Silence in OCR: What could handwritten documents tell us?

Theo Zhang

UCSF Archives and Special Collections Data Science Fellowship

August 16, 2024

0 Table of Contents

1 Intro

My name is Theo Zhang, and I am an incoming senior studying Computer Science at UCLA. In the past year, I have fallen in love with research regarding responsible artificial intelligence and the applications of artificial intelligence, so I was extremely excited to have the opportunity to conduct research over the summer here at the UCSF Archives and Special Collections. My main areas of interest are machine learning and artificial intelligence, specifically taking a humanistic approach to machine learning and ethical artificial intelligence. You will see these considerations in my project as you continue reading!

Before I jump into explaining my project and findings, I would first like to extend a huge amount of gratitude to the people at UCSF who have supported me throughout this internship:

- Lisa Nguyen, my supervisor, whose insights and willingness to be my audience as I bounced many ideas off of her led to many of my “eureka” moments throughout my research. She has been the best, kindest, and most helpful supervisor ever.
- Sean Purcell, who devoted many consultation hours to helping me carve out a meaningful structure to my project, which turned out to be much more difficult than I anticipated.
- Geoffrey Boushey, whose technical experience and expertise on finding new tools to try, helped me expand my project in a way that will hopefully help future researchers and librarians.
- Rebecca Tang, who led many meetings for the interns and helped guide us through the summer.

- Peggy Tran-Le, Kate Tasker, and Rachel Taketa who devoted their time to meeting with me, where they gave me very useful feedback on my project and pep talks that kept me going.
- Gordon Lichtstein, my co-intern for the summer, who was a pleasure to collaborate with and have as my counterpart on the team.
- Everyone from the Industry Document Library (IDL) and Archives and Special Collections (ASC) who watched my presentation, gave me feedback or pieces of encouragement, and helped my project along in any way this summer! I am so blessed to have joined (even for one summer) a wonderful team that loves what they do and gave me the resources to contribute!

My project ended up being much more interesting and comprehensive than I could have imagined at the start, so I will do my best to make my process and findings as interesting as possible. I hope you enjoy the culmination of many, many weeks of my work!

2 Project background

As I had mentioned previously, my interests revolve around a humanistic approach to machine learning, and I am especially interested in unintentional consequences that many of the new tools we use can cause. One tool that has seen great improvement in the past couple of years is optical character recognition (OCR), a process that "extracts and processes text from images automatically" in order for computers to parse images with text on them easier (i.e., a photo of a page out of a book) (Hamad and Kaya 2016). OCR is a very promising tool for organizations like archives, as it allows for the opportunity to unlock more information from the documents they already have. Many documents are "hidden" due to the sheer amount of information that one would have to look through to discover a useful or relevant piece.

For example, UCSF ASC alone has over 20 million documents and IDL has about 50 terabytes of data of digital/digitized data from approximately 350 archival collections. Being able to OCR all of these documents would allow a researcher to search through the documents much quicker and find more information rather than looking through it by hand. However, OCR has important drawbacks that need to be addressed in research that uses OCR-ed material and in the act of OCR-ing itself.

I was drawn to the "No More Silence" dataset, a project from the UCSF Library AIDS History Project that "[extracted] text from digitized archival documents on HIV/AIDS Epidemic" in order to create a "patient-centric view of AIDS/HIV," as it was focused on humanizing a subject through more representative datasets (Macquarie 2021). However, in talking to Geoffrey Boushey, who is the Head of Data Engineering at the UCSF Library, I was made aware that this

dataset focused more on typed documents than handwritten documents due to the OCR limitations in 2019[1].

This leads me to the most pressing drawback of OCR that I will be focusing on in my project: OCR tools previously used at UCSF ASC for handwritten materials is notably subpar compared to typewritten tools according to my examination of the accompanying OCR from the "No More Silence" dataset. Handwritten OCR would be full of random characters and symbols, making it generally completely gibberish while typewritten OCR would be almost completely perfect most of the time. Additionally, in talking to the ASC and IDL teams, the resounding agreement was that researchers are more likely to use typewritten documents over handwritten documents if OCR is used in a project due to the poor quality of handwritten OCR. The goal of the "No More Silence" dataset and broader project was to provide a more "patient-centric view," but it couldn't provide accurate OCR for handwritten documents due to technological limitations. What kind of biases could it have introduced that contradicted the main goal of the project? Was it possible that the dataset then missed important information in handwritten documents that would have furthered the goal of the project? These questions then led me to my main research questions that I spent the next couple of weeks trying to answer.

