic slightly more than half of the time; divided among three topics slightly less than half of the time; and split between five topics once. This analysis presents Twain’s daily dictations as just that: daily, touching on a limited number of topics in each dictation. The exception to this observation is the day following the Carnegie Hall event at which Twain and Ogden were featured speakers. In this dictation Twain addresses five topics. The data described above help us to see the extent to which this exceptional dictation may be tied to an exceptional even in Twain’s life at the time he was working on the text. Noting this connection between Twain’s daily life and the life he recounts in his dictations suggests the autobiography may be closely connected with the times in which he is producing the text, rather than the times he may be narrating. As such, the text may perhaps most properly be taken as a statement about Twain’s final years, rather than as a retrospective of his career or his entire life. Approaching the text in this way reveals it to be no more or less bound by time than many of the other texts where his personal voice is particularly strong: Innocents Abroad, for example, or Following the Equator. In each of these texts a particular journey is the impetus for the work. In the case of the Autobiography the journey is through time, rather than space: the reader is treated to Twain late in life engaged in the composition of yet another travelogue, this time without physically needing to go anywhere.

The extent to which the Autobiography is a text about Twain’s relationship to his contemporary moment, rather than his past, is underscored by comparison of the terms that frequently appear in his dictations and those that appear in the media at the same time. Lists of 4 grams ranked by tf-idf weight constructed from Twain's autobiographical dictations from January 1906 and press coverage from the same time period of time show that Twain's Carnegie Hall appearance with Ogden at an event to raise funds for Booker T. Washington and the Tuskegee Institute features prominently in both collections of text. In the dictations this event ranks in alongside the speech Twain delivered at the expense of Emerson, Longfellow, and Homes as part of a birthday dinner in honor of Atlantic publisher Whittier's seventieth birthday; and the death of John Malone, historian of the Players Club, a New York City social club where Twain was a member, in terms of prominence. In the press coverage Twain's Carnegie appearance is linked not with Malone, but with Albert Patrick, a New York lawyer accused of murder that Twain and a number of other people of high social standing sought to have acquitted. Juxtaposing press coverage with the dictations in this way shows not only how Twain's high social standing is a feature of both his public and his personal life, but also how Twain sought to memorialize that standing in the dictations: emphasizing his involvement at Tuskegee and downplaying his involvement in the trial even though both events were unfolding in a very similar timeframe and attracted similar levels of attention from the press of his day. Tf-idf weights draw attention to Twain's willingness to incorporate popular topics of conversation among his contemporaries into his autobiography when those topics may enhance his reputation and to draw the attention of later generations away from topics of conversation that may be less likely to enhance his reputation. Exposing where Twain’s topics of interest mirror and diverge from those that dominate the press allows us to see where his dictations provide a personal account of the day and where they offer up the spirit of the times. These observations are important because they allow us to see where his autobiographical project becomes a record of what interested him and where it records what interested the public. A divergence between the two spheres of attention is perhaps not unexpected, but it raises questions about Twain’s purported plan to put down events that will continue to fire the imaginations of readers in later generations and reveals the outsized nature of his hubris at this point in time: he is essentially asserting that later generations will be interested in exactly those topics he was interested in, rather than those topics that were of interest to the public at large in his day.

Testing Twain: An Agent-Based Model of the Publication of Mark Twain’s Autobiography

Agent-based modeling provides an environment where we may evaluate the potential impact of Twain’s observations by tracking how topics of conversation perform when the topics Twain addresses in his autobiography are introduced. Modeling the presence and absence of a familiarity with Twain’s autobiography allows us to investigate one of the central features of the Autobiography: the decree that the text be withheld from publication during Twain’s lifetime for the benefit of the public good. Twain repeatedly proclaims his dictations too damaging to be published during his lifetime and for a number of years after. The claim is so prominent that it is used as one of the marketing devices during the release of the UC Press edition of Twain's Autobiography in 2010. Modeling the release of this information into a simulated environment allows us not only to see how Twain's text might spread through that environment, it allows us to investigate the kind of environment that would be required for it to spread as Twain predicted. Simulating an environment where Twain's dictations do in fact shape public discussion on a large scale provides a window into how Twain himself may have seen his contemporaries and his place among them.

Mark Twain worked on an autobiography off and on for a substantial portion of his career. Twain scholars have identified pieces of writing intended for the text beginning as early as 1870 ("[The Tennessee Land]"). All of these early drafts were set aside for one reason or another, but late in life the autobiography received a great deal of Twain's attention. "You will never know how much enjoyment you have lost until you get to dictating your autobiography; then you will realize, with a pang, that you might have been doing it all your life if you had only had the luck to think of it," he wrote to his friend and editor William Dean Howells on January 16, 1906 ("Introduction: Paragraph 77").

Twain's interest in dictation came after years of failing to write an autobiography that met his expectations; and, even dictation was not immediately compelling. He experimented with the approach in small doses for several years before launching into an extensive effort to dictate an autobiography. For a period of three years beginning in January 1906 he met nearly every day with stenographer Josephine Hobby and his official biographer, Albert Bigelow Paine, to reflect on his life and current events. By the time these meetings came to an end more than two hundred and fifty dictations had been produced and the autobiography ran in excess of five hundred thousand words ("Introduction: Paragraph 2"). In its printed form, *Autobiography of Mark Twain: The Complete and Authoritative Edition*, published by UC Press in 2010, contains more than two thousand pages spread over three volumes and consists largely of these dictated texts.

The appearance of the UC Press edition one hundred years after Twain's death in 1910 capitalizes on a convenient marketing angle provided by none other than Twain himself. Twenty-five excerpts from the autobiography were published by Twain in the *North American Review* during his lifetime; each contains an announcement that the work would not be published "in book form" until after his death ("Introduction: Paragraph 6"). What may have been a marketing ploy was also framed as a public benefit. Delaying publication, Twain reasoned, would allow him to produce a more honest account of himself. In 1899 Twain mused in an interview that "a book that is not to be published for a century gives the writer a freedom which he could secure in no other way. In these conditions you can draw a man without prejudice exactly as you knew him and yet have no fear of hurting his feelings or those of his sons or grandsons" ("Introduction: Paragraph 4"). Six years later Twain takes this sentiment to heart, suppressing his dictated autobiography to protect his reputation and benefit future generations:

> “I’d like to see a lot of this stuff in print before I die—but not the bulk of it, on no! I am not desiring to be crucified yet. Howells thinks the Auto will outlive the Innocents Abroad a thousand years, & I know it will. I would like the literary world to see (as Howells says) that the form of this book is one of the most memorable literary interventions of the ages. And so it is. It ranks with the steam engine, the printing press & the electric telegraph. I’m the only person who has ever found out the right way to build an autobiography.” (“Introduction: Paragraph 128”)

Whatever we make of the mix of public and private factors driving Twain’s decision, only the *North American Review* excerpts he approved were released while he was alive. Following his death a broader selection of excerpts appeared, but a complete version of the text was not attempted for a century. The text was known to scholars and circulated in different forms, but the unexpurgated autobiography remained largely hidden from public view. Here I treat Twain’s Autobiography as a forum for testing the effects of withholding information from the public and a window into Twain’s sense of his place in the world. The model I have constructed addresses both of these topics without prioritizing one over the other.

Description of the Model

The model world is established in NetLogo by asking 1089 patches arranged in a square grid to sprout 1 turtle, colored black.[15] The turtles on the grid attempt to communicate with each other once each tick and the results of their interactions are tracked using the variables below:

```
turtles-own [
```

[15] For the NetLogo code for the model, see NETLOGOJAN1906 in Supporting Files for Chapter 3.

```
  topic-of-conversation    ;; the topic of conversation
  speaking   ;; how often a turtle starts a conversation
  listening  ;; how often a turtle listens
  conversationfailed ;; how often a turtle fails to start a conversation
  prior-topic ;; prior topic of conversation
  repeatconvo ;; indicates if turtle is making a second attempt at a conversation
with a topic
]
```

Aggregating the turtle-specific information collected above provides an overview of the conversation dynamics unfolding in the environment.

Once the simulation starts every turtle picks a topic of conversation from a list of available topics. Choices are made in proportion to a probability distribution produced by topic modeling a corpus of newspaper coverage from January 1906. I obtained this distribution by topic modeling a selection of historic newspapers obtained from the Chronicling America Project at the Library of Congress using MALLET via the Software Environment for the Advancement of Scholarly Research (SEASR) to identify 30 topics and their distributions over the month of January.[16] The SEASR flow I designed finds a minimum of 1 and a maximum of 5 active topics in each day of coverage, with an average of 2.48 and a standard deviation of 1.39 active topics per day. The distribution of topics in the corpus is skewed to the right: days with few topics are more common and days with multiple active topics are less common in the data.[17]

Using this approach I obtained a representation of topics present in the news for each day of the month. For example, January 4 and January 6 are each described

[16] For Chronicling America, see “Home”; for SEASR, see “Software Environment”; and, for MALLET, see McCallum.

[17] See Appendix, Figure 13 for a visual summary of the distribution of topics. For the 30 topic model, see 30TOPICSJANUARY1906 in Supporting Files for Chapter 3.

by five active topics; while coverage for January 3 divides across two topics; and, January 2 is associated with a single topic. Looking more closely at an individual day in detail reveals, for instance, that January 1 divides into three topics: Topic1 describes 21.5% of the coverage; Topic3 is associated with 78.4% of the coverage; and Topic8 accounts for less than 1%. January 2 is described by Topic10; and January 3 coverage is 88% Topic21 and 12% Topic13.

Turtles in my model select topics of conversation in accordance with my topic modeling data. The first cycle of the model, for example, sees 21% of turtles take up an interest in Topic1; 78% of turtles are ready to discuss Topic3; and 1% of turtles in the environment are ready to converse with their fellow turtles about Topic8. Turtles also adopt a color associated with their particular topic of interest so that the distribution of topics through the space may be noted by looking at an overview of the model world.

New lists of available topics are introduced into the environment every tick. The rate at which turtles take up these new topics is controlled by the percent-active-readers slider, which determines the number of turtles in the environment that draw their topic of conversation directly from the list of active conversations available at every tick of the model. These turtles may be thought of as people that take an active interest in and keep up with news coverage every day. At 100%, every turtle in the model draws its topic of conversation from the current news coverage made available with each tick of the model. At 0%, turtles take up current news topics only when they are unsuccessful in an attempt to discuss their current topic of conversation.

Results

1089 turtles attempting to converse over 31 ticks of the model produces a total of 33,759 opportunities for discussion. When no active readers are present in the environment, the total number of discussions that take place over the life of the model regularly approaches 18,000 (a success rate near 50%).[18] Tracking the number of conversations initiated by each agent during this run shows 54.62% of conversation attempts were successful during this run ("pctyes" in Appendix, Figure 17). Looking at the performance of individual turtles we see that by the end of the simulation turtles will have between 4 and 30 conversations, with an average of 16.93 (SD of 5.422) and a mode of 18.[19] Increasing the percent-active-readers slider slightly (which increases the percentage of turtles in the environment that keep up with current events) depresses the level of conversation between turtles observed. A 5% increase in the number of turtles that keep up with current events drives the total number of discussions that occur over the life of the simulation down to 13,198 (a success rate of 39.09%).[20] However, a large increase in the percent-active-readers slider raises the level of conversation observed between turtles. A 90% increase in the number of

[18] See Appendix, Figure 14 for a screenshot showing the distribution of successful and failed conversations over the month of January for Random Seed 44 when no turtles take an active interest in discussing current topics from the news.
[19] See Appendix, Figure 15.
[20] See Appendix, Figure 16.

turtles that keep up with current events raises the total number of discussions that occur in over the life of the simulation to 20,388 (a success rate of 60.39%).[21]

Graphing the effects of the change in the percentage of active readers over the course of the simulation shows three different patterns of activity. Conversation levels decrease over the course of the month when turtles don't take an active interest in current events, but more often than not turtles succeed in having a conversation with a neighbor.[22] When the number of turtles that take an interest in current events increases slightly, conversation gives way to an inability to communicate.[23] When the number of turtles that take an interest in current events is substantial, conversation levels dip and then rebound over the course of the month.[24] The point at which conversation levels successfully rebound first occurs when approximately 78% of turtles in the environment take an interest in current events.[25] Above this point conversation between turtles eventually outpaces an inability to communicate; below this point the opposite tends to hold true.

Leaving in place the assumption that 78% of turtles in the simulation have an interest in keeping up with current events, I will now turn my attention to seeing how the introduction of an appetite for Twain's text influences conversation patterns over the course of the simulation.

The percent-twain-readers slider controls the percentage of turtles in the environment that will take up an interest in Twain's text. Unlike the news topics,

[21] See Appendix, Figure 17.
[22] See Appendix, Figure 18.
[23] See Appendix, Figure 19.
[24] See Appendix, Figure 20.
[25] See Appendix, Figure 21.

which potentially change every day, Twain's dictations in January are intermittent. He produced twelve dictations in January: Jan 9-13; Jan 15-19; and Jan 23-24. In the current model I introduce these topics into the environment on those days. Like topics drawn from the news, Twain's topics are immediately available to those that are interested in them and are updated every tick. One way to compensate for the intermittent nature of Twain's texts may be to give them a longer period of activity than topics found in the news, but in the current model both sets of texts are equally transient.

The model shows the number of successful conversations begins decreasing once even a few turtles become interested in discussing Twain. For example, a similar decrease is evident whether 10% or 78% of turtles in the model prefer to discuss Twain instead of current events.[26] Equilibrium returns when around 80% of agents in the model take an interest in discussing Twain.[27] The number of active conversations in the model once again outpaces the number failed attempts at conversation by the time 85% of turtles in the environment prefer Twain over current events.[28]

Conclusions

Benedict Anderson assigns the press a powerful role in fostering a sense of national identity through the creation of "imagined communities" of readers. My

[26] See Appendix, Figure 22 for a record of conversation dynamics produced when 10% of agents have an interest in Twain's Autobiography and Figure 23 for results when 78% of agents have an interest in Twain.
[27] See Appendix, Figure 24
[28] See Appendix, Figure 25

results suggest the limits that must be surpassed before these communities can take hold. When less than 78% of turtles in my model take an interest in the news, conversation between agents appears to be stifled more often than not. This result suggests that the nation building effects Anderson attributes to the mass media must truly be operating at scale before they can be said to be a unifying force in the world.

Twain, for his part, may have understood the power of mass communication and his place within that industry very clearly. Withholding the Autobiography enhances the likelihood that turtles in my model will converse until at least 80% of readers can be assumed to take an interest in the work. Above this level publication of the text also enhances levels of conversation in the model. Even for an influential writer and international celebrity, capturing the attention of 80% of the public seems like an unreasonably high bar to cross. Suppressing the text, in other words, could be said to benefit society. On the other hand, Twain's goals for the Autobiography were not modest: "I intend that this autobiography shall become a model for all future autobiographies when it is published, after my death, and I also intend that it shall be read and admired a good many centuries because of its form and method—a form and method whereby the past and the present are constantly brought face to face, resulting in contrasts which newly fire up the interest all along like contact of flint with steel" ("26 March 1906: Paragraph 26"). A writer seeking to influence the shape of "all future autobiographies" would surely not see capturing the attention of 80% of readers as an unattainable goal. Given a sizable audience, publication of the text could also be said to benefit society.

What may first appear to be an improbable goal begins to seem much less ostentatious if we consider conversation dynamics from the perspective of what we may choose to call “embodied data analysis.” Phrased in the language of NetLogo, embodied data analysis supplements the insights offered by the Observer with the individual perspectives offered by agents as they navigate a simulation. Surveying activity in the model from the perspective of individual agents shows that when turtles are given the ability to be aware of what turtles beyond their immediate neighbors are reading, 80% of turtles in the environment can be exposed to Twain’s work when as few as 30% of turtles in the environment actively read Twain’s writing.[29] As the number of connections between turtles and their neighbors grow the percentage of turtles Twain’s work needs to reach to produce 80% awareness continues to fall. As few as 20% of turtles in the environment reading Twain can reach the 80% mark when turtles are given the ability to keep up with the reading habits of all turtles within a 4 unit radius (roughly 48 turtles or 5% of turtles in the environment).[30]

While these individual data points may be of little value in themselves, together they suggest that as connections between turtles in the environment grow, texts need to reach fewer individuals in the environment to reach comparable levels of popularity. In other words, the more connected turtles are to their contemporaries, the fewer turtles need to be exposed to information in order to give the impression that information is circulating in large numbers. Applying this observation to Twain

[29] See Appendix, Figure 26
[30] See Appendix, Figure 27

and his audience suggests that one factor explaining Twain’s confidence in the popularity of his Autobiography could be the strength of the connection he feels to his audience. Twain’s typical rate of letter production has been noted at between two and ten letters a day; and, Robert Hirst notes, "anybody who wrote him tended to get a reply. He easily wrote 50,000 letters" (Griffith). Another factor driving his confidence in the potential reach of the Autobiography could simply be the size of his social group, irrespective of whether or not he feels connected to it. Modeling draws attention to the ways in which these and other contextual factors may have played a part in influencing Twain’s perception of the reception his text was likely to receive.

Perhaps more valuable than the insights the model offers about Twain are the insights the model offers about the world through which his text moves. Instead of focusing on Twain and his perceptions, my model can also be said to show that high levels of interest in a topic are more easily obtained as audience members in the environment become more connected. Even a small audience, if sufficiently connected, generates the impression that large numbers of turtles maintain an active interest in keeping up with Twain’s work. This observation points to a second reading of Twain’s enthusiasm for his text that has less to do with how he may see the work and more with his perception of the world through which his texts move. Read as a comment on the strength of the social bonds between his contemporaries, Twain’s enthusiasm for the Autobiography signals a belief in a vibrant, social, engaged society that stands in contrast to the bleak worldview more commonly associated with his later years.[31]

Modeling and Literary Studies: A Brief Coda

David Damrosch speaks of the ability to "triangulate" the meaning of a word in an unfamiliar language by compiling a host of translations of the word in question in familiar languages in order to establish the "semantic field" or likely range of possible meanings the word in question is likely to hold (How 71). Agent-based models allow us to conduct similar triangulations by approaching the past from a variety of perspectives and exploring how our understanding of prior events is influenced by the lenses through which we view them. Agent-based models are particularly well suited to the representation of history from multiple perspectives because they are easily reproduced. This is not to overlook the deeply encoded structures that Matthew Kirschenbaum deftly brings to our attention while advocating for the study of digital artifacts on a forensic level (Mechanisms), but simply to argue that on a much more shallow level, high above the physical manifestation of magnetically encoded bits and bytes positioned at very real and definable physical locations on the surface of your storage medium of choice, two (or more) copies of a digital object exhibit a high degree of similarity.

[31] See, for example, the biographical portrait provided by the Mark Twain House & Museum ("Life Lived").

Having multiple copies of the same argument has several potential advantages. Among them: it facilitates the sharing, reproduction, and interaction with arguments in whole and in part. Quotation and paraphrase are two traditional tools by which we may carry these practices out within narrative form. Agent-based models provide an opportunity to build upon these practices in the digital world by borrowing, modifying, and repurposing code. Just as traditional practices of citation situate a text within the contexts of an ongoing conversation, the practices of what Lawrence Lessig calls "remix culture" (Remix) make it possible to situate one model within a wide range of modeling practices. One potential benefit to carrying out this conversation *in silico* (as some ABMers are fond of saying) is the scope and speed with which the conversation may be carried out. Publishing printed texts is a slow method of distributing information that is constrained by material and legal factors, the cost of materials and copyright among them. These factors make it impractical, and often illegal, to reproduce a source text in full within the context of a new argument. To be sure, source code may be susceptible to these same issues. However, the open source movement and the rise of version control and sharing services like Git and GitHub make the practice of sharing much more feasible in the digital world. The move from narrative text (which can be easily shared using these same services) to computer model encourages the use of a framework for evaluating arguments (both prosed and programmed) and their alterations in terms of feature enhancement and bug detection; this language emphasizes a connection to what has come before and provides a feeling for what has been done. The bug, the feature, and the fix (or patch) provide a tripartite approach to the characterization of communication whose impact has been confined to the reach of the computing world. Moving this framework into broader circulation provides an organizational structure around which to develop more than code. Like meter, rhythm, and rhyme; melody and harmony; timing, speed, and power, the combination of bug, feature, and fix provides a simple framework with which we may navigate the development of an argument as it moves through space and time; together these three provide the "minimally sufficient conditions" for evaluating both agent-based models and narrative prose. Critical Code Studies has already begun the work of encouraging us to read code as literature (Marino). The use of agent-based models to conduct literary criticism complements this drive by offering a reminder that literature may also be read as code, an approach to that can only help to preserve and strengthen connections between our printed pasts and their digital doppelgangers as we chart a course for literary criticism in the twenty-first century.

Looking Back and Looking Forward: A Concluding Perspective

This dissertation aspires to contribute to an ongoing conversation between proponents of a version of literary criticism infused with technology, but not texts; and equally fervent calls to defend literary studies from encroaching technologies. Signs of middle ground emerging between these two positions are slowly coming into view, most recently in the form of an announcement at MLA 2018 of a planned special issue of PMLA devoted to digital humanities (“Session 347”).[1] In an effort to further collaboration between digital and non-digital approaches to literary studies, the chapters of this dissertation work together to articulate a potential future where printed and digital artifacts are interwoven to enhance our ability to decipher the production and reception histories of works of literature, a task which Jerome McGann likens to scrutinizing “the textual DNA” of literary studies (New Republic 157).

As a first step toward this goal, the dissertation opens with a discussion of several possible avenues for gathering information about literary texts in a digital world. My discussion begins with an overview of an attempt by the Reading Experience Database to crowdsource information about how texts have been read in the past; and concludes with an exploration of ways in which Wikipedia may be mined for information about contemporary attitudes toward literature at the present

[1] The “Big Tent Digital Humanities” theme selected for the 2011 Digital Humanities Conference held at Stanford marks another prominent attempt to inspire collaboration between digital humanities scholars and scholars working in other fields. Patrik Svensson’s 2012 summary of the debate touched off within digital humanities by the “big tent” branding and call to look “Beyond the Big Tent” are reminders that this high-profile event should not be mistaken for a definitive sign of the emergence of an open, inclusive, collaborative environment that draws upon research from many fields.

moment. In the years that have intervened between the drafting of my first chapter and the construction of these concluding remarks the Reading Experience Database has gained new partners like Memories of Fiction, a group of scholars working to gather contemporary oral histories of “individual and collective memories of reading fiction” (“About”), and inspired new projects like the Listening Experience Database, “an open and freely searchable database that brings together a mass of data about people’s experiences of listening to music of all kinds, in any historical period and any culture” by inviting participants to scour written materials in order to document ways in which people have experienced sound (“Listening”). Efforts to influence content on Wikipedia have grown as well. Art+Feminism, for example, launched a campaign in March 2014 to improve “coverage of cis and transgender women, feminism and the arts on Wikipedia” that is now entering its fifth year and has garnered support from an international array of cultural and educational institutions (“Art+Feminism”). The addition of these projects and others like them to the assortment of digital archives discussed in the opening chapter is an indication that interest in building large digital collections of information about interactions between texts and readers remains strong; and that the development of these collections has potential to influence the ways in which scholars working in fields beyond literary studies approach the development of digital archives.

The second chapter of the dissertation demonstrates one possible approach to reading the holdings of a single digital archive using a blend of text analysis, agent-based modeling, and sonification. The chapter opens by exploring a subset of Mark

Twain’s presence in newspapers digitized as part of the Chronicling America project via an array of increasingly complex text analysis techniques—beginning with determining word frequencies, moving to the calculation of term frequency/inverse document frequency scores, and concluding with topic modeling—and closes by exploring how the data produced by these techniques may be employed to create dynamic visual and sonic representations of public discourses surrounding Twain as he set to work on his autobiography. The visual and sonic models with which the chapter concludes offer alternatives to the reliance on static visualizations that characterizes many projects involving the digital analysis of literary texts in the wake of the continuing influence of Franco Moretti’s call to remake literary studies in the image of graphs, maps, and trees.

Beneath the revolutionary fervor, Moretti’s 2003 vision for the future of literary criticism is grounded in an ocularcentric knowledge-making project that has been traced back at least as far as Aristotle (Jütte 61). Where it fails to diverge from this centuries old tendency to connect vision with knowledge, “distant reading” (as the approach to literary scholarship Moretti has helped to popularize is known) foregoes its revolutionary potential and risks becoming a sterile, homogenizing, lifeless form of literary criticism whose primary contribution is to further buttress the authority of an established mode of research in which visualization features prominently and silence abounds.

As Yanni Loukissas and David Mindell note, “creating a legible visualization of data requires leaving out much that could enrich our understanding of an event but that might not graphically fit" (5-6). In the hands of careful scholars like Loukissas and Mindell these sacrifices to visual clarity are identified as points of departure for future scholarly exploration; the failures of the visual image are, in other words, talked about. In less cautious hands, however, silence reigns as data is pruned, cleaned, transformed, and/or otherwise kneaded into a usable form; and is further shaped by the clean lines of a line graph, the crisp edges of a network graph, or the gradient scale of a word cloud (three of the more common visual layouts through which distant readers speak). When deployed without concern for their limitations, these common forms of visualization help to cement into place what Jonathan Westin calls "the concrete lid put on the interpretation of the past through every visual representation" (30). Challenging this "neatness" and the "idealized cultural stereotypes" it supports, Anna Foka and Viktor Arvidsson argue, "is certainly the first step to making Humanities relevant to the actual study of humanity." "As humanists," Elton Barker, Lorna Hardwick, and Mia Ridge contend, "we need technologies that allow open-endedness, that preserve the ambiguity and nuance of our work, rather than render the data flat and colourless" (190). Johanna Drucker and Patrik Svensson encourage humanists to meet this need by embracing technologies that allow information to be layered and studied from multiple perspectives. The flexibility to include, exclude, enhance, and obscure the presence of information, they argue, is an important step in the development of digital environments that "contain tools for thinking in arguments rather than displays of thought whose production processes disappear in the final view" (Drucker and Svensson).

In Michael Gavin's view, agent-based modeling provides scholars with the kind of argument-driven digital environments championed by Drucker and Svensson by digitally twisting the hermeneutic circle into hermeneutic figure eight. Belinda Roman notes that agent-based modeling "introduces some interesting benefits for the study of culture and its emergence and transformation over time; however, much of this exploration is taking place outside the humanities." The second chapter of the dissertation echoes and amplifies their assessment by documenting a progression from archive to analysis to model that may be of interest to future modelers seeking an overview of one possible approach to uniting the historical fragments available to us into portraits of our past. More importantly, however, the chapter makes a contribution to ongoing efforts to draw the attention of humanists to agent-based modeling in the hopes of broadening discussion of the elaborate, often government and industry led proposals underway to shape contemporary life based on output from efforts like Robert Axtell's attempt to create an agent-based model of the entire world.[2]

With the benefit of hindsight, projects like Axtell's may one day be seen as marking a new stage in the development of what Rey Chow has described as the age of the world-target, a turning point after which our world is shaped more by

[2] A 2011 presentation of Axtell's project is available on YouTube (Axtell, "TEDxUVM"). Axtell concludes a summary of more recent progress on this work, presented at the 2017 Conference on Complexity and Policy Studies, by observing "agent-based modeling/computing is an emerging technology to bring together data and behavior at *full-scale* with the social phenomena being studied" ("Computationally Enabled" 55). Although made available online following the conference, Axtell's presentation was taken down some time in February 2018 ("CAPS 2017"). His slides remain accessible as html automatically generated by Google while crawling the web (Axtell, "Computationally Enabled"). Axtell's project is one example of a body of work being undertaking in the belief that "the scientific community is ready to take on the integration of human modeling and earth system modeling" (Allen xi).

algorithmic operations than it is by human activity. The development of agent-based modeling in the sciences has already led Paul Humphreys to ponder a future where scientists play no significant role in scientific research. Alternatives to the spread of the future feared by Humphreys are understandably hard to come by in an environment where computational approaches are unfamiliar, disparaged, and feared. While Tiziana Terranova's observation that "concerned political minds" should be wary of links between capitalism and computers should not be discounted, Alexander Galloway's call for leaving the field runs too great a risk of leaving those who flee with no place to go (Terranova 80; Galloway 138-143). Rather than joining Terranova and Galloway in their vocal opposition to agent-based modeling and other forms of social simulation, humanists may find it more productive to acknowledge simulation as what we might call, under the guidance of Donna Haraway's cyborg-manifest, the latest update to our cyborg-vision. This acknowledgement may not do much to satisfy critics who charge that agent-based simulation cannot show us anything new, but this argument does not diminish the power of agent-based models to show us what we already know (or think we already know) from perspectives that would be otherwise unattainable.

The final chapter of the dissertation seeks to inspire further discussion of the benefits of viewing familiar information from unfamiliar angles by drawing attention to the new perspectives that emerge when the blend of widely available, open-source software detailed in the prior chapter is deployed to read the first volume of Mark Twain's Autobiography. The chapter opens by drawing upon text analysis techniques to identify prominent topics found in two distinct datasets--Twain's dictations and an archive of media coverage from the same period of time culled from Chronicling America—and closes by modeling the circulation of these topics through a simulated environment. Comparing prominent topics in the media with those in Twain's text draws attention to points where Twain's attention overlaps with and diverges from the interests of his contemporaries. For example, my approach reveals that Twain speaks in the Autobiography of his work alongside Joseph H. Choate and other prominent citizens to raise money for Booker T. Washington's Tuskegee Institute; but his involvement around the same time in another effort by Choate, former president Grover Cleveland, and other high profile individuals to challenge the conviction of Albert T. Patrick, a lawyer found guilty in a widely followed trial of conspiring to murder millionaire businessman William Marsh Rice, goes without mention.

The omission of Twain's involvement with the Patrick case from the Autobiography is curious not only because it flies in the face of his longstanding interest—"almost an obsession" in the view of Twain scholar Daniel Morley McKeithan—in murder trials and the criminal justice system; but also because Twain's advocacy for Tuskegee and Patrick were both widely reported on; both took place within the same elite social contexts; and, both occur at the time when Twain began meeting with Paine and Hobby to dictate his autobiography.[3] Acknowledging his interest in both causes paints a picture of Twain as social activist willing to advocate on behalf of economic elites and people in less financially secure positions.

[3] For Twain's interest in the legal system, see McKeithan 6; and, for a book-length account of the Patrick trial, see Friedland.

Omitting his interest in Patrick's case from the Autobiography simplifies the representation of Twain's activism. Whether or not this obfuscation is intentional, a lapse in memory, or a concession to the fact that dictations are governed by the limits of physical and mental resources as much as the written compositions Twain sought to avoid, the omission of Patrick from the Autobiography draws attention the selection process Twain uses to determine which events to include in his text. Twain addresses this process directly in the text, advising his audience(s) that "this autobiography of mine does not select from my life its showy episodes, but deals merely in the common experiences which go to make up the life of the average human" ("26 March 1906: Paragraph 26"). The omission of the Patrick case raises questions about Twain's conception of "common experiences" and "the life of the average human." The text analysis techniques I deploy in this chapter are one approach to determining how Twain construes these terms; and also how his construction overlaps with and departs from the vision of life available in media accounts from the same period of time.

Modeling the movement of these accounts through a simulated environment provides opportunities to move away from static illustrations of how Twain's interests differ from his contemporaries and toward a dynamic environment where the significance of these differences can be explored. The model constructed to conclude this chapter shows that clear divisions between dominant and less dominant public discussions may arise when media coverage from the period during which Twain was working on his autobiography is subjected to topic-modeling and visualization; but these divisions are not located in the past and do not lead to the "more rational literary history" Moretti and his followers seek ("Graphs, Maps, Trees: Abstract Models for Literary History--1" 68); they are artificially produced by a highly constructed perspective that turns a blind eye to the ways in which topics are distributed through their environment. Viewed through the eyes of agents, any one of which could be Twain, dominant conversation patterns become much more difficult to detect than surveys of the same information conducted from a distance may lead us to believe. As a result, modeling the circulation of topics provides a valuable reminder of the fact that the stories, patterns, and trends computational techniques bring to our attention are ours, rather than artifacts from another time. This corrective stands in contrast to the language of mining and information retrieval that has been a prominent feature of discussions about computational approaches to working with text that stretches back at least as far as the construction of Roberto Busa's Index Thomisticus, and has more recent roots in Gregory Crane's effort to spur readers of the now defunct D-Lib Magazine to consider "What Do You Do with a Million Texts?"[4]

Writing in 2006, Crane encouraged librarians, scholars, and other researchers confronting mountains of digital text to work "to extract from the stored record of humanity useful information in an actionable format for any given human being of any culture at any time and in any place." Elsewhere in the same issue Dan Cohen predicts "repetition and cross-referencing should allow us to create tools for mining the vast information and knowledge that lies within the nearly limitless digital

[4] See Busa for an account of creating the Index, which is available online at "Corpus."

collections we are about to encounter." The enthusiasm shared by Crane and Cohen for identifying and extracting information from digital resources is noticeably more subdued when Tanya Clement, John Unsworth, Sara Steger, and Kirsten Uszkalo turn their attention to the topic two years later. While summarizing the work of this group of scholars before an audience at Harvard in 2008, Unsworth argues: "I think it makes more sense to think of text-mining tools as offering provocations, surfacing evidence, suggesting patterns and structures, or adumbrating trends. Whereas text-mining is usually about prediction, accuracy, and ground truth, in literary study, I think it is more about surprise, suggestion, and negative capability." He continues: "the value of these tools, especially with a large full-text collection, is that they can bring to your attention works that otherwise might be overlooked, they can expose patterns that are so fine-grained that they would otherwise escape notice, and they can allow you to not-read a million books on your way to reading a period, or reading a genre, or even reading a book" (Unsworth et al.).

The open-endedness and uncertainty Unsworth embraces also appears in a 2009 assessment of digital work in Classics co-authored by Crane, which concludes "comprehensive collections of industrially scanned written materials provide historic new instruments with which to better understand and to make intellectually accessible the record of human existence. These comprehensive collections of scanned print materials are, however, not an end in themselves but instead provide the foundation on which new collections, integrating images of writing with machine actionable data, will support a new generation of services for a new generation of intellectual projects" (Crane et al.). Crane's new vision of engagement with digital archives as an opportunity to formulate questions, rather than extract answers, becomes a full-fledged methodology in the hands of Stephen Ramsay by 2010.

Advocating for what he terms a "Screwmenutical Imperative," Ramsay advises "there are so many books. There is so little time. Your ethical obligation is neither to read them all nor to pretend that you have read them all, but to understand each path through the vast archive as an important moment in the world's duration—as an invitation to community, relationship, and play" (9). Ramsay's response, reprinted in 2014 as part of an edited collection targeted at historians interested in digital scholarship, is notable for merging the technical concerns that drew the attention of scholars like Crane and Cohen nearly a decade earlier with more recognizably human aspirations.[5]

My dissertation shares Ramsay's desire to mobilize digital technologies in order to strengthen connections between past, present, and future generations. I have taken up this challenge within the context an engagement with the digitally mediated past that aspires to have more in common with his playful, communal vision of digital humanities than it does with attempts to liken digital analysis of written records to acts of mining and extraction. The simulated environments I construct blur boundaries between mining information from textual records and using technology to commune with the past. The blend of agent-based modeling and sonification I propose creates space where multiple individual perspectives on our written records

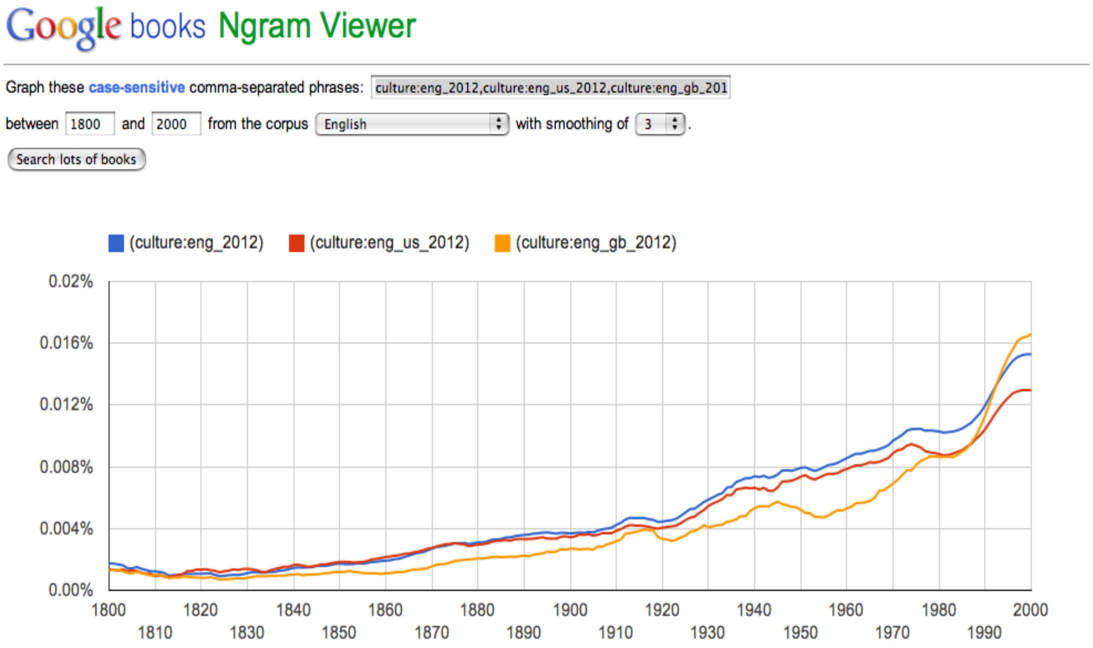
[5] For a brief summary of the response at the 2010 conference where Ramsay offered his advice, see Rockwell; and, see Kee for the 2014 volume in which it later appeared.

are woven together within a single framework; each offering a distinct path through and experience of the information preserved in our historical records. The emphasis I place on experiencing the past from multiple mathematical, visual, and sonic perspectives stands in contrast to a drive that brings scholars as varied as Lev Manovich and the AP U.S. History students who won the K-12 award in the 2015 Chronicling America Data Challenge together on a quest to identify a single, distant overlook from which we can stand back and survey, commandeer, and command the past from a uniform point of view.[6] My dissertation, in contrast, encourages us to use technology to head not to the mountain top, but down into the valleys, out into the plains, and up onto the plateaus; not in an effort to root out the hidden spaces whose names we do not know and mine their secrets, but in order to broaden our appreciation for the existence of other perspectives and other worlds beyond those we already claim for our own.

[6] See "Cultural Analytics" for information about the lab Manovich directs; for information about the challenge see "Chronicling America: Historic American Newspapers Data Challenge"; and for information about the winners see "NEH Announces."

Appendix for Chapter One

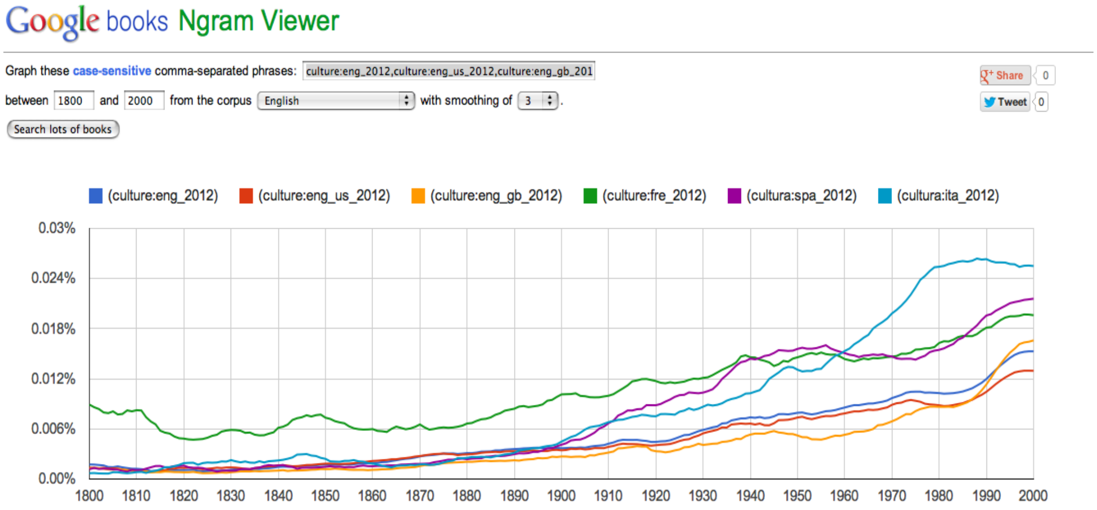
Figure 1 – A comparison of “culture” across the 2012 Google Books corpuses of English language texts; texts published in the United States; and texts published in Great Britain.



Graph these case-sensitive comma-separated phrases: culture:eng_2012,culture:eng_us_2012,culture:eng_gb_201

between 1800 and 2000 from the corpus English with smoothing of 3.

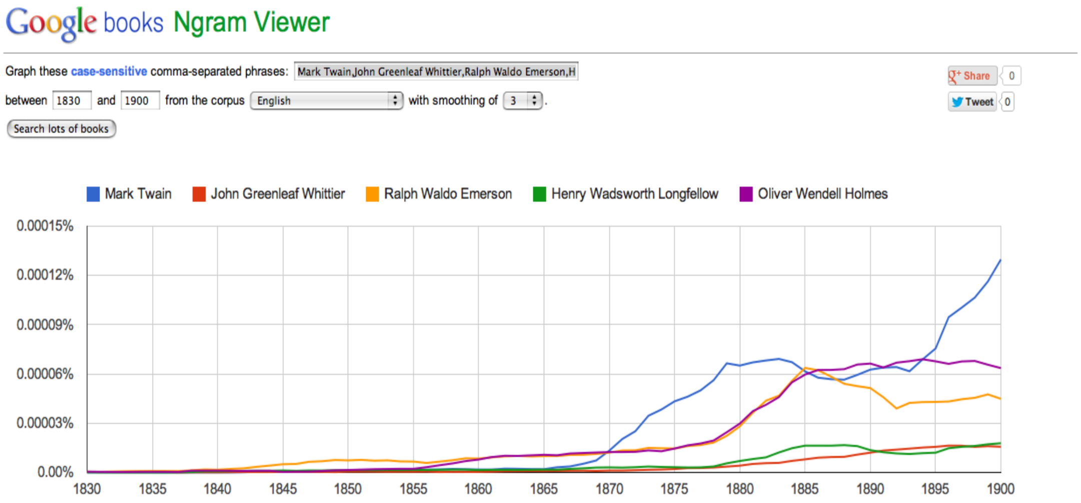
Figure 2 – A comparison of “culture” in the 2012 Google Books corpus of French language texts and “cultura” in the Spanish language and Italian language corpuses.



Graph these case-sensitive comma-separated phrases: culture:eng_2012,culture:eng_us_2012,culture:eng_gb_201

between 1800 and 2000 from the corpus English with smoothing of 3.

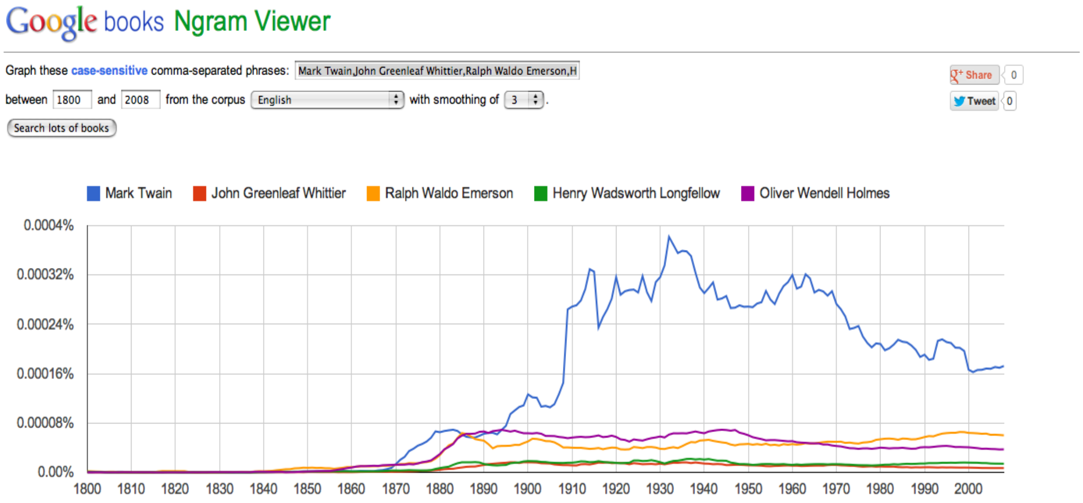
Figure 3 – A comparison of mentions in the English language corpus for some of the attendees at Whittier's Birthday Dinner, held December 17, 1877



Graph these case-sensitive comma-separated phrases: Mark Twain,John Greenleaf Whittier,Ralph Waldo Emerson,H

between 1830 and 1900 from the corpus English with smoothing of 3.

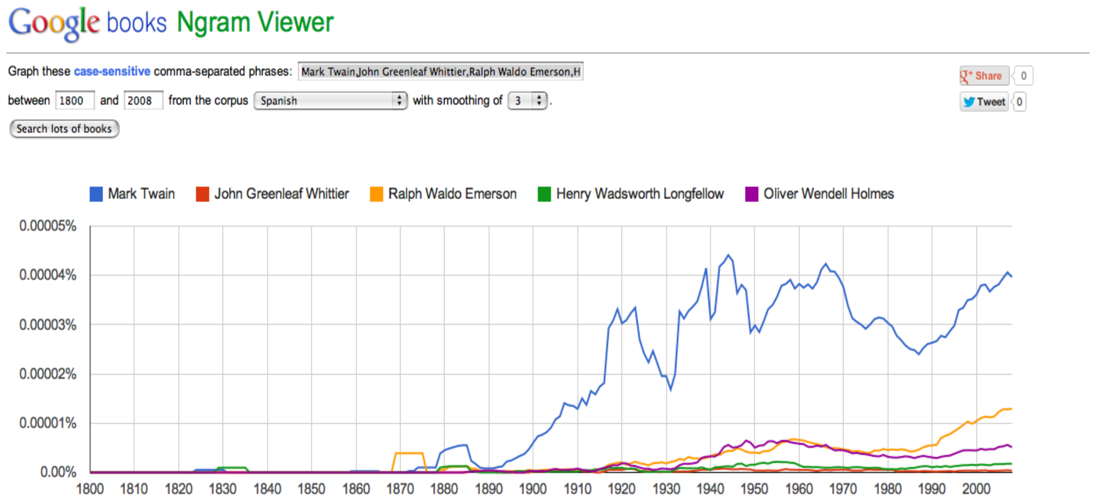
Figure 4 – An extended look at Whittier Birthday Dinner attendees mentioned in the English language corpus



Graph these case-sensitive comma-separated phrases: Mark Twain,John Greenleaf Whittier,Ralph Waldo Emerson,H

between 1800 and 2008 from the corpus English with smoothing of 3.

