Digitizing Virginia Woolf's Diaries and Fragmentary Novels

An Interactive Qualifying Project (IQP) Report
Submitted to the Faculty of
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements
for the Degree of Bachelor of Science in

Computer Science
Mechanical Engineering

By:

Justin Luce,
Cooper Langner

Project Advisor:

Dr. Brigitte Servatius

Date: April 29, 2024

Abstract

This IQP explores the digitization of Virginia Woolf's diaries and fragmentary novels to enhance accessibility and enable more advanced textual analyses. Through collaboration with Dr. Joshua Phillips, an Oxford professor, this project enhances Phillips's existing digital platform, allowing researchers, educators, and the public to engage more effectively with Woolf's fragmented novels. Initially, we investigated OCR (Optical Character Recognition) technologies and their effectiveness in reading Woolf's handwritten diaries, Upon understanding more about Phillips's objectives, we shifted our focus towards converting transcripts of Woolf's work into digital formats, employing vector databases to enable semantic searches and connections between text fragments. This approach not only deepens the engagement with Woolf's literary techniques and themes but also provides innovative ways to interact with her works. By doing so, this IQP contributes to the preservation and exploration of Virginia Woolf's literary heritage, while also reflecting on the broader implications of digitization within the humanities.

Contents

List of Figures

1 Introduction

This IQP paper aims to examine the impact of digitization on the storage of written media through the lens of the digitization of Virginia Woolf's private diaries. For this IQP we worked with an Oxford professor who contacted us to create a road map for a second larger project to assist him with the creation of a digitized archive of Virginia Woolf's diaries, which would allow the user to perform complex textual analysis of the diaries through a variety of means. To accomplish this goal, this project examined the current and historic methods of digitization, as well as investigated a number of methods of cataloging and visualizing text files, with a particular emphasis on graphical displays. In order to understand the decision to focus on Virginia Woolf's writing for this IQP, it is important to understand Woolf's work as both an author and a publisher of literature, as the stylistic elements of digitizing and displaying her diaries present unique challenges because of her experience in the field of printing.

2 Background

2.1 The Digital Anon Project

In the final months of her life, Virginia Woolf started writing a history of English literature she never lived to finish. This project, extant only as a constellation of fragmentary drafts, has come to be known by the dual title of 'Anon' and 'The Reader'. This history is not patterned around the writings of singular-named authors like Chaucer and Shakespeare or Virginia Woolf. Rather, it stems from the 'nameless vitality' of Anon, the unnamed poet-singer whose voice precedes and makes possible literature in English. Dr. Phillips aims to create the first complete edition of this late archive, deploying genetic and digital editorial methodologies in order to best portray the richly generative textuality of this archive.

Source: https://github.com/JoshuaAPhillips/digital-anon/

2.2 Virginia Woolf as a Writer and Publisher

Born on the 25th of January, 1882 as Virginia Stephen, Woolf wrote a total of eight novels, as well as hundreds of published letters and essays(Reid P., 2024). Woolf's most famous works were her novels, particularly Mrs Dalloway and To the Lighthouse, which were published in 1925 and 1927 respectively at the Hogarth Press(Reid P., 2024). Woolf began her work with book publishing at the the age of nineteen, when she would bind her own books as a hobby, but it was not until 1917 when she and her husband Leonard Woolf purchased the Hogarth house and Press, that became fully acquainted with the process. Throughout the operation of the press, from 1917 to 1938, the Woolfs published over 450 titles, most of which were hand-bound and lettered by the Woolfs(Southworth H., 2010).

While the Hogarth Press shared some similarities with other contemporary presses, it was distinguished by its comparatively large output. While many other small English presses only produced an average of five titles a year the Hogarth Press' output significantly outpaced that number producing an average of almost twenty titles a year over its twenty four years of operation by Ms. Woolf and her husband(Southworth H., 2010). The Woolfs also chose to highlight their focus on the interior content of the books published at their press, contrasting this with their peers' emphasis on the binding and face of the books they published. Leonard Woolf stated about their printing that "[they] were interested primarily in the immaterial inside of a book", emphasizing that they were interested in publishing books which would not otherwise be printed by commercial publishers(Southworth, H., 2010).

Virginia Woolf began her work typesetting in 1917 after she and her husband purchased the Hogarth press. Initially, Virginia alone did the work of typesetting for the first works published by the press, as her husband Leonard had tremors in his hands which prevented him from successfully setting type. The first works of the Hogarth press were awash with errors, as the Woolf were both self taught printers, and almost complete beginners at the process. The first publication of the press, Two Stories, contained numerous of these typographic errors, as well as a variety of issues with irregular spacing and ink blotting. These issues did not remain for long in the press's output, as the quality of the Hogarth's published works rapidly improved as the Woolfs became enamored with the process. Virginia remarked that she found the typesetting process "exciting, soothing, ennobling and satisfying" and spent

much of her time experimenting with blocking and shaping the space of letters and words on the pages printed by the press(Jones, M. 2020). Writing about her work typesetting and how she conceptualized written language Woolf stated that "Books are made of tiny little words, which a writer shapes, often with great difficulty, into sentences of different lengths, placing one on top of another, never taking his eye off them, sometimes building them quite quickly, at other times knocking them down in despair, and beginning all over again"(Jones, M. 2020). Her vision of sentences as a series of building blocks in this passage is indicative of the mindset with which she undertook the process of typesetting, placing particular emphasis, not solely on the content of an essay or novel, but also the physical arrangement on the page and how that arrangement colors the reader's interpretation of the media.

2.3 Digitization in Document Preservation

Digitization is the process of conserving physical media by converting it into forms which are more resistant to degradation due to time and weathering. This effort typically consists of scanning a piece of media with a high resolution recording device and then saving those images in a collection somewhere that can be accessed for future reference(Han, Y. et al., 2018). The digitization of media in libraries in the United States began as a method of increasing access to old or damaged books and media, but since the early 2000s, has been endorsed by the Association of Research Libraries as a recommended method of preservation(Caplan, P., 2008). The decision to digitize or otherwise archive a piece of media is motivated by several factors. As previously stated, digitization when it first became popular in the 1990s was utilized mostly as a method of increasing access to media. Consequently, the media which was selected for digitization during this time period was often chosen because of its value for research or its importance to the public(Gertz J., 1999). The decision process for what media should be archived through digitization can also be understood by comparing it with the motivations for more traditional methods of preservation. Traditional preservation methods typically assess an object for preservation based on a variety of factors; the cost of preservation, the value of the object being considered, the viability of the method of preservation for the given object, and the physical integrity of the object are typically considered(Gertz J., 1999). Selection for preservation in the digital age. Library Resources & Technical Services . Comparatively, the criteria for digitization is similar with the added element of the legal implications of the storage and dissemination of digitized material. When considering what materials should be digitized organizations typically consider the value of the material, the viability of digitization as a method of preservation, the cost of the process, the added value of digitization, and the legality of creating, storing, and or disseminating the digitized material(Gertz J., 1999). It is apparent that organizations pursuing both traditional methods of preservation and digitization, consider the cost of the process, the nebulous value of the object being considered, and the effectiveness of the process at actually archiving the object. The most significant difference between these two decision processes is the elements of added value and legal risk associated with digitizing objects. Unlike microfilming, which was typically used to archive materials which were no longer protected by copyright law, digitization is more often used for newer media, both published and unpublished(Gertz J., 1999). This introduces an additional element of consideration for institutions digitizing their collections.