[1] *UC Tech 2023 - Bias and Data Loss in Transcript Generation*. Geoffrey speaks more about OCR and its limitations in this talk. It was also where I got inspiration for examining the loss of information from poor handwritten OCR.

3 Research Questions

My research questions I attempt to answer are:

1. Is there an OCR tool that would be able to OCR both typed and handwritten documents well?
    a. Specifically, this tool should be feasibly used in the digital archival process at UCSF.
2. What are the benefits to using tools that have the best handwriting OCR available to us?
    a. What is the motivation behind using better tools to OCR handwritten documents?
3. What value do handwritten documents add to a research project?
    a. Why should we strive to include handwritten documents in our research despite the potential increase in time and effort and despite the subpar OCR?

4 Project Overview

This project has two parts to it: tool analysis and programmatic analysis/close reading.

4.1 Tool Analysis

The first part of my project is analyzing the OCR tools that UCSF ASC has to their disposal potentially. I utilize my co-intern Gordon's work in this section of my research!

4.2 Programmatic Analysis and Close Reading

The second part of my project is analyzing the actual content of my data in order to grasp if there is a difference in content between the different categories of data. Additionally, I try to determine what consequences those differences can have on a project.

4.3 Project Timeline

Week 1:

- Onboarding, meeting the team
- Completing trainings
- Getting access to various resources

Week 2 & 3:

- Reading and sorting through all the data
- Figuring out what research questions are emerging from the data

Week 4:

- Finalizing research questions

- Forming final dataset
- Starting Tool Analysis part of project

Week 5 & 6:

- Completing Tool Analysis part of project
- Presenting on my work thus far to the Library
- Speaking with librarians and researchers one-on-one about my work and getting feedback on future steps

Week 7 & 8:

- Completing Programmatic Analysis and Close Reading part of project
- Speaking with more librarians and researchers on my progress and getting feedback

Week 9 & 10:

- Wrapping up my data gathering
- Writing my final intern project report/blog post

5 Data

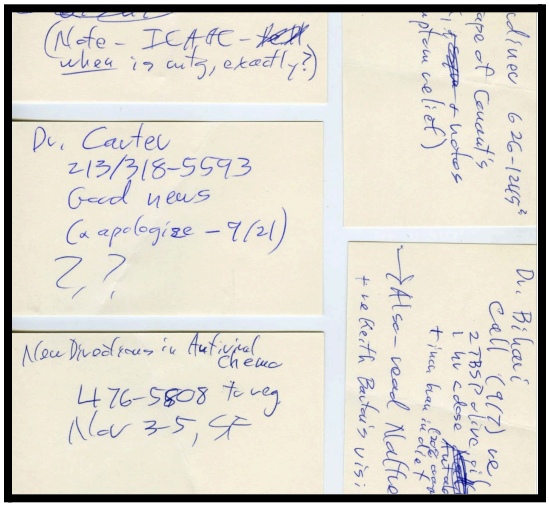
5.1 Sources

I use three sources of data that focus on organizations and prominent figures from the AIDS/HIV epidemic:

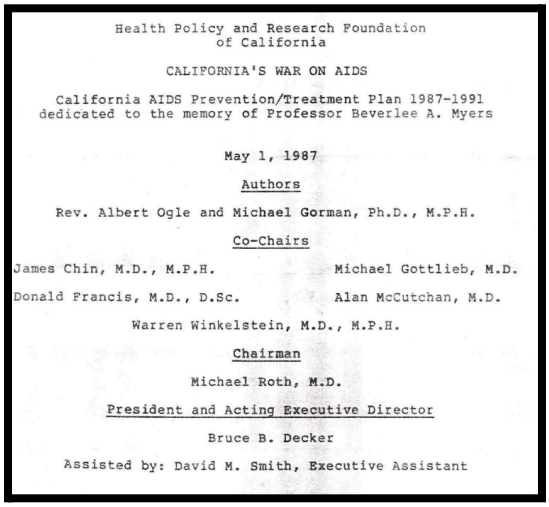
1. "No More Silence", specifically "Donald Francis, 1970-2005, MSS 2015-1"
    a. Frances was an epidemiologist and worked at the United States Centers for Disease Control (CDC) and other public health organizations. This source contains his notes, work, and other documents from his time.
2. AIDS Treatment News Records
    a. This source contains documents such as memos, meeting notes, articles, and more from the AIDS Treatment News Records (ATNR). ATNR was "a publication created by John S. James that investigates and reports on both conventional and experimental treatments for HIV/ AIDS and related social and political issues."
3. San Francisco AIDS Foundation Records
    a. This source similarly contains documents such as memos, meeting notes, forms, and more from the San Francisco AIDS Foundation (SFAF). SFAF was "[A] major resource center for educating the public in order to prevent the transmission of HIV, helping all individuals make informed choices about AIDS-related concerns, and protecting the human rights of those affected by HIV."