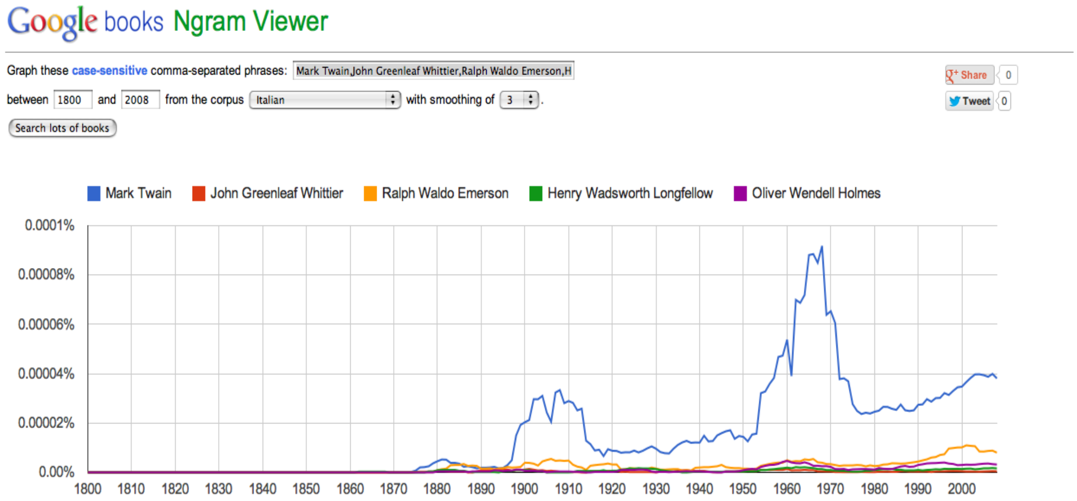
Figure 5 – A comparative look at Whittier Birthday Dinner attendees mentioned in the Spanish language corpus



Graph these case-sensitive comma-separated phrases: Mark Twain,John Greenleaf Whittier,Ralph Waldo Emerson,H

between 1800 and 2008 from the corpus Spanish with smoothing of 3 .

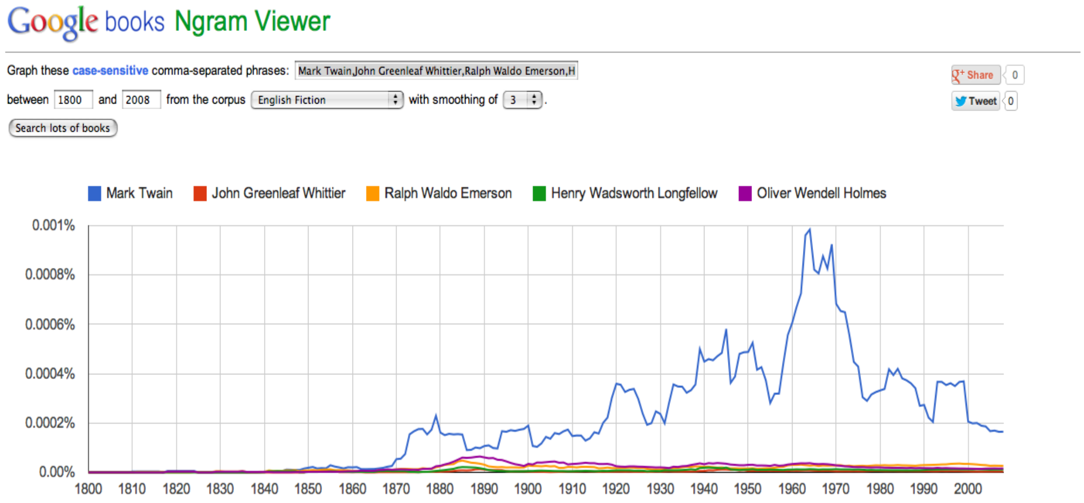
Figure 6 – A comparative look at Whittier Birthday Dinner attendees mentioned in the Italian language corpus



Graph these case-sensitive comma-separated phrases: Mark Twain,John Greenleaf Whittier,Ralph Waldo Emerson,H

between 1800 and 2008 from the corpus Italian with smoothing of 3 .

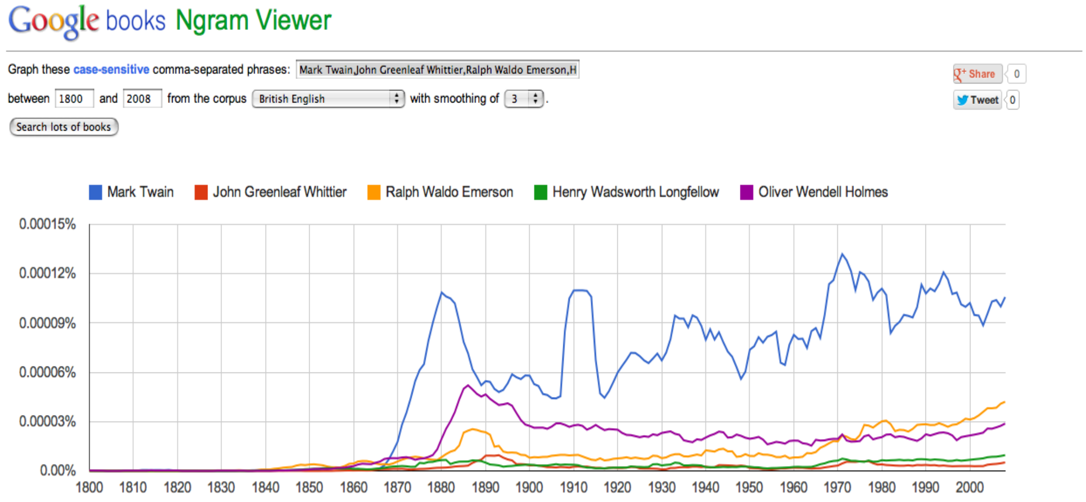
Figure 7 – A comparative look at Whittier Birthday Dinner attendees mentioned in a corpus of works of fiction published in English



Graph these case-sensitive comma-separated phrases: Mark Twain,John Greenleaf Whittier,Ralph Waldo Emerson,H

between 1800 and 2008 from the corpus English Fiction with smoothing of 3 .

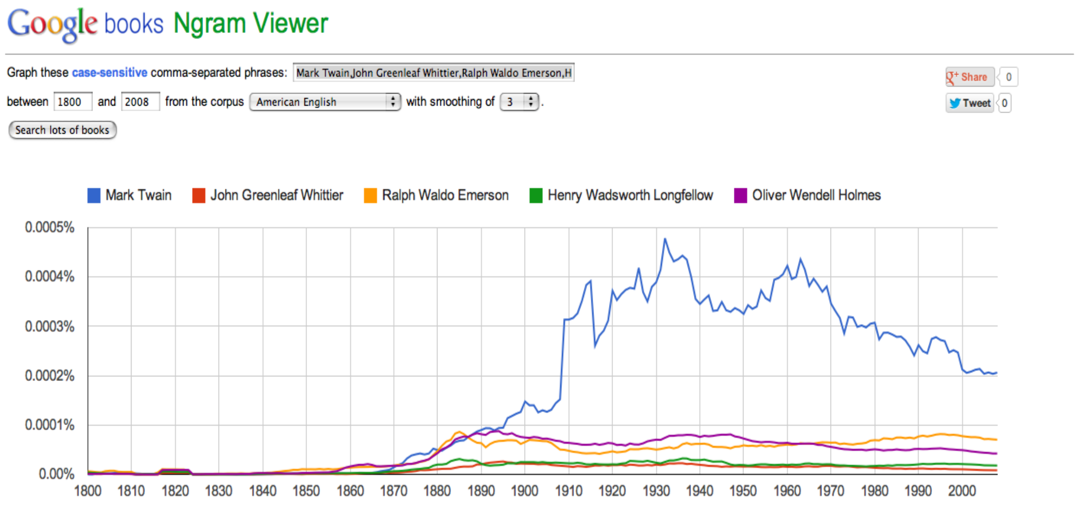
Figure 8 – A comparative look at Whittier Birthday Dinner attendees mentioned in a corpus of books published in England



Graph these case-sensitive comma-separated phrases: Mark Twain,John Greenleaf Whittier,Ralph Waldo Emerson,H

between 1800 and 2008 from the corpus British English with smoothing of 3 .

Figure 9 – A comparative look at Whittier Birthday Dinner attendees mentioned in a corpus of books published in the United States



Google books Ngram Viewer

Graph these case-sensitive comma-separated phrases: Mark Twain,John Greenleaf Whittier,Ralph Waldo Emerson,H

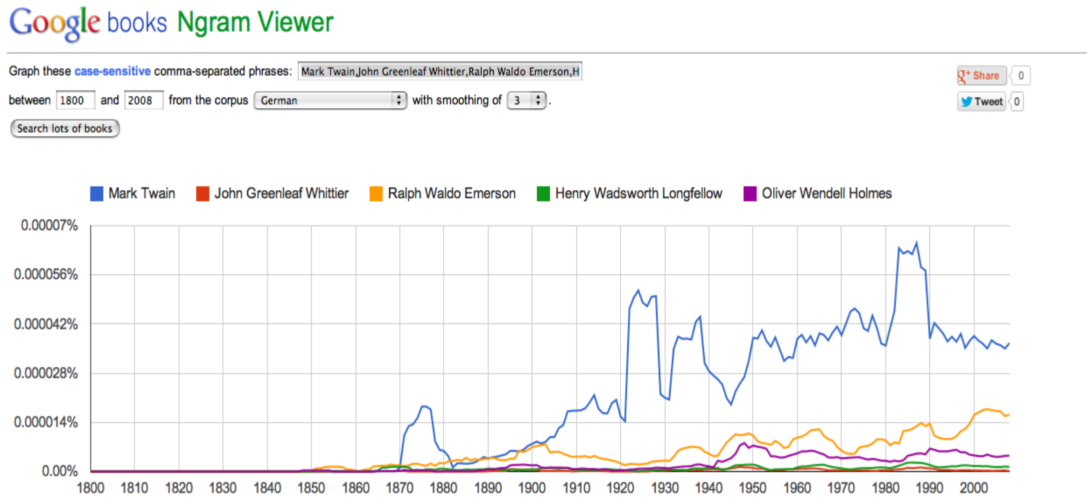
between 1800 and 2008 from the corpus American English with smoothing of 3 .

Search lots of books

Share 0

Tweet 0

Figure 10 – A comparative look at Whittier Birthday Dinner attendees mentioned in the German language corpus



Google books Ngram Viewer

Graph these case-sensitive comma-separated phrases: Mark Twain,John Greenleaf Whittier,Ralph Waldo Emerson,H

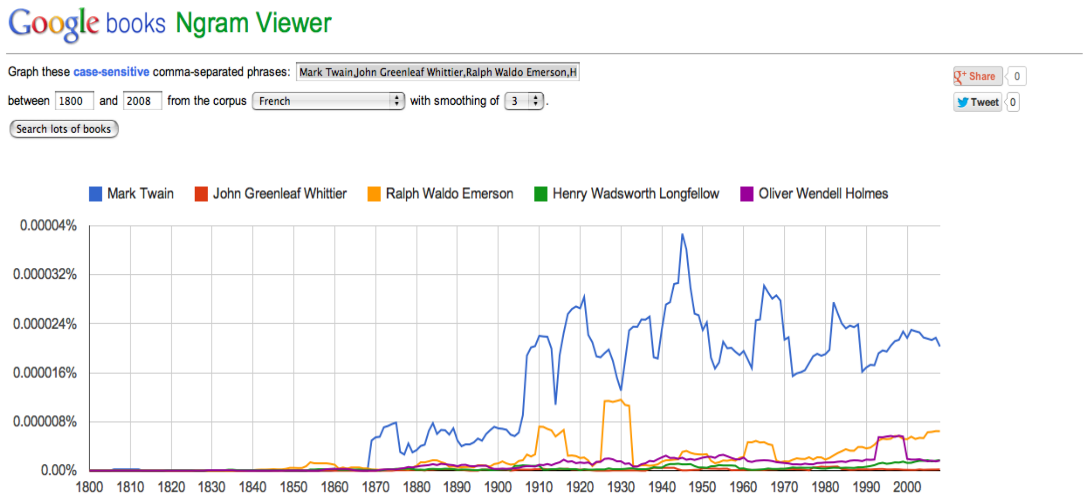
between 1800 and 2008 from the corpus German with smoothing of 3 .

Search lots of books

Share 0

Tweet 0

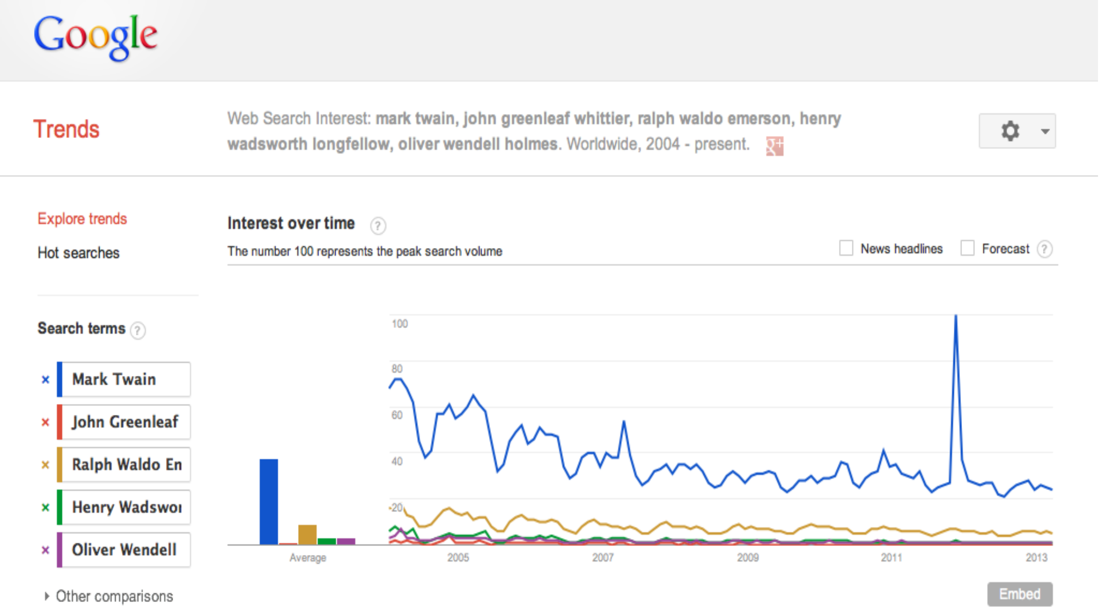
Figure 11 – A comparative look at Whittier Birthday Dinner attendees mentioned in the French language corpus



Graph these case-sensitive comma-separated phrases: Mark Twain,John Greenleaf Whittier,Ralph Waldo Emerson,H

between 1800 and 2008 from the corpus French with smoothing of 3.

Figure 12 – A comparison of Google searches for several Whittier Birthday Dinner attendees



Web Search Interest: **mark twain, john greenleaf whittier, ralph waldo emerson, henry wadsworth longfellow, oliver wendell holmes**. Worldwide, 2004 - present.

Interest over time

The number 100 represents the peak search volume

□ News headlines □ Forecast

Search terms

× Mark Twain

× John Greenleaf

× Ralph Waldo En

× Henry Wadswoı

× Oliver Wendell

Figure 13a – A heat map of searches for “Mark Twain” on Google

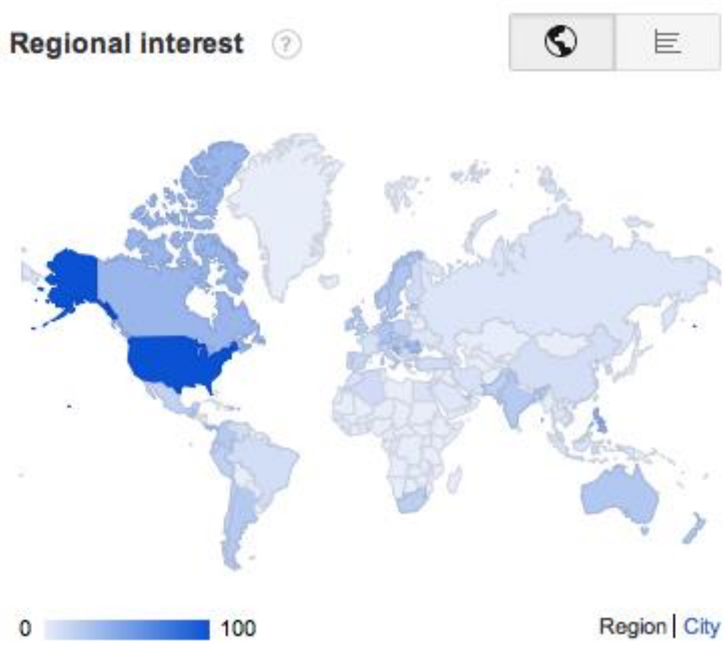
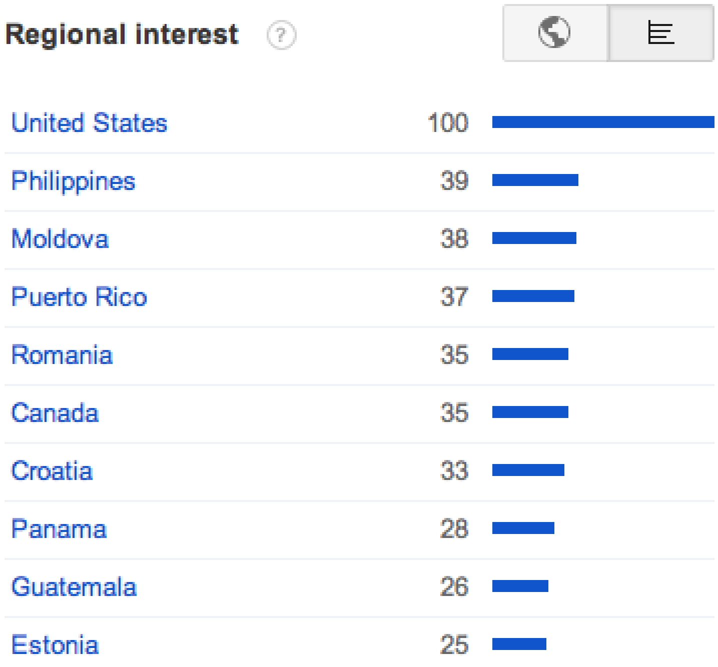


Figure 13b – A partial list of regions searching “Mark Twain” on Google


Regional interest

United States	100	▬
Philippines	39	▬
Moldova	38	▬
Puerto Rico	37	▬
Romania	35	▬
Canada	35	▬
Croatia	33	▬
Panama	28	▬
Guatemala	26	▬
Estonia	25	▬

Figure 14a – A heat map of searches for “Ralph Waldo Emerson” on Google

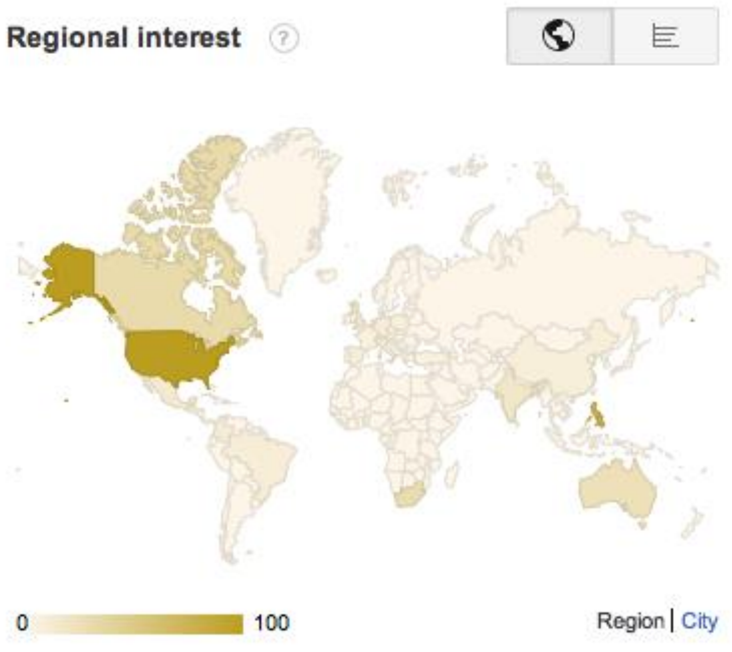
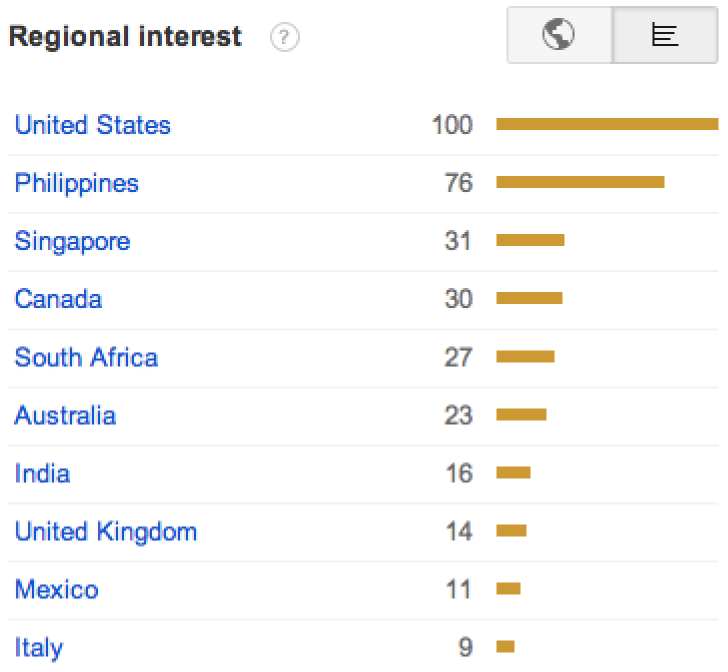


Figure 14b – A partial list of regions searching “Ralph Waldo Emerson” on Google



United States	100	▬
Philippines	76	▬
Singapore	31	▬
Canada	30	▬
South Africa	27	▬
Australia	23	▬
India	16	▬
United Kingdom	14	▬
Mexico	11	▬
Italy	9	▬

Figure 15a – A heat map of searches for “Henry Wadsworth Longfellow” on Google

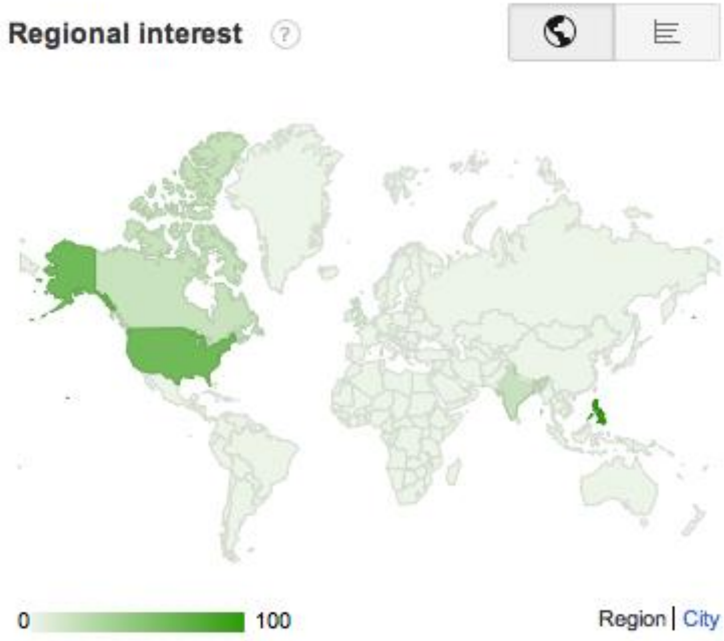
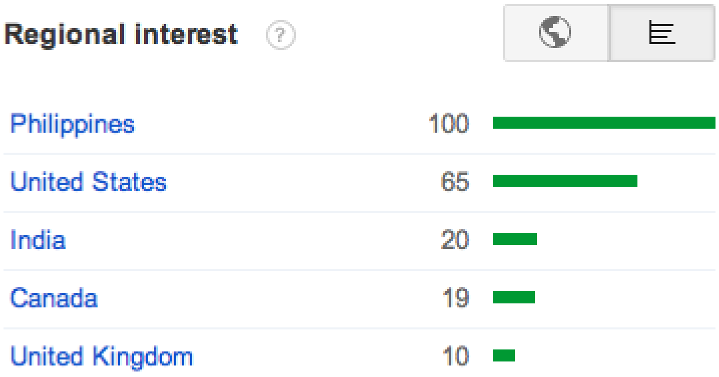


Figure 15b – A partial list of regions searching “Henry Wadsworth Longfellow” on Google



Philippines	100
United States	65
India	20
Canada	19
United Kingdom	10

Figure 16a – A heat map of searches for “Oliver Wendell Holmes” on Google

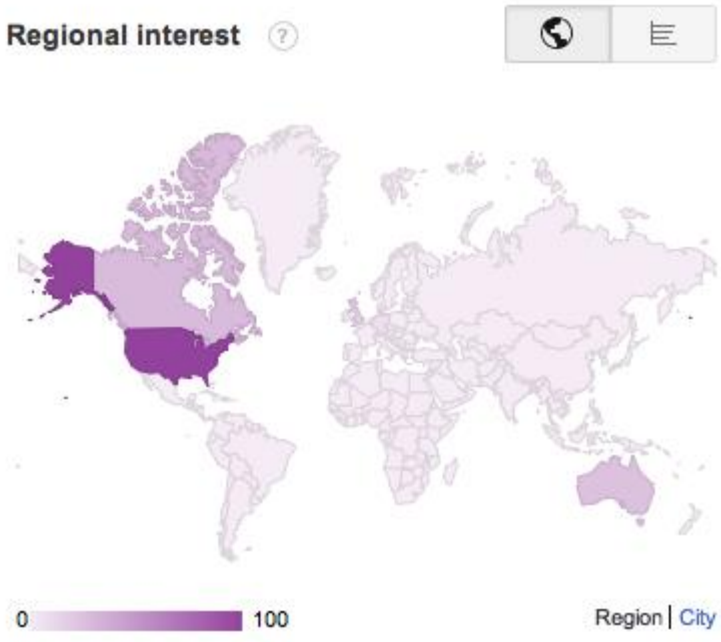
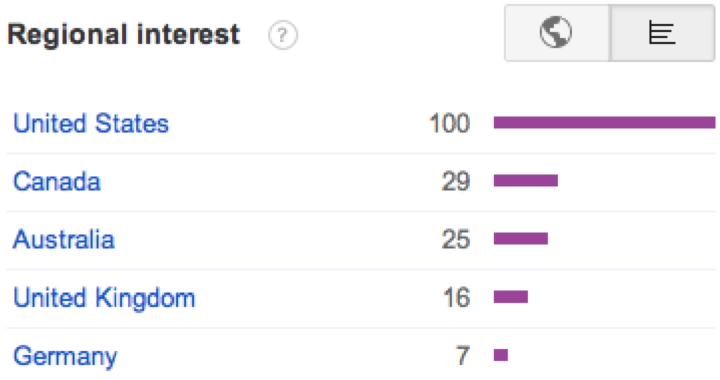


Figure 16b – A partial list of regions searching “Oliver Wendell Holmes” on Google


Regional interest

United States	100
Canada	29
Australia	25
United Kingdom	16
Germany	7

Figure 17a – A heat map of searches for "John Greenleaf Whittier" on Google

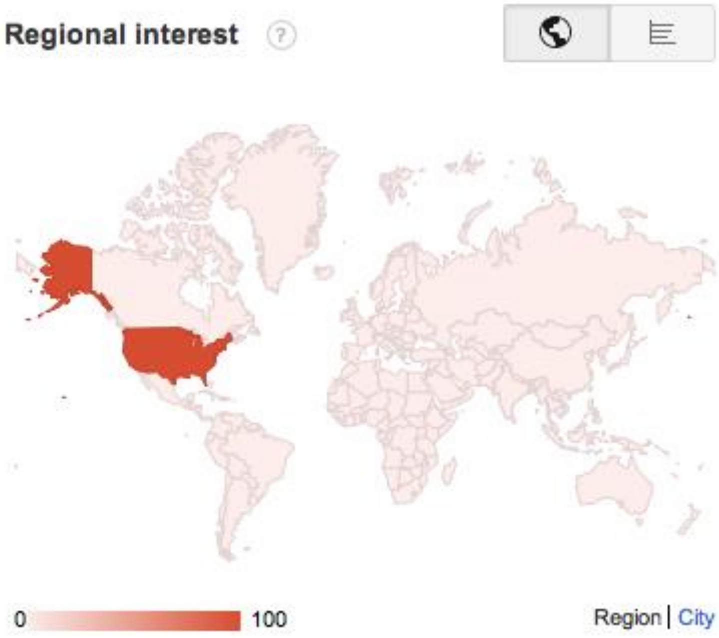
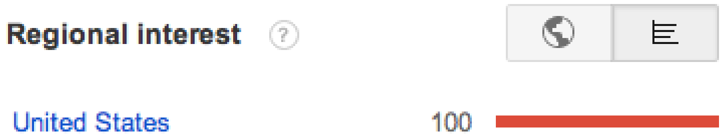


Figure 17b – A partial list of regions searching "John Greenleaf Whittier" on Google



United States 100

Figure 18 – A stacked bar graph of regional interest in attendees at the Whittier Birthday Dinner

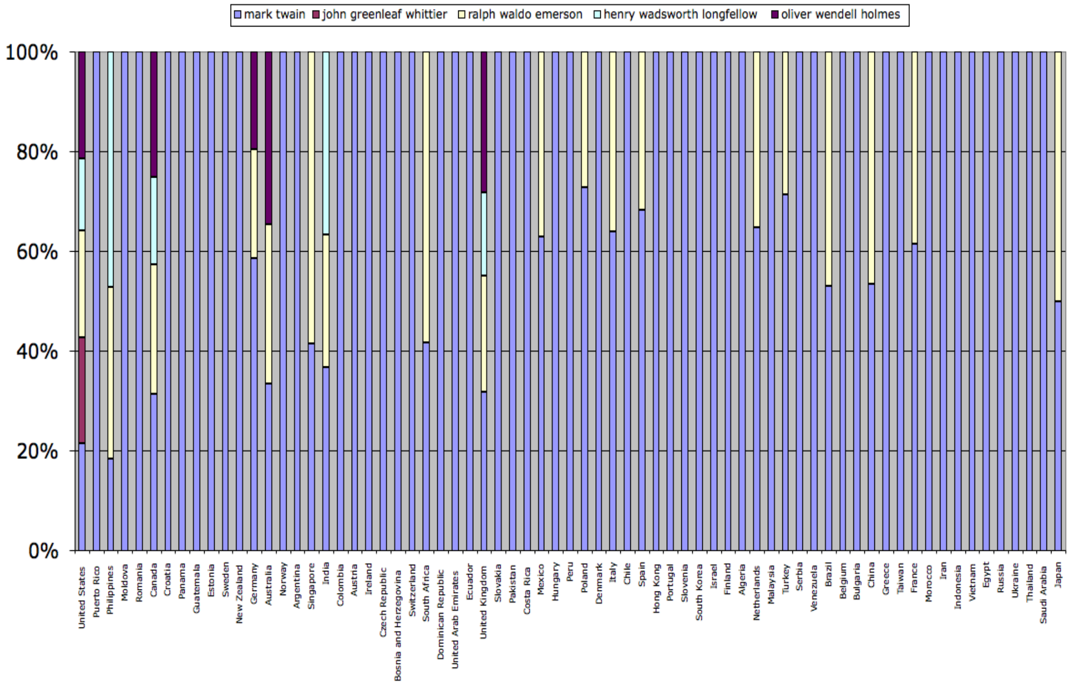


Figure 19 – A comparative look at regional interest in attendees at the Whittier Birthday Dinner

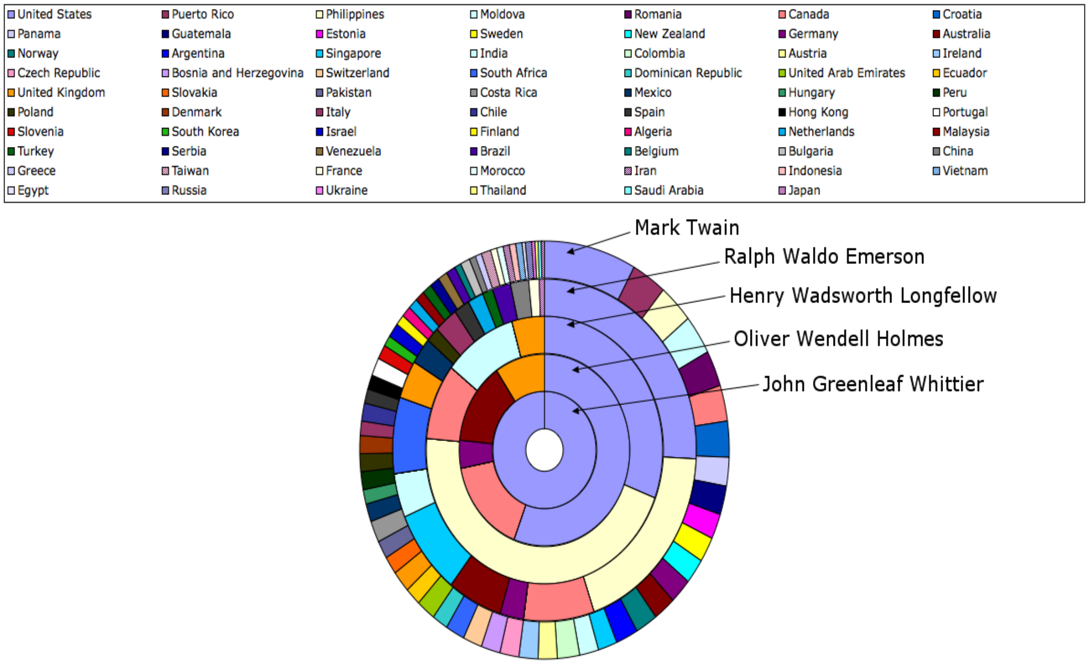


Figure 20 – Searches for “True Jesus Church” on Google

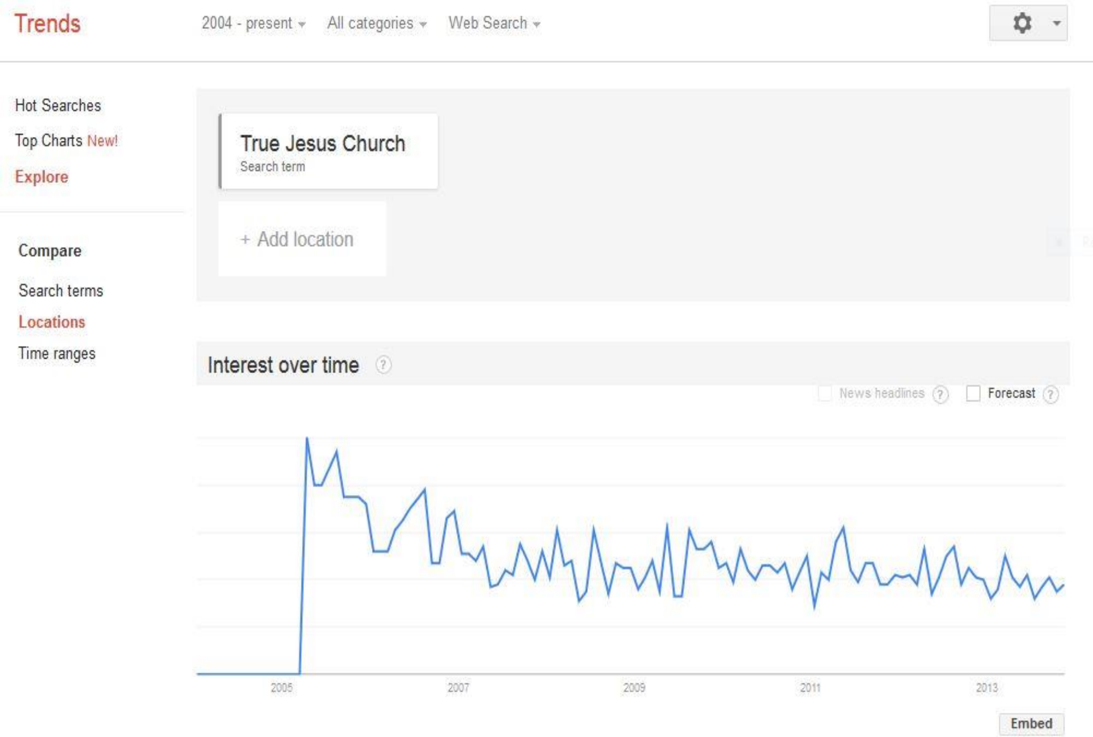
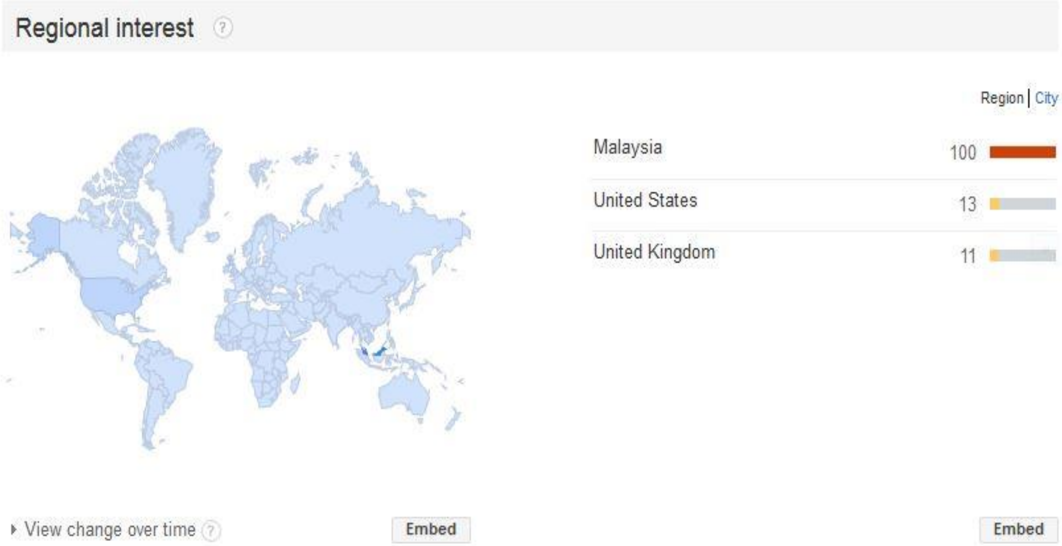
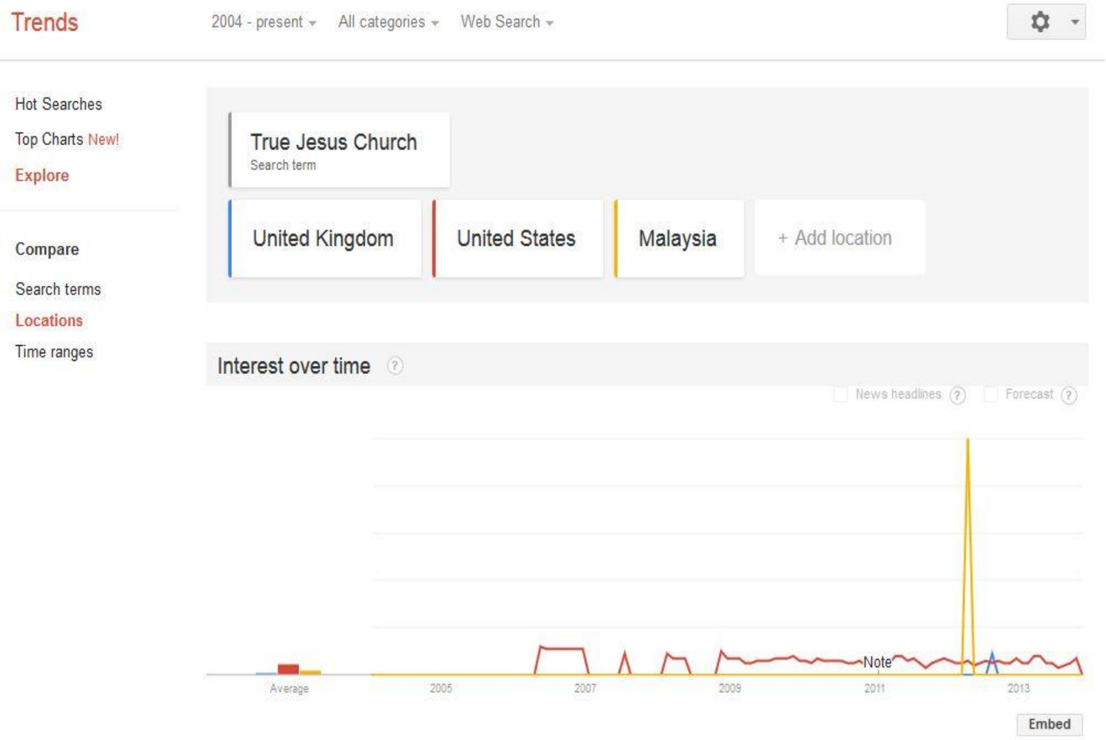


Figure 21 -- A heat map of searches for “True Jesus Church” on Google



Malaysia	100
United States	13
United Kingdom	11

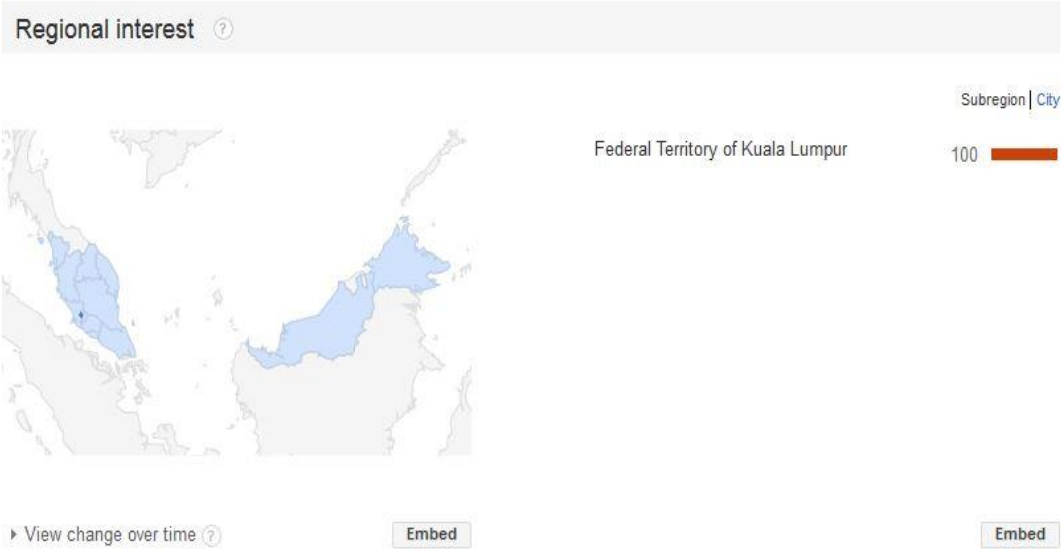
Figure 22 – A comparison of interest in “True Jesus Church”



True Jesus Church
Search term

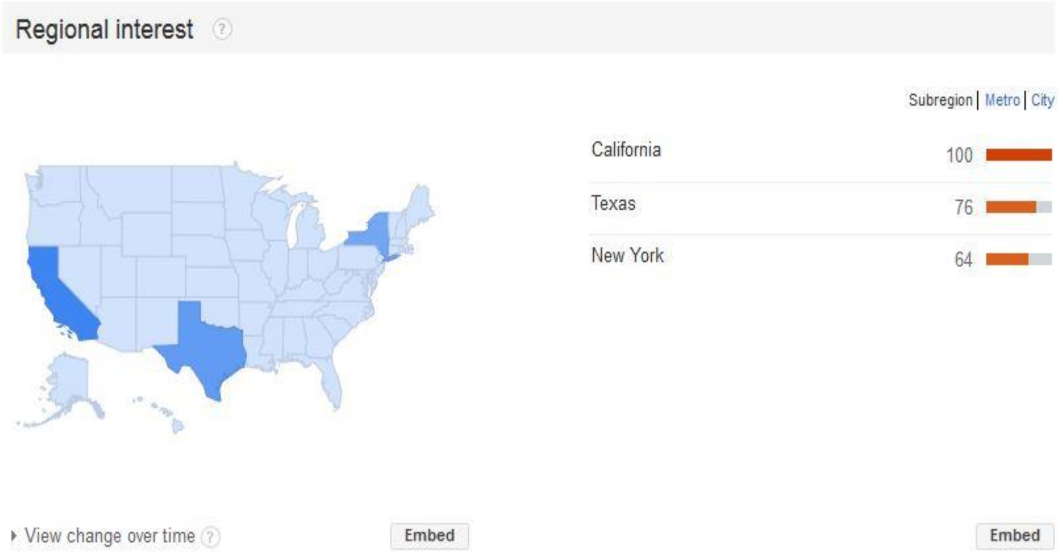
United Kingdom | United States | Malaysia

Figure 23 – A heat map of searches from Malaysia for “True Jesus Church”



Federal Territory of Kuala Lumpur 100

Figure 24 – A heat map of searches from the United States for "True Jesus Church"