The most significant advantage digitization has over traditional preservation is the added value that digitization provides to a piece of media. Digitized media can be searched and queried electronically, shared online, and viewed by anyone with a computer and the necessary access permissions, all of which are difficult or impossible to perform with microfilming. These advantages specifically are what have motivated the transition to digitization for the purpose of preservation as well as distribution.

2.4 Historic Digitization Methods

Current digitization efforts are a product of earlier technological methods of media preservation, which date back to the beginning of the 20th century, when photography was first used for archival purposes(Zaagsma G., 2023). Microfilming and microfiching, both which were common in the United States in the mid to late 20th century, were methods of preservation which consist of storing images of books and newspapers on small photographs approximately four percent of the scale of the original media. These photographs would then be archived and could be viewed with a specially designed microscope for research or reviewing purposes. Microfilming was popular throughout the 20th century as a method of preserving heritage and government documents in post-war Europe and Africa(Zaagsma G., 2023). This method of conservation was effective at reproducing and storing images of books which were resistant to damage, but had the effect of limiting access to the media recorded in microfilms to people with access to the specialized equipment.

2.5 Modern Digitization for Preservation

The current standard for digitizing historical documents and other written media consists of scanning a piece of media with a high quality image recorder and then running that image through a commercially available optical character recognition (OCR) software(Balk H. et al, 2009). This process, when performed successfully, produced a digital document which is completely searchable by text analysis software. The issue with this approach is that the aforementioned OCR software is significantly less effective at reading handwritten or historical fonts, as well as a myriad of other complicating factors which can make optical character recognition an issue for historical document preservation. These issues include but are not limited to problems with the printed material, inking issue, illegible print types, variations in spacing and kerning, and handwritten additions to the text by readers. All of these elements, many of which are common in older texts like Woolf's diaries, cause problems when OCR is attempted on a scanned document, often causing the result to include numerous errors in spelling and erroneous additions of nonsense text(Balk H. et al, 2009).. More recently, OCR methods adapted to older typographic sets have been developed to mitigate these issues and when combined with post-OCR correction, often done manually, historical documents with archaic fonts can be successfully digitized(Balk H. et al, 2009)..

3 Week 1: Investigating OCR Solutions

3.1 Overall Summary

Unfortunately, Professor Phillips was out of office during the first week, and I was unable to receive an overview of his goals and the current state of the project. As such, I decided to get my hands dirty and investigate some potential digitization tools that could assist with extracting text from images of Virginia Woolf's handwritten diaries.

3.2 Methodology Summary

I decided to utilize some older OCR methods as well as some novel methods (such as those utilizing generative AI). Handwriting itself poses a challenge, only amplified by Virginia Woolf's unique writing style. I focused on market leaders in cloud computing, generative AI, as well as a leading open source solution.

It's unlikely that any OCR solution, even tuned for Virginia Woolf's writing style, will yield results that don't require human review. However, if OCR can get close and a human just has to proof-read, that would result in massive cost-savings.

3.3 Preliminary Findings

3.3.1 OCR Performances

- **Python Tesseract:** Python Tesseract is open source (a huge plus as it massively cuts costs), however, it isn't optimized for handwriting, which resulted in unusable performance.
- **GCP Handwriting OCR:** GCP document text extraction yielded the best OCR results. It wasn't perfect, however, with an additional layer of processing, it likely could accelerate human review of the text.
- **AWS OCR:** AWS Textract, despite being praised for its handwriting recognition, provided better results than Python Tesseract but still provided fairly useless results.
- **GPT-4, Claude-3, Gemini:** GPT-4, Claude-3, and Gemini, LLMs that all benchmark quite well, provided the most coherent results. That is, full sentences. However, they were closer to fiction/pure hallucinations than Virginia Woolf's diary.
- **Human Interpretation:** Effectively used as a benchmark. Seemingly quite accurate, but it took a long time.

3.4 Comparative Analysis of Text Results

Pagebreaks have been removed for conciseness.

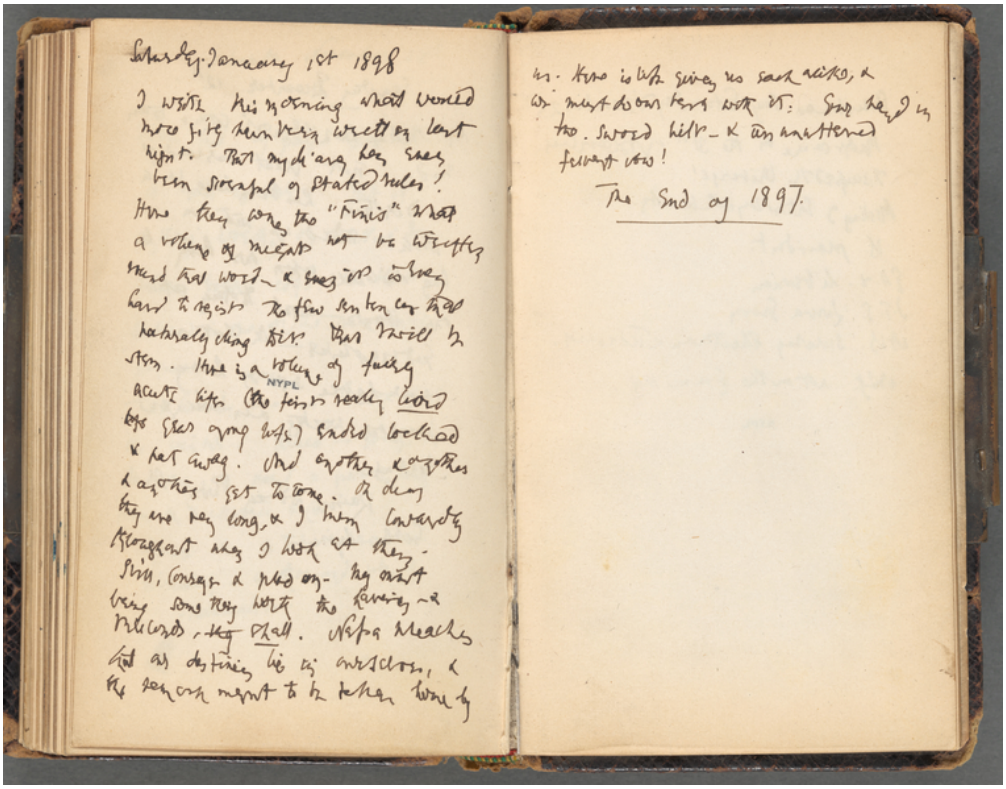

Figure 1: Scan of a Page From Virginia Woolf's Diary
Image source: https://digitalcollections.nypl.org/items/db085950-8ce7-013a-44cf-0242ac110003