5.2 Method of Collection/Creation

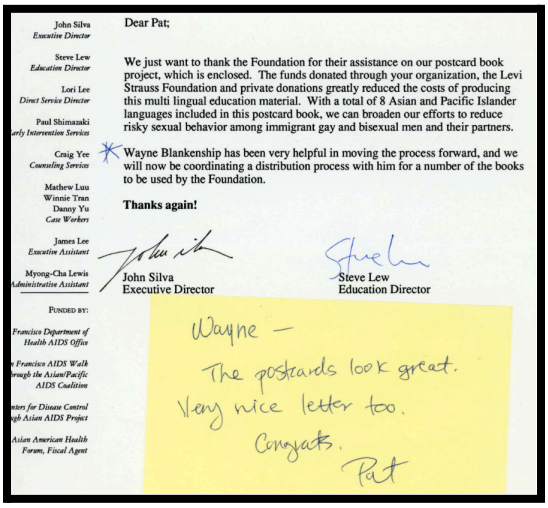
My dataset consists of 30 total one page long documents from the aforementioned sources. There are 3 categories of documents: handwritten, typewritten, and mixed.



Handwritten



Typed



Mixed

Handwritten documents are completely handwritten, and typewritten documents are usually almost completely typed. Some typewritten documents have a signature on it, which I did not consider enough handwriting to categorize it as mixed. Mixed documents have approximately at least 20% handwriting or typewriting on it.

To create this dataset, I had to hand-sort through thousands of documents where I chose documents based on diversity of handwritings, fonts, document layouts, and more. I mainly focused on a diverse and well balanced dataset which is why the dataset is on the small side. Future scholarship could look at this at a larger scale, using the method I've outlined here.

5.3 Data Limitations

Because my internship was only 10 weeks long, I chose to keep the dataset small despite the extremely large source dataset sizes so I could fully analyze the documents in the second part of my project. I was also limited to what has already been scanned and uploaded from the archives, which means I potentially could be missing more context; however, my dataset is small and undoubtedly unable to capture every detail as it is. Additionally, because I am a native English speaker, I limited my documents to English only, but there is a very small subset of the source datasets that had other languages such as Tagalog, Spanish, and Mandarin. Hopefully I can address these limitations in future iterations of this project; more future improvements will be addressed later in this article.

6 OCR Tool Analysis

My code and results can be accessed here. The keyword section's code and results are in the Content Analysis notebook and folder.

6.1 Research Question

The research question I am attempting to answer is "Is there an OCR tool that would be able to OCR both typed and handwritten documents well?" My goal for this portion of the project was to find a tool that can be used for the digital archival process for UCSF ASC or conclude that there are currently no tools on the market that fulfill those specific needs.

Currently, UCSF ASC has two methods of providing materials to researchers and librarians. One method is the site Calisphere, which provides researchers and librarians access to digitized materials. Calisphere does not allow for full text search and OCR, but it does allow for keywords to be attached to materials. Additionally, UCSF ASC creates and distributes datasets, "No More Silence" being one example of such. These two forms of accessing documents will also inform the metrics of evaluation for each tool.

6.2 Tools For Consideration

There are multiple OCR tools available nowadays, but they all vary in quality, price, and more. There are 5 tools I wished to analyze, but I am only including 3 of those tools fully in my final analysis. I initially planned on analyzing Tesseract, Document AI, Versa, Textract, and Doxie but only ended up being able to analyze Tesseract, Document AI, and Textract.

Tesseract is a free, open source OCR tool developed by Google. Document AI is an generative artificial intelligence development platform supported by Google Vertex AI; in my code and resulting spreadsheets, Document AI is referred to as Vertex AI. Textract is a tool developed by Amazon Web Services (AWS), and it is a machine learning OCR tool. From Textract, this project specifically uses the "Detect Document Text" API feature.