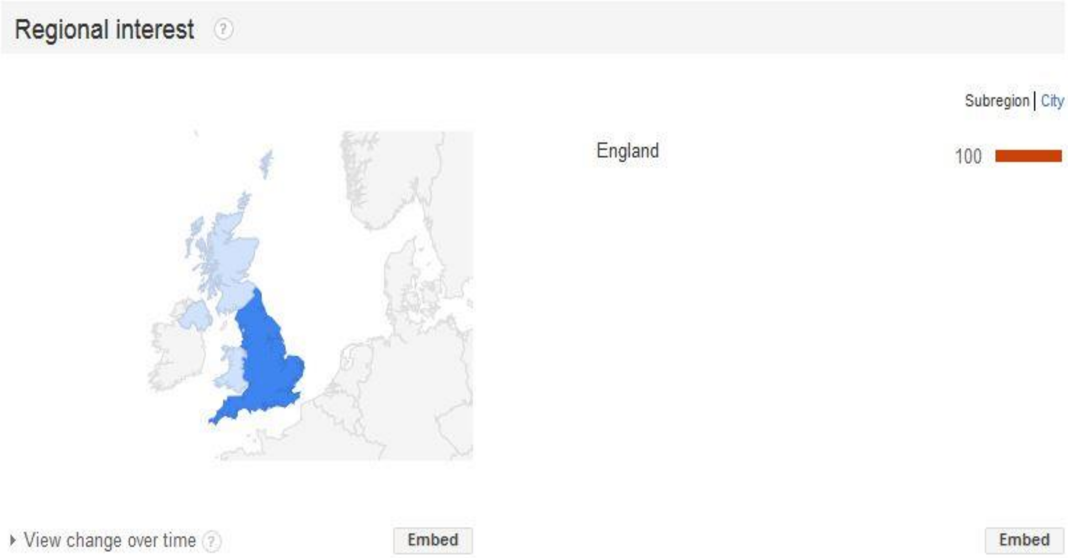


Regional interest

Subregion | Metro | City

California	100
Texas	76
New York	64

Figure 25 – A heat map of searches from the United Kingdom for "True Jesus Church"

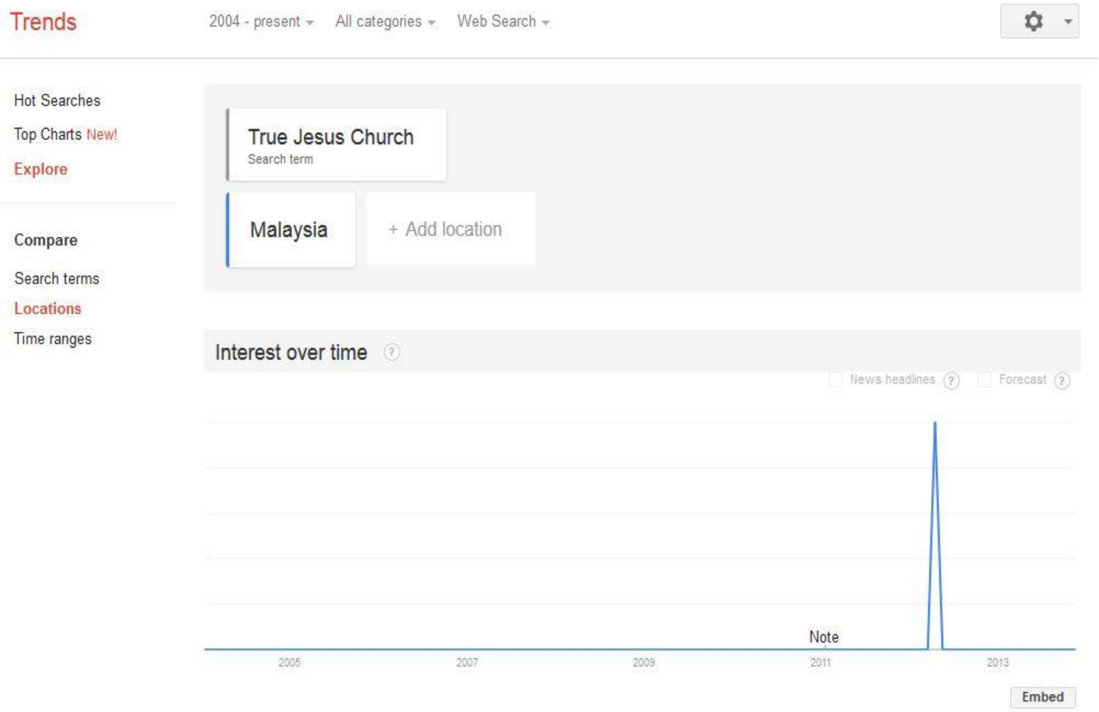


Regional interest

Subregion | City

England	100

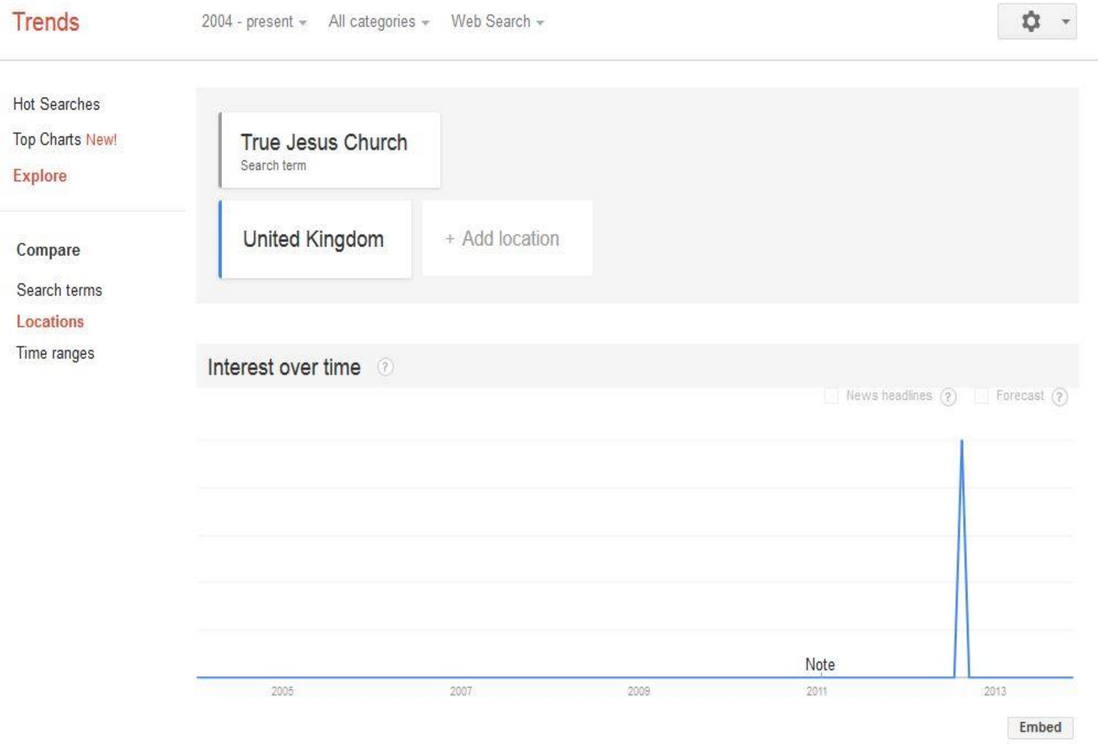
Figure 26 – A timeline of Google searches from Malaysia for “True Jesus Church”



True Jesus Church
Search term

Malaysia

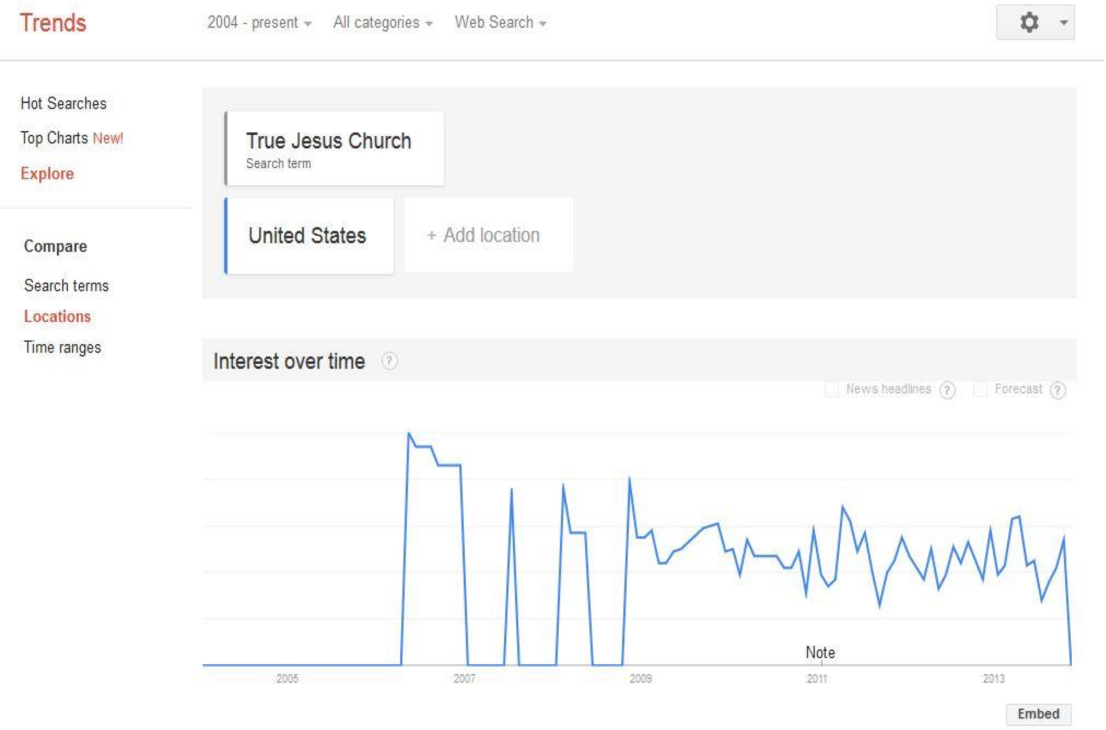
Figure 27 – A timeline of Google searches from the United Kingdom for “True Jesus Church”



True Jesus Church
Search term

United Kingdom

Figure 28 – A timeline of Google searches from the United States for “True Jesus Church”



2004 - present

All categories

Web Search

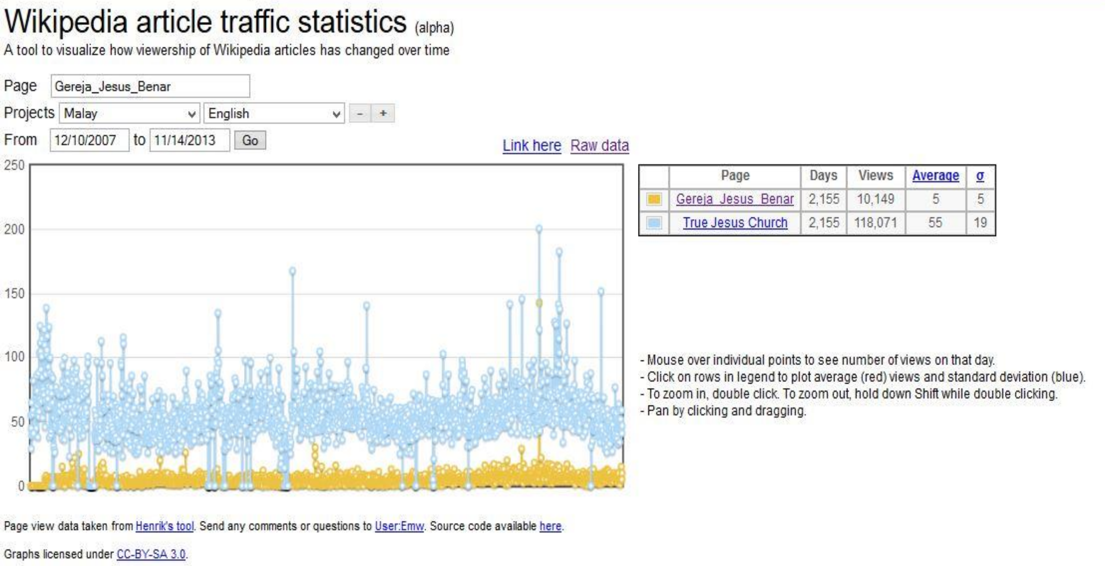
True Jesus Church

Search term

United States

Interest over time

Figure 29 – A comparison of pageviews for True Jesus Church articles on Malay and English Wikipedias



Wikipedia article traffic statistics (alpha)

A tool to visualize how viewership of Wikipedia articles has changed over time

Page: Gereja_Jesus_Benar

Projects: Malay, English

From 12/10/2007 to 11/14/2013

	Page	Days	Views	Average	σ
🟨	Gereja_Jesus_Benar	2,155	10,149	5	5
🟦	True Jesus Church	2,155	118,071	55	19

- Mouse over individual points to see number of views on that day.
- Click on rows in legend to plot average (red) views and standard deviation (blue).
- To zoom in, double click. To zoom out, hold down Shift while double clicking.
- Pan by clicking and dragging.

Page view data taken from Henrik's tool. Send any comments or questions to User:Emw. Source code available here.

Graphs licensed under CC-BY-SA 3.0.

Figure 30 – Edit for the True Jesus Church article on Malay Wikipedia

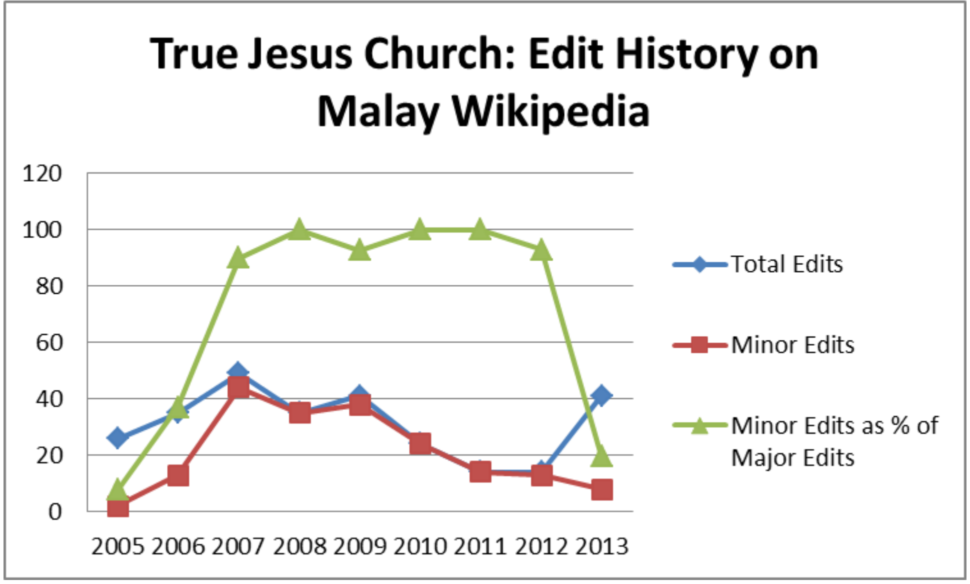


Figure 31 – Edit for the True Jesus Church article on English Wikipedia

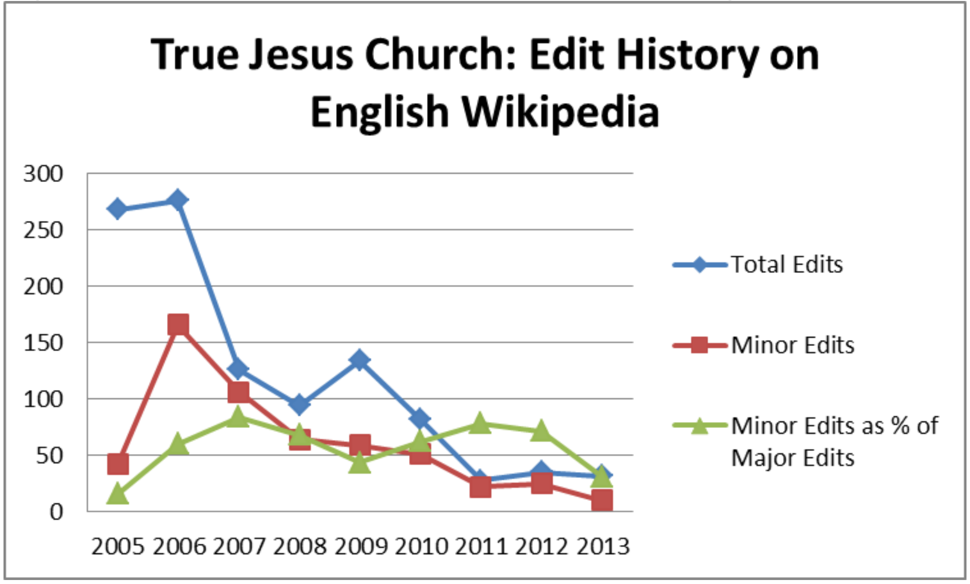


Figure 32 – Sonification of primary date of publication data obtained from HathiTrust by searching for works that list “Mark Twain” as an author

PubDates.midi

Figure 33 – Visualization produced by Sonification Sandbox as part of the sonification of primary date of publication data obtained from HathiTrust by searching for works that list "Mark Twain" as an author

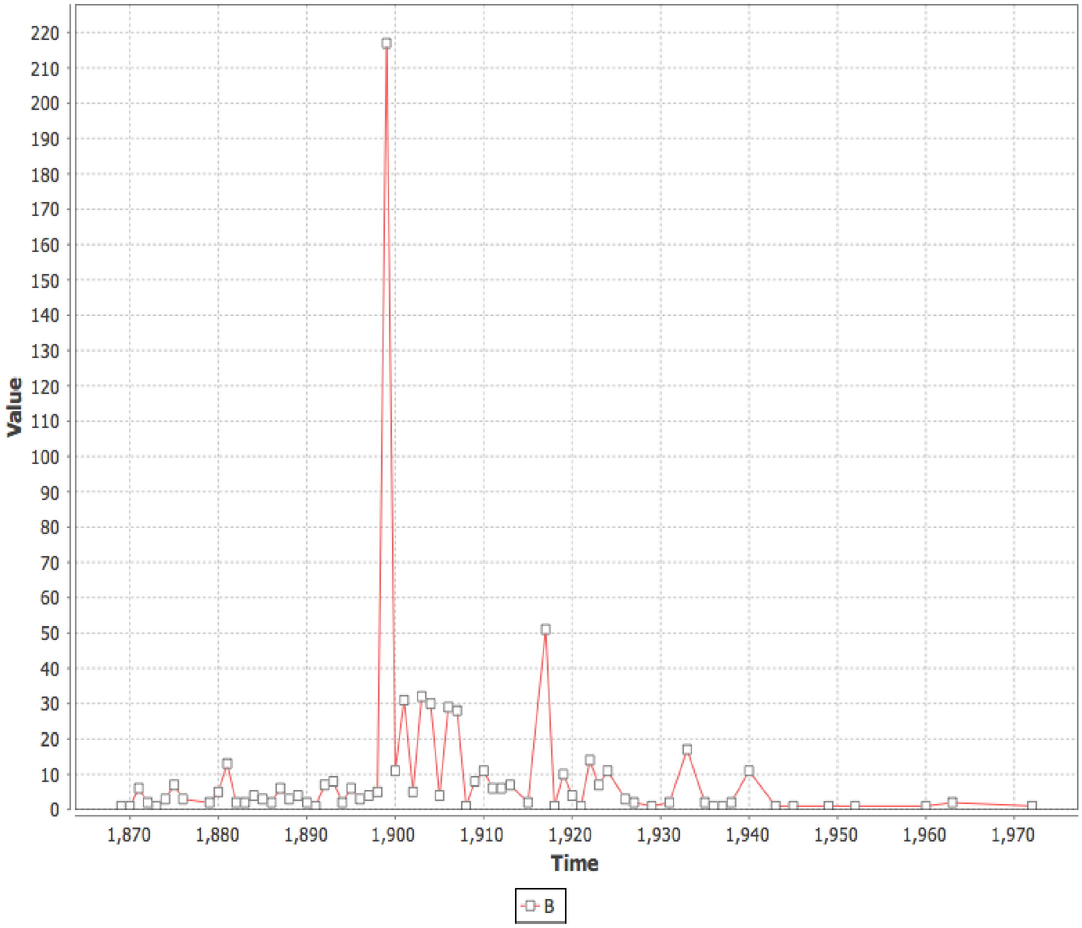
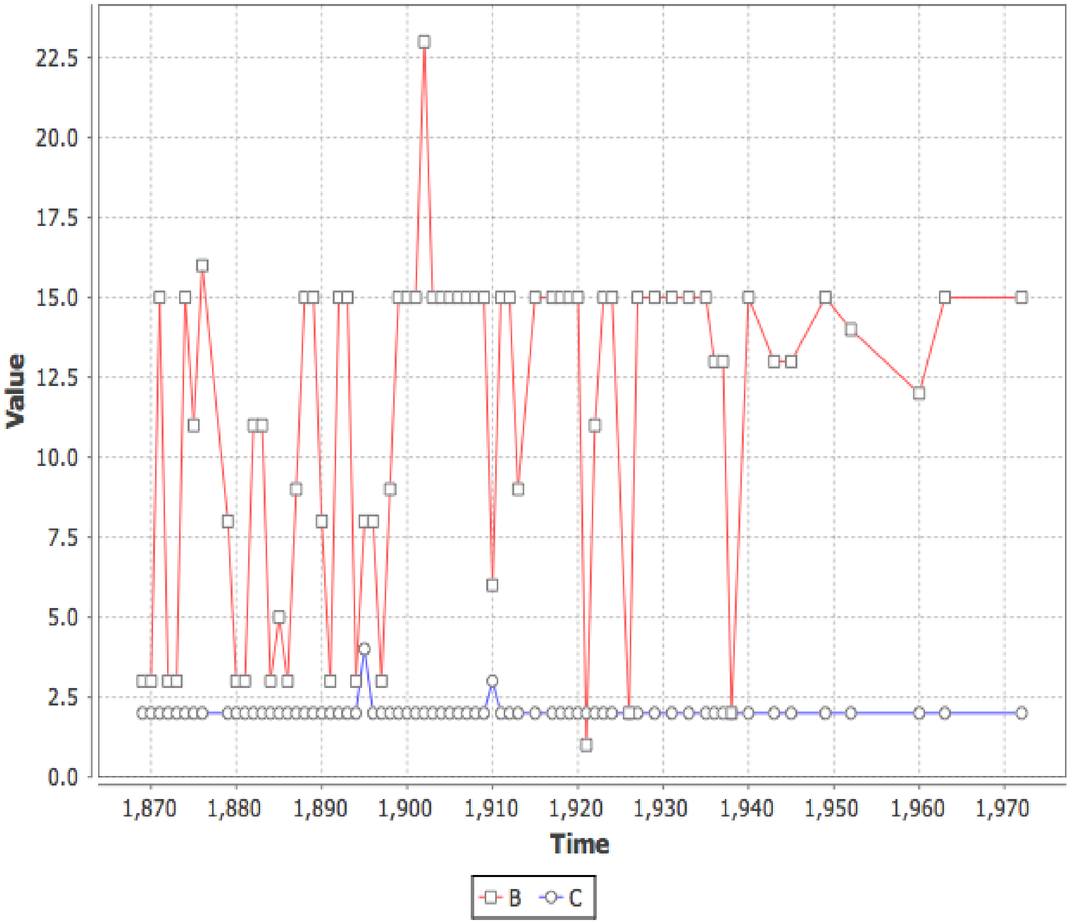


Figure 34 – Sonification of language and place of publication data obtained from HathiTrust by searching for works that list "Mark Twain" as an author

dpl.midi

Figure 35 – Visualization produced by Sonification Sandbox as part of the sonification of language and place of publication data obtained from HathiTrust by searching for works that list “Mark Twain” as an author



Appendix for Chapter Two

Figure 1 – The results of a VCA search of Chronicling America for “Benvenuto Cellini”

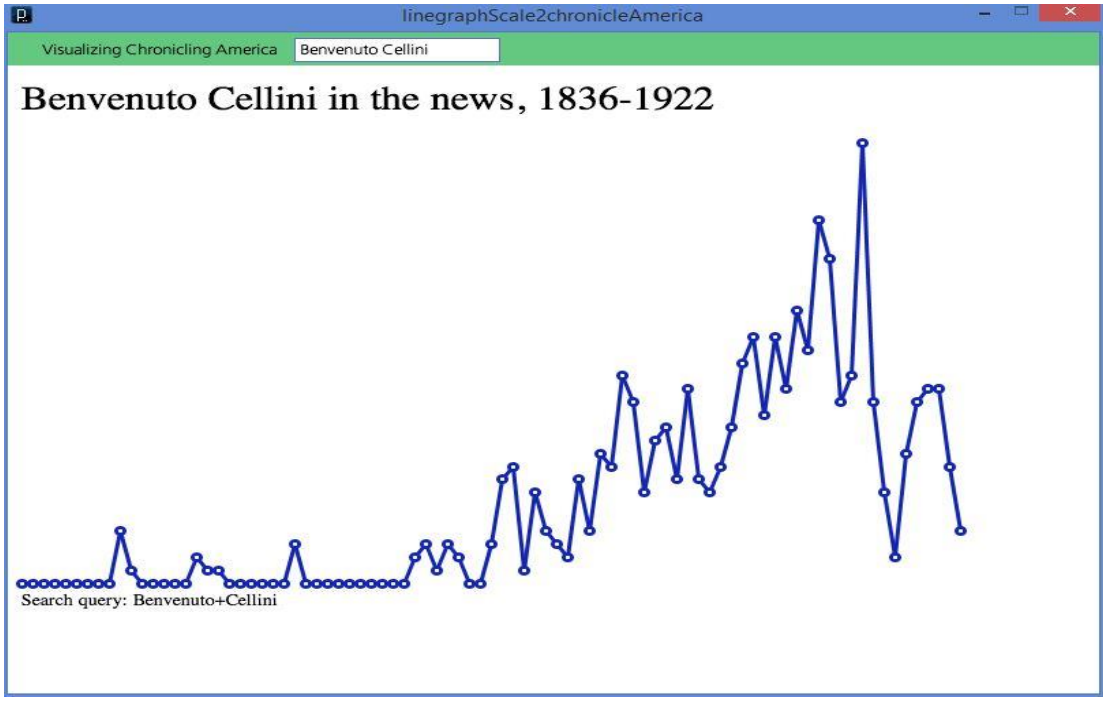

linegraphScale2chronicleAmerica
Visualizing Chronicling America
Benvenuto Cellini
Benvenuto Cellini in the news, 1836-1922
Search query: Benvenuto+Cellini

Figure 2 -- The results of a VCA search of Chronicling America for “Helen Keller”

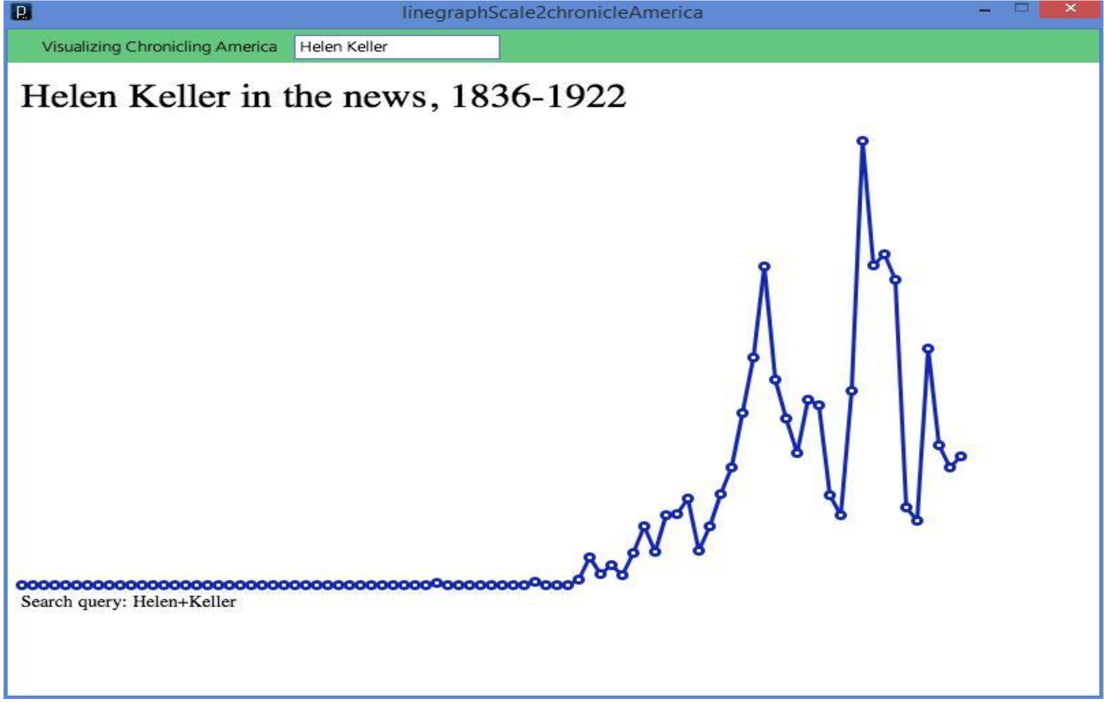
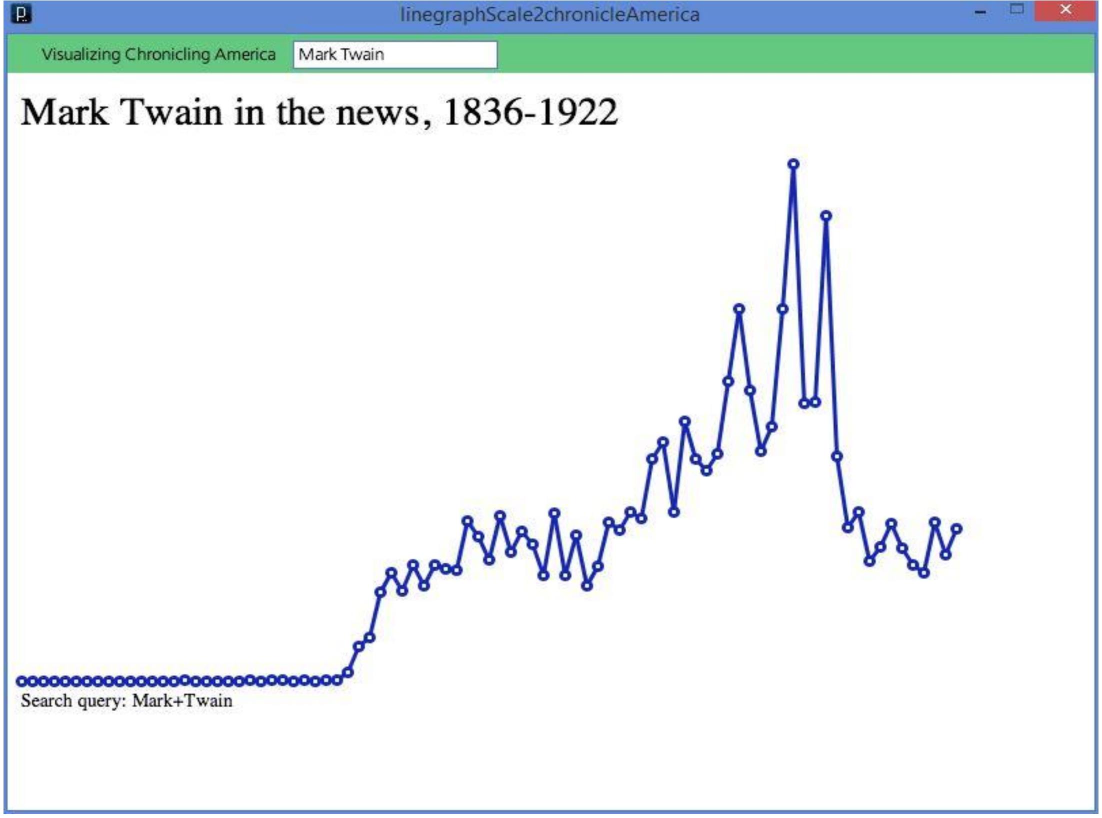

linegraphScale2chronicleAmerica
Visualizing Chronicling America
Helen Keller
Helen Keller in the news, 1836-1922
Search query: Helen+Keller

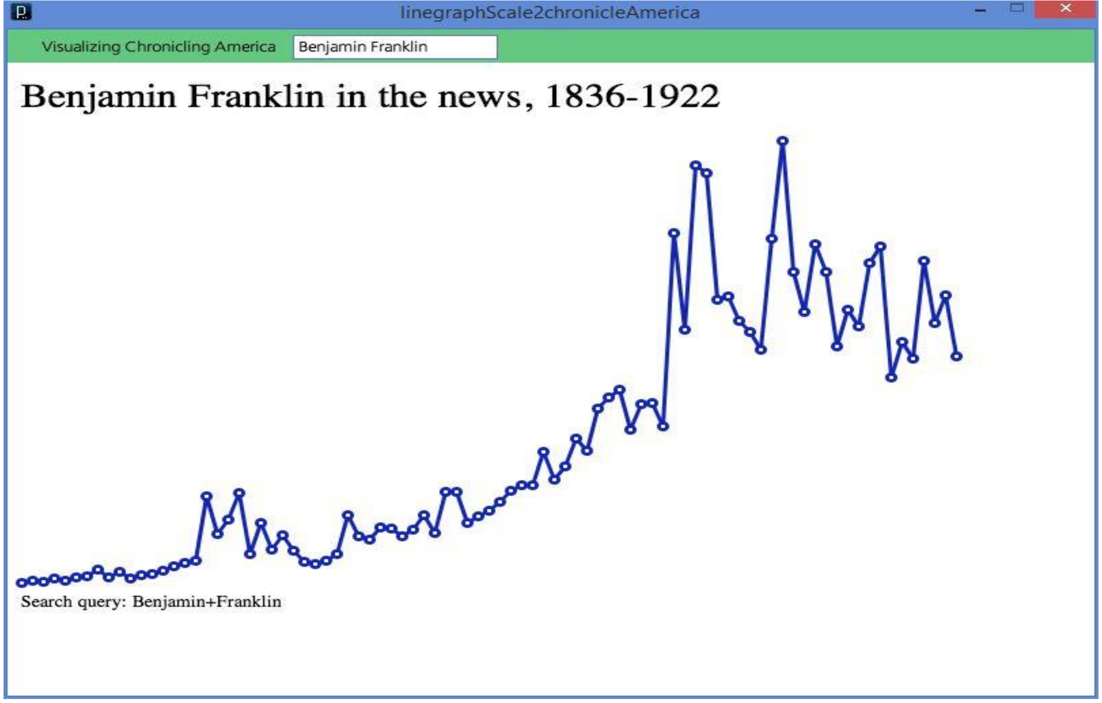
Figure 3 – The results of a VCA search for "Mark Twain"


linegraphScale2chronicleAmerica
Visualizing Chronicling America
Mark Twain

Mark Twain in the news, 1836-1922

Search query: Mark+Twain

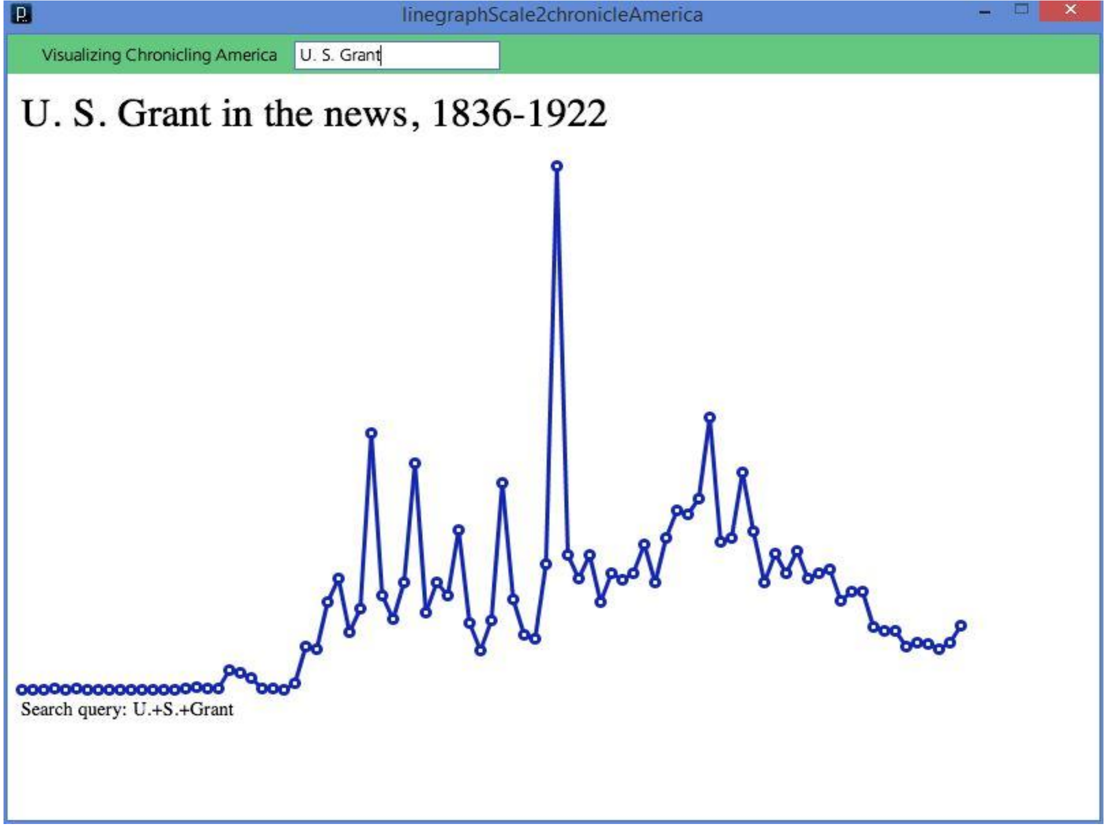
Figure 4 – The results of a VCA search for "Benjamin Franklin"


linegraphScale2chronicleAmerica
Visualizing Chronicling America
Benjamin Franklin

Benjamin Franklin in the news, 1836-1922

Search query: Benjamin+Franklin

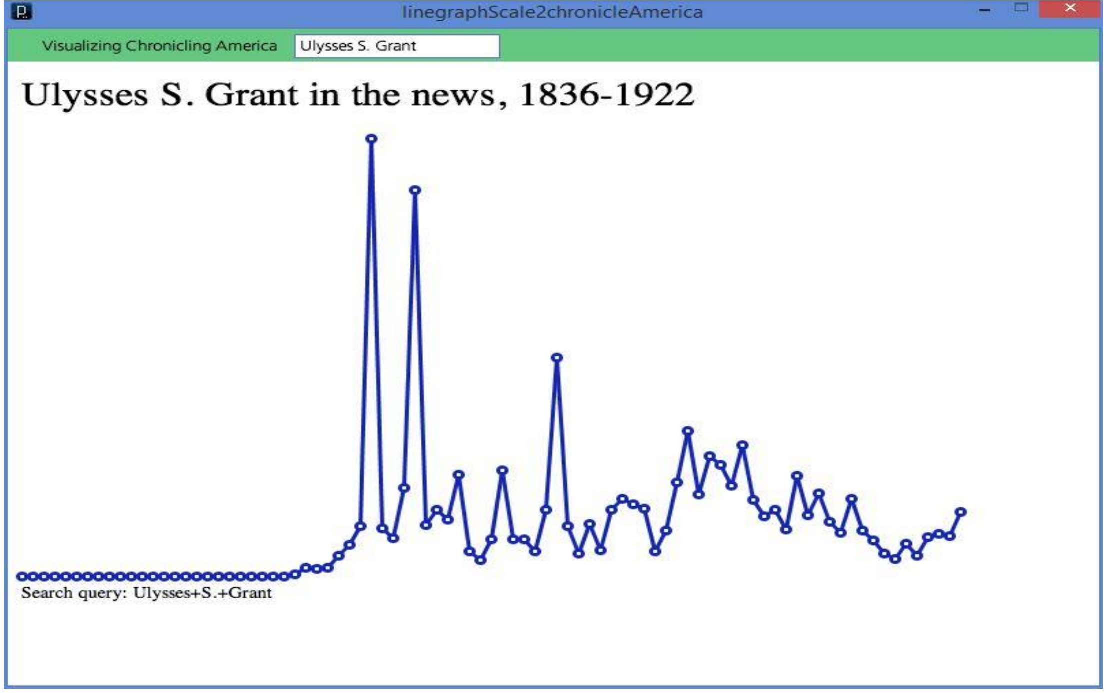
Figure 5 – The results of a VCA search for “U.S. Grant”


linegraphScale2chronicleAmerica
U. S. Grant in the news, 1836-1922

Visualizing Chronicling America U. S. Grant

Search query: U.+S.+Grant

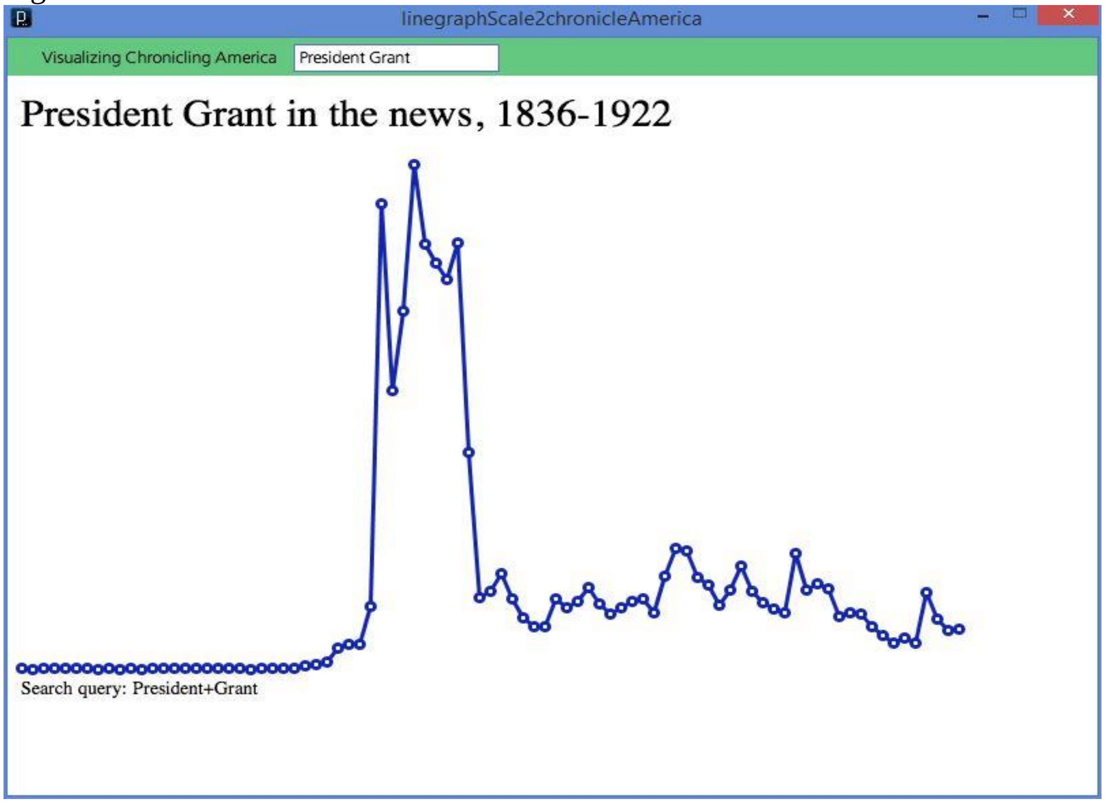
Figure 6 – The results of a VCA search for “Ulysses S. Grant”


linegraphScale2chronicleAmerica
Ulysses S. Grant in the news, 1836-1922

Visualizing Chronicling America Ulysses S. Grant

Search query: Ulysses+S.+Grant

Figure 7 – The results of a VCA search for “President Grant”

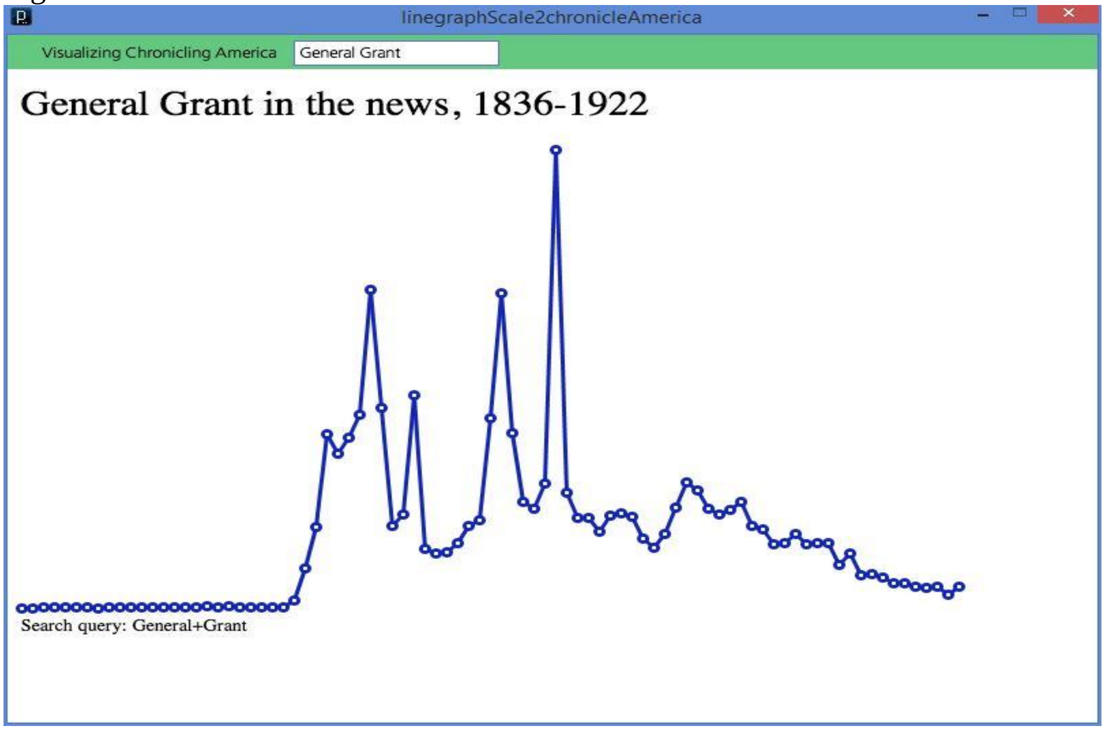

linegraphScale2chronicleAmerica
Visualizing Chronicling America

President Grant

President Grant in the news, 1836-1922

Search query: President+Grant

Figure 8 – The results of a VCA search for “General Grant”