Python Tesseract OCR Results

don bok wT Soy stored) ra ee yell
eae furs veo!
ee)
' i i : _ S

GCP Vision Document Text Extraction Results

Saterdag. January 1st 1898 weite his morning what world more sing hunter written last. hight. That my chang her sher bern diowahal' & stated rules! How they why the "Finis" What a roheme of might not be written mund that word, a suez is whey hard te rezists No für sentences mat haturally ding Dir that will h sts. Here is a blue of fairy Acute life the first really word bago [ras ng ts) Ended locked * her away. And another thes Lathes get to come. On day they are wey long, & I them wwardly Klonggart when I look at they Sims, Consege α pad on my most bring some my worth the Laverie, – Mulords, the shall. Nasa machen. Gut aus destinées lie in weschoss, & the demon ment to be then home by us. Here is life given us sach aciks, & we must do ousters with it. Your hand in the sword helt kan uttered ferment vow! The End of 1897.

AWS Textract Results

January 1st 1898 I WITH his 3 osning what world more fing been very writt on last hight. Than mychiang her Enes been sionapal 9 stated rules ! Him they why the "Finis" was a when I megal not be written must that word x shes in obsy have to regists to flw sen ten by mas naturally ding Air than will h stsm. Hue is a NYPL volume of furry acute life the first realy bord kg Eras going WAS) endrd locked K her way. and another h azthes set To Tone on day they me My long, x I men Cowardy Kiorgeant they I work EA they sin, Consys d read az- my not being some they work the having 2 mucoses, kg shall. Napa Meachs Got as destines The is cursclos, ¡ the ten an mgst to h then were by
in. kno is wh give, as must down term , to sword hilr- furnest vow! The End of

gpt-4-1106-vision-preview Results

Monday, January 1st 1894 I wish the morning would hurry up & come, waited at least two hours after retiring for sleep, but it came only in patches, even then irregular & unsatisfactory. I had taken the "Finis" to the tale of Moses and was waiting and waiting & waiting to see how he might end up, so then to have him actually dying just then raised his story in my interest, & being awake just as the people nearby began their early (morning) tasks, listened & kept awake until perhaps daybreak, if not going to sleep at all, except perhaps catnaps. SST to Tonge, a day bagman, very busy, J. Harrison, Sergeant was to meet me at [Tonge?], comes & goes on my shift. Finis, brought a handbag, & returned my plaid. Night mails and as definite is his instruction, I find upon returning to the House (Vicarage?) un- there is with saving, no sane reason to expect to from first word if. Some leg, by the. Shared wire - K on man[?]
The end of 1897.

Claude-3 Opus Results

January 1st, 1898. I wonder how many "good resolutions" were got down in black & white last night. But after all, they were so regularly written down, and come to nothing that one becomes skeptical about them. And as to "turning a new leaf or making a new start" with a New Year, what is there magical or mystical about the beginning of a new division of time called a year and divided into months, weeks, days, hours? Still, there is a feeling of something new about the breaking up of time into these periods, and a thought like this comes to mind: Where will this year bring me? What shall I think if I am permitted the pleasure to read this a twelvemonths hence? - For after all - a year is much and to know what the future contains would be to live it then. As a matter of fact the little world in which I live moves on by itself (very slowly and surely with the ebbing + flowing natural course of events)[.] A friends visit, an outing, a friend, nothing novel, just a series of accidents, so to speak - + yet nothing is accidental at all.

Gemini Results

Stardejanaary 1st 1898 We must door [tour] ters with it. Sour hand in the [?Sword hilt-kin] an attener
I with his morning what [dawned] noo fing sonvern van [foreign sovereign we] wretton lost been sorafel of [?stated rules] the key wng the [?Finis] Map a veheme of mecats not be written mend that wort & smez od sveg hand theists to for enten les has haterally cling this that wil [hereditary clinging to this that will] Asm. Ithe ga volume of fairy [?Assumed in the large volume of fairy tales] NYPL Acute life the teists sealy word [?Acute life the tastiest sealy word] boys Ghar gong life [?boys gone golfing life] ended locked
her awag [?her away]* Dad another cazothes [?satchels] hajones set to tome [?hay jones set to tome] In deny Lo they are rez long [?they are reason long] & I then warde Kloughout thes [?Clouded out these] Sim, Consgs [?Sims, Consgs] being sons tory worth the foreign Miss Phall. Napa Meaches [?Miss Fall. Napa Peaches] And as destinie lig on wrschon [?And as destiny lies on wershon] & helenor ment to be taken tone ty [?Eleanor meant to be taken tonight]

Human Interpretation

Saturday, January 1st 1898
I write this morning what would more "fitly" have been written last night. But my diary has "ever" been "scorned/scornful" of "skated/stated" rules! Here then comes the "Fines" what a voloume of might not be written around the world – & "" even hard to resist the few sentences that naturally ""cling/sing" ... "it" But "I will/I'll" be stern. Here is a voloume of "fair/fairly" acute life (the first really lived "Year?/time" of my life) ended locked & put away! And another & another & another yet to come. Oh "diary/dear" they/there are very long, I seem cowardly throughout when I look at them – still. Courage & plot on. They must...

Here is life given us each alike, & we must do our best with it: Your hand in the sword hilt - & unuttered fervent vow.
The end of 1897.

3.5 Next Steps

I planned to meet with Professor Phillips the following week to get a better understanding of the project, its current standing, and what we can do to help.

4 Weeks 2-3: Focusing on the Fragmentary Novel Archive

4.1 Feedback from Professor Phillips

One of the goals of this IQP was to assist Dr. Joshua Phillips, a researcher and writer with interests in modernist literature, Virginia Woolf, and textual editing, on his Leverhulme Trust-funded project titled "The Digital 'Anon': A Digital Genetic Edition of Virginia Woolf's Final Essays." Dr. Phillips is a Leverhulme Early Career Fellow at the University of Oxford's Faculty of English and a Junior Research Fellow at Jesus College, Oxford.

After completing the Week 1 update, I had the opportunity to speak with Professor Phillips about his project in more detail. His feedback indicated that the primary goal was not to actually transcribe Virginia Woolf's handwriting but to present her fragmentary novels to users in a meaningful way.

