A Gentle Introduction to Text Analysis with Voyant Tools

Madelynn Dickerson
Research Librarian for Digital Humanities and History
UCI Libraries
mrosed@uci.edu

Digital Humanities Working Group

Founded in 2014 as part of the **Humanities Commons**, the DHWG is made up of students and faculty who have scholarly, pedagogical, and personal interests in the Digital Humanities.

- **Co-Chairs:** Madelynn Dickerson and Dwayne Pack
- **Website:** https://sites.uci.edu/dhworkinggroup/
- **Twitter:** @DH_UCI

Madelynn is the Research Librarian for Digital Humanities and History at UCI Libraries (mrosed@uci.edu)

Dwayne is the Director of Computing at the School of Humanities (Dwayne.pack@uci.edu)

UCI Libraries

UCI School of Humanities
Practical Liberal Arts for the 21st Century

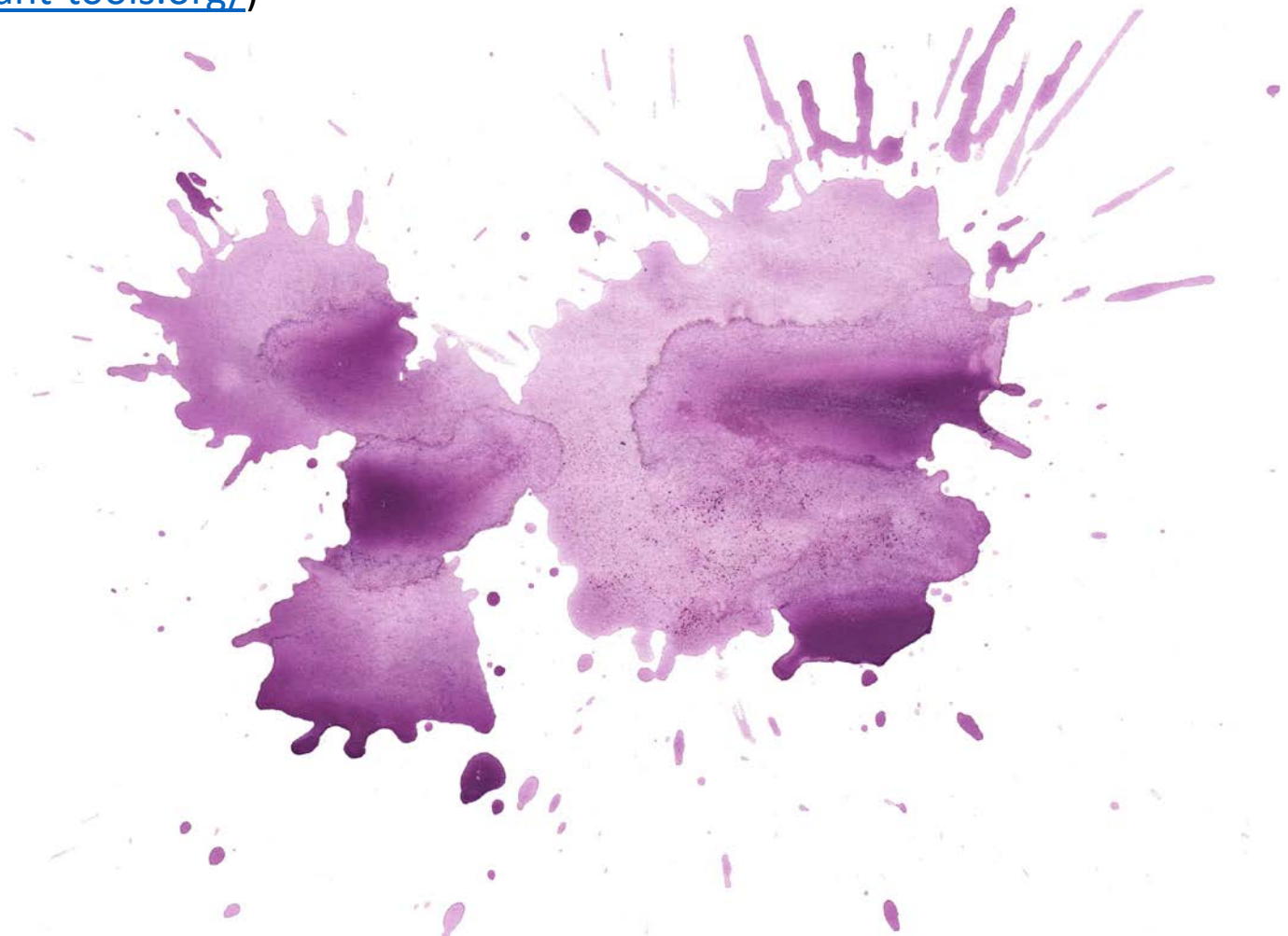
UCI Humanities Commons
SCHOLARSHIP · COLLABORATION · PUBLIC ENGAGEMENT