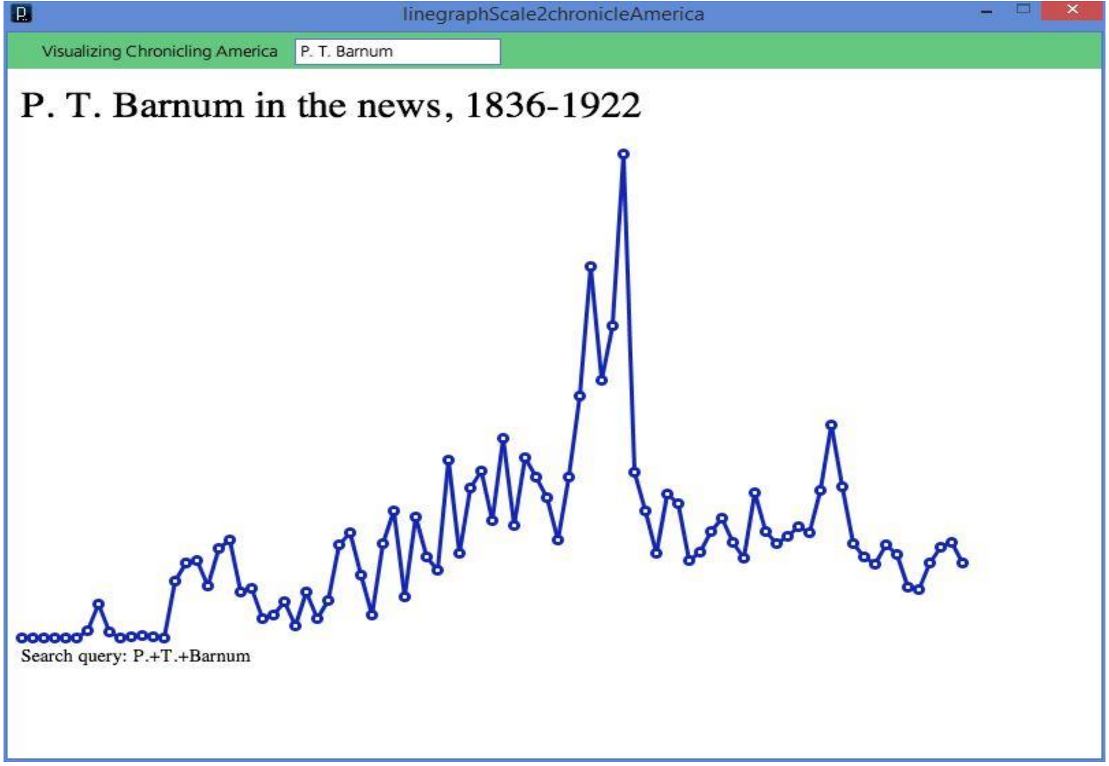

linegraphScale2chronicleAmerica
Visualizing Chronicling America

General Grant

General Grant in the news, 1836-1922

Search query: General+Grant

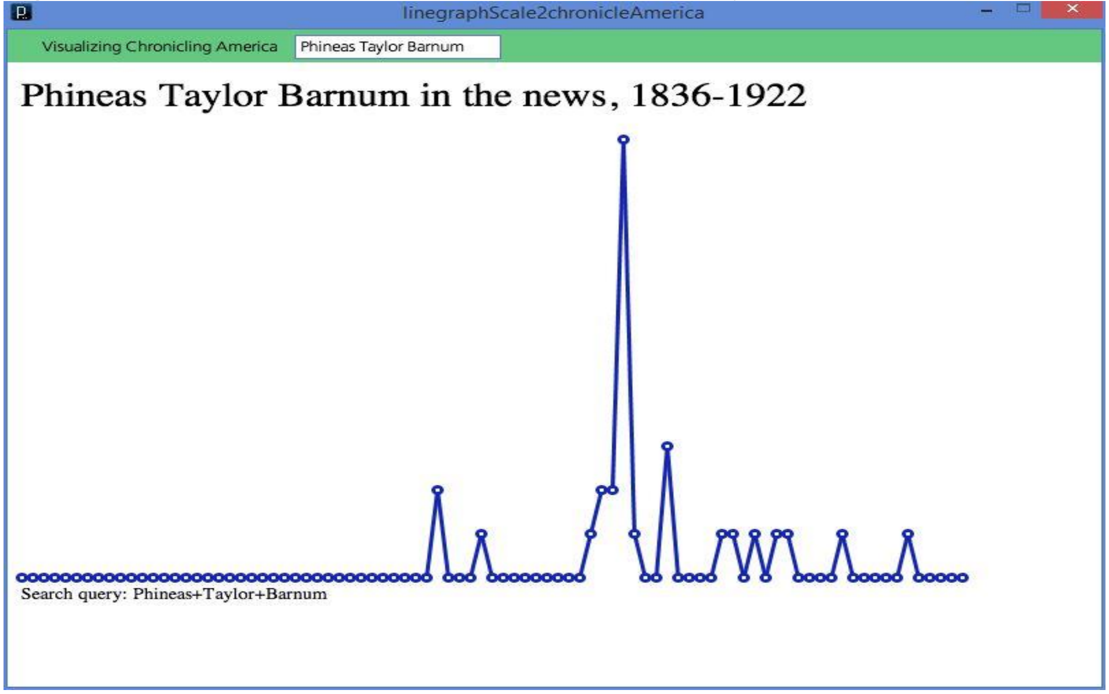
Figure 9 – The results of a VCA search for "P.T. Barnum"


linegraphScale2chronicleAmerica
P. T. Barnum in the news, 1836-1922

Visualizing Chronicling America: P. T. Barnum

Search query: P.+T.+Barnum

Figure 10 – The results of a VCA search for "Phineas Taylor Barnum"


linegraphScale2chronicleAmerica
Phineas Taylor Barnum in the news, 1836-1922

Visualizing Chronicling America: Phineas Taylor Barnum

Search query: Phineas+Taylor+Barnum

Figure 11 – Prototyping a potential future for VCA with Excel: A Comparison of Chronicling America Searches for References to Ulysses S. Grant (raw counts plotted on a standard y-axis)

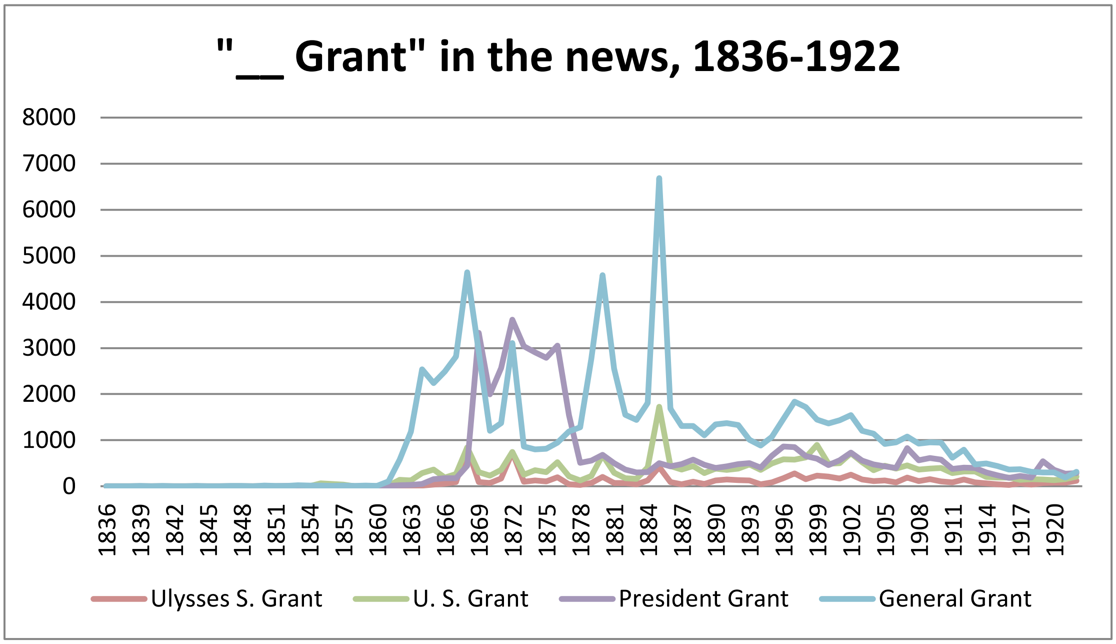


Figure 12 – Prototyping a potential future for VCA with Excel: A Comparison of Chronicling America Searches for References to Ulysses S. Grant (raw counts plotted on a logarithmic y-axis)

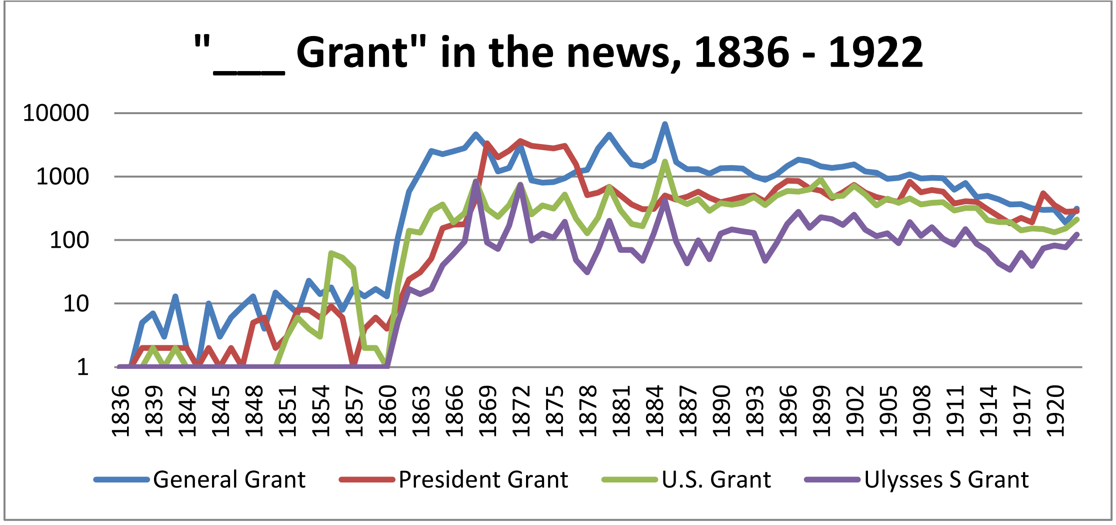


Figure 13 – Appearances of "Mark Twain" in Chronicling America's holdings for 1907

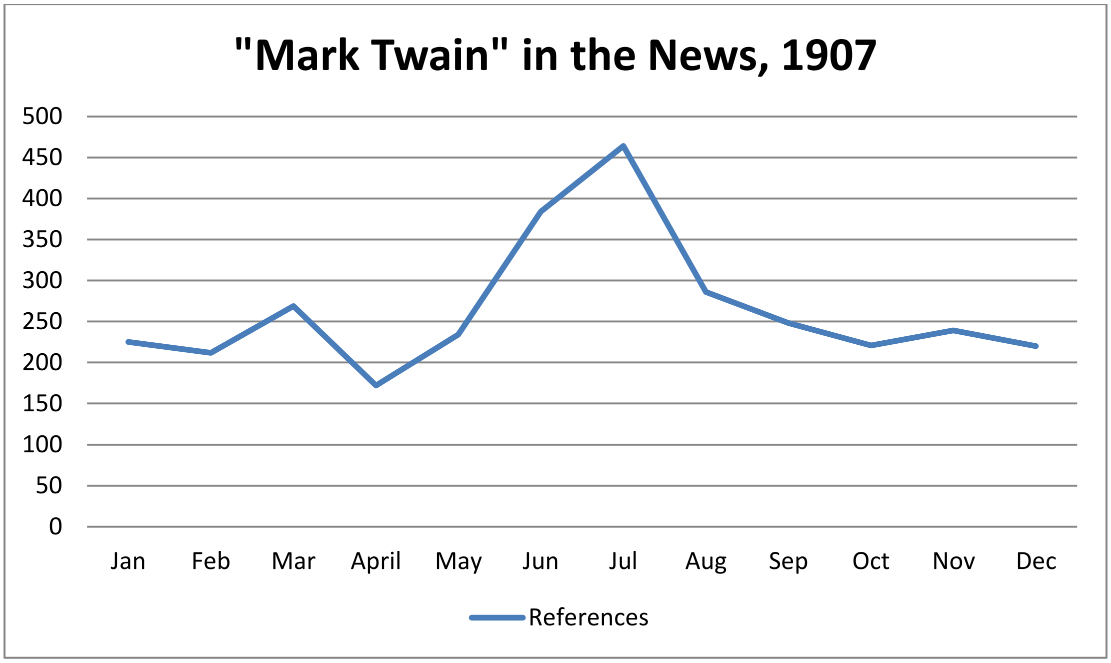
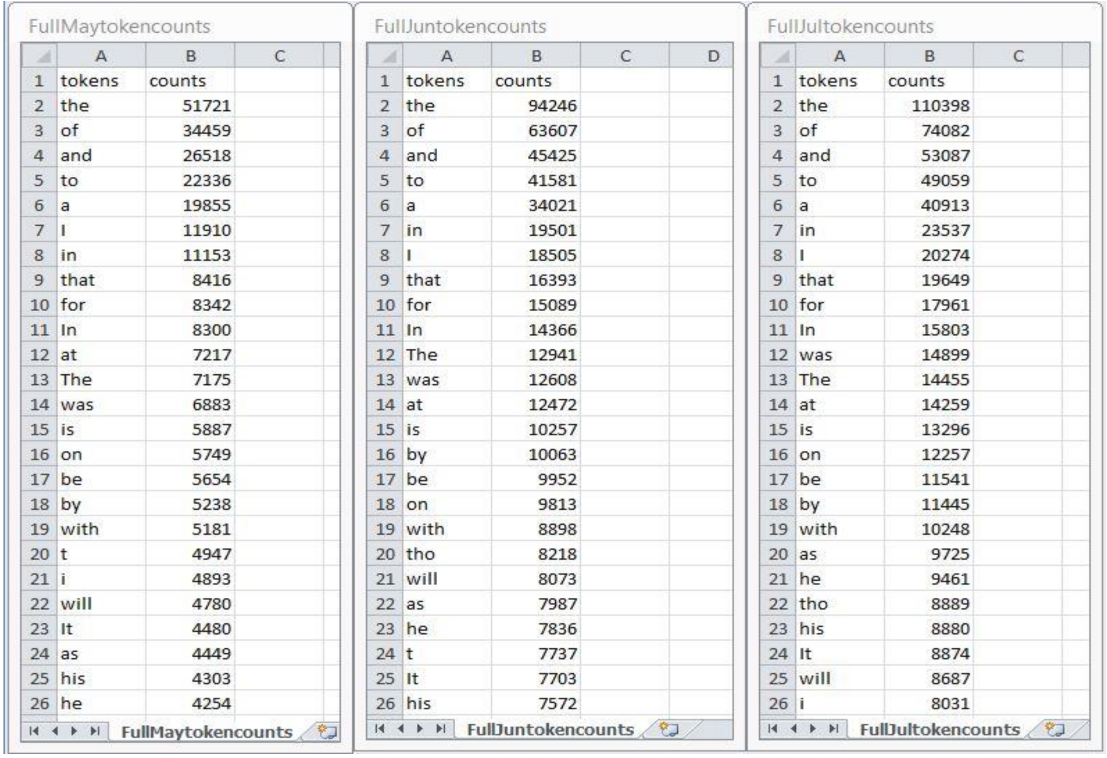


Figure 14 – Most common tokens (excluding punctuation) found in news articles containing "Mark Twain" for May, June, and July of 1907



FullMaytokencounts

	A	B	C
1	tokens	counts	
2	the	51721	
3	of	34459	
4	and	26518	
5	to	22336	
6	a	19855	
7	I	11910	
8	in	11153	
9	that	8416	
10	for	8342	
11	In	8300	
12	at	7217	
13	The	7175	
14	was	6883	
15	is	5887	
16	on	5749	
17	be	5654	
18	by	5238	
19	with	5181	
20	t	4947	
21	i	4893	
22	will	4780	
23	It	4480	
24	as	4449	
25	his	4303	
26	he	4254	

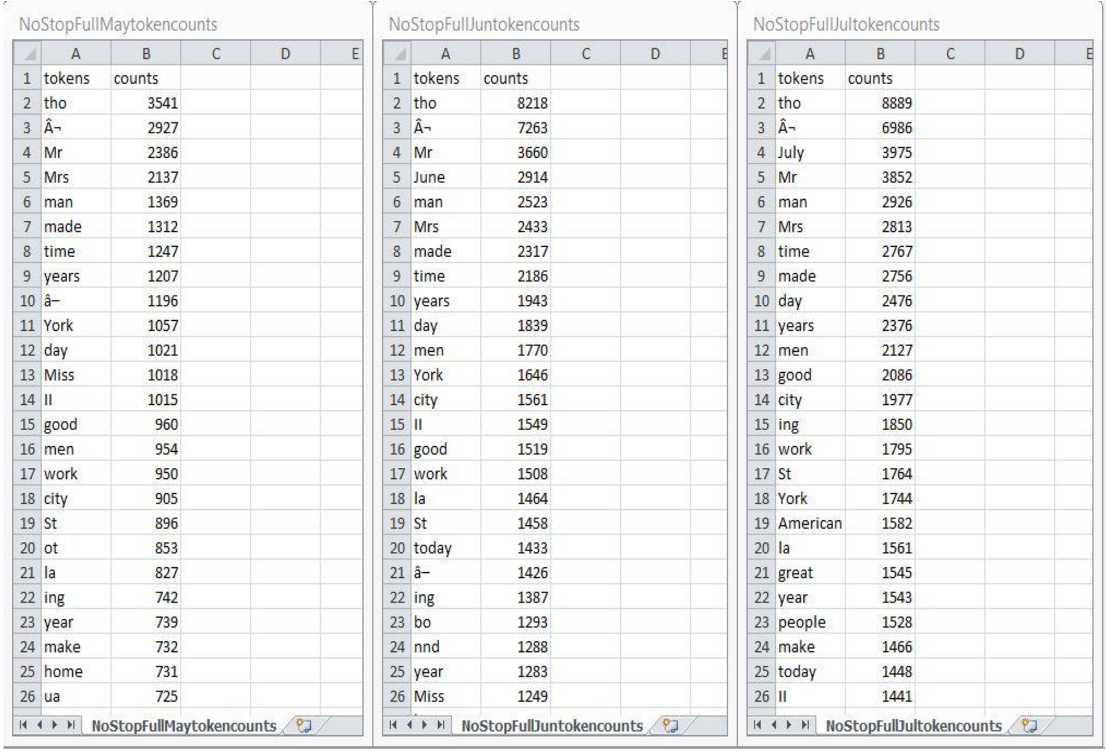
FullJuntokencounts

	A	B	C	D
1	tokens	counts		
2	the	94246		
3	of	63607		
4	and	45425		
5	to	41581		
6	a	34021		
7	in	19501		
8	I	18505		
9	that	16393		
10	for	15089		
11	In	14366		
12	The	12941		
13	was	12608		
14	at	12472		
15	is	10257		
16	by	10063		
17	be	9952		
18	on	9813		
19	with	8898		
20	tho	8218		
21	will	8073		
22	as	7987		
23	he	7836		
24	t	7737		
25	It	7703		
26	his	7572		

FullJultokencounts

	A	B	C
1	tokens	counts	
2	the	110398	
3	of	74082	
4	and	53087	
5	to	49059	
6	a	40913	
7	in	23537	
8	I	20274	
9	that	19649	
10	for	17961	
11	In	15803	
12	was	14899	
13	The	14455	
14	at	14259	
15	is	13296	
16	on	12257	
17	be	11541	
18	by	11445	
19	with	10248	
20	as	9725	
21	he	9461	
22	tho	8889	
23	his	8880	
24	It	8874	
25	will	8687	
26	i	8031	

Figure 15 – Most common tokens (excluding punctuation and stop words) found in news articles containing “Mark Twain” for May, June, and July of 1907



NoStopFullMaytokencounts

	A	B	C	D	E
1	tokens	counts			
2	tho	3541			
3	Â¬	2927			
4	Mr	2386			
5	Mrs	2137			
6	man	1369			
7	made	1312			
8	time	1247			
9	years	1207			
10	â–	1196			
11	York	1057			
12	day	1021			
13	Miss	1018			
14	II	1015			
15	good	960			
16	men	954			
17	work	950			
18	city	905			
19	St	896			
20	ot	853			
21	la	827			
22	ing	742			
23	year	739			
24	make	732			
25	home	731			
26	ua	725			

NoStopFullMaytokencounts

NoStopFullJuntokencounts

	A	B	C	D	E
1	tokens	counts			
2	tho	8218			
3	Â¬	7263			
4	Mr	3660			
5	June	2914			
6	man	2523			
7	Mrs	2433			
8	made	2317			
9	time	2186			
10	years	1943			
11	day	1839			
12	men	1770			
13	York	1646			
14	city	1561			
15	II	1549			
16	good	1519			
17	work	1508			
18	la	1464			
19	St	1458			
20	today	1433			
21	â–	1426			
22	ing	1387			
23	bo	1293			
24	nnd	1288			
25	year	1283			
26	Miss	1249			

NoStopFullJuntokencounts

NoStopFullJultokencounts

	A	B	C	D	E
1	tokens	counts			
2	tho	8889			
3	Â¬	6986			
4	July	3975			
5	Mr	3852			
6	man	2926			
7	Mrs	2813			
8	time	2767			
9	made	2756			
10	day	2476			
11	years	2376			
12	men	2127			
13	good	2086			
14	city	1977			
15	ing	1850			
16	work	1795			
17	St	1764			
18	York	1744			
19	American	1582			
20	la	1561			
21	great	1545			
22	year	1543			
23	people	1528			
24	make	1466			
25	today	1448			
26	II	1441			

NoStopFullJultokencounts

Figure 16a – A chart of the distribution of publication date metadata for 686 MarcXML records obtained by searching the Hathi Trust for “Mark Twain” via the author field.

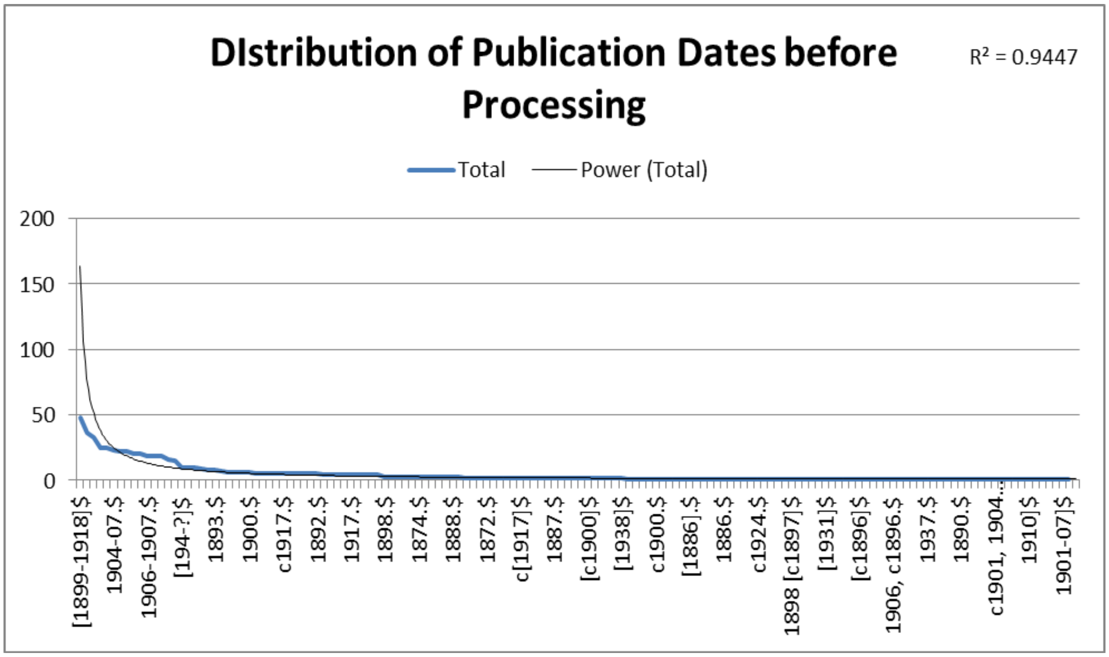


Figure 16b – A chart of the distribution of publication date metadata presented above after it has been subjected to a process of standardization using Open Refine.

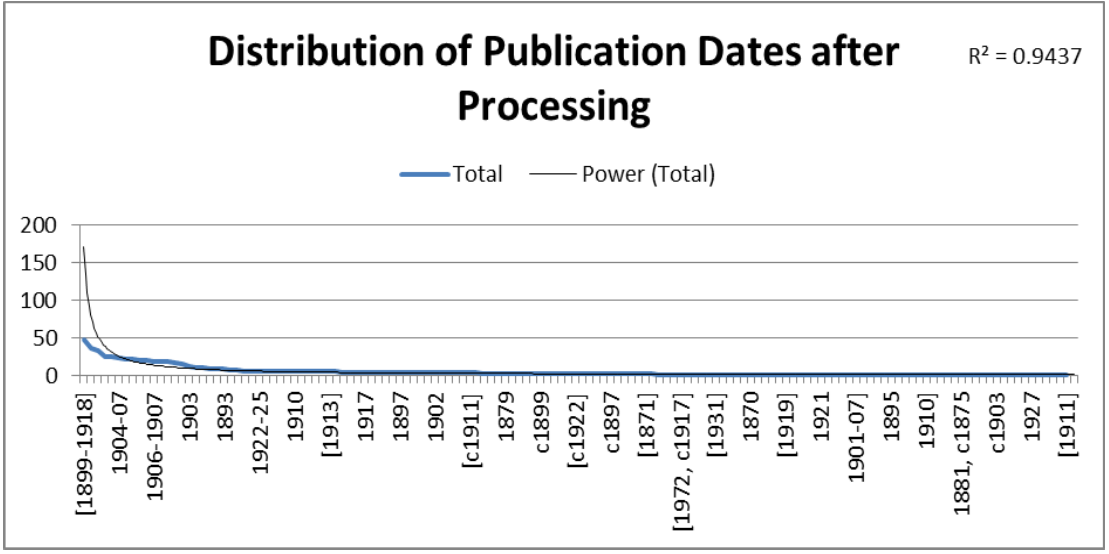


Figure 17a – A chart of the distribution of publisher location metadata for 686 MarcXML records obtained by searching the Hathi Trust for "Mark Twain" via the author field.

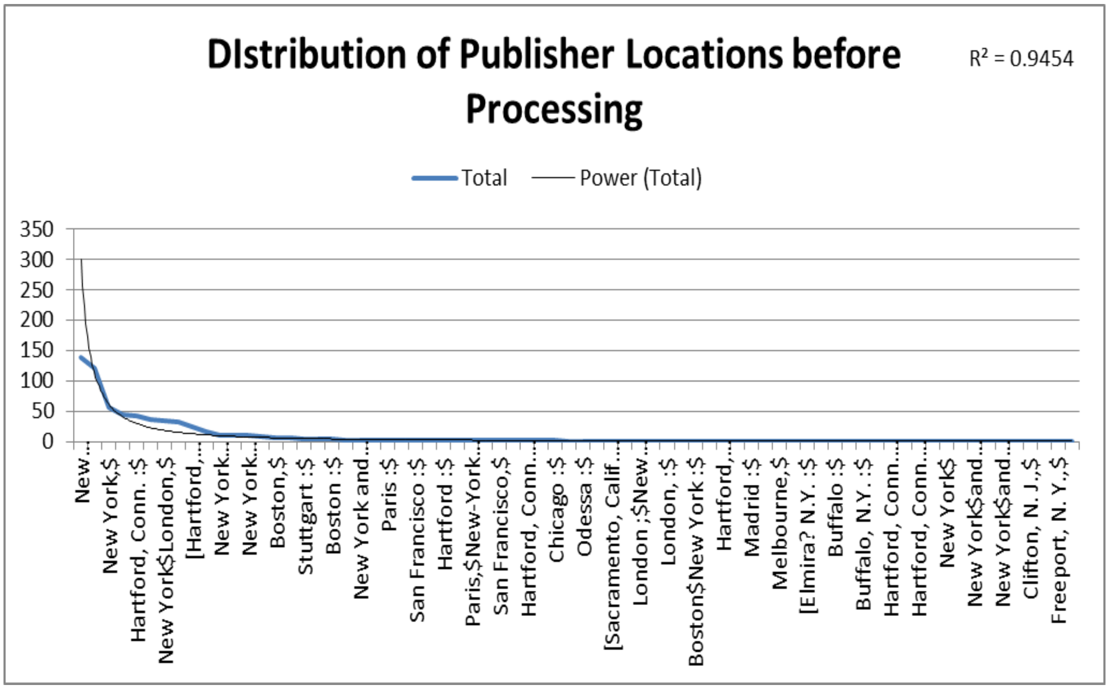


Figure 17b – A chart of the distribution of publisher location metadata presented above after it has been subjected to a process of standardization using Open Refine.

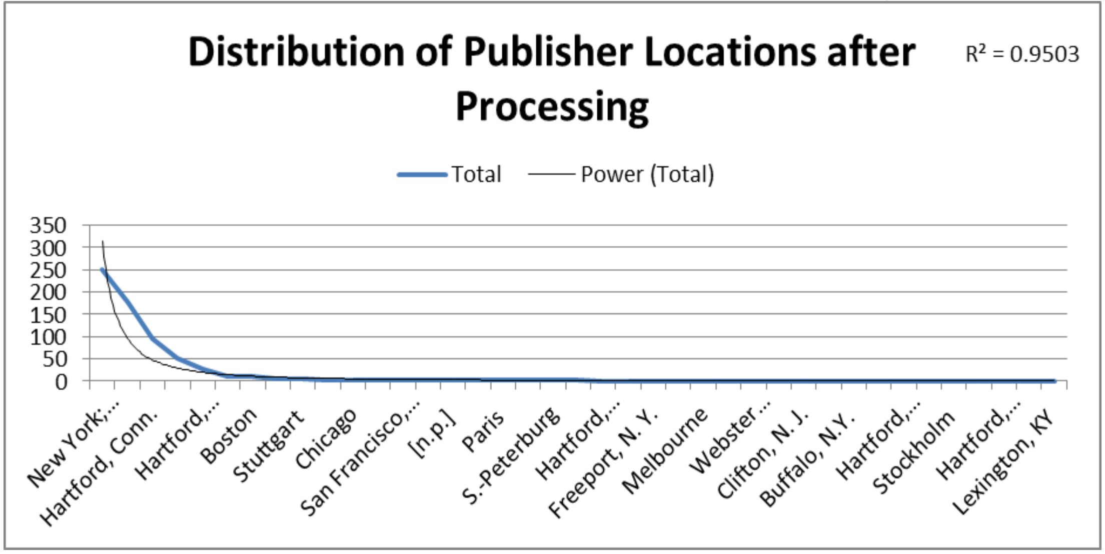


Figure 18a – A chart of the distribution of publisher names metadata for 686 MarcXML records obtained by searching the Hathi Trust for “Mark Twain” via the author field.

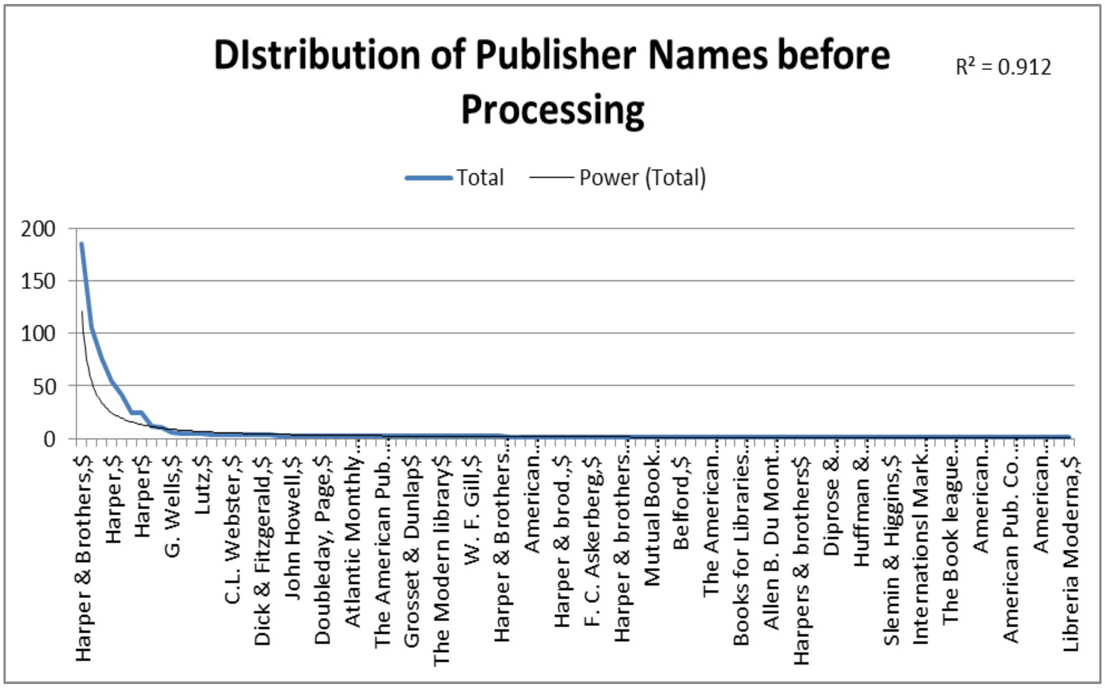


Figure 18b – A chart of the distribution of publisher names metadata presented above after it has been subjected to a process of standardization using Open Refine.

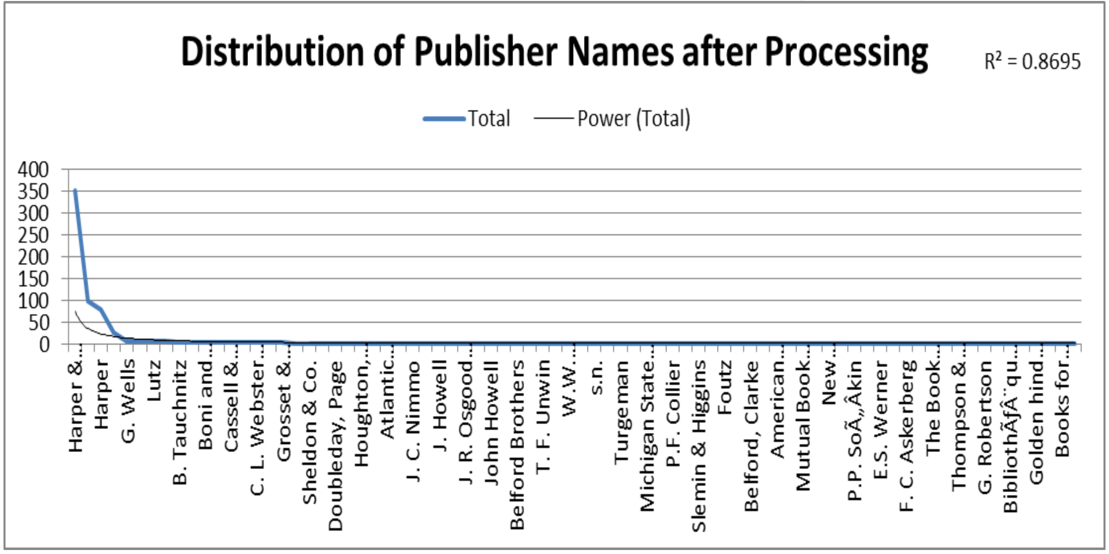


Figure 19a – A chart of the distribution of publisher location metadata for 2998 records from Open Library with “Mark Twain” in the author field.

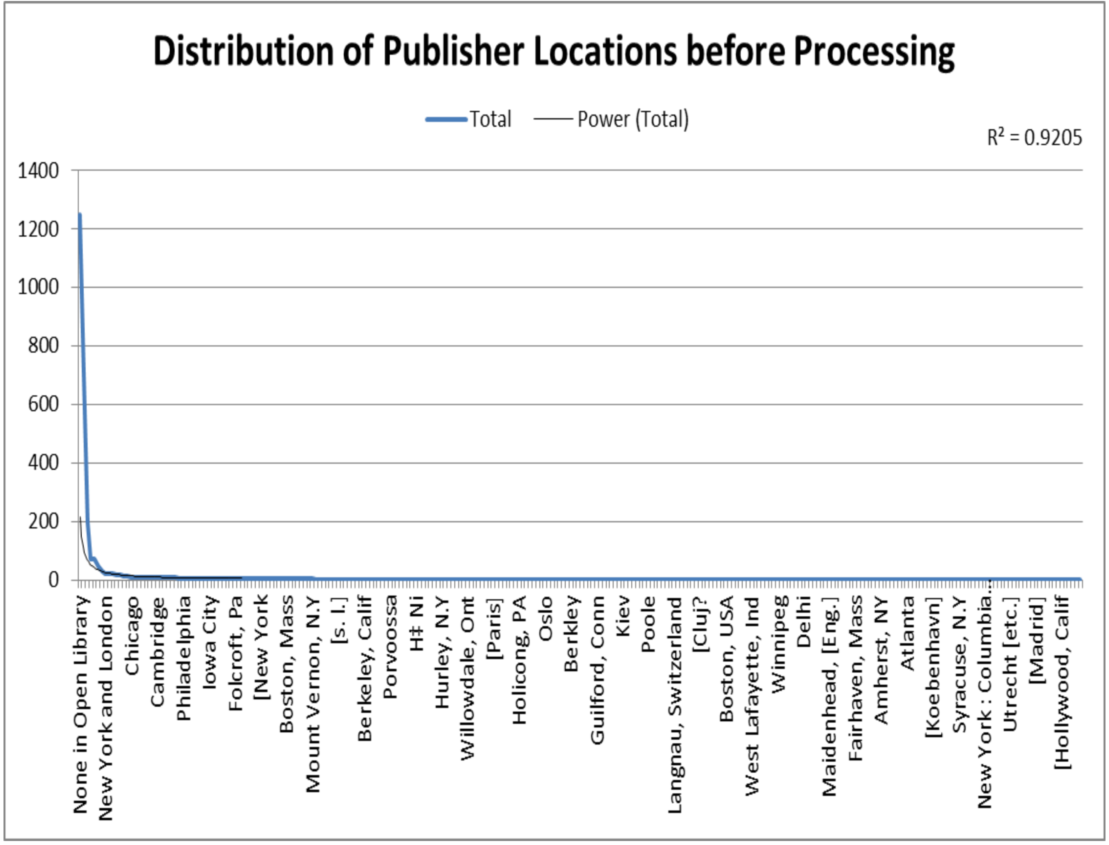


Figure 19b – A chart of the distribution of publisher location metadata presented above after it has been subjected to a process of standardization using Open Refine.

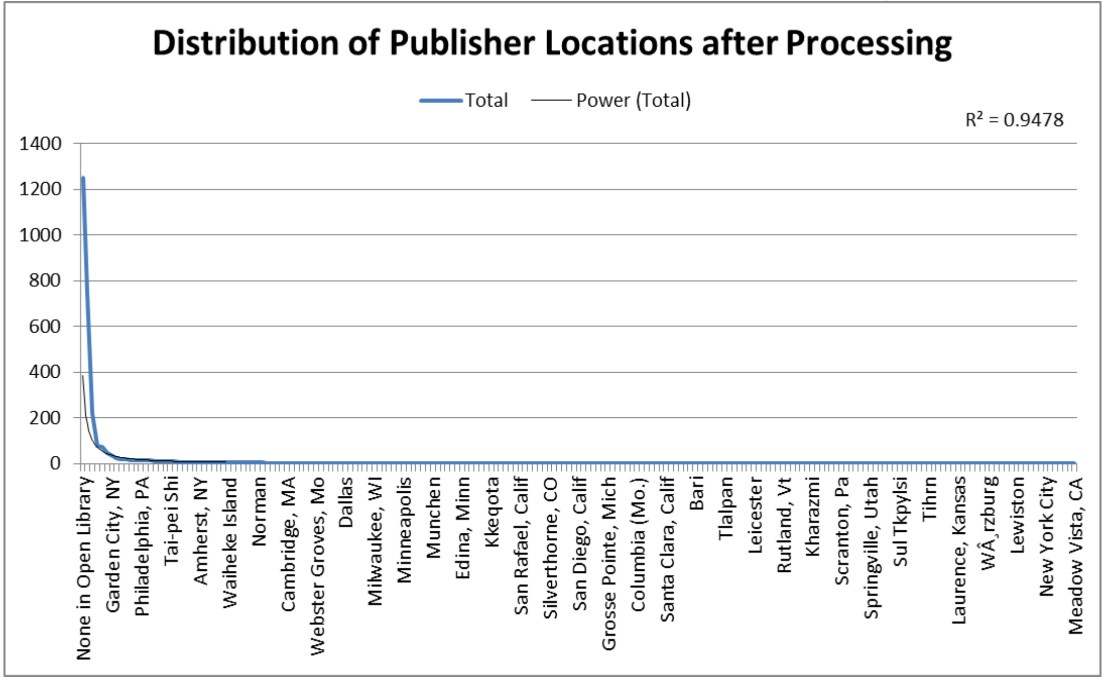


Figure 20a – A chart of the distribution of a subset of the publisher location metadata for a subset of records from Open Library with "Mark Twain" in the author field (records without location data have been removed).

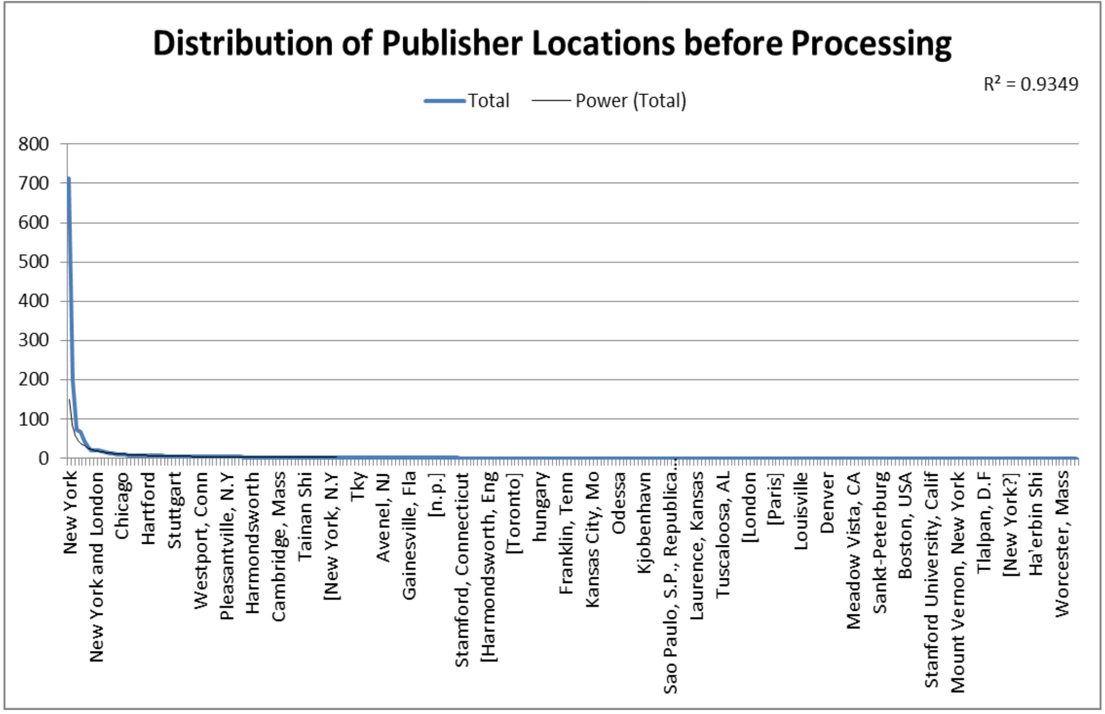


Figure 20b – A chart of the distribution of publisher location metadata presented above after it has been subjected to a process of standardization using Open Refine.

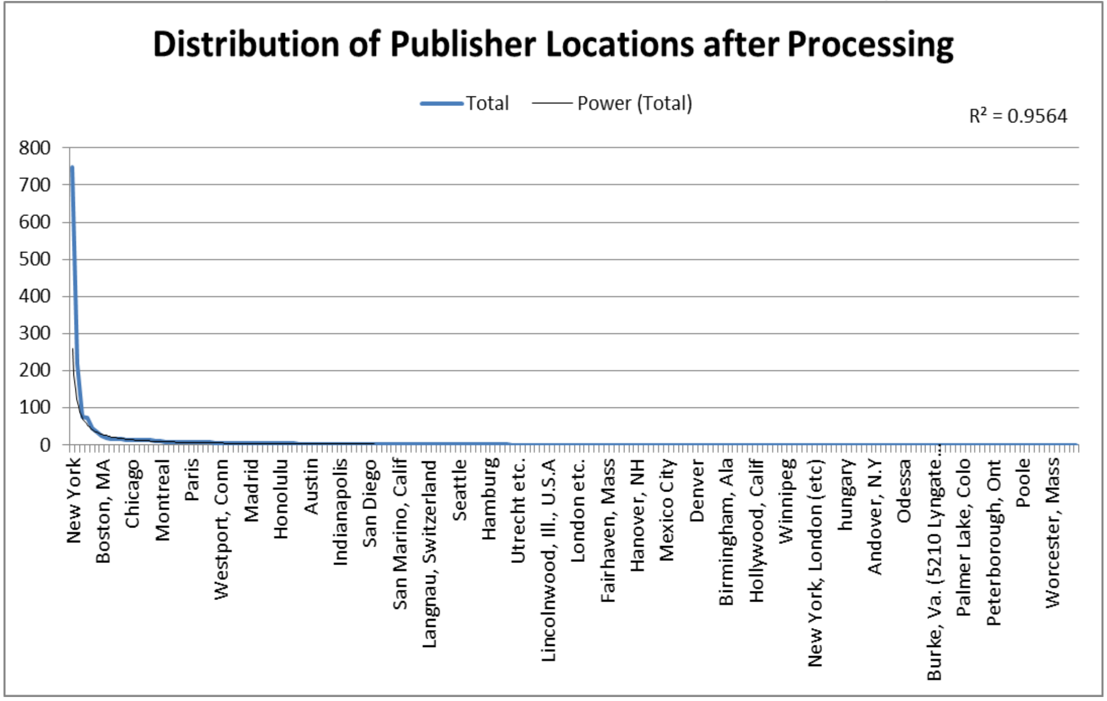


Figure 21a – A chart of the distribution of publication date metadata for 2998 records from Open Library with "Mark Twain" in the author field.