The challenge with these novel fragments is that they don't have a clear narrative connection, making it difficult for students and other interested individuals to make sense of them when reading through the archive. Professor Phillips emphasized the need for a digital platform that could help users navigate and engage with these disjointed texts more effectively.

4.2 Methodology

Based on Professor Phillips' guidance, I shifted my focus in Weeks 2 and 3 to developing tools and digital infrastructure for presenting her fragmentary novels. Here are the key steps I took:

- Created a template and script to convert the existing XML transcripts into a more user-friendly format for web display.
- Developed a Python script to convert the original .TIF image files into .DZI (Deep Zoom Images) format, which is compatible with the OpenSeaDragon viewer. This allows users to explore high-resolution scans of the original manuscript pages.
- Generated .jpg thumbnail images for each manuscript page to be used on the index page of the digital archive.

4.2.1 Index Page Screenshot

Figure 2: Screenshot of the Proposed Index Page of the Digital Archive

4.2.2 Example Manuscript Page Screenshot

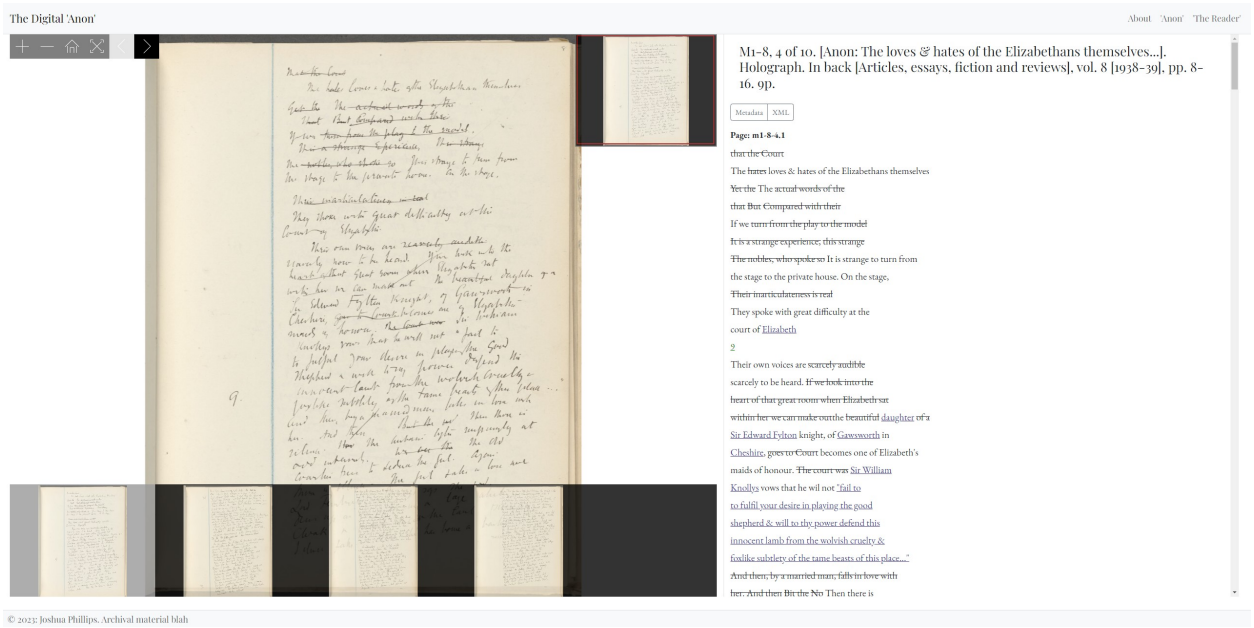

Figure 3: Screenshot of an Example Manuscript Page

4.3 Exploring Vector Databases for Improved Search and Exploration

To further enhance the user experience and help students and researchers navigate the fragmentary novels more effectively, I propose the use of vector databases. Vector databases are designed to store and query embeddings, which are high-density vector representations of data such as text or images. These embeddings capture semantic meaning and enable powerful search and recommendation capabilities.

For text data, like Virginia Woolf's fragmentary novels, vector databases leverage word embeddings or document embeddings. These embeddings are large language models that have been trained on large quantities of text data. By converting the text fragments into high-dimensional vectors, we can capture the semantic relationships between words and sentences, allowing for more relevant search results.

One key advantage of vector databases is their ability to perform similarity search. Instead of relying on exact keyword matching, vector databases can identify text fragments that are semantically similar to a given query. This is particularly useful for exploring Virginia Woolf's novels, where different passages may discuss related themes or ideas without using the exact same words. Similarly, interested individuals may want to search for concepts and not exact keywords.

To implement a vector database for this project, I experimented with using OpenAI's text-embedding-ada-002 model in combination with Pinecone, a hosted vector database platform. The text-embedding-ada-002 model provides 1,536-dimensional embeddings for text.

Figure 4 illustrates a simplified 2D representation of how a vector database can cluster semantically similar text fragments together. The text-embedding-ada-002 model operates in a much higher-dimensional space, enabling more granular and accurate similarity search.

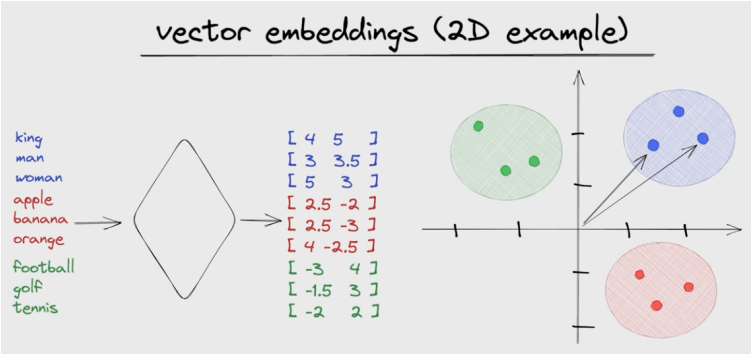

Figure 4: Simplified 2D Representation of a Vector Database
Image source: https://jkfran.com/introduction-vector-embedding-databases.md/

By incorporating a vector database into the digital archive, we can provide users with powerful tools for exploring and discovering connections between the fragmentary novels. This can help students and researchers gain new insights into Virginia Woolf's creative process and the themes that emerge across her unfinished works.

5 The Future of Printed Press in the Digital Age

As part of this IQP, my advisor also asked me to reflect on the future of printed press in the digital age. It's clear that some aspects of physical books and manuscripts can't be fully preserved or replicated in digital formats. For example, one author had her page numbers counting down rather than up, a unique design choice that might require special implementation for every single digital platform the book is offered on, which may not be viable.