Generative artificial intelligence is "a type of machine learning system that generates realistic and credibly human-like content (e.g., text, images, code, audio) in response to an input" (Woodruff et al. 2024). Many OCR tools now use generative artificial intelligence to aid in creating outputs, such as Document AI.

I am unable to include Versa as UCSF was not able to provide me access to it in time. Doxie is also not included, as they provide a custom pipeline to analyze all the documents for an organization. Doxie was only able to provide me with a small sample of their product which was completely accurate, but it was only for one document. Thus, I could not fully test their product and fairly compare it to the other "off the shelf" type products.

Other popular products I will not be considering in this report are eScriptorium, Transkribus, and ABBYY. eScriptorium has heavy collaboration with Tel Aviv University and other Israeli establishments, and in consideration of the current ongoing humanitarian crisis, I chose not to include it. Additionally, the decision to not include eScriptorium is in line with the statements released by other librarians and scholars ("2023 Statement on Gaza – Librarians and Archivists

with Palestine," n.d.). Transkribus does not provide one model that can be used for both typewritten and handwritten material, and the initial exploratory results were not promising and thus was not included. ABBYY is a popular tool as well, but it is not accessible to me as a student researcher at this point.

6.3 Methods and Methodology

Each tool is scored on 3 metrics that are weighted based on importance:

1. Accuracy and quality of OCR
2. Speed
3. Price

Additionally, another metric that is not scored but is considered is keyword analysis.

Because there are 3 OCR tools being evaluated, all 30 documents are run through each OCR tool. This results in 3 different OCRs per document that are then evaluated by category (handwritten, typewritten, and mixed).

6.3.1 Accuracy of OCR

I use Gordon's research on the best method to indicate the quality of OCR for the first metric. His research concluded that the overall most effective method was a tool called Nostril, or Nonsense String Evaluator, which I use in my research to output a percentage that represents the percent of nonsense in the OCR.

The OCR from each document for each tool is evaluated using Nostril, and then averaged for each type of document. There then are 9 resulting scores: one score for each category for each OCR tool.

6.3.2 Speed

Speed is evaluated based on wall time (or the real perceived time) it takes to finish generating the OCRs for all 30 documents. To determine this, the Python3 command "%%time" is utilized.

6.3.3 Price

Price is determined based on the price per page for the API according to each tool's website. API means Application Programming Interface and in this use case, this means that the OCR tool is able to be used through a command in a piece of code rather than having to input the documents into an interface on a website.

6.3.4 Keyword Analysis

From the resulting OCRs, keywords can then be extracted using the Natural Language Toolkit library. Once keywords are extracted, the top 5 keywords for each document for each tool's OCR are examined heuristically (by hand) in order to determine how granular/cohesive the results are.

I chose to analyze the results by hand because I found that assigning a score to the keywords using another tool was too vague and missed issues such as nonsensical words counting as a keyword due to poor OCR results.

6.4 Results

6.4.1 Accuracy of OCR

	Nonsense Percentage per Categories		
Tool	**Handwritten**	**Typed**	**Mixed**
Tesseract	63.73%	35.23%	59.47%
Document AI	36.31%	13.29%	30.45%
Textract	29.88%	7.68%	26.43%

The percentages in the table represent the percentage of nonsense on average for the resulting OCR for each category for each tool. Thus, a higher percentage indicates worse OCR quality. As shown, in ascending order of performance, Tesseract did the worst, Document AI did the second best, and Textract did the best.

Additionally, this result also further proves how current OCR tools struggle on handwritten and mixed documents much more than they do typed. The quality of OCR for handwritten documents is much worse than typewritten documents, which is very clearly demonstrated in the resulting OCR. For example, a handwritten document can be pure gibberish while a typewritten document transcribed by the same tool can almost be a 1:1 transcription.

6.4.2 Speed

Tool	**Time**
Tesseract	3 min 39 sec

Document AI	1 min 5 sec
Textract	1 min 29 sec

For speed, Document AI did the best, followed by Textract, followed by Tesseract. It's worth noting that times do vary when the program is re-run, but not by a degree that impacts the rankings of the final speed or by a large amount (<5 second variations).