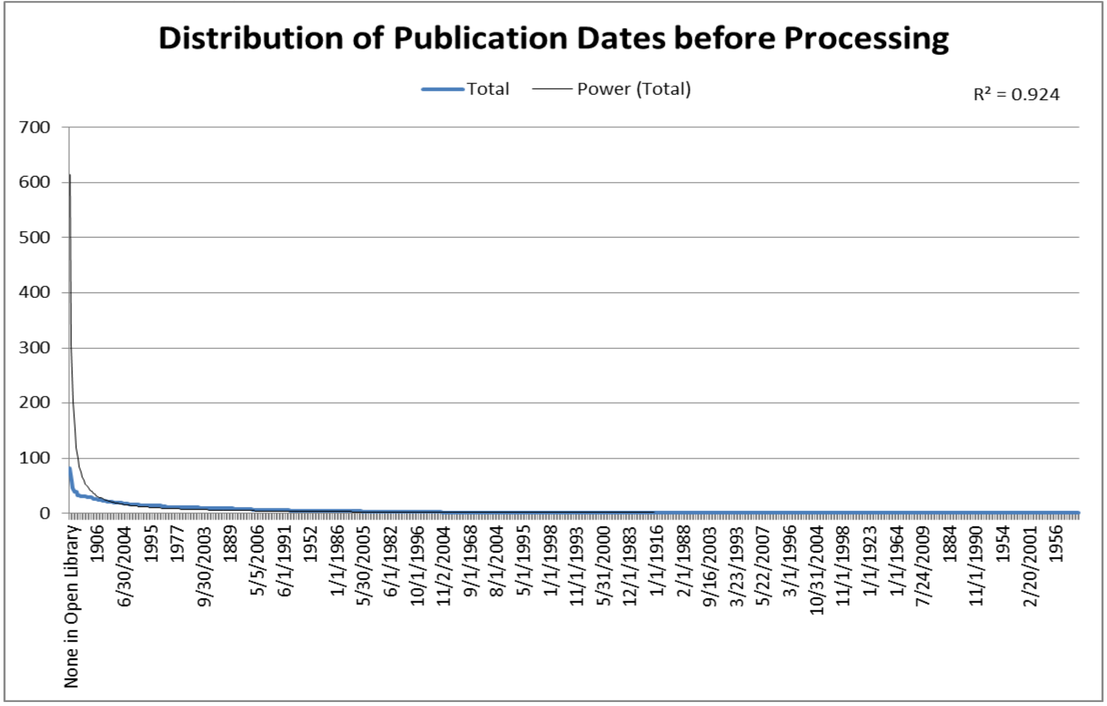


Figure 21b – A chart of the distribution of publication date metadata presented above after it has been subjected to a process of standardization using Open Refine.

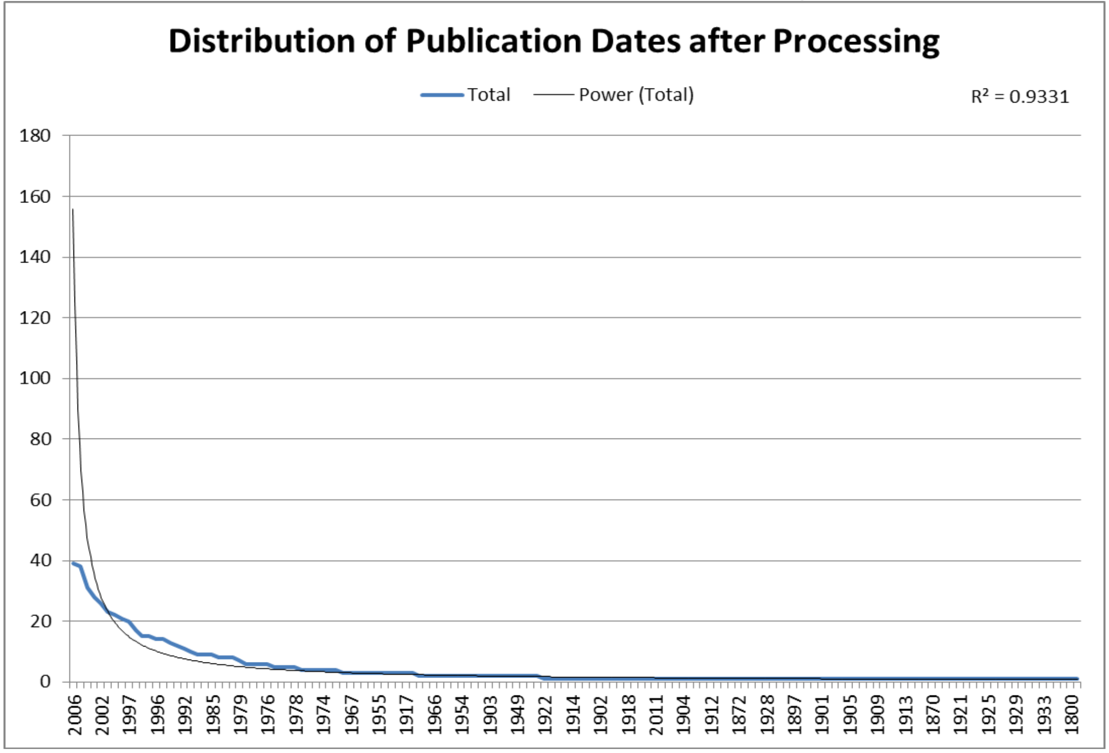


Figure 22a – A chart of the distribution of publisher names metadata for 2998 records from Open Library with “Mark Twain” in the author field.

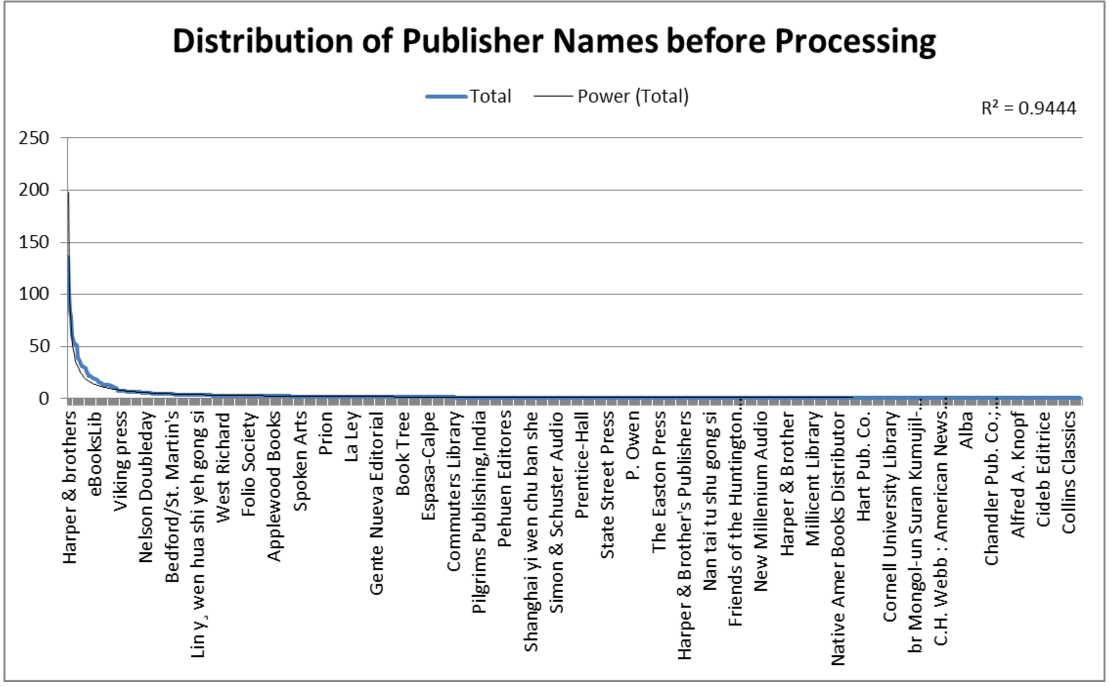


Figure 22b – A chart of the distribution of publisher names metadata presented above after it has been subjected to a process of standardization using Open Refine.

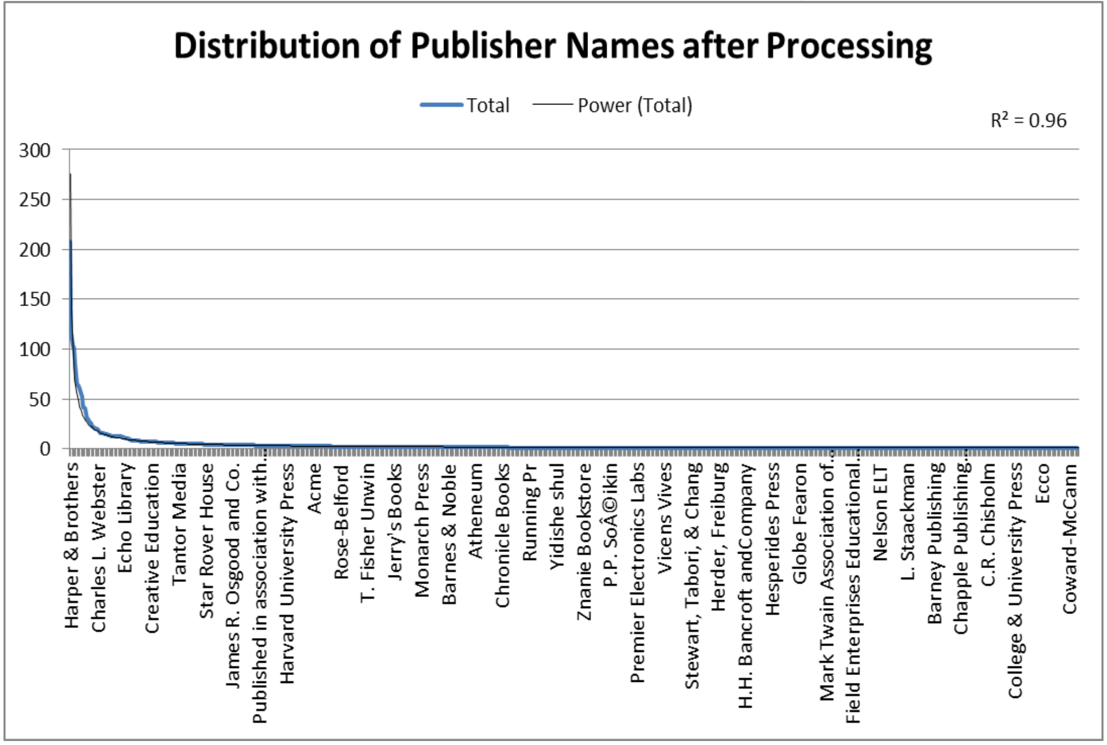


Figure 23 – A view of the model world with an inset close-up showing select agents and their topics

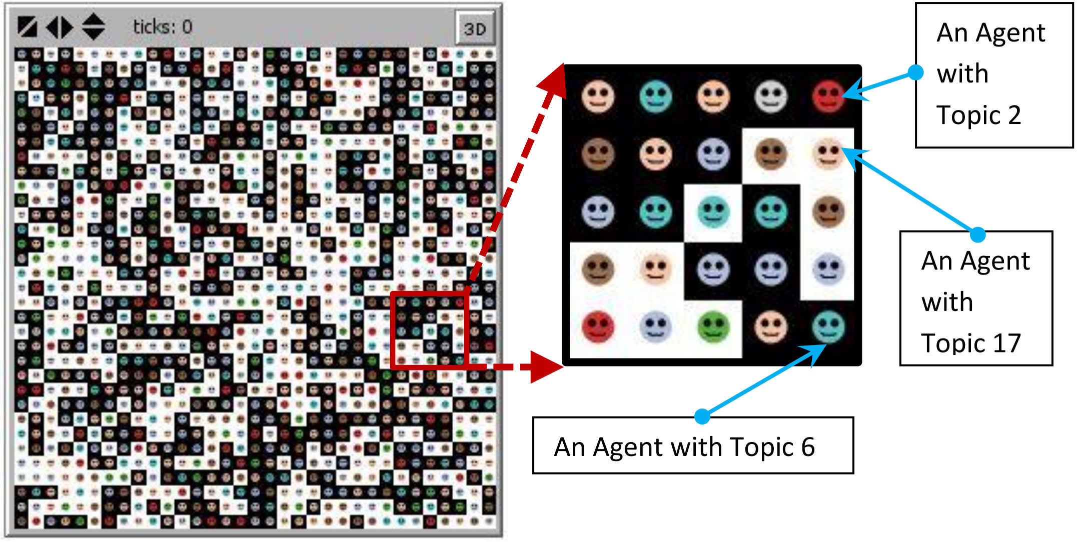


Figure 24 -- % of Turtles Assigned to Each Topic During Ten Runs of the Model

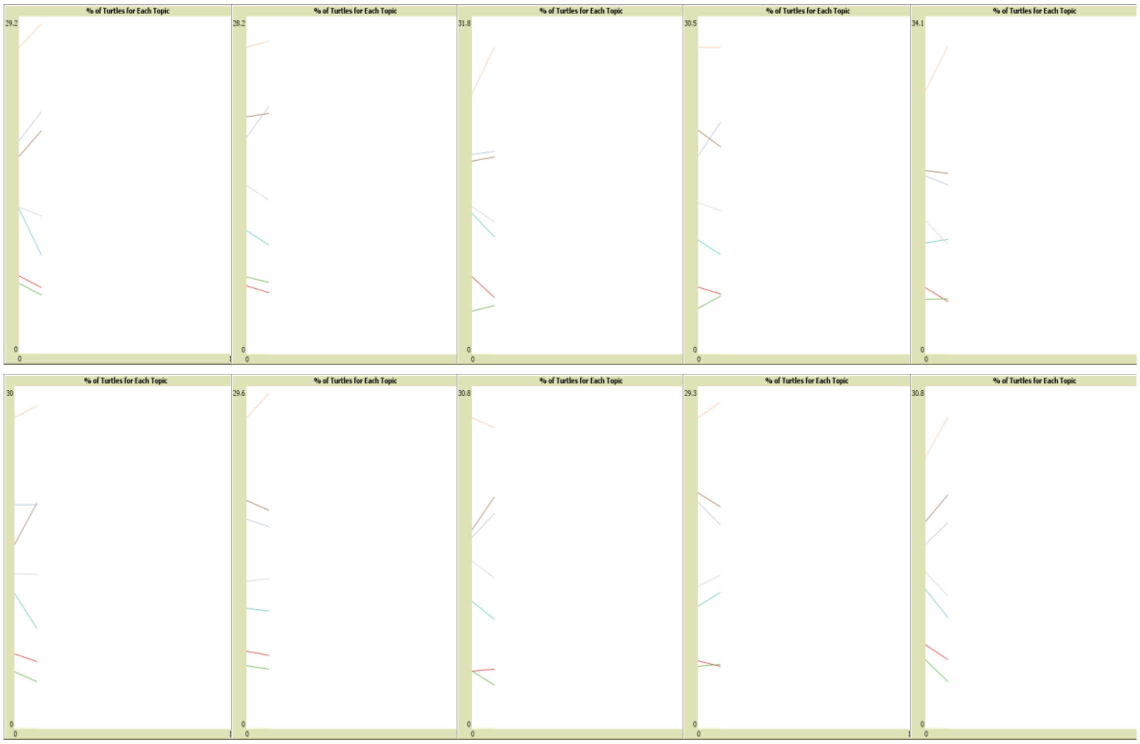

% of Turtles for Each Topic
29.2
0
0
% of Turtles for Each Topic
28.2
0
0
% of Turtles for Each Topic
31.8
0
0
% of Turtles for Each Topic
30.5
0
0
% of Turtles for Each Topic
34.1
0
0
% of Turtles for Each Topic
30
0
0
% of Turtles for Each Topic
29.6
0
0
% of Turtles for Each Topic
30.8
0
0
% of Turtles for Each Topic
29.3
0
0
% of Turtles for Each Topic
30.8
0
0

Figure 25 -- Distribution of Agents Interested in Topic 17 Over 10,000 Runs of the Model

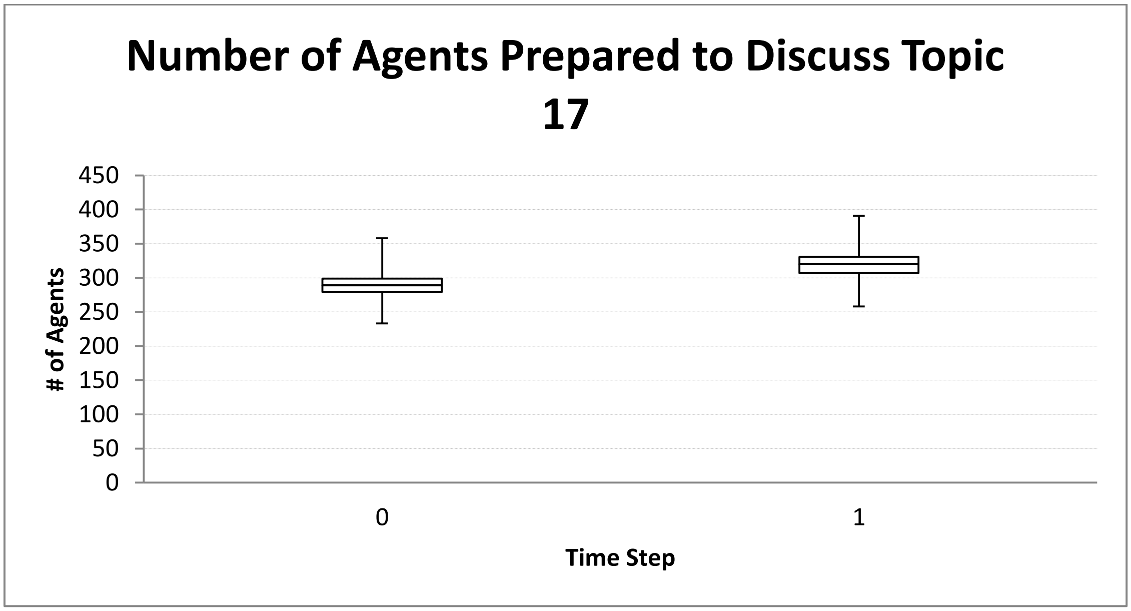


Figure 26 -- Distribution of Decrease in Attention to Topic 17 from Initial Value after

1 Time Step Over 10,000 Runs of the Model

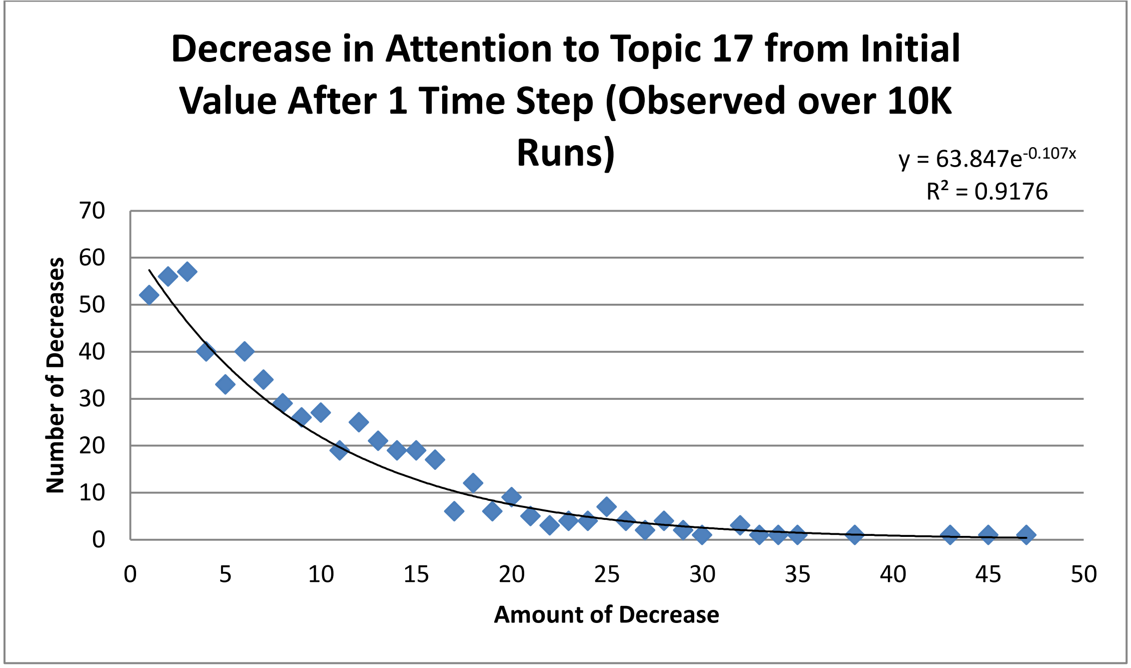


Figure 27 -- Conversation Dynamics Produced by Modeling Topic Data from January 1907 over 1000 time steps (with random seeds 1 to 10, inclusive)

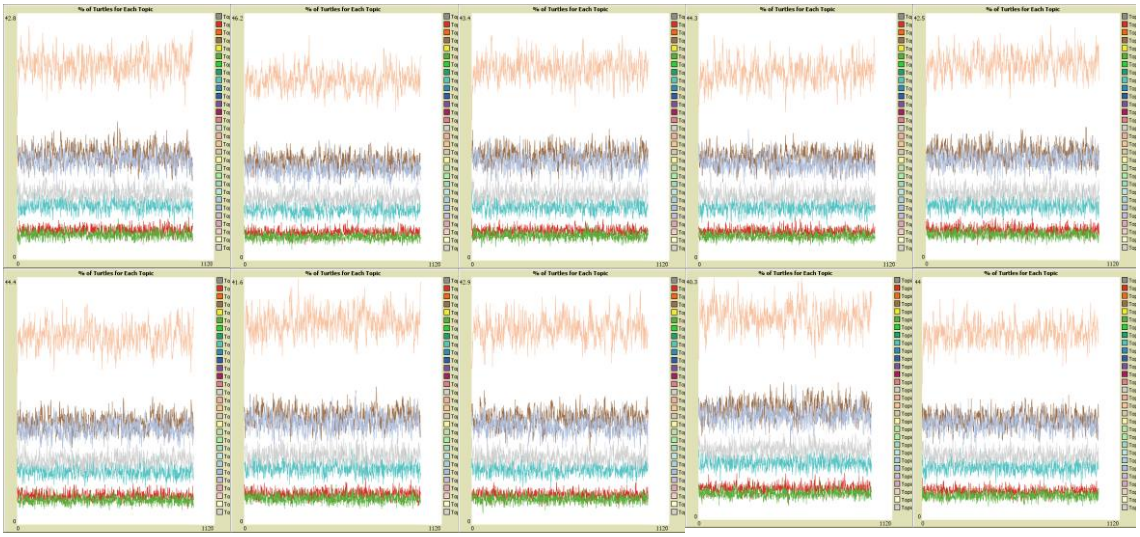


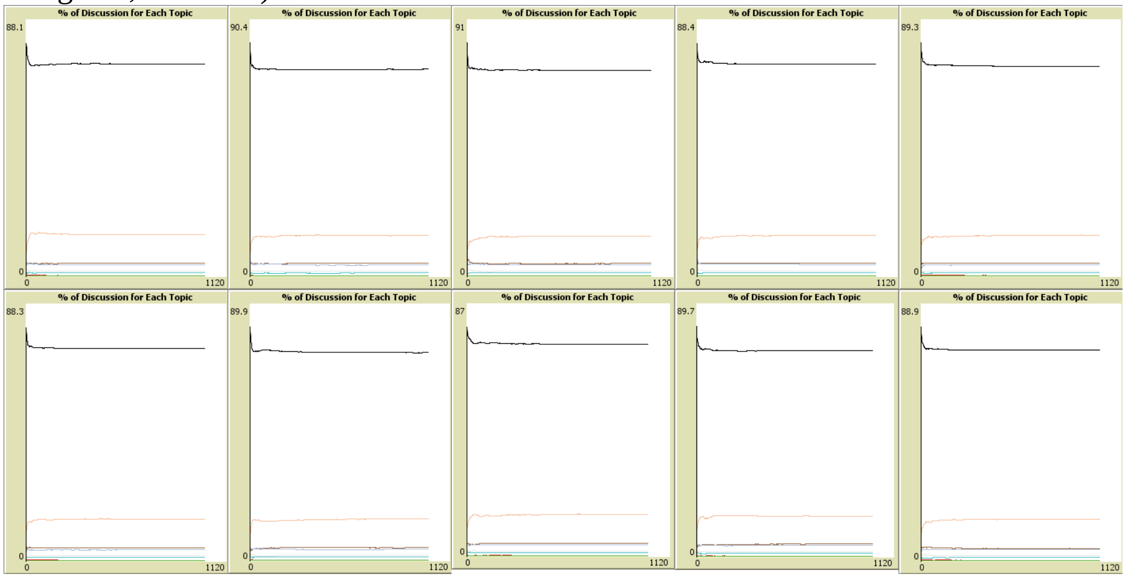
Figure 28 – A Summary of the Distribution of Media Coverage Suggested by Topic Modeling; and, the Minimum, Maximum, and Average % of Agents Discussing Each Topic Over 10 Runs of the Model to 1000 Time Steps (random seeds 1 to 10, inclusive)

	% of Corpus (determined by topic model)	Min % of Agents with Topic	Max % of Agents with Topic	Avg. % of Agents with Topic	Avg. # of Agents with Topic	SD (of pop.)
Topic 0	13.31	7.35	16.35	11.60	126.35	11.98
Topic 2	6.49	2.66	7.81	5.12	55.78	7.64
Topic 4	18.78	12.76	24.06	18.35	199.87	15.71
Topic 6	5.6	2.02	7.16	4.37	47.57	7.05
Topic 9	11.24	6.06	13.22	9.48	103.27	10.86
Topic 17	26.57	24.52	41.97	33.83	368.43	22.08
Topic 25	18.02	12.03	23.51	17.24	187.72	15.17

Figure 29 -- A Summary of Conversation Data Over Ten Runs of the Model to 1000 Time Steps (with random seeds 1 through 10, inclusive)

	% of Corpus (determined by topic model)	Minimum Discussion Level		Maximum Discussion Level		Average Discussion Level		
		% of Total Dis.	Count	% of Total Dis.	Count	% of Total Dis.	Count	SD (of pop.)
Topic 0	13.31	1.89	20532	1.94	21089	1.90	20741.10	154.10
Topic 2	6.49	0.37	4051	0.39	4223	0.38	4143.70	52.32
Topic 4	18.78	4.50	49012	4.70	51130	4.63	50416.70	583.30
Topic 6	5.6	0.27	2897	0.29	3135	0.28	3021.70	80.69
Topic 9	11.24	1.27	13793	1.31	14309	1.28	13956.80	134.01
Topic 17	26.57	14.32	155955	14.67	159804	14.46	157503.90	1113.41
Topic 25	18.02	3.99	43422	4.16	45251	4.09	44544.00	547.42
No Discussion		72.84	793201	73.06	795571	72.97	794672.10	692.03
Total Active Discussion		26.94	293429	27.16	295799	27.03	294327.90	692.03

Figure 30 -- % of the Total Discussion Devoted to Each Topic and to Failed Discussions Over Ten Runs of the Model to 1000 Time Steps (with random seeds 1 through 10, inclusive)



The figure above shows the percentage of the total discussion devoted to each topic and to failed discussions at each time step for ten runs of the model. The dark line at the top of each chart illustrates the high percentage of failed attempts at conversation throughout the space at each time step. Discussion of Topic 17 is clearly visible in the lower quarter of the chart, while the remaining topics cluster near the bottom.

Figure 31 -- % of the Corpus of Media Coverage Attributed to Each Topic by Topic Modeling Compared with the Maximum % of Total Discussion and the Maximum % of Agents Engaged with Each Topic Over Ten Runs of the Model to 1000 Time Steps (with random seeds 1 through 10, inclusive)

	% of Corpus (determined by topic model)	Max. % of Total Dis.	Max % of Agents with Topic
Topic0	13.31	1.94	16.35
Topic2	6.49	0.39	7.81
Topic4	18.78	4.70	24.06
Topic6	5.6	0.29	7.16
Topic9	11.24	1.31	13.22
Topic17	26.57	14.67	41.97
Topic25	18.02	4.16	23.51

Figure 32a -- Maximum % of Agents Reached Per Month by Each Topic over 100 Runs of the Model (at 500 ticks per month)

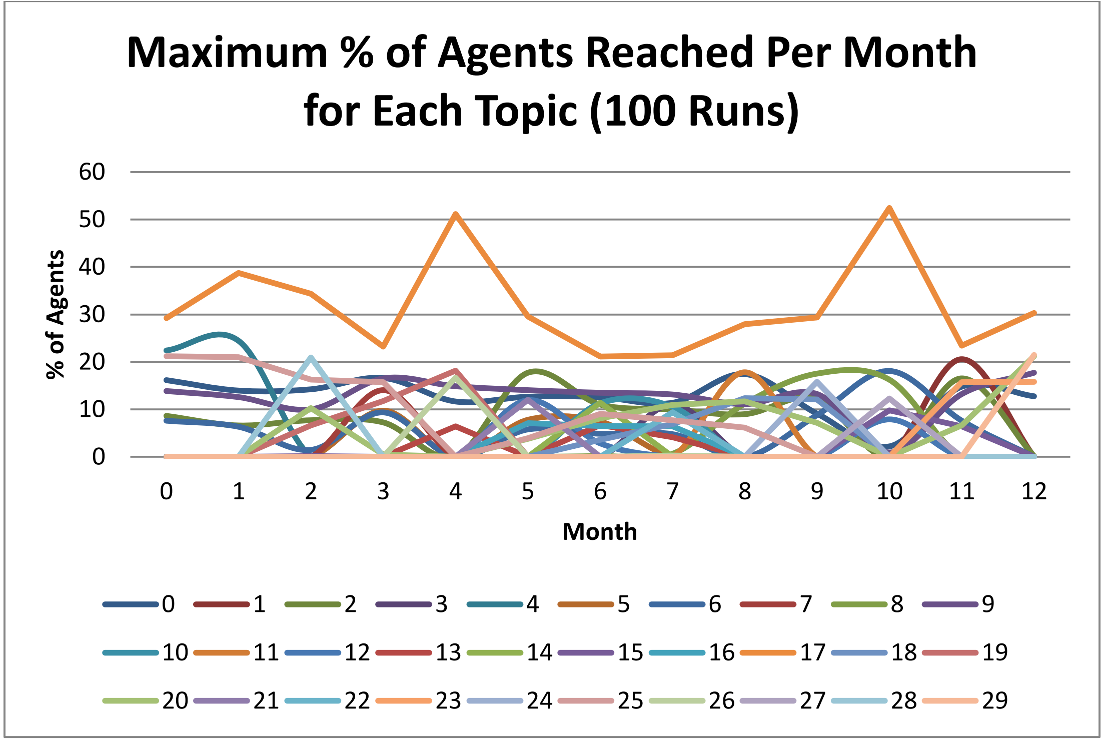


Figure 32b -- % of Total Discussion Devoted to Each Topic Over 100 Runs of the Model (at 500 ticks per month)

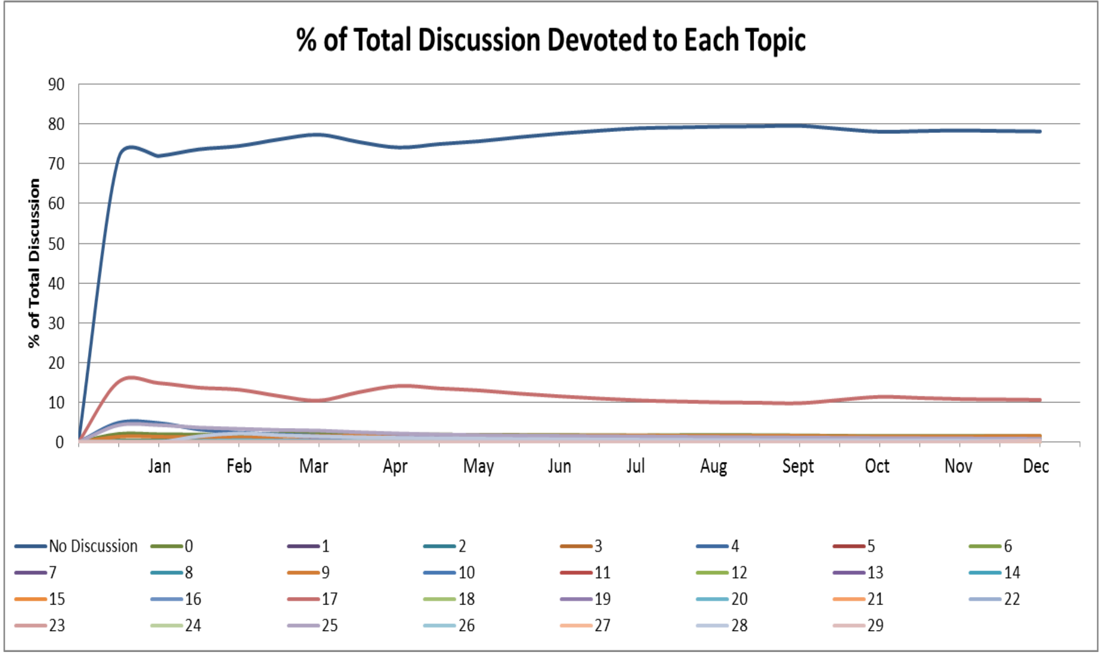


Figure 33a -- % of Turtles for Each Topic in the Environment Over the Course of One Run of the Model (with random seed value of 44)

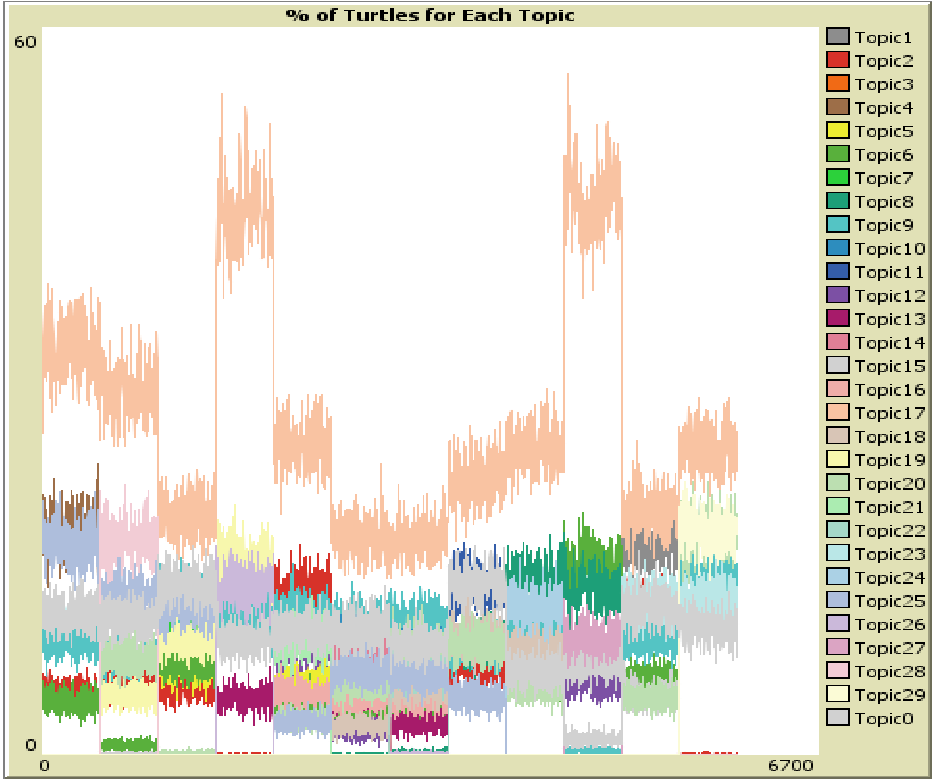


Figure 33b -- % of Discussion for Each Topic Over the Course of One Run of the Model (with random seed value of 44)

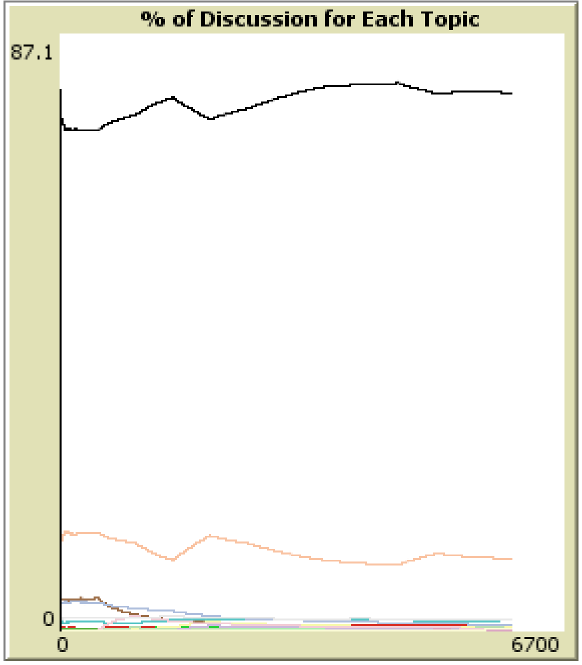


Figure 34 -- Topics in the Environment During the First Three Steps of a Run of the Model (with Random seed 44) as Seen from the Perspective of Three Different Agents

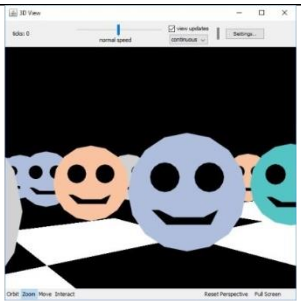
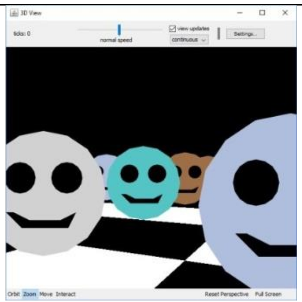
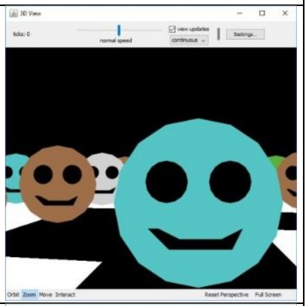
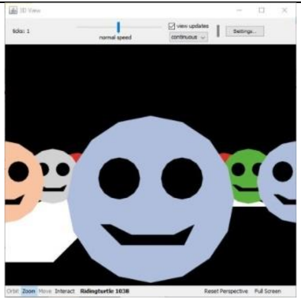
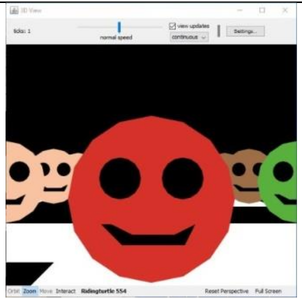
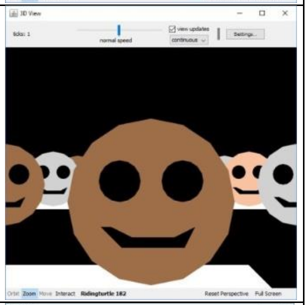
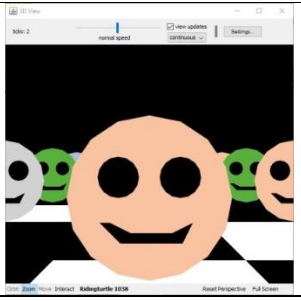
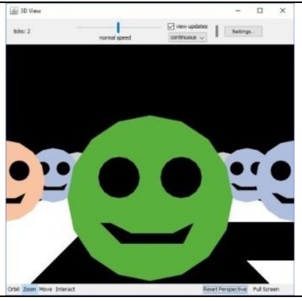
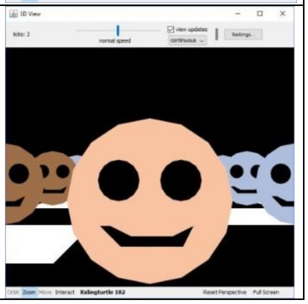
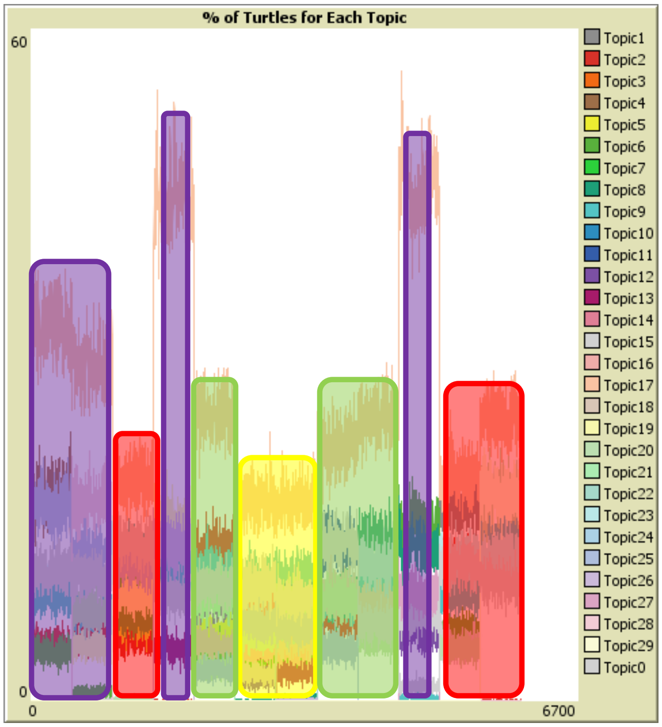
Tick #	Perspective When Traveling with Turtle 1038	Perspective When Traveling with Turtle 554	Perspective When Traveling with Turtle 182
0	 3D View ticks: 0 normal speed ☑ view updates continuous Settings... Orbit Zoom Move Interact Reset Perspective Full Screen	 3D View ticks: 0 normal speed ☑ view updates continuous Settings... Orbit Zoom Move Interact Reset Perspective Full Screen	 3D View ticks: 0 normal speed ☑ view updates continuous Settings... Orbit Zoom Move Interact Reset Perspective Full Screen
1	 3D View ticks: 1 normal speed ☑ view updates continuous Settings... Orbit Zoom Move Interact Ridingturtle 1038 Reset Perspective Full Screen	 3D View ticks: 1 normal speed ☑ view updates continuous Settings... Orbit Zoom Move Interact Ridingturtle 554 Reset Perspective Full Screen	 3D View ticks: 1 normal speed ☑ view updates continuous Settings... Orbit Zoom Move Interact Ridingturtle 182 Reset Perspective Full Screen
2	 3D View ticks: 2 normal speed ☑ view updates continuous Settings... Orbit Zoom Move Interact Ridingturtle 1038 Reset Perspective Full Screen	 3D View ticks: 2 normal speed ☑ view updates continuous Settings... Orbit Zoom Move Interact Reset Perspective Full Screen	 3D View ticks: 2 normal speed ☑ view updates continuous Settings... Orbit Zoom Move Interact Ridingturtle 182 Reset Perspective Full Screen

Figure 35 -- Conversations in the Model Grouped by Sound



Description of Observed Conversation Levels

🟨 Quiet conversations, characterized by low frequency topics (June, July)

🟥 Active conversations, mostly low frequencies (Mar, Nov, Dec)

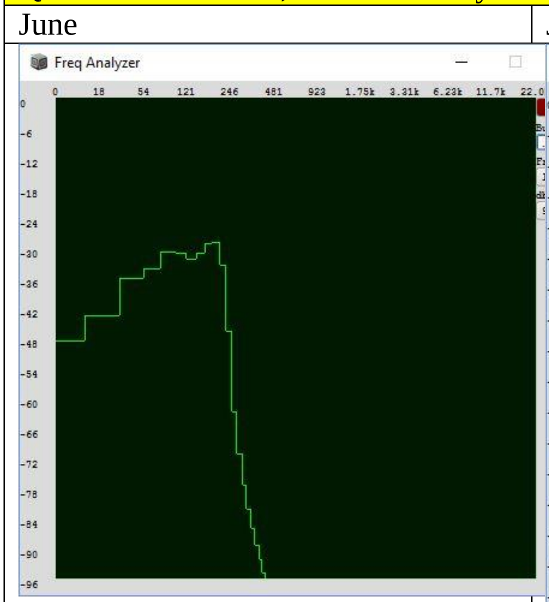
🟩 Active conversations, mostly high frequencies (May, Aug, Sept)

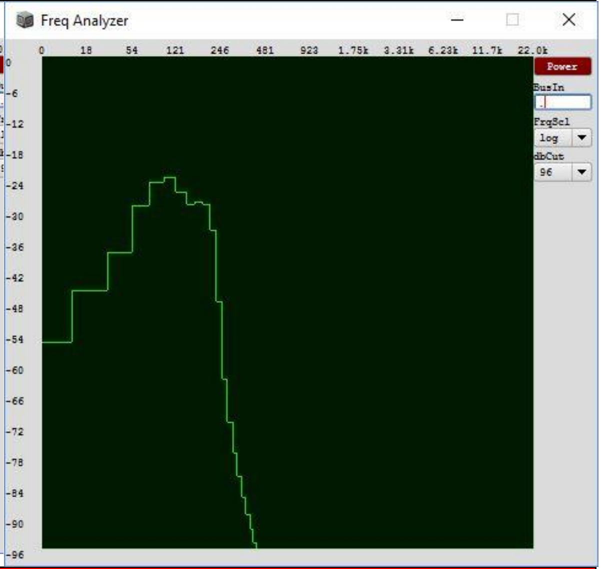
🟪 Loud conversations, characterized by high frequency topics (Jan, Feb, Apr, Oct)

Figure 36 -- Conversation Levels for Each Month as Captured by a Frequency Scope in SuperCollider

Quiet Conversations, characterize by low frequency topics

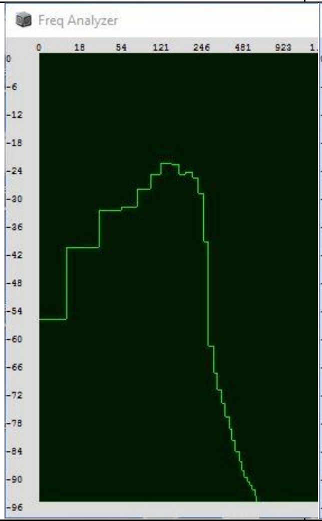
June | July

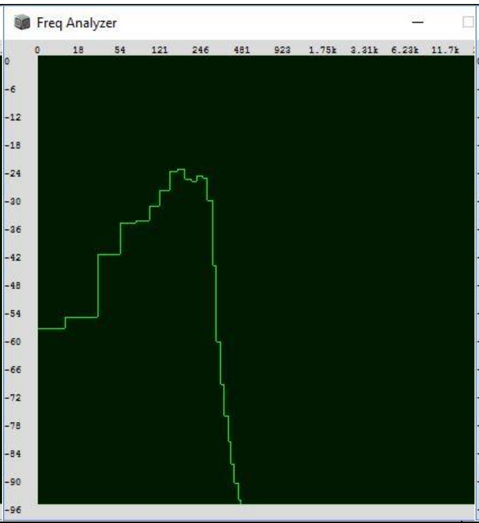


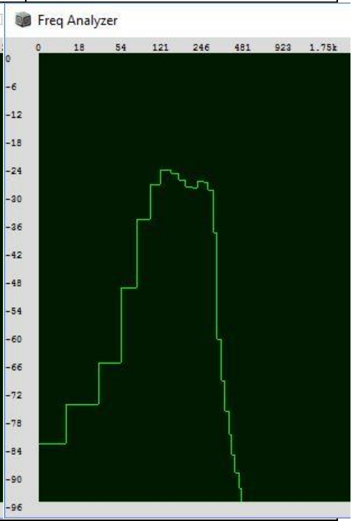


Active Conversations, mostly low frequencies

March | November | December







Active conversations, mostly high frequencies

May | August | September

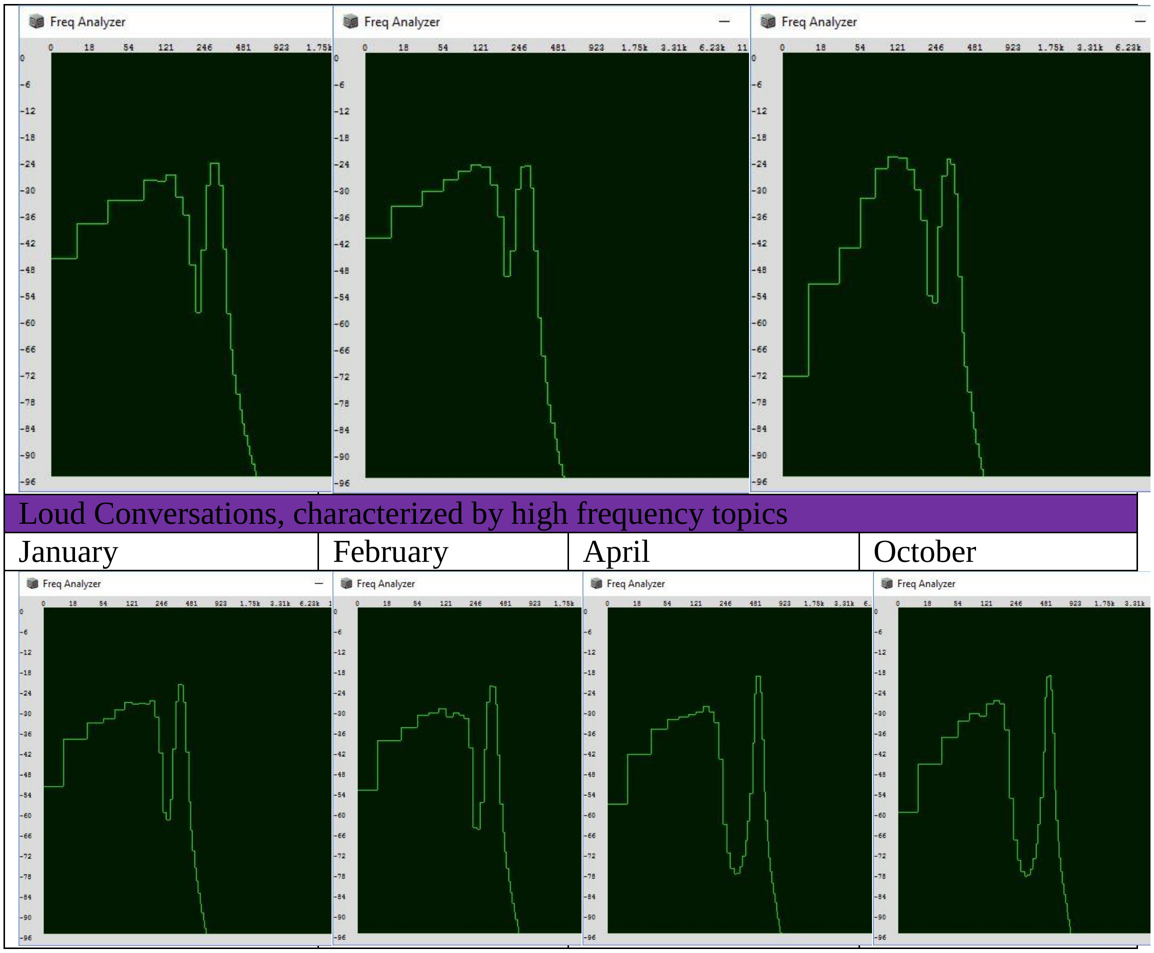


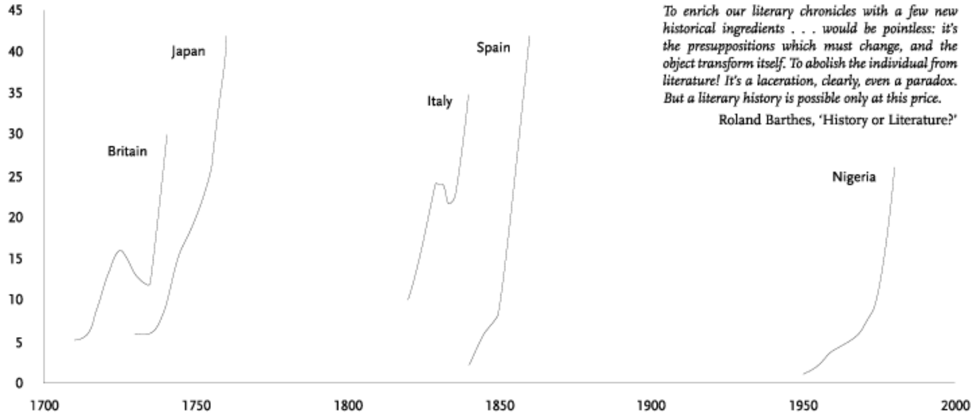
Figure 37 -- Video of the circulation of topics as seen from the perspective of an agent is available in the file below. In this video Turtle 564 navigates through the environment as topics from the month of January are discussed. (This footage can be reproduced by running the model with random seed to 44.)