However, the digital age also presents new opportunities for preserving and engaging with literature. Projects like Dr. Phillips' "The Digital 'Anon'" demonstrate how digital tools and platforms can make fragmented and challenging texts more accessible to a wider audience. By leveraging technologies like high-resolution imaging, XML transcription, and vector databases, we can create digital archives that allow users to explore and make connections between disparate textual fragments in ways that would be difficult or impossible with physical manuscripts alone.

Moreover, digital platforms can enhance the reading experience by providing interactive features, such as annotations, cross-references, and multimedia content. These features can enrich the reader's understanding of the text and provide new avenues for scholarly analysis and interpretation.

It is also important to acknowledge the value of printed books and physical archives. Digital formats cannot fully replicate the experience of handling a physical book, the smell of the pages, and the sense of history embodied in original manuscripts. Additionally, physical archives serve as important backups and long-term preservation solutions, ensuring that literary works can endure even as digital technologies evolve and become obsolete.

The future of printed press in the digital age is to be one of coexistence. While physical books and manuscripts will continue to hold value for their material properties and historical significance, digital platforms will increasingly play a role in preserving and, in many ways, improving our engagement with literature. Both physical and digital media have a place in our future.

6 Conclusion

Through this IQP, we had the opportunity to contribute to Dr. Joshua Phillips' project, "The Digital 'Anon': A Digital Genetic Edition of Virginia Woolf's Final Essays." By investigating OCR solutions, developing tools for converting and presenting the fragmentary novel archive, and exploring the potential of vector databases, we aimed to support Dr. Phillips' goal of creating a digital platform that allows users to meaningfully engage with Virginia Woolf's disjointed final texts.

This project also provided an opportunity to reflect on the broader implications of digitization for the future of printed press. As we continue to develop new technologies and platforms for preserving and engaging with literature, it's important to consider both the challenges and opportunities that arise in the process of translating physical manuscripts into digital formats.

Overall, this IQP has highlighted the potential for collaboration between literary scholars, computer scientists, and other experts in creating digital archives that make complex and fragmented texts more accessible to a wider audience. It's important to continue valuing the unique properties of physical books and manuscripts. However, it's also important to develop digital tools that can bring access to more people and enhance their experience and understanding of the text.

Moving forward, there are several areas where further research and development could enhance the digital archive of Virginia Woolf's fragmentary novels:

1. Refining the vector database implementation to optimize search performance and relevance. 2. Exploring additional natural language processing techniques, such as named entity recognition and sentiment analysis, to provide deeper insights into the content of the fragments. 3. Developing intuitive user interfaces and visualization tools to help users navigate the complex network of relationships between the fragmentary texts. 4. Conducting user studies to evaluate the effectiveness of the digital archive in supporting scholarly research and literary exploration.

A Code Snippets

A.1 Python Script for Converting .TIF to .DZI, Generating Thumbnails, and Converting XML to HTML

```
import os
import deepzoom
from PIL import Image
from lxml import etree

# Set the path to your XML files directory
xml_dir = '../../transcriptions'

# Set the path to your XSLT stylesheet
xslt_path = 'stylesheet.xsl'

# xml to html output directory
page_directory = '../pages/'

# Set the paths for image converstion
input_folder = '../../formatted_imgs'
output_folder = '../images'
thumbnail_folder = '../thumbnails'
thumbnail_size = (256, 256)

# index html file config
template_path = 'template.html'
index_output_path = '../index.html'

# Load the XSLT stylesheet
xslt_doc = etree.parse(xslt_path)
xslt_transformer = etree.XSLT(xslt_doc)

# Array to store the page information
pages = []

# Create the thumbnail folder if it doesn't exist
if not os.path.exists(thumbnail_folder):
    os.makedirs(thumbnail_folder)

# Create the pages folder if it doesn't exist
if not os.path.exists(page_directory):
    os.makedirs(page_directory)

def generate_index_html(pages):
    # Read the template.html file
    with open(template_path, 'r') as file:
        template = file.read()

    # Generate the page cards HTML
    page_cards_html = ''
    for page in pages:
        card = f'''
            <div class="col-lg-3 mb-3">
                <div class="card h-100">
                    <a href="{page['html_path']}" target="_blank">
                        <img src="{page['thumbnail_path']}" class="card-img-top" alt="Thumbnail">
                    </a>
                    <div class="card-body">
                        <h5 class="card-title">{page['name']}</h5>
                    </div>
                </div>
            </div>
        '''
        page_cards_html += card
```

```
    # Replace the placeholder in the template with the generated page cards HTML
    index_html = template.replace('{page_cards}', page_cards_html)

    # Save the generated index.html file
    with open(index_output_path, 'w') as file:
        file.write(index_html)

print('Index page generated successfully.')


#============================================================================
# XML to HTML transformation
#============================================================================

# Iterate over the XML files in the directory
for filename in os.listdir(xml_dir):
    if filename.endswith('.xml'):
        try:
            # Construct the full path to the XML file
            xml_path = os.path.join(xml_dir, filename)

            # Load the XML file
            xml_doc = etree.parse(xml_path)

            # Apply the XSLT transformation
            output_doc = xslt_transformer(xml_doc)

            # Generate the output HTML file path
            output_path = page_directory + filename.split('.xml')[0] + '.html'

            # Save the transformed HTML file
            with open(output_path, 'wb') as f:
                f.write(etree.tostring(output_doc, pretty_print=True, method='html'))

            print(f'Transformed {filename} to {output_path}')

            pages.append({
                'name': filename.split('.xml')[0],
                'html_path': os.path.join('pages', filename.split('.xml')[0] + '.html'),  # Correct path to
    ↪ the HTML page
                'thumbnail_path': os.path.join('thumbnails', filename.split('.xml')[0] + '.jpg')  # Relative
    ↪ path
            })

        except Exception as e:
            print(f'Error transforming {filename}: {e}')


#============================================================================
# IMAGE CONVERSION
#============================================================================

print("Converting images...")
for subdir in os.listdir(input_folder):
    subdir_path = os.path.join(input_folder, subdir)
    if os.path.isdir(subdir_path):
        output_subdir = os.path.join(output_folder, subdir)
        if not os.path.exists(output_subdir):
            os.makedirs(output_subdir)

        first_image = True  # Indicator for the first image in each folder

        for filename in os.listdir(subdir_path):
            if filename.lower().endswith('.tif'):
                tiff_path = os.path.join(subdir_path, filename)
                dzi_filename = os.path.splitext(filename)[0] + '.dzi'
                dzi_path = os.path.join(output_subdir, dzi_filename)
```

```
                try:
                    image = Image.open(tiff_path)
                    creator = deepzoom.ImageCreator(
                        tile_size=256,
                        tile_overlap=2,
                        tile_format="jpg",
                        resize_filter=Image.Resampling.BICUBIC
                    )
                    creator.create(image, dzi_path)

                    # Generate thumbnails for the first image
                    if first_image:
                        thumbnail_path = os.path.join(thumbnail_folder, subdir + '.jpg')
                        thumbnail = image.copy()

                        # Resize while maintaining aspect ratio
                        aspect_ratio = image.width / image.height
                        if aspect_ratio > 1:  # Wide image
                            base_width = thumbnail_size[0]
                            new_height = int(base_width / aspect_ratio)
                            size = (base_width, new_height)
                        else:  # Tall image or square
                            base_height = thumbnail_size[1]
                            new_width = int(base_height * aspect_ratio)
                            size = (new_width, base_height)

                        thumbnail.thumbnail(size)
                        thumbnail.save(thumbnail_path, "JPEG")
                        first_image = False

                    print(f"Generated DZI for {filename}")
                except Exception as e:
                    print(f"Error generating DZI for {filename}: {str(e)}")

print("DZI generation complete.")
print('Thumbnail generation complete.')


# Generate the index.html file
generate_index_html(pages)
print('Index page generated successfully.')
```