Workshop Goals

- Learn **basic concepts** of text analysis and why you might want to do it
- Develop **critical awareness** of Voyant Tools' functionality and value as a potential research tool
- **Feel confident** exploring text analysis further on your own

Workshop Arc

1. Introduction to text analysis
2. Uploading our text to Voyant Tools (https://voyant-tools.org/)
3. Explore the Voyant dashboard
4. Resources for Exploring on Your Own



What is text analysis?

Text analysis:

- is arguably a synonym for **text mining**
- is a process for deriving information from texts, such as novels, monographs, articles, web pages etc.
- generally involves detecting patterns, such as identifying word frequency or associative links between words
- combines a qualitative and quantitative approach to research in the humanities
- is not new.

A (Very) Brief History

- Text analysis has been done for hundreds of years
- Using paper technology, it is extremely labor intensive!
- Well-known examples are:
  - Vulgate Bible Concordance (13th century)
  - Father Roberto Busa's Index Thomisticus (20th century)

Let's look at some (heavy) examples and try it out for ourselves…..

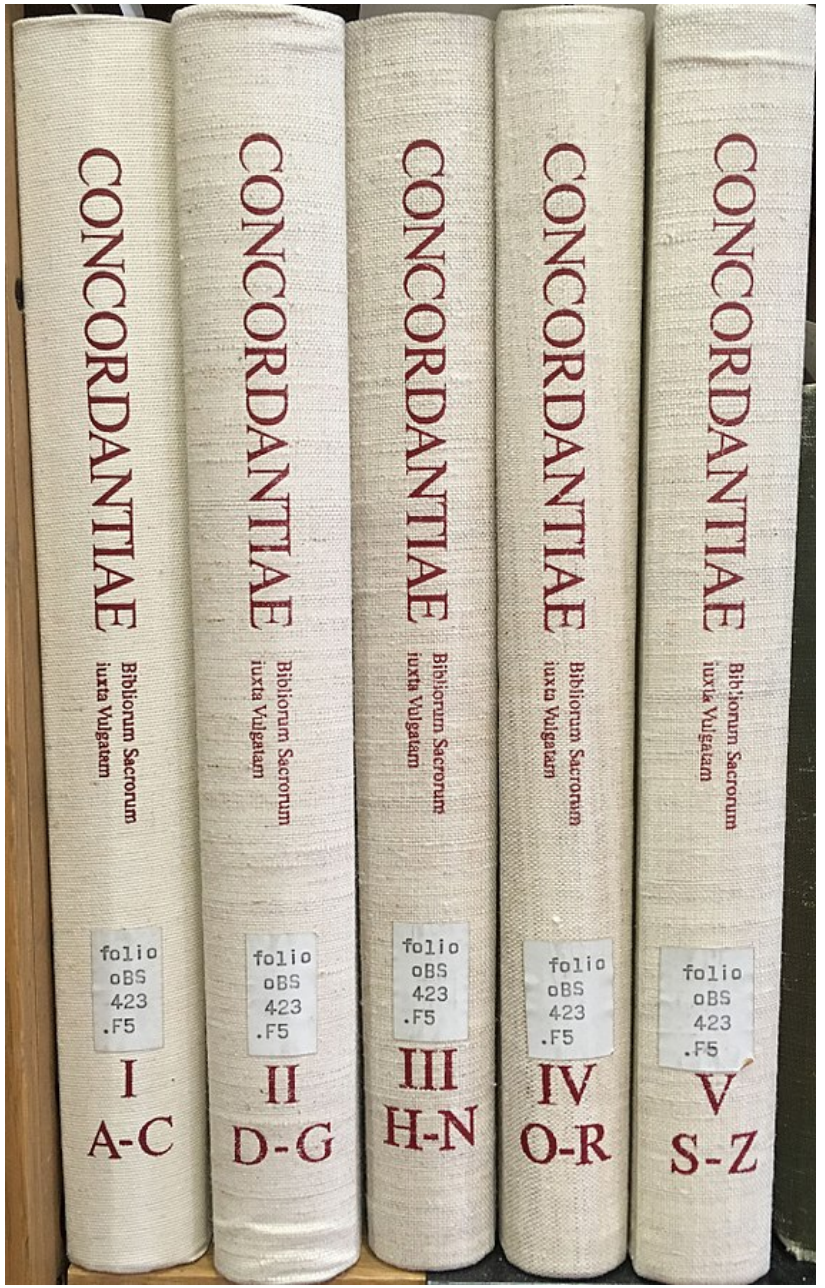


Image: The spines of the 5 volume set of the concordance to the Latin Vulgate Bible. Via Wikimedia Commons.

Voyant Results for Sample Text

(First four paragraphs of Peter Pan)

Summary | Documents | Phrases

This corpus has 1 document with 366 total words and 173 unique word forms. Created about 7 days ago.

Vocabulary Density: 0.473

Average Words Per Sentence: 21.5

Most frequent words in the corpus: wendy (6); darling (4); know (4); got (3); grow (3)

Voyant Results for Sample Text - Comparison

(First four paragraphs of Peter Pan)

Stopwords "On" (Default Setting)

Cirrus | Terms | Links

	Term	Count
1	wendy	6
2	darling	4
3	know	4
4	got	3
5	grow	3
6	kiss	3
7	knew	3
8	mother	3
9	mr	3
10	way	3
11	box	2
12	course	2
13	hand	2
14	house	2
15	like	2
16	loved	2

No Stopwords

Cirrus | Terms | Links

	Term	Count
1	and	16
2	the	16
3	her	13
4	was	11
5	a	8
6	one	8
7	that	8
8	to	8
9	he	7
10	she	7
11	in	6
12	wendy	6
13	of	5
14	they	5
15	up	5
16	all	4

Some Common Jargon in the Text Analysis Universe

API (Application Programming Interface): A specification that allows software applications to communicate with one another. An API allows client programs to access facilities within an application.