6.4.3 Price

Tesseract is free, so it is the most inexpensive tool on this list. Textract (specifically the "Detect Document Text API") and Document AI have the same price of $1.50 per 1000 pages for the first million pages. However, Textract is cheaper because it is then reduced to $0.60 per 1000 pages after using it on one million pages. On the other hand, Document AI only reduces to $0.60 per 1000 pages after using it on five million pages. Thus, for use cases under one million pages, Tesseract is the cheapest and Textract and Document AI are tied. For cases where over one million pages are being OCR-ed, Tesseract is still the cheapest but then Textract is the next cheapest followed by Document AI.

6.4.4 Keyword Analysis

The quality of keywords is inversely directly related to the percentage of nonsense in the OCR; the worse the quality of the keywords, the higher the percentage of nonsense in the OCR. Thus, many of the keywords extracted from poor quality OCRs are pure gibberish.

From the better quality OCRs, the keywords were still rather vague and potentially unhelpful to researchers overall. It varies in specificity and granularity from document to document, so it has potential to be helpful in certain cases.

6.5 Tool Analysis Conclusion

Tool	Accuracy/Quality	Speed	Price	Total
Tesseract	3	3	1	2.6
Document AI	2	1	2	1.8
Textract	1	2	3	1.6
Weight	0.6	0.2	0.2	

Overall, based on all the considerations, Textract would be the most suitable tool out of the three analyzed. In the table above, each tool is ranked from 1-3 (1 being the best, 3 being the worst), and Textract ended up with the best ranking by a small margin, beating out Document AI by 0.2 points. Textract does have the best quality OCR, but supports fewer languages than Document AI which could be an issue that needs to be addressed by introducing new OCR tools to the pipeline.

Additionally, as mentioned earlier, Calisphere does not currently have full-text search capacity, which limits the ability to search OCR-text.. Thus, because the keywords depend on the quality of OCR, Textract may not even be a suitable option for providing OCRs of passable quality for keyword extraction especially for handwritten text for the current distribution platform. Many documents uploaded to Calisphere and datasets created by UCSF ASC also may be hundreds of pages long, meaning a handful of keywords is not specific enough to be useful unless each page is assigned keywords the way this project does it.

Textract thus is an OCR tool that is suitable for the digital archival process for both handwritten and typewritten material, but only for datasets that are not distributed through Calisphere. If UCSF ASC distributes datasets, similar to the “No More Silence” dataset where the full OCR and keywords for each page can be provided, then Textract will work very well. However, for a platform like Calisphere, the keywords generated through Textract’s OCRs may not be enough to be of use to researchers.

Through this section of my project, one can also observe how better OCR results in better keywords which is a clear benefit in using tools that have the best handwriting OCR available. More generally, this supports how better OCR creates cleaner downstream results that use the generated OCR. This is further proven in the next section of this project through analyzing the content of the documents both using the OCR and not using the OCR.

7 Programmatic Analysis and Close Reading

My code and results can be accessed here under all results labeled Content Analysis.

7.1 Research Questions

For this portion of my project, I attempt to answer the following:

- What are the benefits to using tools that have the best handwriting OCR available to us?
- What value do handwritten documents add to a research project?

7.2 Methods and Methodology

Because I am analyzing the content of the dataset, this section will be more focused on qualitative interpretations of results and the data.

7.2.1 Sentiment Analysis

To produce sentiment scores for each OCR, I use a Python library named TextBlob which is free to use. TextBlob's sentiment analysis function is able to return two scores: subjectivity and polarity. Subjectivity is scored in a range of [0,1] where a value closer to 0 indicates a piece of factual information and a value closer to 1 indicates a personal opinion. Polarity is scored in a range of [-1,1] where -1 indicates a highly negative sentiment and 1 indicates a highly positive sentiment.

7.2.2 Topic Modeling

Topic modeling essentially allows a piece of text to be distilled down to a couple of topics chosen from a predefined list. I used another API from Vertex AI named Google Cloud Natural Language API that would take an OCR and return a list of corresponding topics (defined by Google).

Similar to the keyword analysis, the resulting topics are then analyzed by hand to determine accuracy and granularity.

7.2.3 Categorical Details

From the resulting OCRs, categorical details (such as people names, organizations, and dates) can be extracted using spaCy's Named Entity Recognition (NER) system which is free to use. Once categorical details are extracted, they are examined heuristically (by hand) in order to determine how granular/cohesive the results are.