A.2 Template for Converting XML to HTML

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
    ↪ xmlns:tei="http://www.tei-c.org/ns/1.0">
  <xsl:output method="html" indent="yes" encoding="UTF-8"/>

  <xsl:template match="/">
    <html>
      <head>
        <title><xsl:value-of select="//tei:title"/></title>
        <meta charset="utf-8"/>
        <meta name="viewport" content="width=device-width, initial-scale=1"/>
        <link href="https://cdn.jsdelivr.net/npm/bootstrap@5.3.0-alpha1/dist/css/bootstrap.min.css"
    ↪ rel="stylesheet" integrity="sha384-GLhlTQ8iRABdZLl6O3oVMWSktQOp6b7In1Zl3/Jr59b6EGGoI1aFkw7cmDA6j6gD"
    ↪ crossorigin="anonymous"/>
        <link rel="stylesheet" type="text/css" href="../resources/style.css"/>
        <link rel="preconnect" href="https://fonts.googleapis.com"/>
        <link rel="preconnect" href="https://fonts.gstatic.com" crossorigin="anonymous"/>
        <link
    ↪ href="https://fonts.googleapis.com/css2?family=EB+Garamond:ital,wght@0,400;0,500;0,600;0,700;0,800;1,400;1,500;1,600;1,7
    ↪ rel="stylesheet"/>
```

```
    <link rel="preconnect" href="https://fonts.googleapis.com"/>
    <link rel="preconnect" href="https://fonts.gstatic.com" crossorigin="anonymous"/>
    <link
↪ href="https://fonts.googleapis.com/css2?family=Playfair+Display:ital,wght@0,400;0,500;0,600;0,700;0,800;0,900;1,400;1,50
↪ rel="stylesheet"/>
  </head>
  <body>

    <!-- navbar/header -->
    <nav class="navbar navbar-nav bg-body-tertiary navbar-expand-sm border-bottom sticky-top">
      <div class="container-fluid">
        <!-- links -->
        <a class="navbar-brand" href="#">The Digital 'Anon'</a>
        <ul class="navbar-nav">
          <li class="nav-item">
            <a class="nav-link" href="#">About</a>
          </li>
          <li class="nav-item">
            <a class="nav-link" href="#">'Anon'</a>
          </li>
          <li class="nav-item">
            <a class="nav-link" href="#">'The Reader'</a>
          </li>
        </ul>
      </div>
    </nav>

    <div class="container-fluid">
      <div class="row" style="height: 90vh">
        <div class="col-lg-7 border-end">
          <div id="openseadragon1" style="height: 80vh" class="h-100"></div>
        </div>
        <div class="col-lg-5 overflow-auto h-100">
          <h1><xsl:value-of select="//tei:title"/></h1>

          <div class="btn-group btn-group-sm flex" role="group">
            <button class="btn btn-outline-secondary">Metadata</button>
            <a class="btn btn-outline-secondary" href="../../transcriptions/{//tei:idno}.xml"
↪ target="_blank">XML</a>
          </div>

          <xsl:apply-templates select="//tei:body/tei:div"/>
        </div>
      </div>
    </div>

    <!-- modal content -->
    <div class="modal fade" id="annoModal">
      <div class="modal-dialog">
        <div class="modal-content">
          <div class="modal-header">
            <h1 class="modal-title" id="exampleModalLabel">Annotation modal test</h1>
            <button type="button" class="btn-close" data-bs-dismiss="modal" aria-label="Close"></button>
          </div>
          <div class="modal-body">
            ...
          </div>
        </div>
      </div>
    </div>

    <!-- footer -->
    <footer class="footer border-top fixed-bottom">
      <p class="bg-light text-muted p-2"> 2023: Joshua Phillips. Archival material blah</p>
    </footer>

    <!-- js -->
    <script src="https://cdn.jsdelivr.net/npm/bootstrap@5.3.0-alpha1/dist/js/bootstrap.bundle.min.js"
```

```
    ↪ integrity="sha384-w76AqPfDkMBDXo30jS1Sgez6pr3x5MlQ1ZAGC+nuZB+EYdgRZgiwxhTBTkF7CXvN"
    ↪ crossorigin="anonymous"></script>
       <script src="https://cdn.jsdelivr.net/npm/@popperjs/core@2.11.6/dist/umd/popper.min.js"
    ↪ integrity="sha384-oBqDVmMz9ATKxIep9tiCxS/Z9fNfEXiDAYTujMAeBAsjFuCZSmKbSSUnQlmh/jp3"
    ↪ crossorigin="anonymous"></script>
       <script src="https://code.jquery.com/jquery-3.6.3.js"
    ↪ integrity="sha256-nQLuAZGRRcILA+6dMBOvcRh5Pe310sBpanc6+QBmyVM=" crossorigin="anonymous"></script>
       <script src="../resources/modal.js"></script>
       <script src="../resources/openseadragon/openseadragon.min.js"></script>

<!-- OpenSeadragon -->
<script type="text/javascript">
  var xmlFileName = '<xsl:value-of select="//tei:idno"/>';
  var tileSources = [];

  // Function to retrieve the list of JPG images from the respective subfolder
  function getdziImages(folderPath, callback) {
    var xhr = new XMLHttpRequest();
    xhr.open('GET', folderPath, true);
    xhr.responseType = 'text';

    xhr.onload = function() {
      if (xhr.status === 200) {
        var parser = new DOMParser();
        var htmlDoc = parser.parseFromString(xhr.responseText, 'text/html');
        var links = htmlDoc.getElementsByTagName('a');

        var dziImages = [];
        for (var i = 0; i &lt; links.length; i++) {
          var href = links[i].getAttribute('href');
          if (href.toLowerCase().endsWith('.dzi')) {
            dziImages.push(href);
          }
        }

        callback(dziImages);
      }
    };

    xhr.send();
  }

  // Get the list of JPG images and initialize OpenSeadragon
  getdziImages('../images/' + xmlFileName + '/', function(dziImages) {
    tileSources = dziImages;

    var viewer = OpenSeadragon({
      id: "openseadragon1",
      prefixUrl: "../resources/openseadragon/images/",
      tileSources: tileSources,
      sequenceMode: true,
      showReferenceStrip: true,
      showNavigator: true,
      visibilityRatio: 1
    });
    viewer();
  });
</script>
      </body>
    </html>
  </xsl:template>