ScreenCapture_2-2-2016 11.08.54 PM.wmv

Appendix for Chapter Three

Figure 1 – Franco Moretti on the Rise of the Novel

FIGURE 1: *The rise of the novel, 18th to 20th century*



To enrich our literary chronicles with a few new historical ingredients . . . would be pointless: it's the presuppositions which must change, and the object transform itself. To abolish the individual from literature! It's a laceration, clearly, even a paradox. But a literary history is possible only at this price.

Roland Barthes, 'History or Literature?'

New novels per year, by 5-year average. Sources: For Britain: W. H. McBurney, *A Check List of English Prose Fiction, 1700–39*, Cambridge, MA 1960, and J. C. Beasley, *The Novels of the 1740s*, Athens, GA 1982; both partly revised by James Raven, *British Fiction 1750–70: A Chronological Check-List of Prose Fiction Printed in Britain and Ireland*, London 1987. For Japan: Jonathan Zwicker, 'Il lungo Ottocento del romanzo giapponese', in *Il romanzo*, vol. III, *Storia e geografia*, Torino 2002. For Italy: Giovanni Ragone, 'Italia 1815–70', in *Il romanzo*, vol. III. For Spain: Elisa Marti-Lopez and Mario Santana, 'Spagna 1843–1900', *Il romanzo*, vol. III. For Nigeria: Wendy Griswold, 'Nigeria 1950–2000', *Il romanzo*, vol. III.

Figure 2 – Underwood and Goldstone: Two Perspectives on Topic Modeling PMLA

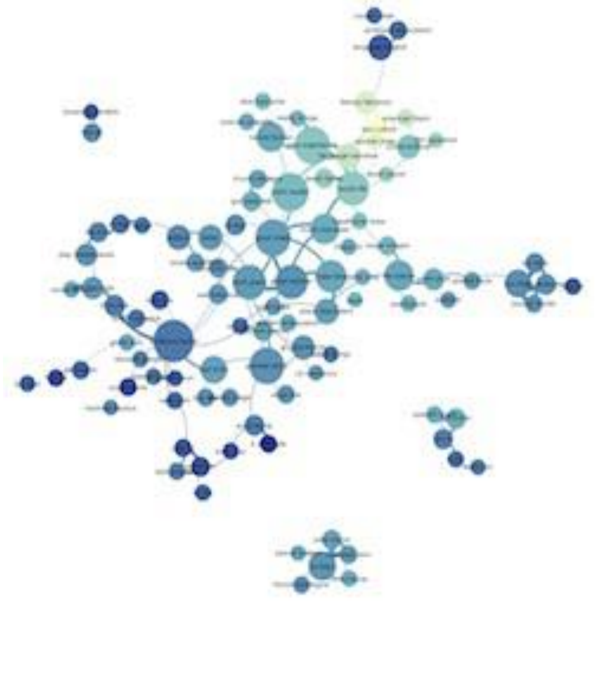
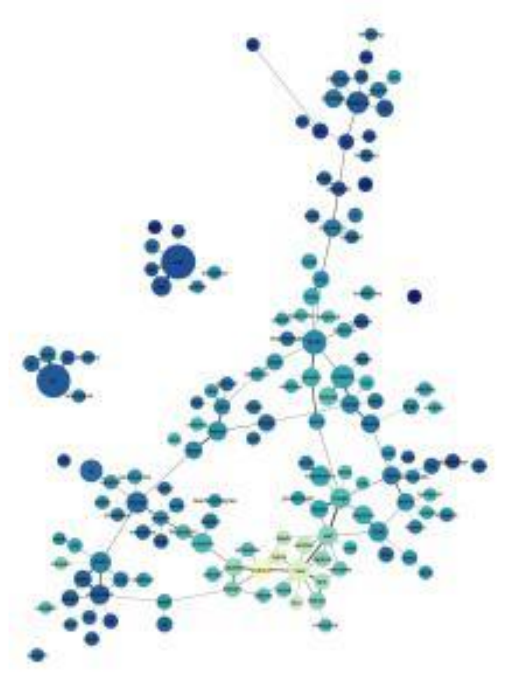
	
Goldstone's 100 topic model	Underwood's 150 topic model

Figure 3 -- Jockers and Mimno Observe Female Fashion

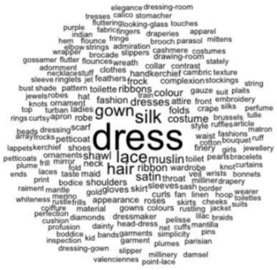
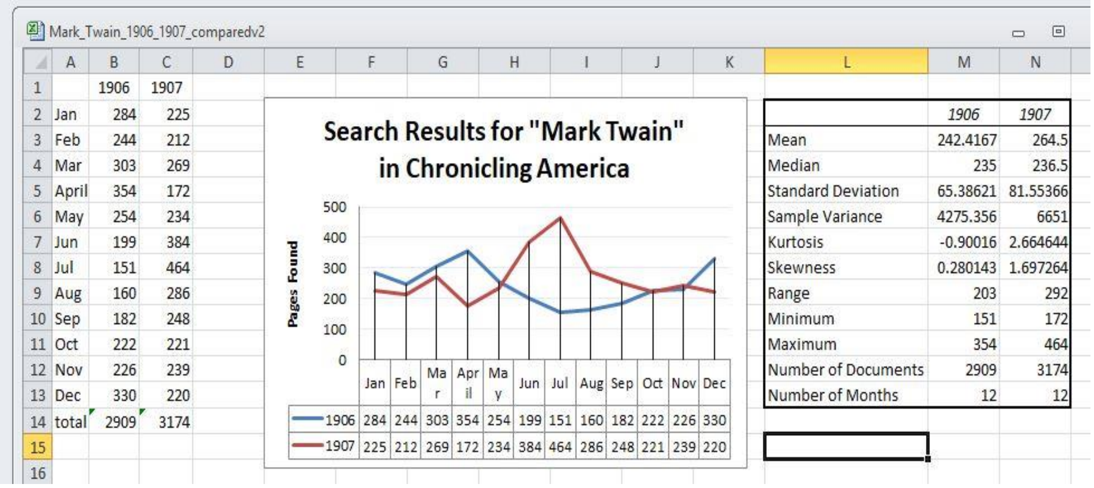

elegance dressing-room
tresses calico stomacher
fluttering looking-glass touches
purple fabric fingers draperies apparel
indian
hem flounce fringe brooch parasol mittens
elbow strings admiration
wrapper brocade slippers cashmere costumes
drawing-room
gossamer flutter flounces wreath stately
adornment collar contrast
necklace stuff clothes handkerchief cambric texture
sleeve ringlets jet feathers frock complexion stockings string
bust shade pattern toilette ribbons gauze suit plaits
jewels robes hat train colour
knots ornament fashion dresses attire front embroidery
top turban ladies gown silk folds crape silks perfume
rings curtsy apron robe costume brussels tulle
style ruffles article
beads dressing scarf matron
array frocks petticoat dress waist fashions ruff
cotton bouquet
lappets kerchief shoes finery girls jewellery
ornaments
petticoats shawl lace muslin toilet pearls bracelets
frill mirror neck
plume fold knot curtains
ends laces taste maid hair ribbon wardrobe veil wrists bonnets
print satin throat milliner drapery
bodice shoulders
raiment mantle gold gloves skirt sleeves sash border
merino curls fan linen hoop wearer
whiteness rustle frills appearance roses skirts cheeks toilettes
coiffure material gowns colours rustling jacket suits
perfection diamonds dressmaker lilac braids
cushion pelisse
profusion dainty head-dress net cuffs mantilla
boddice kid bands garments simplicity pins
inspection garment plumes parisian
dressing-gown slipper millinery damsel
valenciennes point-lace

Figure 3: Word cloud of topic labeled "Female Fashion."

Figure 4 – A Comparison of the Distribution of Search Results for "Mark Twain" over 1906 and 1907 in Chronicling America

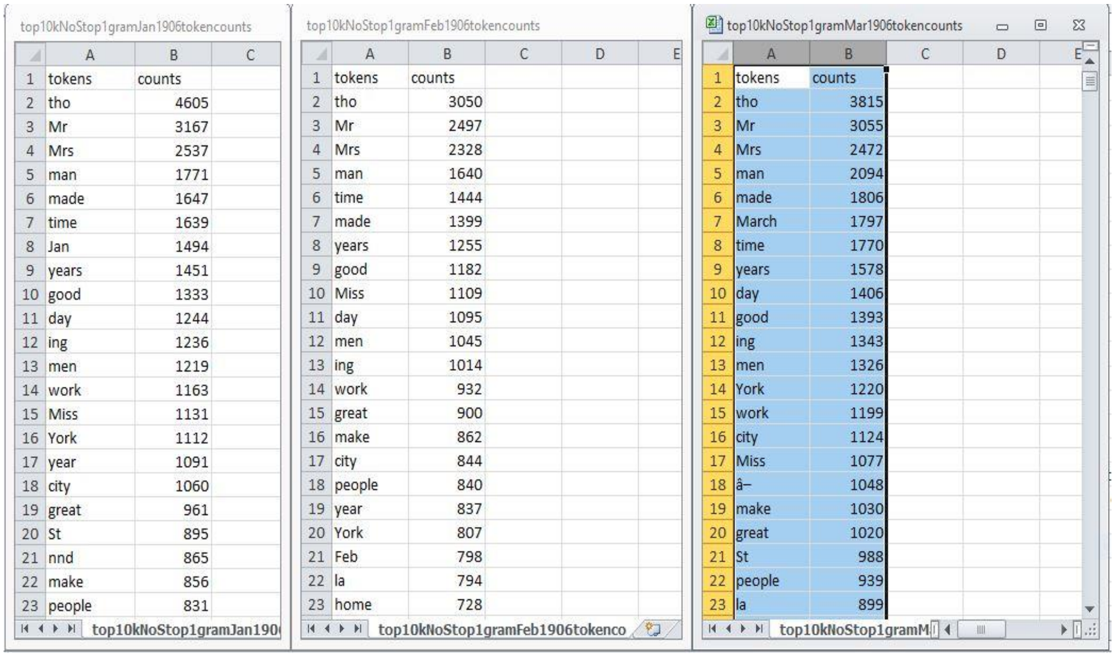

Mark_Twain_1906_1907_comparedv2
Search Results for "Mark Twain" in Chronicling America
Pages Found
500
400
300
200
100
0
Jan
Feb
Mar
April
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
1906
1907

	A	B	C
1		1906	1907
2	Jan	284	225
3	Feb	244	212
4	Mar	303	269
5	April	354	172
6	May	254	234
7	Jun	199	384
8	Jul	151	464
9	Aug	160	286
10	Sep	182	248
11	Oct	222	221
12	Nov	226	239
13	Dec	330	220
14	total	2909	3174

	Jan	Feb	Mar	April	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1906	284	244	303	354	254	199	151	160	182	222	226	330
1907	225	212	269	172	234	384	464	286	248	221	239	220

	1906	*1907*
Mean	242.4167	264.5
Median	235	236.5
Standard Deviation	65.38621	81.55366
Sample Variance	4275.356	6651
Kurtosis	-0.90016	2.664644
Skewness	0.280143	1.697264
Range	203	292
Minimum	151	172
Maximum	354	464
Number of Documents	2909	3174
Number of Months	12	12

Figure 5 – Most common 1-grams in media coverage related to "Mark Twain" during the first three months of 1906



top10kNoStop1gramJan1906tokencounts

	A	B	C
1	tokens	counts	
2	tho	4605	
3	Mr	3167	
4	Mrs	2537	
5	man	1771	
6	made	1647	
7	time	1639	
8	Jan	1494	
9	years	1451	
10	good	1333	
11	day	1244	
12	ing	1236	
13	men	1219	
14	work	1163	
15	Miss	1131	
16	York	1112	
17	year	1091	
18	city	1060	
19	great	961	
20	St	895	
21	nnd	865	
22	make	856	
23	people	831	

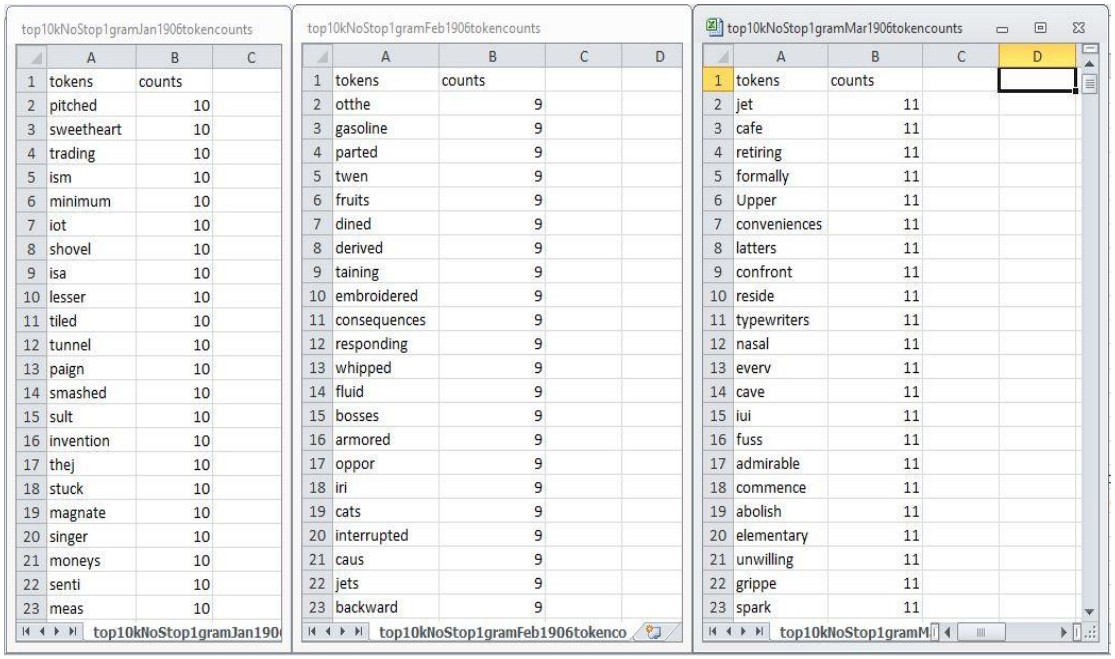
top10kNoStop1gramFeb1906tokencounts

	A	B	C	D	E
1	tokens	counts			
2	tho	3050			
3	Mr	2497			
4	Mrs	2328			
5	man	1640			
6	time	1444			
7	made	1399			
8	years	1255			
9	good	1182			
10	Miss	1109			
11	day	1095			
12	men	1045			
13	ing	1014			
14	work	932			
15	great	900			
16	make	862			
17	city	844			
18	people	840			
19	year	837			
20	York	807			
21	Feb	798			
22	la	794			
23	home	728			

top10kNoStop1gramMar1906tokencounts

	A	B	C	D	E
1	tokens	counts			
2	tho	3815			
3	Mr	3055			
4	Mrs	2472			
5	man	2094			
6	made	1806			
7	March	1797			
8	time	1770			
9	years	1578			
10	day	1406			
11	good	1393			
12	ing	1343			
13	men	1326			
14	York	1220			
15	work	1199			
16	city	1124			
17	Miss	1077			
18	â–	1048			
19	make	1030			
20	great	1020			
21	St	988			
22	people	939			
23	la	899			

Figure 6 – Less common 1-grams in media coverage related to "Mark Twain" during the first three months of 1906



top10kNoStop1gramJan1906tokencounts

	A	B	C
1	tokens	counts	
2	pitched	10	
3	sweetheart	10	
4	trading	10	
5	ism	10	
6	minimum	10	
7	iot	10	
8	shovel	10	
9	isa	10	
10	lesser	10	
11	tiled	10	
12	tunnel	10	
13	paign	10	
14	smashed	10	
15	sult	10	
16	invention	10	
17	thej	10	
18	stuck	10	
19	magnate	10	
20	singer	10	
21	moneys	10	
22	senti	10	
23	meas	10	

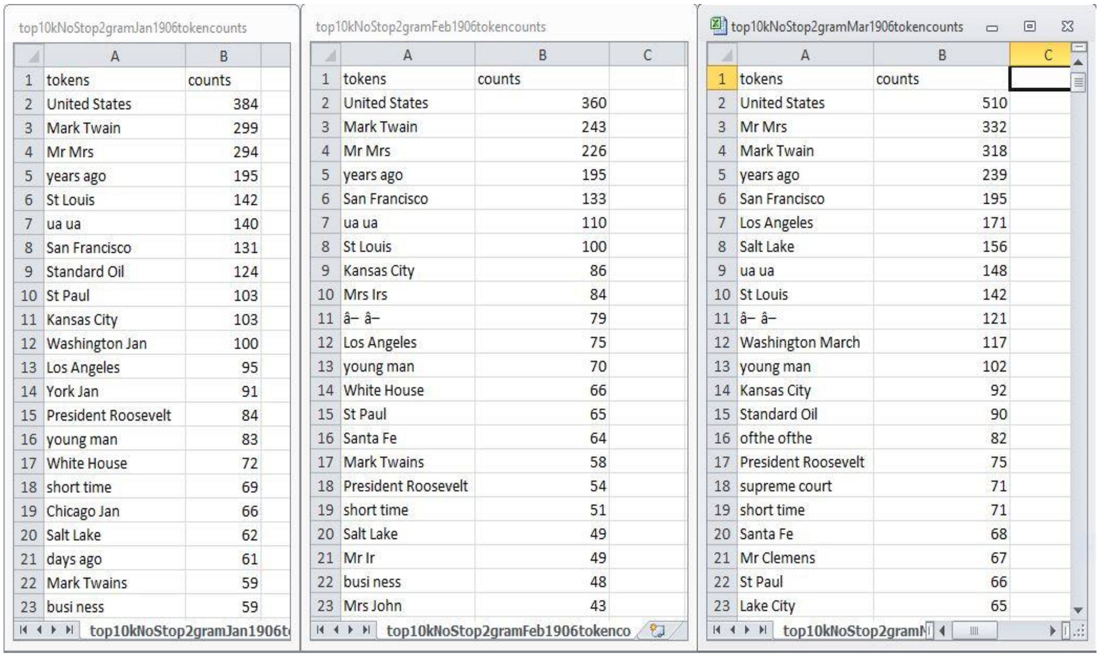
top10kNoStop1gramFeb1906tokencounts

	A	B	C	D
1	tokens	counts		
2	otthe	9		
3	gasoline	9		
4	parted	9		
5	twen	9		
6	fruits	9		
7	dined	9		
8	derived	9		
9	taining	9		
10	embroidered	9		
11	consequences	9		
12	responding	9		
13	whipped	9		
14	fluid	9		
15	bosses	9		
16	armored	9		
17	oppor	9		
18	iri	9		
19	cats	9		
20	interrupted	9		
21	caus	9		
22	jets	9		
23	backward	9		

top10kNoStop1gramMar1906tokencounts

	A	B	C	D
1	tokens	counts		
2	jet	11		
3	cafe	11		
4	retiring	11		
5	formally	11		
6	Upper	11		
7	conveniences	11		
8	latters	11		
9	confront	11		
10	reside	11		
11	typewriters	11		
12	nasal	11		
13	everv	11		
14	cave	11		
15	iui	11		
16	fuss	11		
17	admirable	11		
18	commence	11		
19	abolish	11		
20	elementary	11		
21	unwilling	11		
22	grippe	11		
23	spark	11		

Figure 7 -- Most common 2-grams in media coverage related to “Mark Twain” during the first three months of 1906



top10kNoStop2gramJan1906tokencounts

	A	B
1	tokens	counts
2	United States	384
3	Mark Twain	299
4	Mr Mrs	294
5	years ago	195
6	St Louis	142
7	ua ua	140
8	San Francisco	131
9	Standard Oil	124
10	St Paul	103
11	Kansas City	103
12	Washington Jan	100
13	Los Angeles	95
14	York Jan	91
15	President Roosevelt	84
16	young man	83
17	White House	72
18	short time	69
19	Chicago Jan	66
20	Salt Lake	62
21	days ago	61
22	Mark Twains	59
23	busi ness	59

top10kNoStop2gramJan1906t

top10kNoStop2gramFeb1906tokencounts

	A	B	C
1	tokens	counts	
2	United States	360	
3	Mark Twain	243	
4	Mr Mrs	226	
5	years ago	195	
6	San Francisco	133	
7	ua ua	110	
8	St Louis	100	
9	Kansas City	86	
10	Mrs Irs	84	
11	â– â–	79	
12	Los Angeles	75	
13	young man	70	
14	White House	66	
15	St Paul	65	
16	Santa Fe	64	
17	Mark Twains	58	
18	President Roosevelt	54	
19	short time	51	
20	Salt Lake	49	
21	Mr Ir	49	
22	busi ness	48	
23	Mrs John	43	

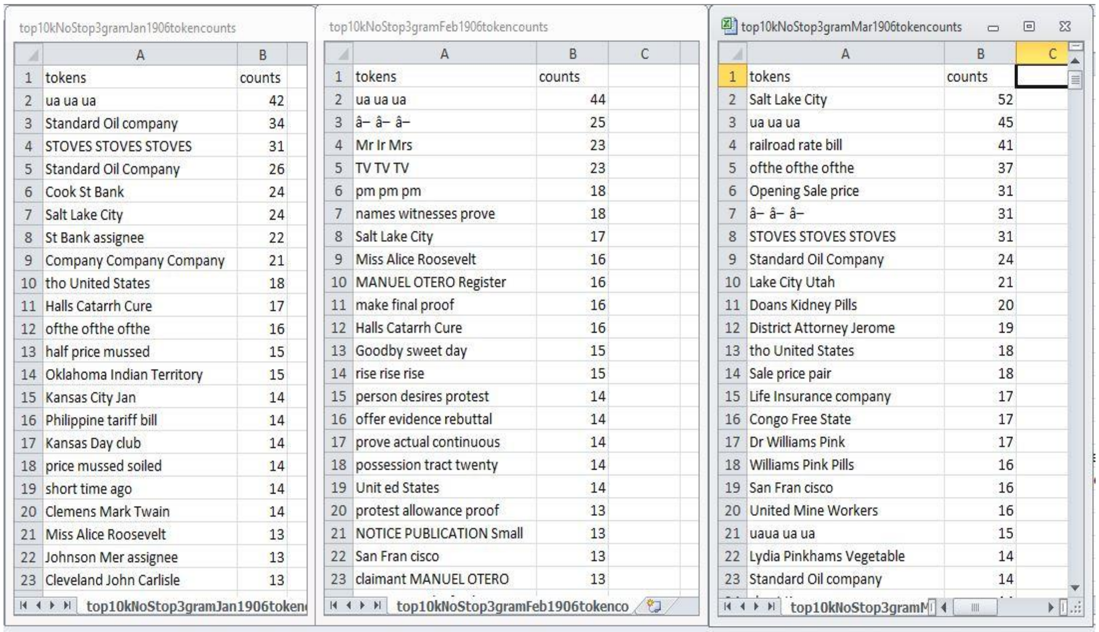
top10kNoStop2gramFeb1906tokenco

top10kNoStop2gramMar1906tokencounts

	A	B	C
1	tokens	counts	
2	United States	510	
3	Mr Mrs	332	
4	Mark Twain	318	
5	years ago	239	
6	San Francisco	195	
7	Los Angeles	171	
8	Salt Lake	156	
9	ua ua	148	
10	St Louis	142	
11	â– â–	121	
12	Washington March	117	
13	young man	102	
14	Kansas City	92	
15	Standard Oil	90	
16	ofthe ofthe	82	
17	President Roosevelt	75	
18	supreme court	71	
19	short time	71	
20	Santa Fe	68	
21	Mr Clemens	67	
22	St Paul	66	
23	Lake City	65	

top10kNoStop2gramM

Figure 8 -- Most common 3-grams in media coverage related to “Mark Twain” during the first three months of 1906



top10kNoStop3gramJan1906tokencounts

	A	B
1	tokens	counts
2	ua ua ua	42
3	Standard Oil company	34
4	STOVES STOVES STOVES	31
5	Standard Oil Company	26
6	Cook St Bank	24
7	Salt Lake City	24
8	St Bank assignee	22
9	Company Company Company	21
10	tho United States	18
11	Halls Catarrh Cure	17
12	ofthe ofthe ofthe	16
13	half price mussed	15
14	Oklahoma Indian Territory	15
15	Kansas City Jan	14
16	Philippine tariff bill	14
17	Kansas Day club	14
18	price mussed soiled	14
19	short time ago	14
20	Clemens Mark Twain	14
21	Miss Alice Roosevelt	13
22	Johnson Mer assignee	13
23	Cleveland John Carlisle	13

top10kNoStop3gramJan1906token

top10kNoStop3gramFeb1906tokencounts

	A	B	C
1	tokens	counts	
2	ua ua ua	44	
3	â– â– â–	25	
4	Mr Ir Mrs	23	
5	TV TV TV	23	
6	pm pm pm	18	
7	names witnesses prove	18	
8	Salt Lake City	17	
9	Miss Alice Roosevelt	16	
10	MANUEL OTERO Register	16	
11	make final proof	16	
12	Halls Catarrh Cure	16	
13	Goodby sweet day	15	
14	rise rise rise	15	
15	person desires protest	14	
16	offer evidence rebuttal	14	
17	prove actual continuous	14	
18	possession tract twenty	14	
19	Unit ed States	14	
20	protest allowance proof	13	
21	NOTICE PUBLICATION Small	13	
22	San Fran cisco	13	
23	claimant MANUEL OTERO	13	

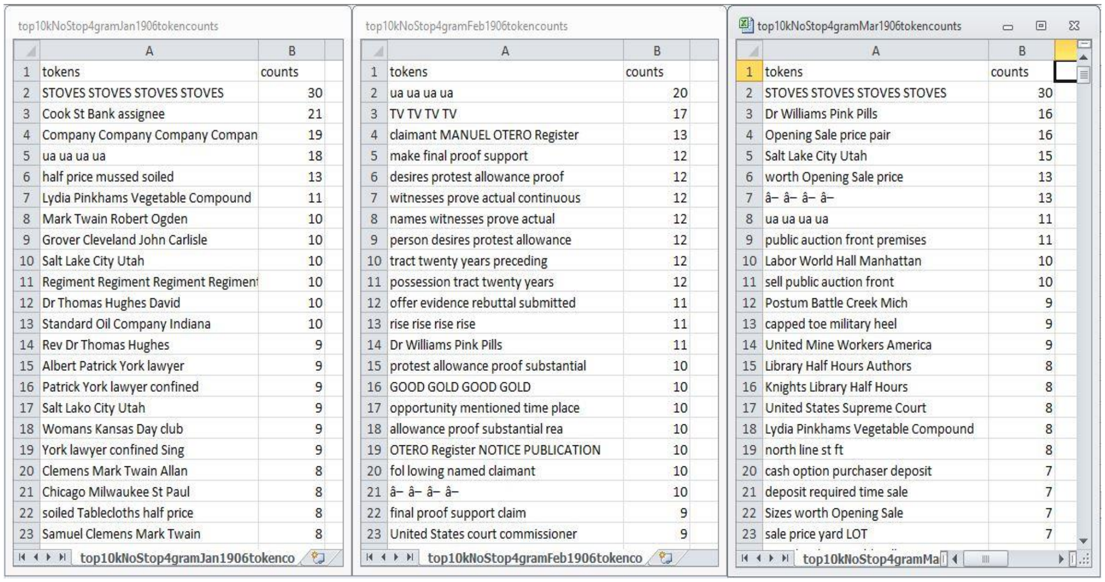
top10kNoStop3gramFeb1906tokenco

top10kNoStop3gramMar1906tokencounts

	A	B	C
1	tokens	counts	
2	Salt Lake City	52	
3	ua ua ua	45	
4	railroad rate bill	41	
5	ofthe ofthe ofthe	37	
6	Opening Sale price	31	
7	â– â– â–	31	
8	STOVES STOVES STOVES	31	
9	Standard Oil Company	24	
10	Lake City Utah	21	
11	Doans Kidney Pills	20	
12	District Attorney Jerome	19	
13	tho United States	18	
14	Sale price pair	18	
15	Life Insurance company	17	
16	Congo Free State	17	
17	Dr Williams Pink	17	
18	Williams Pink Pills	16	
19	San Fran cisco	16	
20	United Mine Workers	16	
21	uaua ua ua	15	
22	Lydia Pinkhams Vegetable	14	
23	Standard Oil company	14	

top10kNoStop3gramM

Figure 9 -- Most common 4-grams in media coverage related to “Mark Twain” during the first three months of 1906



top10kNoStop4gramJan1906tokencounts

	A	B
1	tokens	counts
2	STOVES STOVES STOVES STOVES	30
3	Cook St Bank assignee	21
4	Company Company Company Compan	19
5	ua ua ua ua	18
6	half price mussed soiled	13
7	Lydia Pinkhams Vegetable Compound	11
8	Mark Twain Robert Ogden	10
9	Grover Cleveland John Carlisle	10
10	Salt Lake City Utah	10
11	Regiment Regiment Regiment Regiment	10
12	Dr Thomas Hughes David	10
13	Standard Oil Company Indiana	10
14	Rev Dr Thomas Hughes	9
15	Albert Patrick York lawyer	9
16	Patrick York lawyer confined	9
17	Salt Lako City Utah	9
18	Womans Kansas Day club	9
19	York lawyer confined Sing	9
20	Clemens Mark Twain Allan	8
21	Chicago Milwaukee St Paul	8
22	soiled Tablecloths half price	8
23	Samuel Clemens Mark Twain	8

top10kNoStop4gramJan1906tokenco

top10kNoStop4gramFeb1906tokencounts

	A	B
1	tokens	counts
2	ua ua ua ua	20
3	TV TV TV TV	17
4	claimant MANUEL OTERO Register	13
5	make final proof support	12
6	desires protest allowance proof	12
7	witnesses prove actual continuous	12
8	names witnesses prove actual	12
9	person desires protest allowance	12
10	tract twenty years preceding	12
11	possession tract twenty years	12
12	offer evidence rebuttal submitted	11
13	rise rise rise rise	11
14	Dr Williams Pink Pills	11
15	protest allowance proof substantial	10
16	GOOD GOLD GOOD GOLD	10
17	opportunity mentioned time place	10
18	allowance proof substantial rea	10
19	OTERO Register NOTICE PUBLICATION	10
20	fol lowing named claimant	10
21	â– â– â– â–	10
22	final proof support claim	9
23	United States court commissioner	9

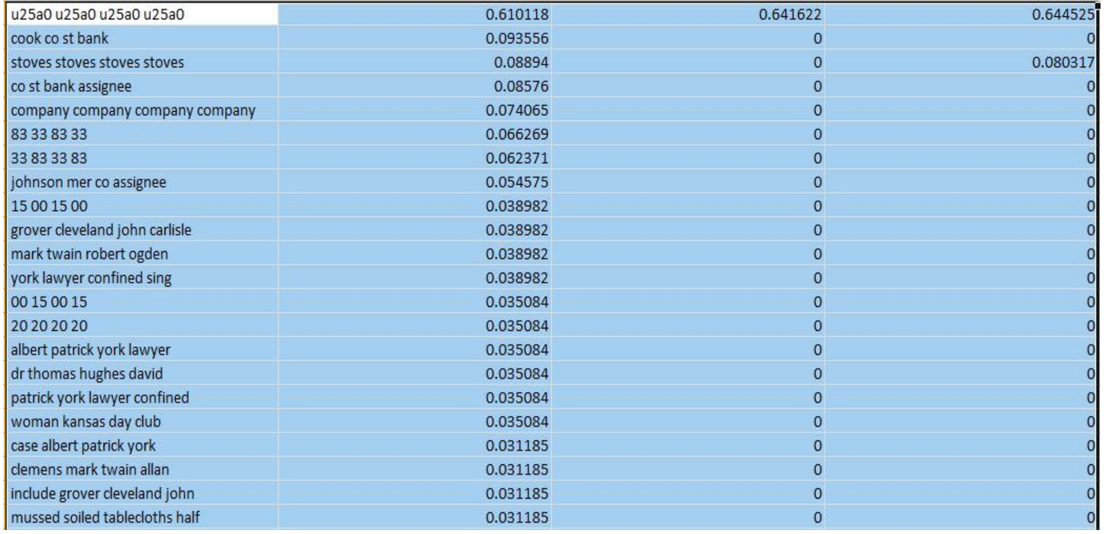
top10kNoStop4gramFeb1906tokenco

top10kNoStop4gramMar1906tokencounts

	A	B
1	tokens	counts
2	STOVES STOVES STOVES STOVES	30
3	Dr Williams Pink Pills	16
4	Opening Sale price pair	16
5	Salt Lake City Utah	15
6	worth Opening Sale price	13
7	â– â– â– â–	13
8	ua ua ua ua	11
9	public auction front premises	11
10	Labor World Hall Manhattan	10
11	sell public auction front	10
12	Postum Battle Creek Mich	9
13	capped toe military heel	9
14	United Mine Workers America	9
15	Library Half Hours Authors	8
16	Knights Library Half Hours	8
17	United States Supreme Court	8
18	Lydia Pinkhams Vegetable Compound	8
19	north line st ft	8
20	cash option purchaser deposit	7
21	deposit required time sale	7
22	Sizes worth Opening Sale	7
23	sale price yard LOT	7

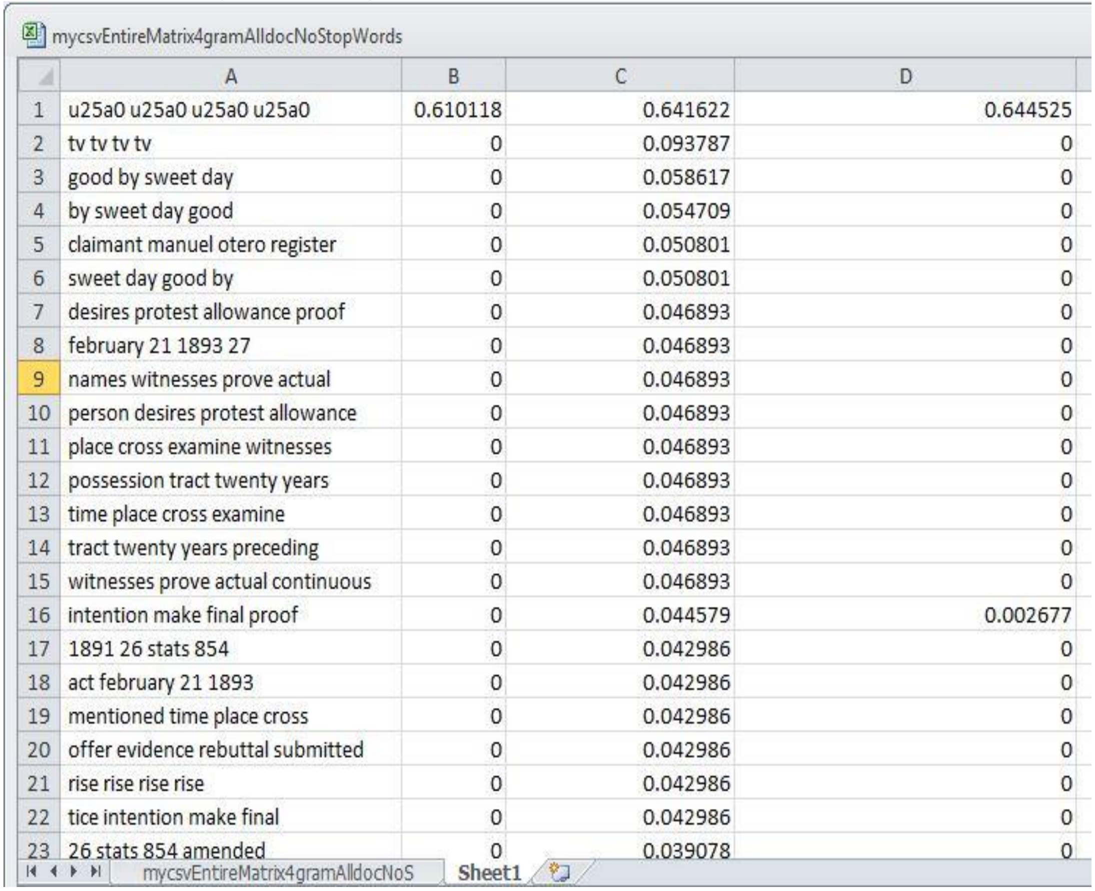
top10kNoStop4gramMa

Figure 10 – 4-grams found in January 1906 media coverage related to “Mark Twain” ranked by tf-idf weight



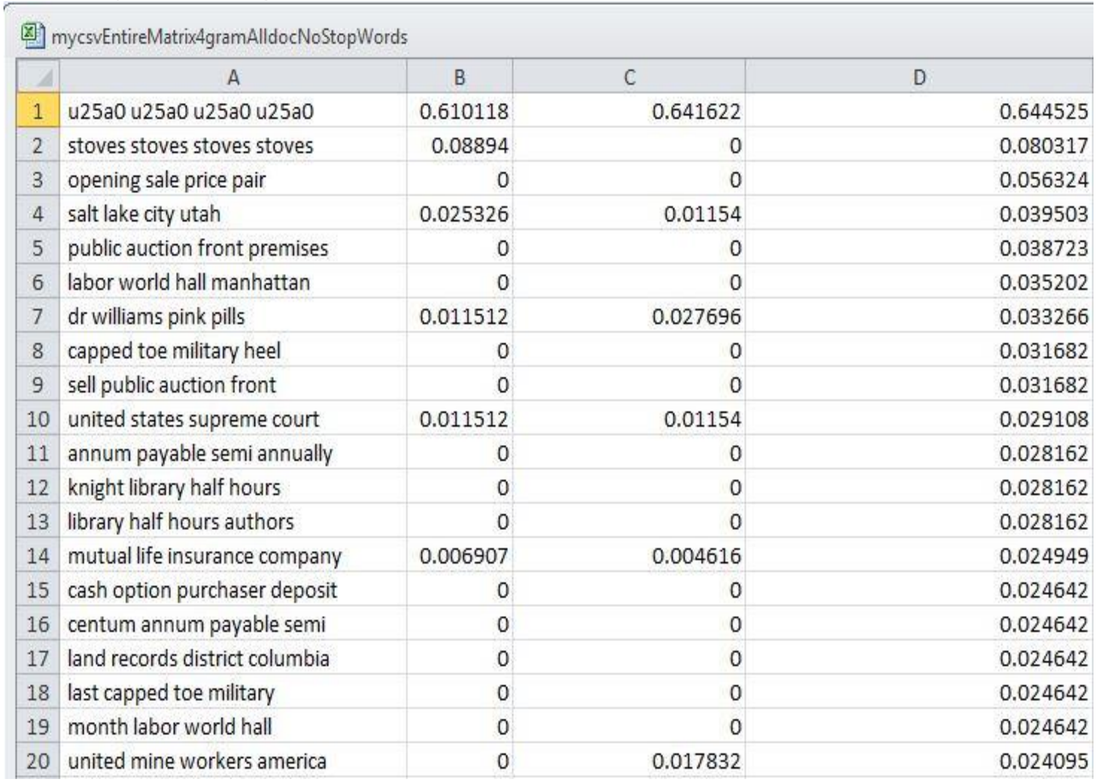
u25a0 u25a0 u25a0 u25a0	0.610118	0.641622	0.644525
cook co st bank	0.093556	0	0
stoves stoves stoves stoves	0.08894	0	0.080317
co st bank assignee	0.08576	0	0
company company company company	0.074065	0	0
83 33 83 33	0.066269	0	0
33 83 33 83	0.062371	0	0
johnson mer co assignee	0.054575	0	0
15 00 15 00	0.038982	0	0
grover cleveland john carlisle	0.038982	0	0
mark twain robert ogden	0.038982	0	0
york lawyer confined sing	0.038982	0	0
00 15 00 15	0.035084	0	0
20 20 20 20	0.035084	0	0
albert patrick york lawyer	0.035084	0	0
dr thomas hughes david	0.035084	0	0
patrick york lawyer confined	0.035084	0	0
woman kansas day club	0.035084	0	0
case albert patrick york	0.031185	0	0
clemens mark twain allan	0.031185	0	0
include grover cleveland john	0.031185	0	0
mussed soiled tablecloths half	0.031185	0	0