7.2.4 Close reading

Close reading involves reading all 30 documents one by one and noting down the main topic, the overall mood, and the granularity of the document.

The granularity of a document is defined in three levels:

1. The most granular, this document involves specific people, internal policies, and internal meetings.

2. Getting more general, this document is concerned with the broader organization (San Francisco Aids Foundation, AIDS Treatment News, or other) such as how an organization interacts with the public as an entity.
3. Lastly, the broadest level categorizes a document that concerns the AIDS/HIV topic as a whole, such as drug treatments, government policies, and activist movements.

Once all of the documents have been close read, they are then analyzed by hand between categories to note any differences across data types.

7.3 Results

7.3.1 Sentiment Analysis

Tool	Format	Subjectivity	Polarity
Tesseract	handwritten	0.1062878788	0.03196969697
	typed	0.3039428764	0.05501719441
	mixed	0.3230204753	0.06408692854
Document AI	handwritten	0.3176355984	0.1256177959
	typed	0.3061420117	0.03126203602
	mixed	0.4236769685	0.1243974403
Textract	handwritten	0.285746152	0.1110529101
	typed	0.3322950697	0.05897745969
	mixed	0.4366945076	0.1320777778

The first important result this table shows us is that OCR quality greatly impacts both subjectivity and polarity scores. From the first part of the project, Tesseract's OCRs have the highest amount of nonsense whereas Document AI and Textract perform better and more

similarly. Tesseract's handwritten document OCR results also have roughly twice the amount of nonsense compared to the handwritten document OCR results of the other two tools.

Now, when we take a look at sentiment analysis results, typed and mixed documents score similarly across all three tools in both subjectivity and polarity. However, specifically for handwritten documents, the two scores for Tesseract's results are significantly lower compared to the more similar scores from Document AI and Textract's results. Tesseract's results for handwritten documents are lower because for the OCRs that were essentially complete nonsense (and Tesseract produced more of those comparatively), those got a score of zero for both subjectivity and polarity.

It is then clear that the lower quality an OCR is, the more likely the sentiment analysis tools will mark it as closer to neutral or zero. Thus, the quality of an OCR significantly impacts the results of sentiment analysis, making lower quality OCRs produce sentiment analysis scores that are potentially more inaccurate or untrustworthy. Sentiment analysis is a tool that is commonly and increasingly used to extract information from a dataset, and handwritten documents are more likely to have lower quality OCR, proven in the first part of this project. As a result, handwritten documents are then more likely to have inaccurate sentiment analysis results compared to other types of documents, resulting in cascading negative impacts on a research project that might use both OCR and sentiment analysis.

Even though the OCRs from Textract and Document AI have significantly less overall nonsense compared to Tesseract, both tools produced handwritten OCRs that contain roughly triple the

amount of nonsense compared to their typewritten OCRs. As a result, these sentiment scores may not be very accurate. I am thus choosing to not directly interpret the implications of the actual sentiment scores between document categories here and instead will be analyzing the content through close reading later on in the project.

7.3.2 Topic Modeling

	Topic Count per Category		
Tool	**Handwritten**	**Typed**	**Mixed**
Tesseract	5	46	13
Document AI	7	44	24
Textract	6	45	21

	Documents per Category With >0 Topics		
Tool	**Handwritten**	**Typed**	**Mixed**
Tesseract	2	9	5
Document AI	4	10	6
Textract	3	10	6

The list of topics that Google Cloud Natural Language API drew from was relatively vague, so the resulting topics are not very promising. Many of the OCRs resulted in topics such as “Health” or “People & Society,” which may be too vague to be very useful for researchers. The majority of the handwritten OCRs across tools did not even result in topics, even for the tools that had better OCR like Textract and Document AI and when they did get topics assigned, the amount of topics were far fewer than typed or mixed documents received. Once again, the quality of OCR has a significant effect on the results of a downstream task.

However, if a custom, more specific list could be used in this case, being able to generate topics could be immensely helpful for documents that are hundreds of pages long. A general list of topics for those longer documents may not be helpful on Calisphere, but it can be utilized for datasets that UCSF ASC creates and distributes outside of Calisphere. Along with an OCR for those datasets, an index could be generated for those documents with page numbers and topics or keywords to make the data more accessible to researchers.