Corpus: Pl. Corpora, a collection of written texts, particularly the entire body of work on a subject or by a specific creator; a collection of written or spoken material in machine-readable form, assembled for the purpose of studying linguistic structures, frequencies, etc.

OCR (optical character recognition): The use of computer technologies to convert scanned images of typewritten, printed, or handwritten text into machine-readable text.

Text mining: The process of automatically deriving previously unknown information from written texts using computational techniques. Text mining tools facilitate researchers' discovery of patterns within structured data.

*From Folgerpedia's Glossary of Digital Humanities Terms

Some Common Voyant Jargon Explained

Stopwords: words filtered out before or after processing of natural language data (text), usually words with little meaning such as “and,” “the,” “a,” “an”

Vocabulary density: a measurement of vocabulary usage in comparison to the length of a document. Think of how many words will be read on average before a new word is encountered. (For Moby Dick, a new word appears every 12 words!)

Distinctive words: High frequency words that are relatively unique to a particular document (Only appears in Voyant when comparing multiple documents).

Correlation coefficient: calculated by comparing the relative frequencies of terms. A coefficient that approaches 1 indicates that values correlate positively, that they rise and fall together. Coefficients that approach 0 indicate little correlation. Approaching -1, terms correlate negatively (as one term rises, the other falls).

Voyant Tools (https://voyant-tools.org/)

- Developed by Stéfan Sinclair (McGill University) and Geoffrey Rockwell (University of Alberta)
- Great for getting started with text analysis
- Open-source web-based reading and analysis environment
- Large, robust user community and consistently upgraded and supported infrastructure
- Website appears simple, but there is a lot going on – there are over 20 tools!
- Options for more advanced researchers
- Good (but sometimes outdated) documentation can help you along the way
  - http://docs.voyant-tools.org/start/

Primary Dashboard Tool Names in Voyant

Cirrus: a kind of word cloud showing the most frequent terms

Reader: a view into the corpus that fetches segments of text as you scroll

Trends: a distribution graph showing terms across the corpus (or terms within a document)

Summary: a tool that provides a simple, textual overview of the current corpus

Contexts: a concordance that shows each occurrence of a keyword with a bit of surrounding context