  <xsl:template match="tei:div">
    <h3>Page: <xsl:value-of select="@xml:id"/></h3>
    <xsl:apply-templates/>
  </xsl:template>

  <xsl:template match="tei:p">
```

```
    <p>
      <xsl:apply-templates/>
    </p>
  </xsl:template>

  <xsl:template match="tei:l">
    <p>
      <xsl:apply-templates/>
    </p>
  </xsl:template>

  <xsl:template match="tei:del">
    <span class="del"><xsl:apply-templates/></span>
  </xsl:template>

  <xsl:template match="tei:add">
    <span class="add-{@place}"><xsl:apply-templates/></span>
  </xsl:template>

  <xsl:template match="tei:rs">
    <a href="{@ref}" data-bs-toggle="modal" class="{@rend}"><xsl:apply-templates/></a>
  </xsl:template>

  <xsl:template match="tei:quote">
    <xsl:choose>
      <xsl:when test="@rend='block'">
        <a href="{@source}" data-bs-toggle="modal" class="block"><xsl:apply-templates/></a>
      </xsl:when>
      <xsl:otherwise>
        <a href="{@source}" data-bs-target="#annoModal" data-bs-toggle="modal"
    ↪ class="inline"><xsl:apply-templates/></a>
      </xsl:otherwise>
    </xsl:choose>
  </xsl:template>

  <xsl:template match="tei:choice/tei:sic">
    <span data-title="{.}"><xsl:apply-templates select="parent::tei:choice/tei:corr"/></span>
  </xsl:template>

  <xsl:template match="tei:choice/tei:corr">
    <i class="corr"><xsl:apply-templates/></i>
  </xsl:template>

</xsl:stylesheet>
```

A.3 Python Script for Vector Search

```
import os
from lxml import etree as ET
from pinecone import Pinecone, ServerlessSpec
from openai import OpenAI
import requests

openai_api_key = '************************************'
folder_path = 'digital-anon/transcriptions'

# Initialize OpenAI
client = OpenAI(api_key=openai_api_key)

# Initialize Pinecone
pc = Pinecone(api_key="****************************************")
index_name = "novel-fragments"

# Create or connect to Pinecone index
if index_name not in str(pc.list_indexes()):
```

```
    pc.create_index(index_name, dimension=1536, metric="cosine", spec=ServerlessSpec(
        cloud="aws",
        region="us-east-1"
    ))

index = pc.Index(index_name)

def parse_xml(file_path):
    """ Parses an XML file using lxml and extracts concatenated text from <l> tags inside <p> tags, handling
    ↪ namespaces. """
    namespace = {'tei': 'http://www.tei-c.org/ns/1.0'}
    parser = ET.XMLParser(recover=True)
    texts = []
    try:
        tree = ET.parse(file_path, parser)
        root = tree.getroot()
        for paragraph in root.xpath('.//tei:div/tei:p', namespaces=namespace):
            raw_xml = ET.tostring(paragraph, encoding='unicode', method='xml')
            text = "".join([elem.text for elem in paragraph.xpath('.//tei:l', namespaces=namespace) if
    ↪ elem.text])
            if text:  # Ensure text is not empty
                texts.append((text.strip(), raw_xml.strip()))
        return texts
    except ET.XMLSyntaxError as e:
        print(f"Error parsing {file_path}: {e}")
        return []

def text_to_embedding(text):
    """Converts text to a vector using OpenAI embeddings."""
    return client.embeddings.create(input = [text], model='text-embedding-ada-002').data[0].embedding

def index_data(texts, file_id):
    """Indexes the given list of texts with their embeddings and corresponding XML."""
    items_to_index = []

    for i, (text, xml) in enumerate(texts):
        try:
            embedding = text_to_embedding(text)
            metadata = {'text': text, 'xml': xml}
            vector_item = {
                "id": f"{file_id}-{i}",
                "values": embedding,
                "metadata": metadata
            }
            items_to_index.append(vector_item)
        except Exception as e:
            print(f"Error embedding text for {file_id}-{i}: {e}")

    # Perform the upsert operation
    if items_to_index:
        index.upsert(vectors=items_to_index)
    else:
        print(f"No items to index for {file_id}")

def search_fragments(query, top_k=5):
    """ Searches the Pinecone index for the query and returns the top_k results with metadata. """
    query_vector = text_to_embedding(query)
    results = index.query(vector=query_vector, top_k=top_k, include_metadata=True)
    return results

def main():
    # Index all XMLs in the folder
    # Only has to be called once
    for filename in os.listdir(folder_path):
        if filename.endswith(".xml"):
            file_path = os.path.join(folder_path, filename)
            texts = parse_xml(file_path)
            # print(texts)
```

```
            index_data(texts, filename)
            print(f"Indexed {filename}")

    # Example search
    query = "Playhouse"
    search_results = search_fragments(query, top_k=10)
    print(search_results)