7.3.3 Categorical Details

A numerical representation of the number of total categorical details extracted would be inaccurate because similar to the keyword extraction, some details are nonsense. However, by sorting through the results by hand, there are some promising results. Overall, many important categorical details such as names and dates were able to be captured which again can be used in an index for datasets. Following the prior results, due to the lower quality of handwritten OCRs, the extracted details for handwritten documents are much more likely to be gibberish.

Even though the results are not perfect, the tool does work very well when the OCR is readable. Thus, if OCR technology improves in the next couple of years (which it very likely will), this and the keyword extractor may have more potential as another way to generate indexes or key details without having to resort to more monetarily, resource, and environmentally expensive generative artificial intelligence tools.

7.3.4 Close reading

Category	**Average Granularity**
Handwritten	1.5

Typewritten	2.7
Mixed	1.6

On average, handwritten documents were more granular than typewritten documents and about the same granularity as mixed documents. For this dataset and for the original sources, the different levels of granularity are emphasized by a rough breakdown of different types of documents that emerged through close reading:

- Typewritten:
    - Official documents and forms
    - Formal correspondence
    - Published or near finished articles
    - Official meeting notes
- Handwritten/mixed:
    - Day-to-day operations and memos of an organization
    - Informal correspondence
    - First drafts or unpublished articles
    - Informal meeting notes that had more (perhaps unnecessary) details compared to the typed, official versions

Through this breakdown, it is clear that handwritten and typewritten documents contain different types of information that concern different levels of an organization and movement, specifically AIDS/HIV organizations during the 1980's and 1990's.

I initially hypothesized that handwritten documents would have more emotion in them because it was possible that people would write down more personal opinions or thoughts before typing out

an official version and editing those out. However, I actually found that handwritten documents contained mostly scribbled notes that might have contained more factual information, but generally did not contain many emotional or personal opinions. Typewritten documents, such as news articles or strongly worded letters, actually contained more emotional aspects to them, which then could implicate how factual notes, such as hastily written down meeting notes, then turn into politicized pieces of news articles.

7.4 Programmatic Analysis and Close Reading Conclusion

Through analyzing the content of the documents using an algorithmic approach like sentiment analysis, topic modeling, and categorical data extraction, the benefits to using tools that generate the best handwriting OCR available to us become clear. Simply, the better the OCR, the more reliable and accurate the results from these methods are. Due to the fact that OCR for handwritten documents are at a much lower quality than typewritten documents, these tools are then shown to not be the most effective way to analyze a group of mixed-type documents. The lower accuracy for handwritten documents could introduce biases into a research project if those results are taken at face value without closer examination through techniques such as close reading by hand.

In a broader research context, the results of the close reading imply that handwritten documents provide a deeper context to a research project, specifically concerning details that may not ever show up in typewritten documents. By excluding or lowering the scrutiny for handwritten documents, researchers could miss more intimate detail in their research subjects and important contextual information, especially about the people of an organization, how an organization was

run, and what tensions were there between different groups within and between organizations. Those details may never be officially documented in typed documents, especially in the time period of this project's focus when computers and the internet had yet to be all encompassing and typing notes was more commonplace.

8 Conclusion

It is important to note that my project is focused on sources from notable AIDS/HIV organizations and organizers around the 1980's to early 1990's. Nowadays, typing notes and communicating with others through digital means is much more common than handwritten forms, and a century ago, handwritten materials might have been the most common. It just happens that my project encompasses a time period where handwritten and typewritten forms of communication were both essential and commonplace which presents an interesting look into how the two differ. Thus, this project's conclusions may slightly differ in different time periods and with different source materials; perhaps the conclusions to how handwritten and typewritten documents differed would be different or handwritten documents could be completely absent from the sources.

What this project proves is that no matter the type of document or source a research project utilizes, excluding certain types of materials can lead to biases that could change any conclusions drawn from the research. Exclusions can happen organically as well – they are not necessarily a conscious decision when collecting data. The bias could be generated by the tools at hand themselves unknowingly, proving the importance of also critically evaluating methods one may assume as unbiased or trustworthy. This is especially important as tools like generative AI and machine learning become more and more popular; those tools do not produce easily interpretable results, yet are often treated as producing infallible truths.