Figure 11 – 4-grams found in February 1906 media coverage related to “Mark Twain” ranked by tf-idf weight


mycsvEntireMatrix4gramAlldocNoStopWords
mycsvEntireMatrix4gramAlldocNoS
Sheet1

	A	B	C	D
1	u25a0 u25a0 u25a0 u25a0	0.610118	0.641622	0.644525
2	tv tv tv tv	0	0.093787	0
3	good by sweet day	0	0.058617	0
4	by sweet day good	0	0.054709	0
5	claimant manuel otero register	0	0.050801	0
6	sweet day good by	0	0.050801	0
7	desires protest allowance proof	0	0.046893	0
8	february 21 1893 27	0	0.046893	0
9	names witnesses prove actual	0	0.046893	0
10	person desires protest allowance	0	0.046893	0
11	place cross examine witnesses	0	0.046893	0
12	possession tract twenty years	0	0.046893	0
13	time place cross examine	0	0.046893	0
14	tract twenty years preceding	0	0.046893	0
15	witnesses prove actual continuous	0	0.046893	0
16	intention make final proof	0	0.044579	0.002677
17	1891 26 stats 854	0	0.042986	0
18	act february 21 1893	0	0.042986	0
19	mentioned time place cross	0	0.042986	0
20	offer evidence rebuttal submitted	0	0.042986	0
21	rise rise rise rise	0	0.042986	0
22	tice intention make final	0	0.042986	0
23	26 stats 854 amended	0	0.039078	0

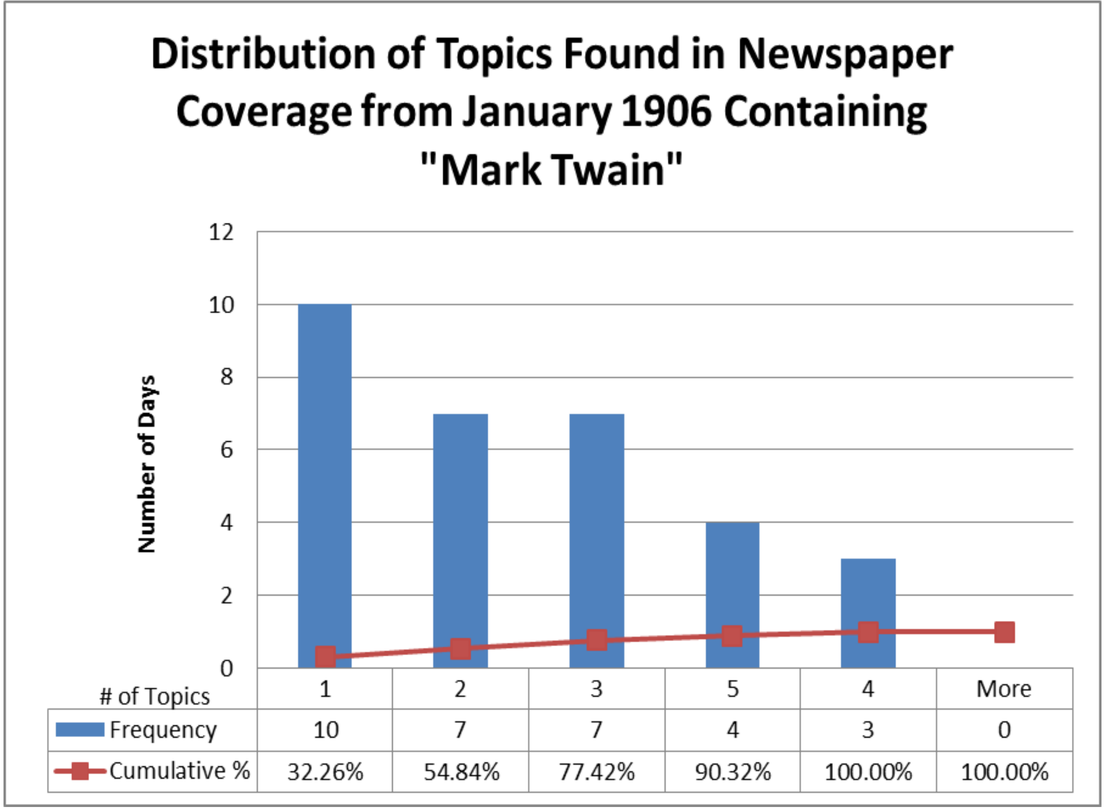
Figure 12 – 4-grams found in March 1906 media coverage related to “Mark Twain” ranked by tf-idf weight



mycsvEntireMatrix4gramAlldocNoStopWords

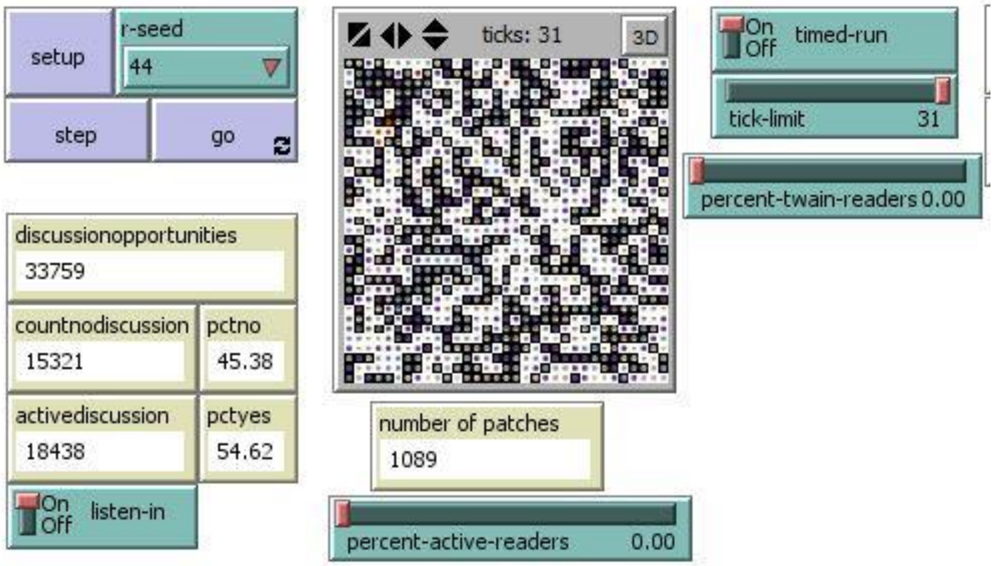
	A	B	C	D
1	u25a0 u25a0 u25a0 u25a0	0.610118	0.641622	0.644525
2	stoves stoves stoves stoves	0.08894	0	0.080317
3	opening sale price pair	0	0	0.056324
4	salt lake city utah	0.025326	0.01154	0.039503
5	public auction front premises	0	0	0.038723
6	labor world hall manhattan	0	0	0.035202
7	dr williams pink pills	0.011512	0.027696	0.033266
8	capped toe military heel	0	0	0.031682
9	sell public auction front	0	0	0.031682
10	united states supreme court	0.011512	0.01154	0.029108
11	annum payable semi annually	0	0	0.028162
12	knight library half hours	0	0	0.028162
13	library half hours authors	0	0	0.028162
14	mutual life insurance company	0.006907	0.004616	0.024949
15	cash option purchaser deposit	0	0	0.024642
16	centum annum payable semi	0	0	0.024642
17	land records district columbia	0	0	0.024642
18	last capped toe military	0	0	0.024642
19	month labor world hall	0	0	0.024642
20	united mine workers america	0	0.017832	0.024095

Figure 13 – Summary of the distribution of topics produced by my topic model



# of Topics	1	2	3	5	4	More
Frequency	10	7	7	4	3	0
Cumulative %	32.26%	54.84%	77.42%	90.32%	100.00%	100.00%

Figure 14 – The distribution of successful and failed conversations over the month of January for Random Seed 44 when no turtles take an active interest in discussing current topics from the news.



r-seed 44

ticks: 31

tick-limit 31

percent-twain-readers 0.00

discussionopportunities 33759

countnodiscussion 15321

pctno 45.38

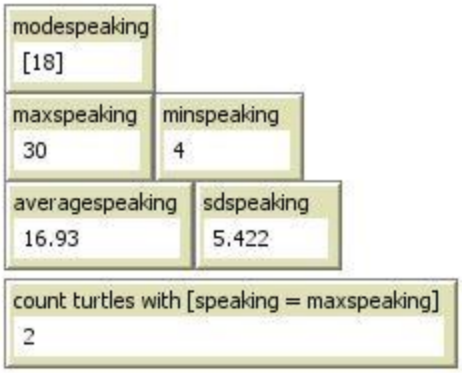
activediscussion 18438

pctyes 54.62

number of patches 1089

percent-active-readers 0.00

Figure 15 – A summary of the performance of individual turtles over the course of a single run of the simulation



modespeaking: [18]

maxspeaking: 30

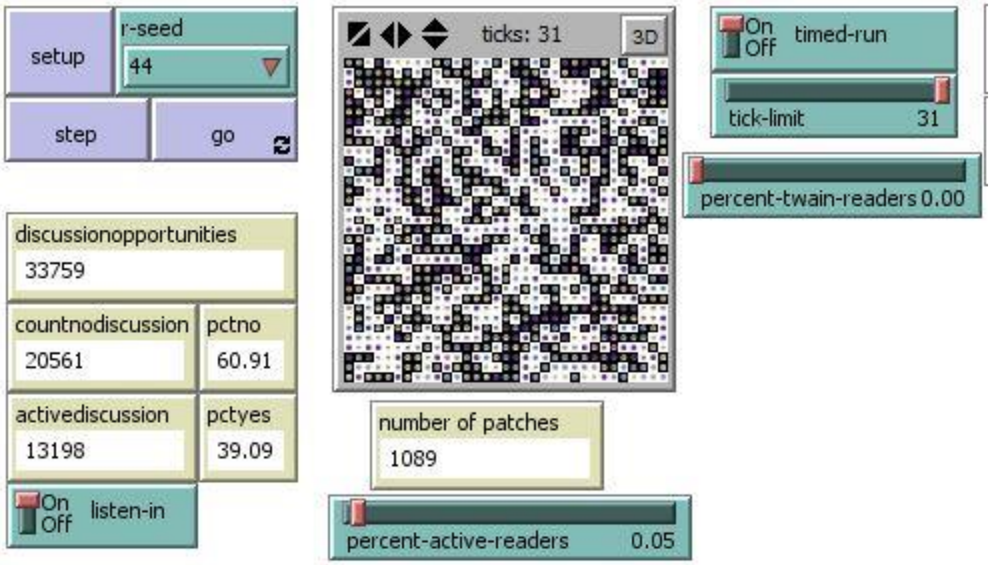
minspeaking: 4

averagespeaking: 16.93

sdspeaking: 5.422

count turtles with [speaking = maxspeaking]: 2

Figure 16 – Impact on conversation dynamics in the model that results from a 5% increase in the number of agents in that keep up with current events



r-seed: 44

ticks: 31

timed-run

tick-limit 31

percent-twain-readers 0.00

discussionopportunities: 33759

countnodiscussion: 20561

pctno: 60.91

activediscussion: 13198

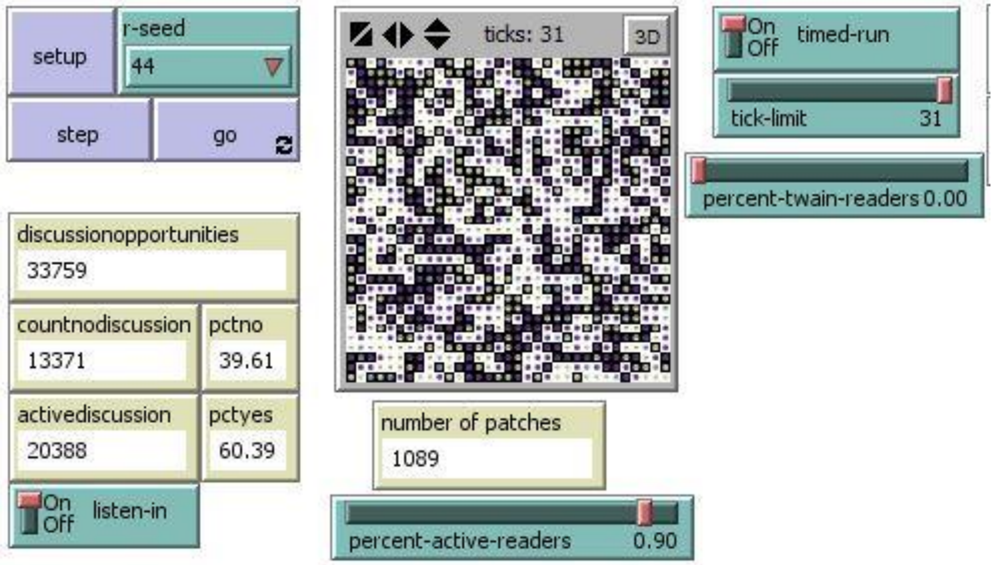
pctyes: 39.09

listen-in

number of patches: 1089

percent-active-readers 0.05

Figure 17 – Impact on conversation dynamics in the model that results from a 90% increase in the number of agents in that keep up with current events



r-seed 44

ticks: 31

timed-run On Off

tick-limit 31

percent-twain-readers 0.00

discussionopportunities 33759

countnodiscussion 13371

pctno 39.61

activediscussion 20388

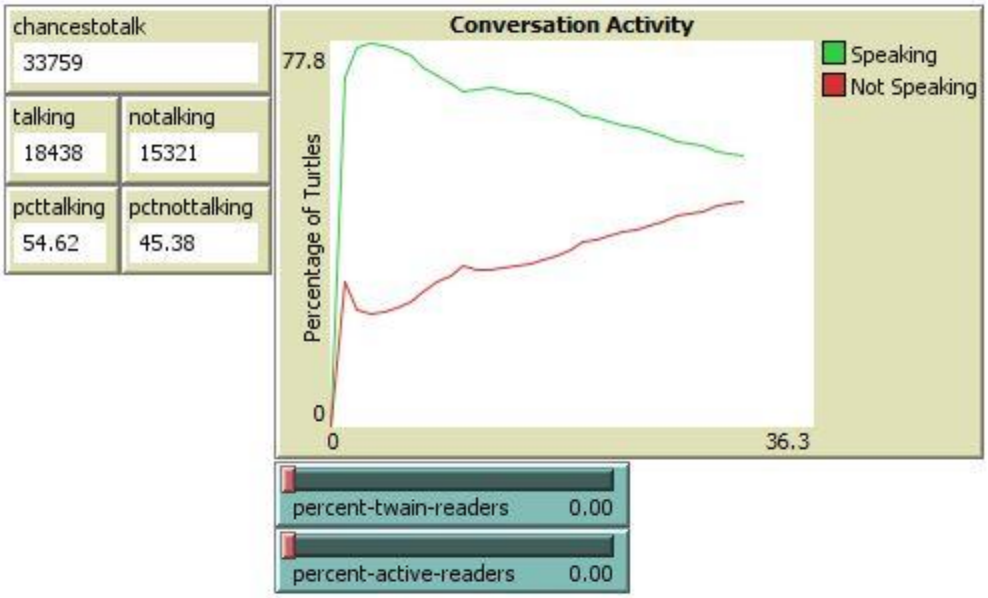
pctyes 60.39

listen-in On Off

number of patches 1089

percent-active-readers 0.90

Figure 18 – A Graph of conversation levels in the model when agents don't take an active interest in current events



chancestotalk 33759

talking 18438

notalking 15321

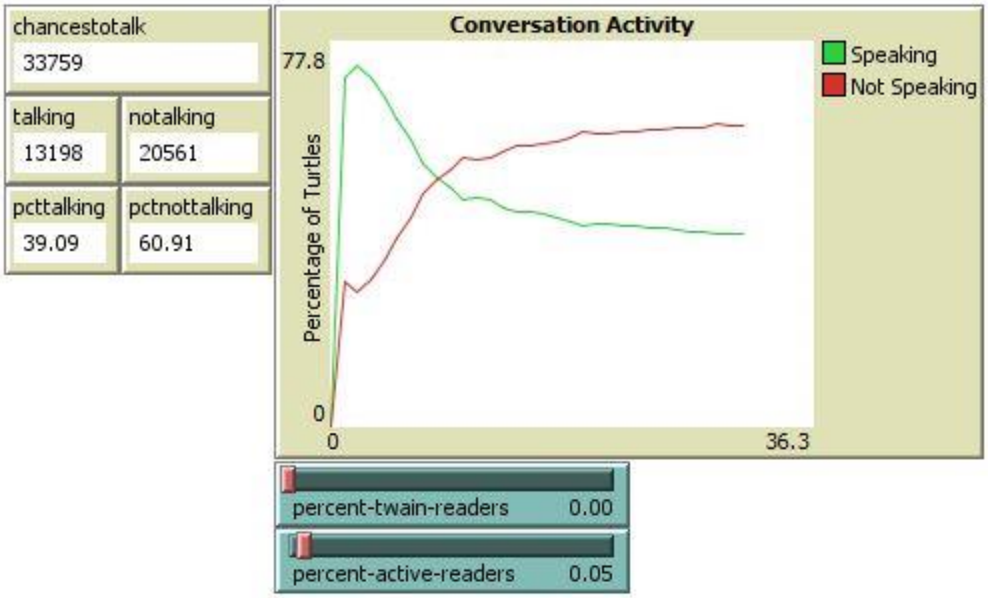
pcttalking 54.62

pctnottalking 45.38

percent-twain-readers 0.00

percent-active-readers 0.00

Figure 19 – A Graph of conversation levels in the model when 5% of agents take an active interest in current events



chancestotalk 33759

talking 13198

notalking 20561

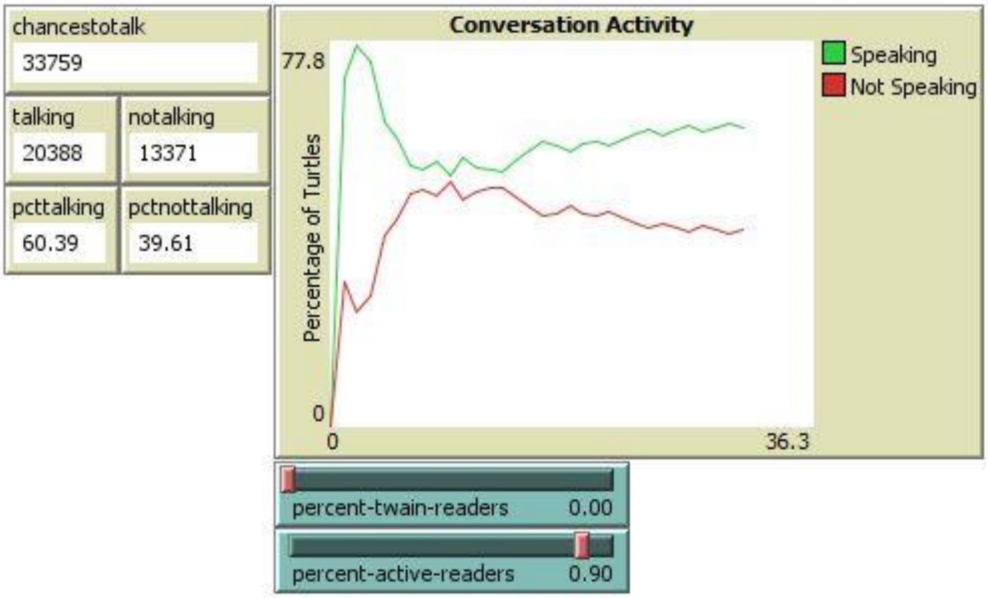
pcttalking 39.09

pctnottalking 60.91

percent-twain-readers 0.00

percent-active-readers 0.05

Figure 20 – A Graph of conversation levels in the model when 90% of agents take an active interest in current events



chancestotalk 33759

talking 20388

notalking 13371

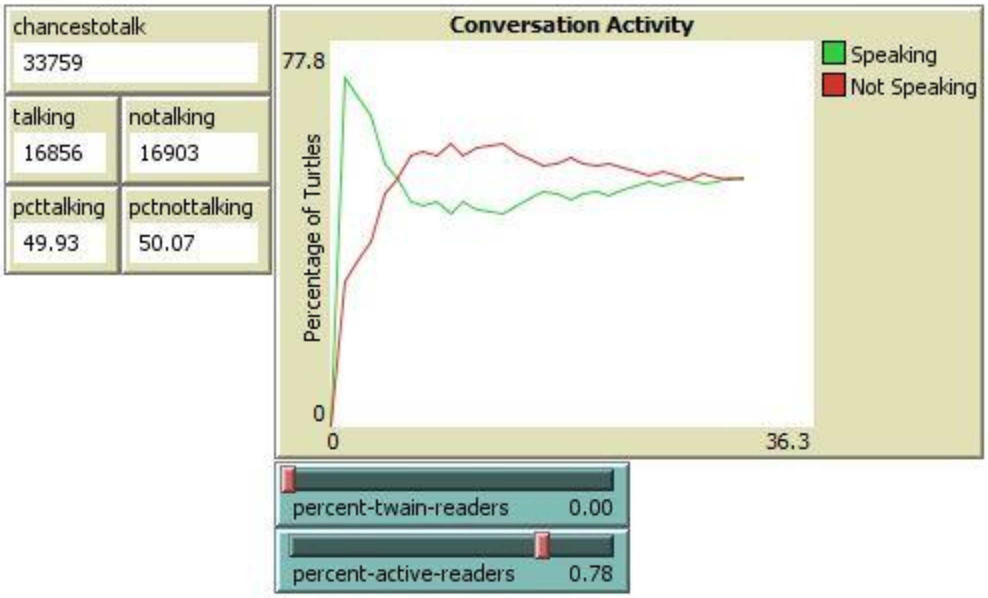
pcttalking 60.39

pctnottalking 39.61

percent-twain-readers 0.00

percent-active-readers 0.90

Figure 21 – A Graph of conversation levels in the model when 78% of agents take an active interest in current events



chancestotalk 33759

talking 16856

notalking 16903

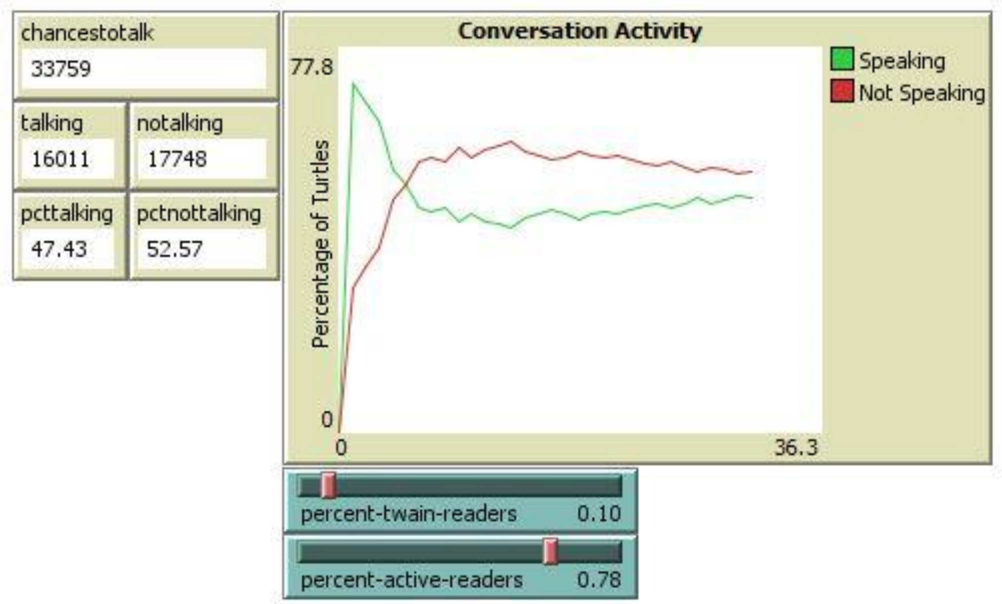
pcttalking 49.93

pctnottalking 50.07

percent-twain-readers 0.00

percent-active-readers 0.78

Figure 22 – Results obtained when 10% of turtles in the environment are interested in Twain's Autobiography and 78% are interested in current events



chancestotalk 33759

talking 16011

notalking 17748

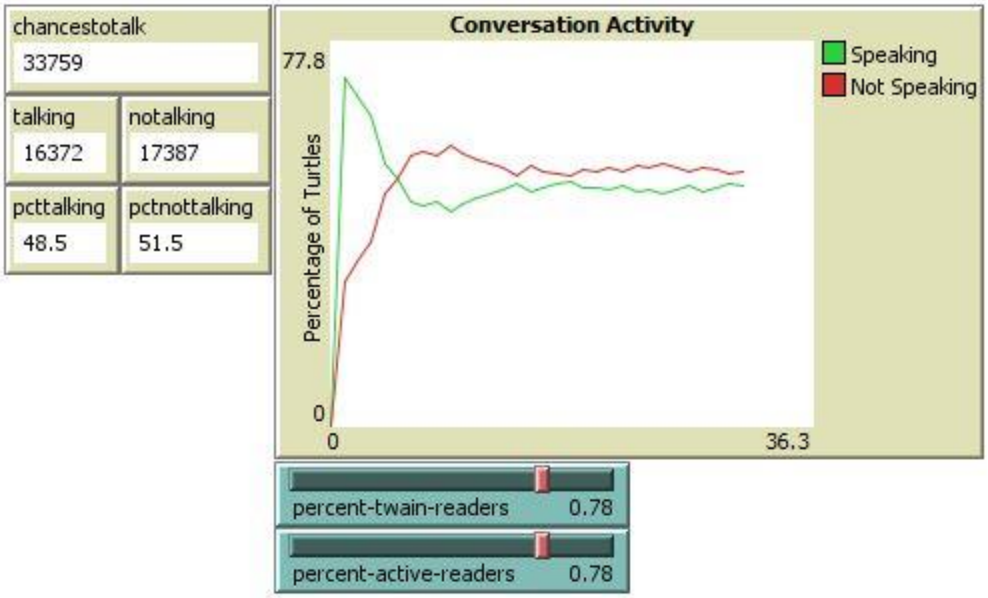
pcttalking 47.43

pctnottalking 52.57

percent-twain-readers 0.10

percent-active-readers 0.78

Figure 23 – Results obtained when 78% of turtles in the environment are interested in Twain's Autobiography and 78% are interested in current events



chancestotalk 33759

talking 16372

notalking 17387

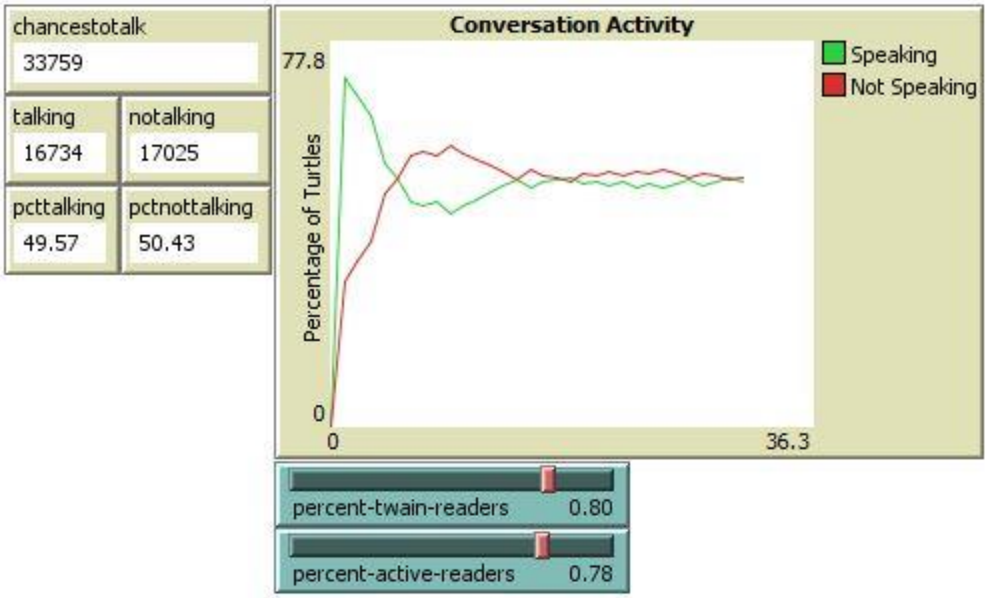
pcttalking 48.5

pctnottalking 51.5

percent-twain-readers 0.78

percent-active-readers 0.78

Figure 24 -- Results obtained when 80% of turtles in the environment are interested in Twain's Autobiography and 78% are interested in current events



chancestotalk 33759

talking 16734

notalking 17025

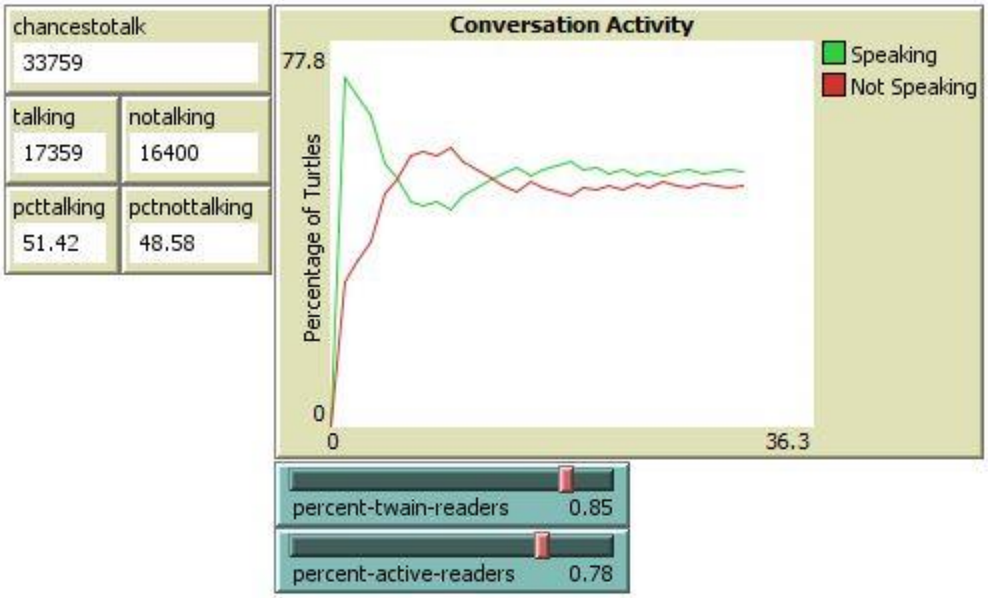
pcttalking 49.57

pctnottalking 50.43

percent-twain-readers 0.80

percent-active-readers 0.78

Figure 25 -- Results obtained when 85% of turtles in the environment are interested in Twain's Autobiography and 78% are interested in current events



chancestotalk 33759

talking 17359

notalking 16400

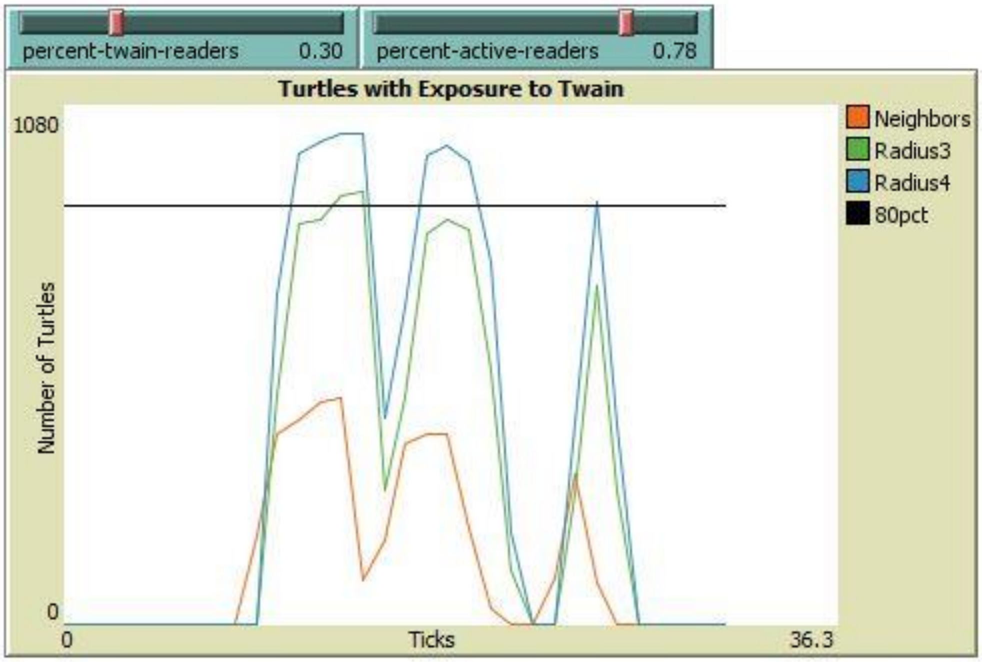
pcttalking 51.42

pctnottalking 48.58

percent-twain-readers 0.85

percent-active-readers 0.78

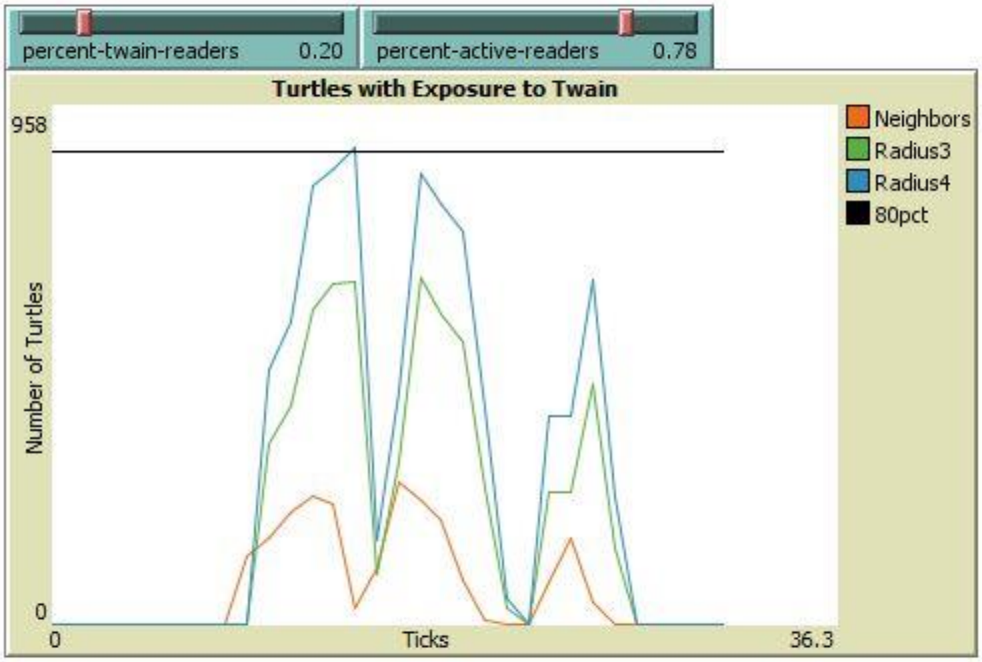
Figure 26 – Comparison of conversation dynamics in the model when turtles connect with increasing numbers of neighbors and have a 30% interest in keeping up with Twain's work.



percent-twain-readers 0.30

percent-active-readers 0.78

Figure 27 – Comparison of conversation dynamics in the model when turtles connect with increasing numbers of neighbors and have a 20% interest in keeping up with Twain's work.



percent-twain-readers 0.20

percent-active-readers 0.78

List of Supplemental Files

Supporting Files for Chapter One

Data exported from Google Trends for figures at the Whittier birthday dinner (note 11)
WhittierDinner.csv

Data on edits to True Jesus Church page on English and Malay Wikipedias (note 24)
TJC_Histories.xlsx

Processing code for Visualizing Chronicling America (VCA) (note 27)
linegraphScale2chronicleAmerica

686 MarcXML files for “Mark Twain” downloaded from HathiTrust (note 29)
myInputfiles

An Excel file of the data I extracted from the MarcXML files above (note 29)
HathiXMLDataforTwain.xlsx

An Excel file of data for the American literature article on multiple Wikipedias (note 35)
ChartComparingAmLitOnWikis.xls

Format of interwiki language links on the American Literature page on English Wikipedia (note 36)
AmLitLangs.txt

An Excel file of language codes extracted from interwiki language links on the American Literature page on English Wikipedia (note 36)
AmLit01212013.xls

An archived version of the data provided by Wikipedian Aka’s Wikipedia Page History Statistics tool on January 22, 2013 for the American literature article on Arabic Wikipedia. (Webarchive files are produced by the Safari browser.) (note 40)
أمري كي أدب - statistics.webarchive

A HTML file produced by converting the webarchive file above (note 40)
HTML_conversion_of_statistics_webarchivefile

An Excel file of data for the American literature article on multiple Wikipedias (note 42)
ChartComparingAmLitOnWikis.xls

An Excel file of language codes for Wikipedias with a Mark Twain page (note 44)
TwainHomePages011713.xls

An Excel file of data for the Mark Twain article on multiple Wikipedias (note 45)
MarkTwain020313.xls

An Excel file of data for the American literature article collected on multiple Wikipedias. The data in this file is slightly different in some cases from the prior file of American literature article data noted above (ChartComparingAmLitOnWikis) because it was collected at a later date. For example the file shows one more edit to the American literature page on English Wikipedia (note 45)
AmLit020313.xls

An Excel file comparing summaries of the data collected for the Mark Twain (MarkTwain020312 above), American literature (AmLit020313 above) articles and the Huckleberry Finn article (HuckFinn020312 below) (note 45)
StatsCompared.xls

An Excel file of data for the Huckleberry Finn article on multiple Wikipedias (note 46)
HuckFinn020313.xls

Supporting Files for Chapter Two

CONTENTS OF VCADATA folder (footnote 4)

Data harvested from Chronicling America using VCA for "Benvenuto Cellini" is available at
BenvenutoCellini.xlsx

Data harvested from Chronicling America using VCA for "Helen Keller" is available at
HelenKeller.xlsx

Data harvested from Chronicling America using VCA for "Mark Twain" is available at
MarkTwain.xlsx
Data harvested from Chronicling America using VCA for "Benjamin Franklin" is available at
BenjaminFranklin.xlsx

Data harvested from Chronicling America using VCA for "U.S. Grant" is available at
U_S_Grant.xlsx

Data harvested from Chronicling America using VCA for "Ulysses S. Grant" is available at

UlyssesSGrant.xlsx

Data harvested from Chronicling America using VCA for "President Grant" is available at

PresidentGrant.xlsx

Data harvested from Chronicling America using VCA for "General Grant" is available at

GeneralGrant.xlsx

Data harvested from Chronicling America using VCA for "P.T. Barnum" is available at

P_T_Barnum.xlsx

Data harvested from Chronicling America using VCA for "Phineas Taylor Barnum" is available at

PhineasTaylorBarnum.xlsx

A comparison of the four VCA searches involving "Grant" (footnote 12)

All_Grant.xlsx

CONTENTS OF TWAINCHRONICLINGAMERICA1907 folder (footnote 14)

An Excel chart of “Mark Twain” in the holdings of Chronicling America for 1907

summaryfileCharted.xlsx

A Python script used to produce the summary data above by processing the JSON below. The script distributes the JSON corpus of media coverage over twelve files, one for each month of the year.

extractjson

For summary data produced by a more refined tool, compare with the summary data produced by AntConc

summaryfileChartedwAntConc.xlsx

A JSON corpus of holdings in Chronicling America for "Mark Twain" in 1907 is available

As four files at

Twain1907json

OR

As one large file at

all1907Twain.txt

CONTENTS OF 1GRAMS (footnote 15)

Excel files of term frequency counts for media coverage from May, June, and July 1907 (note 15)

May

FullMaytokencounts.csv

June

FullJuntokencounts.csv

July

FullJultokencounts.csv

Stopwords in the term frequency counts for May

checkingStopwordCountsForMay1907.xlsx

Stopwords in the term frequency counts for June

checkingStopwordCountsForJune1907.xlsx

Stopwords in the term frequency counts for July

checkingStopwordCountsForJuly1907.xlsx

CONTENTS OF NOSTOPWORDS (footnote 16)

Excel files of term frequency counts for media coverage from May, June, and July 1907 with stopwords removed (note 16)

May

NoStopFullMaytokencounts.csv

June

NoStopFullJuntokencounts.csv

July

NoStopFullJultokencounts.csv

An Excel file comparing tokens from May, June and July 1907. Unique values (tokens that appear in only one month) are highlighted in red. Tokens that appear in more than one month receive no color. (note 17)

CompareMayJunJuly1907.xlsx

CONTENTS OF NGRAMSWITHTFIDF (footnote 20)

1gramMatrix1907

2gramMatrix1907

3gramMatrix1907

4gramMatrix1907

Voyant query for “roarer” in media coverage from May 1907 (footnote 22)

voyant for Roarer.docx

CONTENTS OF HATHITRUST (footnote 23)

MarcXML data from HathiTrust – publishing data before and after cleaning

HathiTrust

CONTENTS OF OPENLIBRARY (footnote 27)

Open Library data before processing

TwainOpenLibrarydata022012.xls

Dates after processing

TwainOpenLibrarydata022012withCleanDates (Autosaved).xls

Names after processing

NAMESTwainOL022012-csv.xlsx

Place of Publication after processing

PLACESTwainOL022012-csv.xlsx

CONTENTS OF MAY1907TOPICMODELS (footnote 32)

5TA10B_1
10TA5B_1
20TA2_5B_1
30TA1_67B_1
60TA0_83B_1
90TA0_56B_1
150TA0_33B_1

CONTENTS OF MODELCODE (footnote 37)

NetLogo model
TripleConversationv4RunFullYearRseedwColorswTurtlesw30Sound2FINAL2.nlogo

SuperCollider file to add sound to the model
my30sounds4mixto1soundFINAL2RawPlus20Compensated.scd"

Supporting Files for Chapter Three

Comparison of 1906 and 1907 Chronicling America search results for “Mark Twain” (footnote 8)

Mark_Twain_1906_1907_comparedv2.xlsx

CONTENTS OF NGRAMS (footnotes 9, 10, and 11)

Excel files for 1-grams in media coverage related to “Mark Twain” during the first three months of 1906 (note 9)

1grams_noStopWords

Excel files for 2-grams in media coverage related to “Mark Twain” during the first three months of 1906 (note 10)

2grams_noStopwords

Excel files for 3-grams in media coverage related to “Mark Twain” during the first three months of 1906 (note 11)
3grams_noStopwords

Excel files for 4-grams in media coverage related to “Mark Twain” during the first three months of 1906 (note 11)
4grams_noStopwords

CONTENTS OF TFIDF (footnote 12)
Excel files for n-grams with tf-idf weights
04Twain1906_Ch3_tf_idf_noStopWords

30 topic model of 1906 media coverage (footnote 13)
Twain_1906_Jan_30T_A1_67_B_1

Excel file comparing topic models of 1906 media coverage (footnote 14)
ApproachesToTopicModeling1906.xlsx

CONTENTS OF NETLOGOJAN1906 (footnote 15)
NetLogo code for discussion of January 1906 media coverage
TripleConversationv4RunFullYearRseedwColorswTurtlesw30Sound2FINAL2forAuto1906newsVTwainminimalv6ForDistractedReadingWRadiusNeighbors2.nlogo

CONTENTS OF 30TOPICSJANUARY1906 (footnote 17)
For the 30 topic model from January 1906 media coverage
Twain_1906_Jan_30T_A1_67_B_1

Excel file of data used to produce the visual summary of output from the model (also found in the folder above)
doc_topics_analysis.csv

Bibliography

Aaron, Daniel. The Unwritten War: American Writers and the Civil War. U of Alabama P, 2003, https://books.google.com/books?id=CSB6CgAAQBAJ&pg=PA179&lpg=PA179&dq=%22By+chance+I+had+been+comparing+the+memoirs+with+Caesar%27s+commentaries%22&source=bl&ots=sF8jY9tqSH&sig=ZT78MYeh9dvWgSxK0pbqur0CRY0&hl=en&sa=X&ved=0ahUKEwix6dmHi9vZAhVJ-lQKHdpkDacQ6AEIMDAD#v=onepage&q=Caesar&f=false 7 March 2018.

“About.” Digging into Data Challenge, Trans-Atlantic Platform for the Social Sciences and Humanities, https://diggingintodata.org/about 4 March 2018.

“About.” Memories of Fiction: An Oral History. Arts & Humanities Research Council, https://memoriesoffiction.org/about/ 13 March 2018.

“About.” OCLC. https://www.oclc.org/en/about.html 4 March 2018

“About Google Books.” Google Books. Google, https://books.google.com/intl/en/googlebooks/about/index.html. 4 March 2018

“About Omnipedia.” CollabLab, Northwestern University, http://omnipedia.northwestern.edu/, 5 March 2018.

“Actions.” Project Gutenberg, http://copy.pglaf.org/ 4 March 2018

Aka. “Wikipedia Page History Statistics, a Wikipedia tool by [[user:aka]].” http://vs.aka-online.de/cgi-bin/wppagehiststat.pl 5 March 2018.

Alanen, Arnold Robert. Morgan Park: Duluth, U.S. Steel, and the Forging of a Company Town. U of Minnesota P, 2007. https://books.google.com/books?id=CMcedXXe7ZIC&lpg=PP1&pg=PP1#v=onepage&q&f=false 7 March 2018.

Allen, Melissa, et al. Workshop on Human Activity at Scale in Earth System Models. Oak Ridge National Laboratory, US Department of Energy, 26 January 2017, doi:10.2172/1343540, https://www.osti.gov/biblio/1343540 15 March 2018.

Allington, Daniel, et al. “Neoliberal Tools (and Archives): A Political History of Digital Humanities.” LA Review of Books, 1 May 2016, https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/ 10 March 2018.

“American literature.” Wikipedia, The Free Encyclopedia. https://web.archive.org/web/20130108221253/https://en.wikipedia.org/wiki/American_literature 5 March 2018.

"Analytics/AQS/Legacy Pagecounts." Wikitech, 12 Apr 2017, https://wikitech.wikimedia.org/w/index.php?title=Analytics/AQS/Legacy_Pagecounts&oldid=1756305 6 March 2018.

"Analytics/AQS/Pageviews." Wikitech, 14 Feb 2018, https://wikitech.wikimedia.org/w/index.php?title=Analytics/AQS/Pageviews&oldid=1782762 6 March 2018.