Uploading the Sample Text

Peter Pan [Peter and Wendy] by J.M. Barrie
https://www.gutenberg.org/files/16/16-h/16-h.htm

STEPS

1. Visit Project Gutenberg website using link
2. “Select all” (Ctrl A) and copy (Ctrl C)the Peter Pan text
3. Open new Microsoft Word Document
4. Paste (Ctrl V) the Peter Pan text into the blank document
5. Delete the text that appears before the first chapter heading
6. Delete the text that appears after THE END
7. Select all text again (Ctrl A), and copy your new text (Ctrl C)
8. Open the Voyant Tools website on a separate tab
9. Paste (Ctrl V) the text into the “Add Texts” box
10. Click “REVEAL”

Now let’s look at the Voyant Tools Dashboard!

Word Cloud (Cirrus, Terms, Links)

1. What words are most prominent? How do you know?

2. Spend some time exploring:

    1. What happens when you hover your mouse over different parts of the Cirrus widget?

    2. What happens when you click on a word?

    3. What hidden buttons can you find?

    4. Find the “Options” button and edit or view the list of “Stopwords.” Does this change your results?

3. What do we learn about the text from this Word Cloud?

Trends

1. How are words in the “trends” widget chosen for inclusion?

2. What kind of information to the trend lines, x- and y- axes convey?

3. What does the graph tell us about the story?

4. Bonus Question: What is the difference between **relative** and **raw frequencies** and how do you find that information?

5. Extra Bonus Question: Is this a helpful? What might affect the value of this visualization?

Choose your own tool and explore

Find the drop-down men and choose a new tool we haven’t looked at yet.

1. What is it called?
2. How do you interact with it?
3. What do you learn from it?
4. What is confusing about it?
5. How can you save or export the data?
6. Where can you find help understanding how to use it?

Comparing Corpora

1. Return to Voyant Tools homepage

2. Click “Open”

3. Choose “Shakespeare’s Plays” from drop-down menu

Let’s talk: How is this display different than our Peter Pan dashboard?

Further discussion: For what kinds of projects might you envision using Voyant Tools or similar research methods?

Conclusion: Critical Questions for Text Analysis

We must constantly review and question technology and methodology:

- How complete is the text being analyzed? What is the quality?
- What Optical Character Recognition (OCR) process was used and how accurate was it?
- How was the tool being used created, by whom and for what purpose?
- Is large-scale text analysis the most effective way to draw meaning from words?
- How do we balance “close” and “distant” reading methods to avoid simplistic interpretations?

Resources for Exploring on Your Own: Tools

- **Wordle**
  - http://www.wordle.net/
- **Google Ngram Viewer**
  - https://books.google.com/ngrams
- **Voyant Tools**
  - https://voyant-tools.org/
- **HathiTrust Research Center Analytics**
  - https://analytics.hathitrust.org/
- **Natural Language Toolkit:**
  - https://www.nltk.org/
- **MALLET**
  - http://mallet.cs.umass.edu/

DiRT Directory: https://dirtdirectory.org/

Resources for Exploring on Your Own: Corpora for Analysis

- **Project Gutenberg**
  - https://www.gutenberg.org/
- **HathiTrust Digital Library**
  - https://www.hathitrust.org/
- **JSTOR Data for Research**
  - https://www.jstor.org/dfr/

And **MANY** more, including government documents, social media data, and even historical magazine and newspaper archives licensed by UCI Libraries. Please reach out before starting a large data scraping project!

Check out UC Berkeley's guide to Text Mining and Copyright:
http://www.lib.berkeley.edu/scholarly-communication/publishing/copyright/text-mining

I am working on curating a resource list for our own DH Research Guide.

Voyant (and other DH-related) Tutorials and Learning Resources

Voyant Tools Tutorial Screencasts

- https://www.youtube.com/playlist?list=PLDCADF35691404F54

Programming Historian

- https://programminghistorian.org/

Intro to Digital Humanities (DH 101) from UCLA

- http://dh101.humanities.ucla.edu/

Stay in touch!

Digital Humanities Working Group
https://sites.uci.edu/dhworkinggroup/

Twitter
@DH_UCI

Email
Madelynn Dickerson mrosed@uci.edu
Dwayne Pack Dwayne.pack@uci.edu