Tools like OCR, artificial intelligence/machine learning, and automated information extraction are now being utilized more in the archival field, and it is very likely that this will have a positive

impact on making data more accessible to researchers and faster to process. However, in the archival field, being able to trust the tools you use is especially important because misrepresenting history can lead to extreme consequences, such as adverse political and social repercussions. In order for these artificial intelligence systems for archives to continue to be usefully, ethically and properly implemented, they will have to "recognize and maintain the nature and trustworthiness of archival material and its context(s); provide trusted access to archives that respects privacy rights; document and make transparent the provenance of material derived from different sources and combined in ways different from its original purpose" (Rogers 2024).

Even if these artificial intelligence systems are designed with that framework in mind, many of these tools are still in their nascent stages with opaque processes and inaccurate results and thus not entirely to be trusted. Ensuring that a human is checking over the work of any generated output (machine learning or not) is then also important in the archival field, as I found out in this project with inaccurate OCRs, sentiment scores, and keyword extractions. This is called a "human in the loop" (HITL) system, which can be used to review the work of automated systems. The issue with the HITL system is when humans overestimate the abilities of these tools; to combat this, it is important for researchers to actively seek and address these tools' limitations to make sure artificial intelligence systems remain appropriate and useful for the archival field (Woodruff et al. 2024). These biases and limitations may shift depending on different temporal, spatial, demographical, and other focuses of a particular project. Being cognizant of how one's research focus can also create new biases within these tools is paramount to maintaining a balance between automated and human work.

Ultimately, archival work and work in any field is enjoyable, unique, and reliable because of the human input and experience with the material. My project highlights and reflects the growing relationship between technology and archival research processes; by keeping a clear-eyed focus on the methods available at hand, the joy and art in the field and processes can thus be maintained.

9. Future Inquiries

While this is the end of my project for now, in the future, I would like to address some shortcomings and other paths of inquiry.

Firstly, in my project specifically, security is not a concern due to the fact that all the data I am working with is publicly available and does not include patient health information. However, I would like to examine the OCR tools and other methods I use for security and privacy concerns.

Next, increasing the size of my dataset would be a great way to see these trends on a larger scale as well as applying it to other datasets. I would like to draw from more sources and include more documents in general, especially from more organizations and organizers during this era.

I also would like to continue to explore what other tools might be more accurate for my work, such as other keyword extraction tools or personalized topic lists/visualizations for topic modeling. Improved tools may not be so affected by the poor OCR quality and show less bias against handwritten documents as a result. Additionally, I was not granted Versa AI access (UCSF's "secure generative artificial intelligence (AI) platform") in time for this internship, but because it is a platform geared towards specific privacy concerns for UCSF, it would be a very convenient tool for UCSF ASC to use if it worked well. Versa AI then would be another tool I would like to test in the future.

Lastly, a diversity of languages would be an important addition to the dataset and evaluation. For archives and many research projects, there is a diversity of language and being able to have a

new section for accuracy of non-English OCR would be very useful and enlightening. There may be even more nuances to the biases non-English OCR could show.

10 Bibliography

"2023 Statement on Gaza – Librarians and Archivists with Palestine." n.d. Accessed August 15, 2024. https://librarianswithpalestine.org/2023-statement-on-gaza/.

Geoffrey Boushey. 2023. "UC Tech 2023 - Bias and Data Loss in Transcript Generation." Presented at the UC Tech 2023, Berkeley Marina, July 18. https://www.youtube.com/watch?v=sNNrx1i96wc.

Hamad, Karez Abdulwahhab, and Mehmet Kaya. 2016. "(PDF) A Detailed Analysis of Optical Character Recognition Technology." September 3, 2016. https://www.researchgate.net/publication/311851325_A_Detailed_Analysis_of_Optical_Character_Recognition_Technology.

Macquarie, Charlie. 2021. "No More Silence: Opening the Data of the HIV/AIDS Epidemic at the UCSF Library." July 6, 2021. https://ucsf.app.box.com/file/1586310333174.

Rogers, Corinne. 2024. "AI in the Archives: Wise Oracle or Master Manipulator?" Presented at the Archives Association of British Columbia - Association of Records Managers and Administrators International, Online, May 29. https://interparestrustai.org/assets/public/dissemination/Rogers-AIintheArchives-AABC-ARMA2024.pdf.

Woodruff, Allison, Renee Shelby, Patrick Gage Kelley, Steven Rousso-Schindler, Jamila Smith-Loud, and Lauren Wilcox. 2024. "How Knowledge Workers Think Generative AI Will (Not) Transform Their Industries." In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–26. CHI '24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3613904.3642700.