Anderson, Benedict. Imagined Communities. Verso, 1983.

Appadurai, Arjun. “Disjuncture and Difference in the Global Cultural Economy.” Public Culture 2, no. 2, Spring 1990, pp. 1–24 https://doi.org/10.1215/08992363-2-2-1 10 March 2010.

“Appfest.” DPLA Appfest, Chattanoga Public Library, November 8-9, 2012. http://web.archive.org/web/20160316195922/http://dp.la/wiki/Appfest#View_the_Apps 5 March 2018.

“Art+Feminism.” http://www.artandfeminism.org/ 13 March 2018.

Axtell, Robert. “TEDxUVM 2011 - Rob Axtell - Modeling the Economy with 150 Million Agents.” YouTube, uploaded by UVMcomplexity, 7 February 2012, https://www.youtube.com/watch?v=c-sieJVR5TI 15 March 2018.

---. “Computationally Enabled Public Policy Analysis.” The Conference on Complexity and Policy Studies, CAPS 2017, April 12-14, http://webcache.googleusercontent.com/search?q=cache:cf2VCfehGF0J:capsconference.org/wp-content/uploads/2017/04/S7-Axtel_Apr13_2017.pdf+&cd=1&hl=en&ct=clnk&gl=us&client=firefox-b-1 15 March 2018.

Bao, Patti, et al. “Omnipedia: Bridgint the Wikipedia Language Gap.” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2012): 1075-1084. https://dl.acm.org/citation.cfm?doid=2207676.2208553 5 March 2018.

Barker, Elton, et al. “Colloquium: Digital technologies: Help or hindrance for the humanities?” Arts and Humanities in Higher Education, vol. 11, no. 1-2, 2012, pp. 185-200, http://ahh.sagepub.com/content/11/1-2/185 15 March 2018.

Bassnett, Susan, and André Lefevere. "General editors' preface." The Translator's Invisibility, by Lawrence Venuti, Routledge, 2004. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.475.4973&rep=rep1&type=pdf 10 March 2018.

Benjamin, Walter. "On the Concept of History." Translated by Dennis Redmond, Marxists Internet Archive, https://www.marxists.org/reference/archive/benjamin/1940/history.htm 10 March 2018.

---. "The Task of the Translator." Walter Benjamin: Selected Writings, Volume 1, 1913-1926, edited by Marcus Bullock and Michael W. Jennings, Harvard, 1996. http://users.clas.ufl.edu/burt/deconstructionandnewmediatheory/WalterBenjaminTaskTranslator.pdf 10 March 2018.

Betjemann, Peter. Talking Shop: The Language of Craft in an Age of Consumption. U of Virginia P, 2011, https://books.google.com/books?id=3-EW_C2QBdMC&lpg=PT262&ots=0IU64JdN0Z&dq=Benvenuto%20Cellini%20nineteenth%20century%20america&pg=PT18#v=onepage&q=Benvenuto%20Cellini%20nineteenth%20century%20america&f=false 7 March 2018.

Bosman, Julie, and Simon Romero. "Vargas Llosa Takes Nobel in Literature." The New York Times, 7 October 2010, http://www.nytimes.com/2010/10/08/books/08nobel.html?pagewanted=all&_r=0, 6 March 2018.

Burdick, Anne, et al. Digital_Humanities. MIT Press, 2012. https://mitpress.mit.edu/books/digitalhumanities 10 March 2018.

Busa, Roberto. "The Annals of Humanities Computing: The Index Thomisticus." Computers and the Humanities 14 (1980), pp. 83-90, https://www.jstor.org/stable/30207304?seq=1#page_scan_tab_contents 16 March 2018.

"CAPS 2017 Presentations." Journal for Policy and Complex Systems, CAPS 2017, April 12-14, http://web.archive.org/web/20170501074120/http://capsconference.org:80/caps-2017/caps-2017-presentations/ 15 March 2018.

"Charles Klein: An Inventory of His Plays in the Manuscript Collection at the Harry Ransom Humanities Research Center." Harry Ransom Center, UT Austin, http://norman.hrc.utexas.edu/fasearch/findingAid.cfm?eadid=00335 8 March 2018.

Chronicling America. Library of Congress, https://chroniclingamerica.loc.gov/ 26 March 2018.

"Chronicling America: Historic American Newspapers Data Challenge." Challenge.gov, https://www.challenge.gov/challenge/chronicling-america-historic-american-newspapers-data-challenge/ 16 March 2018.

Clements, Jacquelyn. "In Support of The Digital Humanities." 3 May 2016, http://www.jacquelynclements.com/blog/in-support-of-the-digital-humanities 10 March 2018.

Clifford, James. Routes: Travel and Translation in the Late Twentieth Century. Harvard UP, 1997.

Cohen, Daniel. "From Babel to Knowledge: Data Mining Large Digital Collections." D-Lib Magazine, March 2006. http://www.dlib.org/dlib/march06/cohen/03cohen.html 16 March 2018.

"Corpus Thomisticum." http://www.corpusthomisticum.org/it/index.age 16 March 2018.

Crane, Gregory. "What Do You Do with a Million Books?" D-Lib Magazine, Vol. 12, Num. 3, March 2006, http://www.dlib.org/dlib/march06/crane/03crane.html 16 March 2018.

Crane, Gregory, et al. "Classics in the Million Book Library." DHQ 3.9 (2009) http://www.digitalhumanities.org/dhq/vol/3/1/000034/000034.html 16 March 2018.

Crowley, John W. The Dean of American Letters: The Late Career of William Dean Howells. U of Massachusetts P, 1999. http://www.umass.edu/umpress/title/dean-american-letters

"Cultural Analytics Lab." http://lab.culturalanalytics.info/ 16 March 2018.

"Culturomics." Culturomics.org, http://www.culturomics.org/ 4 March 2018

Damrosch, David. *How to Read World Literature*. Wiley-Blackwell, 2009.

---. *What is World Literature*? Princeton UP, 2003.

Darnton, Robert. “The Library in the New Age.” New York Review of Books, 12 June 2008, http://www.nybooks.com/articles/archives/2008/jun/12/the-library-in-the-new-age/?pagination=false 20 March 2018.
“Data mining.” Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Data_mining&oldid=829478447 10 March 2018.

Davidson, Cathy. “My Commencement Address: School of Information and Library Science, UNC.” CathyDavidson.com, 14 May 2012, https://www.cathydavidson.com/blog/my-commencement-address-school-of-information-and-library-science-unc/ 20 March 2018.

---. “What’s It Like To Be On an Author Tour in the Digital Age? (Just ask Parson Weems!)” CathyDavidson.com, 11 October 2011, https://www.cathydavidson.com/blog/whats-it-like-to-be-on-an-author-tour-in-the-digital-age-just-ask-parson-weems/ 20 March 2018.

Davidson, Cathy, and David Theo Goldberg. The Future of Learning Institutions in a Digital Age. MIT Press, 2009, https://dmlcentral.net/wp-content/uploads/files/Future_of_Learning.pdf 20 March 2018.

Dean, Jeffrey, et al. “Understanding Anasazi Culture Change Through Agent-Based Modeling.” SFI Working Paper: 1998-10-094. Santa Fe Institute. https://www.santafe.edu/research/results/working-papers/understanding-anasazi-culture-change-through-agent 10 March 2018.

“Depositing with the University of Oxford Text Archive.” University of Oxford Text Archive, http://www.ota.ox.ac.uk/about/deposit.xml 4 March 2018

Derrida, Jacques. Specters of Marx, the state of the debt, the Work of Mourning, & the New International. Translated by Peggy Kamuf. Routledge, 1994.

Dittenbach, Michael. “Scoring and Ranking Techniques – tf-idf term weighting and cosine similarity.” Information Retrieval Facility, 31 March 2010, https://www.ir-facility.org/scoring-and-ranking-techniques-tf-idf-term-weighting-and-cosine-similarity 7 March 2018.

Drucker, Johanna. SpecLab. Chicago: U Chicago P, 2009.

Drucker, Johanna, and Patrik Svensson. “The Why and How of Middleware.” DHQ 10.2 (2016). http://www.digitalhumanities.org/dhq/vol/10/2/000248/000248.html 15 March 2018.

Dunne, Nora. “Nobel Prize in Literature: Which Latin American writers have won?” Christian Science Monitor, 8 October 2010, http://nldunne.wordpress.com/2010/10/08/nobel-prize-in-literature-which-latin-american-writers-have-won/ 6 March 2018.

Eagleman, David. “Sensory Substitution.” Eagleman.com, http://www.eagleman.com/research/sensory-substitution 8 March 2018.

---. “VEST: A Sensory Substitution Neuroscience Project.” Kickstarter.com, https://www.kickstarter.com/projects/324375300/vest-a-sensory-substitution-neuroscience-project 8 March 2018.

Eakin, Emily. “Studying Literature By the Numbers.” New York Times, 10 January 2004, http://www.nytimes.com/2004/01/10/books/studying-literature-by-the-numbers.html 10 March 2018.

Eder, Maciej. “Mind Your Corpus: Systematic Errors in Authorship Attribution.” Literary and Linguistic Computing (2013) 28.4: 603-614. http://llc.oxfordjournals.org/content/28/4/603 8 March 2018.

“Eligibility and Agreements.” HathiTrust Digital Library. https://www.hathitrust.org/eligibility_agreements 4 March 2018.

Eliot, Simon. “The Reading Experience Database; or, what are we to do about the history of reading?” The Reading Experience Database (RED), 1450–1945, The Open University, http://www.open.ac.uk/Arts/RED/redback.htm 5 March 2018

Ely, Alexander. “Woman’s Diplomacy.” Amador Ledger, Jackson, California, 17 May 1907. https://cdnc.ucr.edu/cgi-bin/cdnc?a=d&d=AL19070517.2.58 8 March 2018.

Emw. “Wikistats.” https://web.archive.org/web/20130111123354/http://toolserver.org/~emw/wikistats/ 6 March 2018.

“English Main Page.” Wikipedia, The Free Encyclopedia, http://web.archive.org/web/20100623194907/http://en.wikipedia.org/wiki/Main_Page 6 March 2018.

Epstein, Joshua and Robert Axtell. Growing Artificial Societies. Cambridge: MIT Press, 1996.

Foka, Anna, and Viktor Arvidsson. “Experiential Analogies: A Sonic Digital Ekphrasis as a Digital Humanities Project.” DHQ, 10.2 (2016). http://www.digitalhumanities.org/dhq/vol/10/2/000246/000246.html 15 March 2018.

“Folger Digital Texts.” Folger Shakespeare Library, edited by Barbara Mowat, Paul Werstine, Michael Poston, and Rebecca Niles. Folger Shakespeare Library, http://www.folgerdigitaltexts.org/ 10 March, 2018.

Folsom, Ed, and Kenneth M. Price. Re-Scripting Walt Whitman. Blackwell Publishing, 2005.

Frader, Laura Levine. “Workers, Socialists, and the Winegrowers’ Revolt of 1907.” Peasants and Protest: Agricultural Workers, Politics, and Unions in the Aude, 1850-1914. U of California P, 1991. http://ark.cdlib.org/ark:/13030/ft900009sf/ 7 March 2018.

Friedland, Martin. The Death of Old Man Rice: A True Story of Criminal Justice in America. NYU Press, 1994.

Fukuyama, Francis. “The End of History?” The National Interest, no. 16, 1989, pp. 3–18. JSTOR, www.jstor.org/stable/24027184.

Galloway, Alexander. The Interface Effect. Polity Press, 2012.

Gargan, Edward. “Shen Congwen, 85, a Champion of Freedom for Writers in China.” New York Times, 13 May 1988, http://www.nytimes.com/1988/05/13/obituaries/shen-congwen-85-a-champion-of-freedom-for-writers-in-china.html 6 March 2018.

Garrity, Patrick. “Classics Review: The Personal Memoirs of U.S. Grant.” Claremont Review of Books, Claremont Institute, 16 January 2013, http://www.claremont.org/crb/basicpage/classics-review-the-personal-memoirs-of-us-grant/ 7 March 2018.

Gavin, Michael. “Agent-Based Modeling and Historical Simulation.” DHQ 8.4 (2014). http://www.digitalhumanities.org/dhq/vol/8/4/000195/000195.html 10 March 2018.

Genette, Gérard. Paratexts: Thresholds of Interpretation. Cambridge UP, 1997.

“Getting Content Into HathiTrust.” HathiTrust Digital Library. https://www.hathitrust.org/ingest 4 March 2018.

Gilbert, Nigel. Agent-Based Models. Los Angeles: SAGE, 2008.

Gillman, Susan. Dark Twins: Imposture and Identity in Mark Twain's America. U of Chicago P, 1989.

---. "Humboldt's American Mediterranean," American Quarterly, Spec. issue "Las Américas Quarterly," 66: 3 (September 2014): 505-528. http://www.academia.edu/8454288/_Humboldts_American_Mediterranean_American_Quarterly_Spec._issue_Las_Am%C3%A9ricas_Quarterly_66_3_September_2014_505-528 11 March 2018.

Git. https://git-scm.com/ 11 March 2018.

GitHub. https://github.com/ 11 March 2018.

Glazer, Nathan, and Cynthia R. Field. The National Mall: Rethinking Washington's Monumental Core. JHU Press, 2008.

Griffith, Martin. "Twain: A Man of 50,000 Letters." LA Times, 2 December 2001, http://articles.latimes.com/2001/dec/02/local/me-10527 11 March 2018.

Hamm, Richard F. "Girl Whose Story Caused Father to Kill." Murder, Honor, and Law: Four Virginia Homicides from Reconstruction to the Great Depression. U of Virginia P, 2003, http://books.google.com/books?id=GXsYnI97XSAC&lpg=PP1&pg=PA97#v=onepage&q&f=false 7 March 2018.

Haraway, Donna. "A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century." Simians, Cyborgs and Women: The Reinvention of Nature, Routledge, 1991, pp. 149-181.

Harrison, Les. The Temple and the Forum: The American Museum and Cultural Authority in Hawthorne, Melville, Stowe, and Whitman, U of Alabama P, 2007.

"The Hardy Boys." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/The_Hardy_Boys 4 March 2018.

"Help:Minor edit." Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Help:Minor_edit 5 March 2018.

"Help:Page Information." Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Help:Page_information 6 March 2018.

Henrik. “Wikipedia article traffic statistics.” https://web.archive.org/web/20170721102933/http://stats.grok.se/ 5 March 2018.

---. “Frequent questions.” Wikipedia article traffic statistics, https://web.archive.org/web/20170719105506/http://stats.grok.se:80/about 5 March 2018.

Hermann, Thomas, et al. “Introduction.” *The Sonification Handbook*. Logos Publishing House, 2011. http://sonification.de/handbook/download/TheSonificationHandbook-chapter1.pdf 5 March 2018.

Hirst, Robert. “Textual Editing at the Mark Twain Projet: A Brief Account.” The Mark Twain Project Online, http://www.marktwainproject.org/about_hirst_essay.shtml 4 March 2018

“Home.” Chronicling America. Library of Congress, https://chroniclingamerica.loc.gov/ 5 March 2018.

Hoover, David. “Quantitative Analysis and Literary Studies.” A Companion to Digital Literary Studies, edited by Susan Schreibman and Ray Siemens, Blackwell, 2008, http://www.digitalhumanities.org/companionDLS/ 20 March 2018.

“Hornblow, Arthur 1865-1942.” WorldCat Identities, OCLC, http://www.worldcat.org/identities/lccn-no96001535/ 8 March 2018.

“How library book scanning works.” Books Help. Google. https://support.google.com/books/partner/faq/3396243?visit_id=1-636558156588851242-3769835259&hl=en&rd=1 4 March 2018.

Howe, Lawrence. Mark Twain and the Novel: The Double-Cross of Authority. Cambridge: Cambridge UP, 1998.

Howells, William Dean. “Editor’s Easy Chair.” Harper’s Monthly Magazine, Edited by Henry Mills Alden, Thomas Bucklin Wells, and Lee Foster Hartman, Vol 126, January 1913, pp. 310-312, http://books.google.com/books?id=1Fc2AQAAMAAJ&pg=PA312

Humphreys, Paul. Extending Ourselves: Computational Science, Empiricism, and Scientific Method. Oxford UP, 2004.

“Independent States in the World.” U.S. Department of State, https://www.state.gov/s/inr/rls/4250.htm 4 March 2018

“Information for ‘Adventures of Huckleberry Finn.’” Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=Adventures_of_Huckleberry_Finn&action=info 6 March 2018.

“Inside WorldCat,” WorldCat. OCLC, https://www.oclc.org/en/worldcat/inside-worldcat.html 4 March 2018

"Introduction: Paragraph 2," in Autobiography of Mark Twain, Volume 1. 2010 http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=fr0011#d1e710 11 March 2018.

"Introduction: Paragraph 4," in Autobiography of Mark Twain, Volume 1. 2010 http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=fr0011#d1e719 11 March 2018.

"Introduction: Paragraph 6," in Autobiography of Mark Twain, Volume 1. 2010 http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=fr0011#d1e737 11 March 2018.

"Introduction: Paragraph 77," in Autobiography of Mark Twain, Volume 1. 2010 http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=fr0011#d1e1426 11 March 2018.

"Introduction: Paragraph 128," in Autobiography of Mark Twain, Volume 1. 2010 http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=fr0011#d1e1761 11 March 2018.

Jakobson, Roman. “On Linguistic Aspects of Translation.” The Translation Studies Reader. Edited by Lawrence Venuti, Routledge, 2004, pp. 138-143. http://culturalstudiesnow.blogspot.com/2011/10/roman-jakobson-on-linguistic-aspects-of.html 10 March 2018.

“Japan Society Timeline: Celebrating a Century, 1907-2007.” Japan Society, http://www.japansociety.org/resources/content/9/7/8/documents/CentBook.97-112.pdf 7 March 2018.

Jockers, Matthew, and David Mimno. “Significant Themes in 19th-Century Literature.” [pre-print], August 2012. http://digitalcommons.unl.edu/englishfacpubs/105/ 10 March 2018.

Jütte, R. A History of the Senses: From Antiquity to Cyberspace. Polity, 2005.

Kaku, Michio. The Future of the Mind: The Scientific Quest to Understand, Enhance, and Empower the Mind. Doubleday, 2014.

Kakutani, Michiko. “A Storyteller Enthralled by the Power of Art.” New York Times, 7 October 2010, http://www.nytimes.com/2010/10/08/books/08appraisal.html 6 March 2018.

Kaplan, Justin. Introduction. A Connecticut Yankee at King Arthur’s Court, by Mark Twain, 1889, Penguin, 1971, pp. 9-23.

Kee, Kevin, editor. Pastplay: Teaching and Learning History with Technology. U of Michigan P, 2014, http://dx.doi.org/10.3998/dh.12544152.0001.001 16 March 2018.

Kirschenbaum, Matthew. Mechanisms: New Media and the Forensic Imagination. MIT Press, 2007.

Klein, Charles, and Arthur Hornblow. “The Lion and the Mouse.” The Washington times. (Washington [D.C.]), 19 May 1907. Chronicling America: Historic American Newspapers. Lib. of Congress. http://chroniclingamerica.loc.gov/lccn/sn84026749/1907-05-19/ed-1/seq-50/ 8 March 2018.

Koktsidis, Alexandra. “Computers and Literature: ‘Distant Reading’ Might Be Close At Hand.” WBUR 90.9, 27 February 2015, http://artery.wbur.org/2015/02/27/digital-humanities 10 March 2018.

“Kuroki in New York.” San Francisco Call, Vol. 101, Num. 172, May 21, 1907, http://cdnc.ucr.edu/cgi-bin/cdnc?a=d&d=SFC19070521.2.31 7 March 2018.

Lessig, Lawrence. Remix: Making Art and Commerce Thrive in the Hybrid Economy. Bloomsbury, 2008, https://archive.org/details/LawrenceLessigRemix 11 March 2018.

Levin, David Michael. “A New Ontology of Sight.” Modernity and the Hegemony of Vision. UC Press, 1993, https://books.google.com/books?id=FXeNbyhrlooC&lpg=PA160&ots=qYlfSvQRaL&dq=visual%20hegemony&pg=PA163#v=onepage&q=visual%20hegemony&f=false 11 March 2018.

“A Life Lived in a Rapidly Changing World: Samuel L. Clemens, 1835-1910.” The Mark Twain House & Museum, https://marktwainhouse.org/about/mark-twain/biography/ 11 March 2018.

Linder, Douglas. “Thaw (White Murder) Trials (1907-08).” Famous Trials, UMKC School of Law, 1995-2018, http://www.famous-trials.com/Thaw 7 March 2018.

“List of Wikipedias.” Wikimedia.org, https://web.archive.org/web/20130113055446/http://meta.wikimedia.org/wiki/List_of_Wikipedias 5 March 2018.

“Listening Experience Database.” Open University, https://led.kmi.open.ac.uk/ 13 March 2018.

Loukissas, Yanni, and Mindell, D. “Visual Apollo: A Graphical Exploration of Computer-Human Relationships.” Design Issues, vol. 20, no. 2, 2014, pp. 4-16, http://hdl.handle.net/1721.1/88420 13 March 2018.

Lowry, Richard. “Littery Man”: Mark Twain and Modern Authorship. Oxford UP, 1996.

Lystra, Karen. Dangerous Intimacy: The Untold Story of Mark Twain’s Final Years. U of California P, 2006.

The MARC 21 Format for Bibliographic Data. Library of Congress, Network Development and MARC Standards Office, 1999-2017, http://www.loc.gov/marc/bibliographic/ 5 March 2018.

Marino, Mark. “Critical Code Studies.” Electronic Book Review, 4 December 2006, http://www.electronicbookreview.com/thread/electropoetics/codology 11 March 2018.

“Mark Twain’s 176th Birthday.” Google Doodle, Google.com, 30 November 2011, https://www.google.com/doodles/mark-twains-176th-birthday 6 March 2018.

“The Mask.” Publishers’ Weekly, 24 May 1913. The Publishers’ Weekly, Vol. 83, January-June 1913, New York, 1913. p. 1933 https://books.google.com/books?id=ZiYzAQAAMAAJ&lpg=PA1933&ots=o5-wEoaUWv&dq=MR.%20HORNBLOW'S%20NOVEL.%3B%20THE%20END%20OF%20THE%20GAME.&pg=PA1933#v=onepage&q&f=false 8 March 2018.

McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." 2002. http://mallet.cs.umass.edu 8 March 2018.

McGann, Jerome. “Marking Texts of Many Dimensions.” A Companion to Digital Humanities, edited by Susan Schreibman, Ray Siemens, John Unsworth, Blackwell, 2004, http://www.digitalhumanities.org/companion/ 7 March 2018.

---. A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction, Harvard UP, 2014, https://books.google.com/books?id=JMruAgAAQBAJ&lpg=PA157&ots=sG9AuOnRPE&dq=jerome%20mcgann%20dna%20of%20literary%20studies&pg=PA157#v=onepage&q=jerome%20mcgann%20dna%20of%20literary%20studies&f=false 13 March 2018.

---. The Textual Condition. Princeton: Princeton UP, 1991.

McGill, Meredith L., and Andrew Parker. “The Future of the Literary Past.” PMLA 125.4 (2010): 959-967.

McKeithan, Daniel Morley. Court Trials in Mark Twain and other Essays. Martinus Nijhoff, 1958, https://books.google.com/books?id=lKhDBQAAQBAJ&printsec=frontcover#v=onepage&q=obsession&f=false 16 March 2018.

Michel, Jean-Baptiste, et al. “Quantitative Analysis of Culture Using Millions of Digitized Books.” Science. Published Online Ahead of Print: 12/16/2010. DOI: 10.1126/science.1199644

Michelson, Bruce. “Technology.” *The Oxford Companion to Mark Twain*. Edited by Gregg Camfield, Oxford UP, 2003.

Moretti, Franco. “Conjectures On World Literature.” New Left Review 1, January-February 2000, https://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature 20 March 2018.

---. “Graphs, Maps, Trees: Abstract Models for Literary History--1.” New Left Review 24, November-December 2003, https://newleftreview.org/II/24/franco-moretti-graphs-maps-trees-1 10 March 2018.

---. “Graphs, Maps, Trees: Abstract Models for Literary History--2.” New Left Review 26, March-April 2004, https://newleftreview.org/II/26/franco-moretti-graphs-maps-trees-2 10 March 2018.

---. “Graphs, Maps, Trees: Abstract Models for Literary History--3.” New Left Review 28, July-August 2004, https://newleftreview.org/II/28/franco-moretti-graphs-maps-trees-3 10 March 2018.

Mueller, Martin. “Digital Shakespeare, or Towards a Literary Informatics.” Shakespeare 4.3, 2008, pp. 284-301, http://dx.doi.org/10.1080/17450910802295179 20 March 2018.

MZMcBride. “Hello.” https://web.archive.org/web/20140219145440/http://toolserver.org/~mzmcbride/watcher/ 6 March 2018.

---. “Watcher.” https://web.archive.org/web/20121025135250/http://toolserver.org:80/~mzmcbride/watcher/ 6 March 2018.

“NEH Announces the Winners of the Chronicling America Data Challenge.” National Endowment for the Humanities, 27 July 2016, https://www.neh.gov/news/press-release/2016-07-25 16 March 2018.

NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, 1999, http://ccl.northwestern.edu/netlogo/ 25 March 2018.

Nielsen, Finn Årup, “Wikipedia Research and Tools: Review and Comments.” SSRN, February 6, 2012, http://dx.doi.org/10.2139/ssrn.2129874 5 March 2018.

Open Library. The Internet Archive, https://openlibrary.org/ 22 March 2018.

OpenRefine. http://openrefine.org/ 22 March 2018.

Open Translation Tools. FLOSS Manuals, 2009, http://archive.flossmanuals.net/open-translation-tools/ch029_wikipedia 5 March 2018.

“Our Partnership.” HathiTrust Digital Library. https://www.hathitrust.org/partnership 4 March 2018.

The Oxford Companion to Mark Twain. Edited by Gregg Camfield, New York: Oxford UP, 2003.

“Page History.” Wikimedia Foundation Labs, https://xtools.wmflabs.org/articleinfo 5 March 2018.

“Pageviews Analysis.” Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Wikipedia:Web_statistics_tool 5 March 2018.

“Page view statistics for Wikimedia projects.” Wikimedia.org, https://dumps.wikimedia.org/other/pagecounts-raw/ 5 March 2018.

Paine, Albert Bigelow. Mark Twain: A Biography. Vol. 3, Harper & Brothers, 1912. http://books.google.com/books?id=uf8QAAAAYAAJ&dq=editions%3Ac-Swrv_1ic8C&pg=PA1150#v=onepage&q=advancing%20years%20did%20little&f=false

Perry, Mark. Grant and Twain: The Story of a Friendship that Changed America. Random House, 2004, https://books.google.com/books?id=d6KqdlzWy2EC&lpg=PA234&ots=2f49Ao1-71&dq=The%20most%20remarkable%20work%20of%20its%20kind%20since%20the%20Commentaries%20of%20Julius%20Caesar&pg=PA234#v=onepage&q=Julius%20Cesar&f=false 7 March 2018.

“Perseus Digital Library.” Department of Classics, Tufts University, edited by Gregory Crane, http://www.perseus.tufts.edu/hopper/ 10 March 2018.

Petzold, T., et al. "A World Map of Knowledge in the Making: Wikipedia's Inter-Language Linkage as a Dependency Explorer of Global Knowledge Accumulation." Leonardo 45.3 (2012): 284. Project MUSE, 27 Feb. 2013. https://muse.jhu.edu/article/475466/pdf 5 March 2018.

"Photographs," in *Autobiography of Mark Twain, Volume 1*. 2010 http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=bd0012 7 March 2018.

Possner, Roger. The Rise of Militarism in the Progressive Era, 1900-1914. McFarland, 2009, 119-120, https://books.google.com/books?id=ejNhpt4B-2MC&lpg=PA119&ots=0_3MY-2PX2&dq=admiral%20sakamoto&pg=PA119#v=onepage&q=admiral%20sakamoto&f=false 7 March 2018.

Pratt, Mary Louise. Imperial Eyes: Travel Writing and Transculturation. Routledge, 1992.

Price, Kenneth. “Electronic Scholarly Editions.” A Companion to Digital Literary Studies, edited by Susan Schreibman and Ray Siemens. Blackwell, 2008. http://www.digitalhumanities.org/companionDLS/ 20 March 2018.

“Processing.” Processing Foundation, https://processing.org/, 5 March 2018.

Ramsay, Stephen. “The Hermeneutics of Screwing Around; or What You Do with a Million Books.” Playing with History. The History Education Network/Histoire et Éducation en Réseau. Niagara-on-the-Lake, Canada. April 29-30, 2010. https://web.archive.org/web/20121028010854/http://www.playingwithhistory.com/wp-content/uploads/2010/04/hermeneutics.pdf 16 March 2018.

“Record Number: 5044.” Reading Experience Database. The Open University, http://www.open.ac.uk/Arts/reading/UK/record_details.php?id=5044, 5 March 2018. Rose, Jonathan. The Intellectual Life of the British Working Classes, New Haven, 2001, p. 373.

“Record Number: 5279.” Reading Experience Database, The Open University, http://www.open.ac.uk/Arts/reading/UK/record_details.php?id=5279, 5 March 2018. Rose, Jonathan. The Intellectual Life of the British Working Classes, New Haven, 2001, p. 376.

“Record Number: 5314.” Reading Experience Database, The Open University, http://www.open.ac.uk/Arts/reading/UK/record_details.php?id=5314, 5 March 2018. Rose, Jonathan. The Intellectual Life of the British Working Classes, New Haven, 2001, p. 378.

“Record Number: 11428.” Reading Experience Database, The Open University, http://www.open.ac.uk/Arts/reading/UK/record_details.php?id=11428, 5 March 2018. Gamble, Rose. *Chelsea Child*, London, 1979, p. 143.

“Record Number: 29001.” Reading Experience Database, The Open University, http://www.open.ac.uk/Arts/reading/UK/record_details.php?id=29001, 5 March 2018. Conrad, Jessie. Joseph Conrad and His Circle. London, 1935, p. 252-3.

“Re-Scripting Walt Whitman: An Introduction to His Life and Work.” Commentary. Walt Whitman Archive, http://whitmanarchive.org/criticism/current/anc.00152.html 4 March 2018

“Revision History.” أمريكي_أدب, Wikipedia, The Free Encyclopedia, https://ar.wikipedia.org/w/index.php?title=أمريكي_أدب&dir=prev&action=history 5 March 2018.

Rheingold, Howard. “Social Media.” Rheingold.com, http://rheingold.com/social-media/ 5 March 2018.

Rockwell, Geoffrey. “Playing With History.” https://philosophi.ca/pmwiki.php/Main/PlayingWithHistory 16 March 2018.

Roman, Belinda. “An Agent-based Model for the Humanities.” DHQ 7.1 (2013). http://digitalhumanities.org:8081/dhq/vol/7/1/000142/000142.html 15 March 2018.

Sawada, Mitziko. Tokyo Life, New York Dreams: Urban Japanese Visions of America, 1890-1924. U of California P, 1996, https://books.google.com/books?id=Rr_m9lDbwnQC&lpg=PP1&pg=PA17#v=onepage&q&f=false 25 March 2018.

“SEASR Stopwords.” School of Information Sciences, U of Illinois at Urbana-Champaign, http://repository.seasr.org/Datasets/Text/common_words.txt 30 March 2018.

"Second Preface, in Three Parts: Paragraph 3," in Autobiography of Mark Twain, Volume 1. 2010, 2008. http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=d1e9128#d1e9140

Selden, Daniel. “Text Networks,” in Ancient Narrative, Volume 8, 2010, pp 1-23. http://www.academia.edu/1330782/Text_Networks 7 March 2018.

“Session 347: Varieties of Digital Humanities.” MLA2018NYC, https://mla.confex.com/mla/2018/meetingapp.cgi/Session/2517 13 March 2018.

Shelden, Michael. Mark Twain: Man in White: The Grand Adventure of His Final Years. Random House, 2010.

“SLC and OLC to Pamela A. Moffett, 12 June 1870, Buffalo, N.Y. (UCCL 00479).” In Mark Twain’s Letters, 1870–1871. Edited by Victor Fischer, Michael B. Frank, and Lin Salamo. Mark Twain Project Online. Berkeley, Los Angeles, London: University of California Press. 1995, 2007. http://www.marktwainproject.org/xtf/view?docId=letters/UCCL00479.xml;style=letter;brand=mtp

“SLC to Mortimer D. Leggett ..., 6 Oct 1871, Hartford, Conn. (UCCL 00656).” In Mark Twain’s Letters, 1870–1871. Edited by Victor Fischer, Michael B. Frank, and Lin Salamo. Mark Twain Project Online. Berkeley, Los Angeles, London: University of California Press. 1995, 2007. http://www.marktwainproject.org/xtf/view?docId=letters/UCCL00656.xml;style=letter;brand=mtp

“SLC to Orion Clemens, 11 Aug 1872, New Saybrook, Conn. (UCCL 00791).” In Mark Twain Project Online. Berkeley, Los Angeles, London: University of California Press. 2007.

http://www.marktwainproject.org/xtf/view?docId=letters/UCCL00791.xml;style=letter;brand=mtp

Smith, J. Harvey. “Agricultural Workers and the French Wine-Growers’ Revolt of 1907.” Past & Present, No. 79 (May, 1978), pp. 101-125. http://www.jstor.org/stable/650250 7 March 2018.

“Software Environment for the Advancement of Scholarly Research (SEASR).” School of Information Sciences, U of Illinois at Urbana-Champaign, https://ischool.illinois.edu/research/projects/software-environment-advancement-scholarly-research-seasr 11 March 2018.

The Sonification Handbook. Edited by Thomas Hermann, Andy Hunt, and John Neuhoff. Logos Publishing House, 2011. http://sonification.de/handbook/ 5 March 2018.

“Spanish Main Page." Wikipedia, The Free Encyclopedia, http://web.archive.org/web/20101009140815/http:/es.wikipedia.org/wiki/Wikipedia:Portada 6 March 2018.

Spivak, Gayatri Chakravorty. Death of a Discipline. Columbia UP, 2003.

Strange, Carolyn et al. “Mining for Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers.” DHQ 8.1 (2014). http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html 8 March 2018.

SuperCollider. https://supercollider.github.io/ 25 March 2018.

Svensson, Patrik. “Beyond the Big Tent.” Debates in the Digital Humanities. 2012 Print Edition, http://dhdebates.gc.cuny.edu/debates/text/22 13 March 2018.

tedunderwood [Ted Underwood]. “Category Archives: Topic Modeling.” The Stone and the Shell, https://tedunderwood.com/category/methodology/topic-modeling/ 8 March 2018.

---. “What can topic models of PMLA teach us about the history of literary scholarship?” The Stone and the Shell, 14 December 2012, https://tedunderwood.com/2012/12/14/what-can-topic-models-of-pmla-teach-us-about-the-history-of-literary-scholarship/ 8 March 2018.

Terranova, Tiziana. Network Culture: Politics for the Information Age. Pluto Press, 2004.

Thorp, Jer. "7 Days of Source Day #2: NYTIMES 365/360." blprnt.blg, 12 October 2009, http://blog.blprnt.com/blog/blprnt/7-days-of-source-day-2-nytimes-36536 5 March 2018.

---. "NYTimes Visualizations." Flickr, https://www.flickr.com/photos/blprnt/sets/72157613381549987/ 5 March 2018.

Throne, Jeremy. "An Agent-Based Approach to Reading in 'N-Dimensions.'" Complexity and the Human Experience. Pan Stanford Publishing, 2014, pp. 105-119. https://www.taylorfrancis.com/books/e/9789814463270 10 March 2018.

"Topic Modeling." Machine Learning for LanguagE Toolkit (MALLET), U Mass – Amherst, http://mallet.cs.umass.edu/topics.php 8 March 2018.

"topic-modeling-tool – TopicModelingTool.wiki." Google Code Archive, https://code.google.com/archive/p/topic-modeling-tool/wikis/TopicModelingTool.wiki 8 March 2018.

Trombley, Laura Skandera. Mark Twain's Other Woman: The Hidden Story of His Final Years. Vintage, 2011.

Tuckey, John. Mark Twain's Which Was the Dream? And Other Symbolic Writings of the Later Years. U of California P, 1967.

Twain, Mark. Mark Twain: The Complete Interviews. Edited by Gary Scharnhorst, U of Alabama P, 2006, https://books.google.com/books?id=dOD7ilN1R38C&printsec=frontcover#v=onepage&q&f=false 7 March 2018.

---. Mark Twain's Civil War. Edited by David Rachels, UP of Kentucky, 2010, https://books.google.com/books?id=6q1KjnGbGeUC&lpg=PA86&ots=dMW584u7dW&dq=%22this%20is%20the%20simple%20soldier%2C%20who%2C%20all%20untaught%20of%20the%20silken%20phrase-makers&pg=PA86#v=onepage&q=%22this%20is%20the%20simple%20soldier,%20who,%20all%20untaught%20of%20the%20silken%20phrase-makers&f=false 7 March 2018.

"UN Member States | On the Record." UN.org, http://www.un.org/depts/dhl/unms/whatisms.shtml 4 March 2018

Unsworth, John, et al. "How Not to Read a Million Books." 19 October 2008, Harvard University, Cambridge, MA, http://people.virginia.edu/~jmu2m/hownot2read.html 16 March 2018.

"User talk:Emw." Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=User_talk:Emw&oldid=828527229 6 March 2018

---."3 2010 3.5 Statistics." User talk:Emw, Wikipedia, The Free Encyclopedia, 12 January 2013, https://en.wikipedia.org/w/index.php?title=User_talk:Emw&oldid=828527229#Statistics 6 March 2018.

---."4 2011 4.1 Year Wikipedia traffic." User talk:Emw, Wikipedia, The Free Encyclopedia, 12 February 2011, https://en.wikipedia.org/w/index.php?title=User_talk:Emw&oldid=828527229#Year_Wikipedia_traffic 6 March 2018.

---."4 2011 4.8 wikistats on toolserver." User talk:Emw, Wikipedia, The Free Encyclopedia, 29 July 2011, https://en.wikipedia.org/w/index.php?title=User_talk:Emw&oldid=828527229#wikistats_on_toolserver 6 March 2018.

"User talk:Henrik." Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=User_talk:Henrik&oldid=814602427 6 March 2018.

"User talk:MZMcBride." Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=User_talk:MZMcBride&oldid=828124120 6 March 2018.

---. "94 Number of watchers." User talk:MZMcBride, Wikipedia, The Free Encyclopedia, 12 June 2013, https://en.wikipedia.org/w/index.php?title=User_talk:MZMcBride&oldid=828124120#Number_of_watchers 6 March 2018.

---. "107 Watcher tool count." User talk:MZMcBride, Wikipedia, The Free Encyclopedia, 3 July 2013, https://en.wikipedia.org/w/index.php?title=User_talk:MZMcBride&oldid=828124120#Watcher_tool_count 7 March 2018.

van der Heide, Edwin and Maarten Lamers. "Course Material 'Perceptualization.'" Media Technology MSc Programme, Universiteit Leiden, Fall 2016, http://mediatechnology.leiden.edu/openaccess/course-perceptualization 5 March 2018.

Venuti, Lawrence. The Translator's Invisibility. Routledge, 2004. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.475.4973&rep=rep1&type=pdf 10 March 2018.

"Volunteers' FAQ." Project Gutenberg, http://www.gutenberg.org/wiki/Gutenberg:Volunteers'_FAQ 4 March 2018

Walker, Bruce. Sonification Sandbox. GT Sonification Lab, Georgia Institute of Technology, http://sonify.psych.gatech.edu/research/sonification_sandbox/ 5 March 2018.

Walker, Bruce, and Joshua Cothran. "Sonification Sandbox: A Graphical Toolkit for Auditory Graphs." Proceedings of the International Conference on Auditory Display, Boston, MA, 6-9 July 2003. http://sonify.psych.gatech.edu/publications/pdfs/2003ICAD-WalkerCothran-Sandbox.pdf 5 March 2018.

Walker, Daniel D., et al. "Evaluating models of latent document semantics in the presence of OCR errors." In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 240-250. http://dl.acm.org/citation.cfm?id=1870658.1870682 8 March 2018.

Warncke-Wang, Morten, et al. "In search of the ur-Wikipedia: universality, similarity, and translation in the Wikipedia inter-language link network." Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration. ACM, 2012. http://www.grouplens.org/system/files/p3wikisym2012.pdf 5 March 2018.

Weingarten, Scott. "dhconf." The Scottbot Irregular, http://www.scottbot.net/HIAL/index.html@tag=dhconf.html 10 March 2018.

---. "Submissions to Digital Humanities 2015 (pt. 1)." The Scottbot Irregular, 6 November 2014, http://www.scottbot.net/HIAL/index.html@p=41041.html 10 March 2018.

"Welcome to RED." Reading Experience Database, The Open University, http://www.open.ac.uk/Arts/reading/ 5 March 2018

Westin, Jonathan. Negotiating 'Culture', Assembling a Past. U of Gothenburg Press, 2012. http://hdl.handle.net/2077/30093 15 March 2018.

Whalen, Terence. "Introduction: P.T. Barnum and the Birth of Capitalist Irony." P.T. Barnum: The Life of P.T. Barnum, Written by Himself. U of Illinois P, 2000, http://www.press.uillinois.edu/books/catalog/83gnd8zd9780252069024.html 7 March 2018.

Whitman, Walt. "Song of Myself." Leaves of Grass, 1891. The Walt Whitman Archive, https://whitmanarchive.org/published/LG/1891/poems/27 4 March 2018.

"Why is Mario Vargas Llosa popular right now?" Whatthetrend.com, https://web.archive.org/web/20101011112640/https://whatthetrend.com/trend/Mario+Vargas+Llosa/detail 6 March 2018.

"Wikibu." http://www.wikibu.ch/ 5 March 2018.

"Wikimedia Statistics." Wikimedia, https://stats.wikimedia.org/ 5 March 2018.

"Wikipedia:New pages patrol." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Wikipedia:New_pages_patrol 5 March 2018.

"Wikipedia:Pageview statistics." Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Wikipedia:Pageview_statistics 5 March 2018.

"Wikipedia Statistics Article Count (official)" Wikimedia, https://web.archive.org/web/20130128003703/https://stats.wikimedia.org/EN/TablesArticlesTotal.htm 5 March 2018.

"Wikipedia:Translation." Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Wikipedia:Translation 6 March 2018.

"Wikipedia: Web statistics tool." Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Wikipedia:Web_statistics_tool 5 March 2018.

Wilensky, U. NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, 1999, http://ccl.northwestern.edu/netlogo/ 25 March 2018.

Williams, Raymond. "Culture." Keywords. Oxford UP, 1983, pp. 87-93. https://web.archive.org/web/20160601232919/http://pubpages.unh.edu:80/~dml3/880williams.htm#N_1_ 5 March 2018.

Wilson, Edmund. Patriotic Gore: Studies in the Literature of the American Civil War. Norton, 1994, https://books.google.com/books?id=TPpkiosd5MgC&pg=PA132&lpg=PA132&dq=%22most+remarkable+work+of+its+kind+since+the+commentaries+of+Julius+Caesar%22&source=bl&ots=cyaqXDjG2e&sig=WG-GyrIzU81G55q1kzVpkWFKwAQ&hl=en&sa=X&ved=0ahUKEwimlKWzhNvZAhUR-58KHb4xDloQ6AEINDAD#v=onepage&q=%22most%20remarkable%20work%20of%20its%20kind%20since%20the%20commentaries%20of%20Julius%20Caesar%22&f=false 7 March 2018.

Winger, Howard W. Review of The Brothers Harper: A Unique Publishing Partnership and Its Impact upon the Cultural Life of America from 1817 to 1853 by Eugene Exman. The Library Quarterly: Information, Community, Policy, Vol. 36, No. 1, January 1966, pp. 61-63, http://www.jstor.org/stable/4305622 8 March 2018.

“WorldCat.org frequently asked questions.” OCLC Support & Training. OCLC, http://www.oclc.org/support/questions/worldcatorg/default.htm 4 March 2018

Yang, Tze-I, et al. “Topic modeling on historical newspapers.” In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH '11), Portland, OR, USA, 24 June 2011. Association for Computational Linguistics, 2011, pp. 96-104. http://dl.acm.org/citation.cfm?id=2107636.2107649 8 March 2018.

“008 – Fixed-Length Data Elements-General Information (NR).” The MARC 21 Format for Bibliographic Data. Library of Congress, Network Development and MARC Standards Office, 1999-2017, http://www.loc.gov/marc/bibliographic/bd008.html 5 March 2018.

"1 June 1906: Paragraph 3," in *Autobiography of Mark Twain, Volume 2*. 2013, 2008. http://www.marktwainproject.org/xtf/view?docId=works/MTDP10363.xml;style=work;brand=mtp;chunk.id=dv0016#pa000534 7 March 2018.

"10 January 1906: Paragraph 6," in *Autobiography of Mark Twain, Volume 1*. 2010 http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=d1e9759#d1e9777 7 March 2018.

“11 January 1906," in Autobiography of Mark Twain, Volume 1. 2010 http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=d1e9807 5 March 2018.

"16 January 1906: Paragraph 3," in Autobiography of Mark Twain, Volume 1. 2010, 2008. http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=d1e10370#d1e10379 11 March 2018.

"23 January 1906: Paragraph 72," in *Autobiography of Mark Twain, Volume 1*. 2010 http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=d1e10768#d1e10957 5 March 2018.

"26 March 1906: Paragraph 26," in *Autobiography of Mark Twain, Volume 1*. 2010 http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=d1e13987#d1e14052 7 March 2018.

"26 March 1906: Paragraph 27," in *Autobiography of Mark Twain, Volume 1*. 2010, 2008. http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=d1e13987#d1e14054 7 March 2018.

“403: User Account Expired.” Toolserver.org, https://web.archive.org/web/20140204005017/http://toolserver.org/~emw/wikistats/ 5 March 2018.

“7.7 Wikistats Tool.” User talk:Emw, Wikipedia, The Free Encyclopedia, 5 April 2014, https://en.wikipedia.org/w/index.php?title=User_talk:Emw&oldid=828527229 5 March 2018.

"[The Tennessee Land]," in Autobiography of Mark Twain, Volume 1. 2010 http://www.marktwainproject.org/xtf/view?docId=works/MTDP10362.xml;style=work;brand=mtp;chunk.id=d132371 11 March 2018.