if __name__ == "__main__":
    main()
```

A.4 Index HTML Template

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Index Page</title>
    <!-- Bootstrap CSS -->
    <link href="https://cdn.jsdelivr.net/npm/bootstrap@5.3.0-alpha1/dist/css/bootstrap.min.css"
    ↪ rel="stylesheet" integrity="sha384-GLhlTQ8iRABdZLl6O3oVMWSktQOp6b7In1Zl3/Jr59b6EGGoI1aFkw7cmDA6j6gD"
    ↪ crossorigin="anonymous">
    <link rel="stylesheet" type="text/css" href="../resources/style.css">
    <link rel="preconnect" href="https://fonts.googleapis.com">
    <link rel="preconnect" href="https://fonts.gstatic.com" crossorigin="anonymous">
    <link
    ↪ href="https://fonts.googleapis.com/css2?family=EB+Garamond:ital,wght@0,400;0,500;0,600;0,700;0,800;1,400;1,500;1,600;1,7
    ↪ rel="stylesheet">
    <link rel="preconnect" href="https://fonts.googleapis.com">
    <link rel="preconnect" href="https://fonts.gstatic.com" crossorigin="anonymous">
    <link
    ↪ href="https://fonts.googleapis.com/css2?family=Playfair+Display:ital,wght@0,400;0,500;0,600;0,700;0,800;0,900;1,400;1,50
    ↪ rel="stylesheet">
    <!-- Custom CSS -->
    <style>
        body {
            background-color: #f8f9fa;
            padding-bottom: 100px;
        }
        .footer {
            background-color: #f8f9fa;
        }
        .card {
            transition: transform 0.3s;
        }
        .card:hover {
            transform: translateY(-5px);
        }
    </style>
</head>
<body>
    <!-- navbar/header -->
    <nav class="navbar navbar-nav bg-body-tertiary navbar-expand-sm border-bottom sticky-top">
        <div class="container-fluid">
            <!-- links -->
            <a class="navbar-brand" href="#">The Digital 'Anon'</a>
            <ul class="navbar-nav">
                <li class="nav-item">
                    <a class="nav-link" href="#">About</a>
                </li>
                <li class="nav-item">
                    <a class="nav-link" href="#">'Anon'</a>
                </li>
```

```
                <li class="nav-item">
                    <a class="nav-link" href="#">'The Reader'</a>
                </li>
            </ul>
        </div>
    </nav>

    <!-- HTML Page Cards -->
    <div class="container mt-4">
        <div class="row">
            {page_cards}
        </div>
    </div>

    <!-- footer -->
    <footer class="footer border-top fixed-bottom">
        <p class="bg-light text-muted p-2">(c) 2023: Joshua Phillips. Archival material blah</p>
    </footer>

    <!-- Bootstrap JS -->
    <script src="https://cdn.jsdelivr.net/npm/bootstrap@5.3.0-alpha1/dist/js/bootstrap.bundle.min.js"
    ↪ integrity="sha384-w76AqPfDkMBDXo30jS1Sgez6pr3x5MlQ1ZAGC+nuZB+EYdgRZgiwxhTBTkF7CXvN"
    ↪ crossorigin="anonymous"></script>
    <script src="https://cdn.jsdelivr.net/npm/@popperjs/core@2.11.6/dist/umd/popper.min.js"
    ↪ integrity="sha384-oBqDVmMz9ATKxIep9tiCxS/Z9fNfEXiDAYTujMAeBAsjFuCZSmKbSSUnQlmh/jp3"
    ↪ crossorigin="anonymous"></script>
    <script src="https://code.jquery.com/jquery-3.6.3.js"
    ↪ integrity="sha256-nQLuAZGRRcILA+6dMBOvcRh5Pe310sBpanc6+QBmyVM=" crossorigin="anonymous"></script>
</body>
</html>
```

Vector Search for "Interest in the theater" Console Output

```
{'matches': [{'id': 'rn2.xml-37',
              'metadata': {'text': 'When The theatre for those who cant '
                                   'read.brought in [pit & gallery?]. But did '
                                   "thenobles 'go to the play'?",
                           'xml': '<p xmlns="http://www.tei-c.org/ns/1.0">\n'
                                  '          <l>When <del>did</del> was the '
                                  'first theatre built?</l>\n'
                                  '          <l>The theatre for those who cant '
                                  'read.</l>\n'
                                  '          <l>brought in [pit &amp;amp; '
                                  'gallery?]. But did the</l>\n'
                                  "          <l>nobles 'go to the play'?</l>\n"
                                  '        </p>'},
              'score': 0.840461731,
              'values': []},
             {'id': 'm113.xml-5',
              'metadata': {'text': 'We can suppose that the reading public '
                                   'came into existencesome four hundred years '
                                   'ago when the playhouse was shut.But the '
                                   'reading public must have been a small one, '
                                   'composedof scholars and aristocrats. The '
                                   'lack of \n'
                                   '            accounts for the long pause '
                                   'between in \n'
                                   '            \n'
                                   '            were open on \n'
                                   '            the detached, the \n'
                                   '            public and its specialised '
                                   'nature accounts for the scarcityof \n'
                                   '            criticism there was. Both with '
                                   'the \n'
                                   '            ...'
                           'xml': '...'
                          }}, ...]}
```

The results for the query "Interest in the theater" showcase the power of semantic/vector search. The results include passages mentioning playhouses even though the exact query phrase wasn't used. This is because the word/vector embeddings capture the semantic meaning and context of words, allowing the search to understand that playhouses are closely related to "theater" in this context. This is extraordinarily useful for fragmented text, where searching for exact queries is likely to return few results.

B Bibliography

1. Reid, P. (2024, April 16). Virginia Woolf. Encyclopedia Britannica. https://www.britannica.com/biography/Virginia-Woolf

2. Southworth, H. (Ed.). (2010). *Leonard and Virginia Woolf, The Hogarth Press and the Networks of Modernism*. Edinburgh University Press. http://www.jstor.org/stable/10.3366/j.ctt1r23tc

3. Jones, M. (2020). *Hogarth House, Richmond*. HOGARTH PRESS – virginiawoolf.ca. https://virginiawoolf.ca/horgarth-press/

4. Han, Y., & Wan, X. (2018). Digitization of Text Documents Using PDF/A. *Information Technology and Libraries*, *37*(1), 52–64. https://doi.org/10.6017/ital.v37i1.9878

5. Caplan, Priscilla. "Chapter 1: What Is Digital Preservation?" *Library Technology Reports* 44, no. 2 (February/March 2008): via https://libguides.ala.org/libpreservation/digitization

6. Gertz J. (1999). Selection for preservation in the digital age. Library Resources & Technical Services

7. Gerben Zaagsma, Digital History and the Politics of Digitization, *Digital Scholarship in the Humanities*, Volume 38, Issue 2, June 2023, Pages 830–851

8. Balk, H., & Ploeger, L. (2009). IMPACT: working together to address the challenges involving mass digitization of historical printed text. *OCLC Systems & Services: International digital library perspectives*, *25*(4), 233-248.

9. Boroş, E., Hamdi, A., Pontes, E. L., Cabrera-Diego, L. A., Moreno, J. G., Sidere, N., & Doucet, A. (2020, November). Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th conference on computational natural language learning* (pp. 431-441).

10. *Collections care*. Preservation Guidelines for Digitizing Library Materials - Collections Care (Preservation, Library of Congress). (n.d.). https://www.loc.gov/preservation/care/scan.html

11. Svendsen, J. (2012). *Hogarth Press*. Modernism Lab. https://campuspress.yale.edu/modernismlab/hogarth-press/

12. Henry W. and Albert A. Berg Collection of English and American Literature, The New York Public Library. (1897 - 1898). [Diary] Holograph notebook. Retrieved from https://digitalcollections.nypl.org/items/db085950-8ce7-013a-44cf-0242ac110003

C Disclaimer

During the development of the digital archive for Virginia Woolf's fragmentary novels, ChatGPT, an AI language model developed by OpenAI, was utilized to assist in various tasks, including the ideation, creation, and debugging of the python, HTML, and XSL snippets.