University
of Colorado
Boulder

University of Colorado Boulder

Latent Semantic Indexing

Student :
Divya Nallawar
Ainsley Braunscheidel
Reza Naimen
Luke Petet

Teacher :
Prof. James H. Curry

December 15, 2024

Contents

1 Introduction to LSI

1.1 Background

- **Overview of text analysis and LSI**: Text analysis converts unstructured text into structured data to extract insights, including tasks like sentiment analysis and document classification. Latent Semantic Indexing (LSI) is a technique used in text analysis to uncover hidden relationships between words and documents, while trying to address possible synonymy and polysemy. LSI uses Singular Value Decomposition (SVD) to reduce a term-document matrix into a smaller, more meaningful representation, capturing latent semantic dimensions. LSI is applied in information retrieval, document clustering, recommendation systems, and text summarization, improving the relevance and accuracy of text-based analysis.

- **Importance of analyzing semantic relationships in text**: Analyzing semantic relationships in text is crucial for understanding meaning and context beyond exact word matches. It improves information retrieval, handles synonymy and polysemy, and enables tasks like sentiment analysis, topic modeling, and summarization. By capturing relationships between words and their meanings, it enhances search engines, AI applications, and decision-making systems, making them more precise and context-aware. This approach is foundational for advanced tasks like machine translation, recommendation systems, and personalized content delivery, ensuring that systems interpret and respond to text in a human-like, meaningful way.

1.2 Objective

The purpose of this report is to identify the latent semantic structures in a corpus of ten distinct novels covering a range of genres and topics by using Latent Semantic Indexing (LSI). This includes exploring the ways in which various terms contribute to broad subjects and the relationships between books in the reduced semantic space.

Through this analysis, the report aims to achieve the following:

- Identify the primary latent topics or themes in the books.
- Highlight relationships between books based on semantic similarities.
- Demonstrate the utility of LSI in reducing dimensionality while preserving the essence of textual content.

1.3 Scope

For the books we have chosen, we have selected various genres in order to have a wide variety of words and topics in our LSI analysis. The breakdown is as follows: we have selected 2 comedy books, 2 historical books, 2 science fiction books, 2 adventure novels, and 2 fantasy novels. We have tried to select authors from around the world, as well as from a variety of backgrounds and time periods, to increase the diversity for comparisons within our dataset. The themes of these texts range from books that are appreciative and considerate of the surrounding society (Life on the Mississippi, Alice in Wonderland) to books that are antagonistic (A Pickle for the Knowing Ones, Heart of Darkness). The goal is then to compare how themes, especially positive and negative themes, are shared across the novels, and to see how both similar themes and similar genres compare, and which of the two have the most similarities in vocabulary.

1.4 Structure of the Report

This report is organized into several sections to provide a comprehensive overview of the study:

1. Introduction: Offers a background on LSI, the objectives of the analysis, and the scope of the study.
2. Dataset: Describes the dataset, source of the dataset and little bit about each book.
3. Methodology: Describes the preprocessing steps and the technical details of implementing LSI.
4. Analysis: Explores the results of applying LSI, including visualizations of latent topics and semantic similarities.
5. Results: Summarizes quantitative and qualitative findings from the analysis.
6. Discussion: Interprets the results in the context of the books, comparing the insights gained through LSI with traditional analysis methods.
7. Conclusion: Concludes the report by summarizing findings, discussing practical applications, and proposing future work.
8. References and Appendices: Provides references for books, tools, and techniques used, along with supplementary materials like visualizations and code.

Latent Semantic Indexing, or LSI as it will be referred to throughout this project, is a technique designed to help computers have a better understanding of the context and connotations associated with words given to it by a user. The process of LSI is heavily reliant on Singular Value Decomposition, or SVD, to break down the data given and use it to make various estimates of the users' intentions. LSI has been used in the application of many different processes such as search engine optimization, recommendation systems, document classification, language understanding, etc.

It does have some flaws however, the two largest ones being the existence of synonyms, which are many different words that can describe the same thing, and polysemy, which are words and phrases that have more than one meaning. Computers and statistical software have a difficult time interpreting human languages, often not understanding the contexts or connotations of certain requests and questions. LSI was patented in 1988 and at the time was a very innovative technique for information retrieval. However, as time has passed and technology has continued to evolve, it has become relatively antiquated and been rendered almost obsolete by the new approaches and more advanced hardware.

1.5 Comparison with Traditional Analysis

In comparison with traditional word frequency analysis, LSI is able to detect patterns in words that lead to thematic sequencing - for instance, in word frequency analysis, positive words like "good" and "great" might come up often, making you think a body of text could have a positive thematic connotation. However, if the word “not” is frequently used before the positive words, the thematic nature of the text might be more negative than word frequency allows for detection of. Therefore, LSI allows for deeper analysis than traditional word frequency analysis would. Additionally, LSI uses SVD to filter out much of the noise present within text - as the high dimensionality of a work is hard to process, SVD is used to reduce noise and content that has little semantic meaning - this allows for irrelevant information to be processed out in a way that word frequency analysis does not account for. Reduction of dimensionality also increases computational efficiency, making LSI a viable alternative to traditional analysis methods.

2 SVD in LSI

2.1 SVD

LSI is heavily reliant on Singular Value Decomposition. Given an $m \times n$ matrix (i.e. word to document matrix) A, it can be broken down by SVD into three parts,

$$A_{m\times n} = U\Sigma V^T \tag{1}$$

Where $U^TU = I$ and $V^TV = I$ are symmetric matrices and Σ is a diagonal matrix with

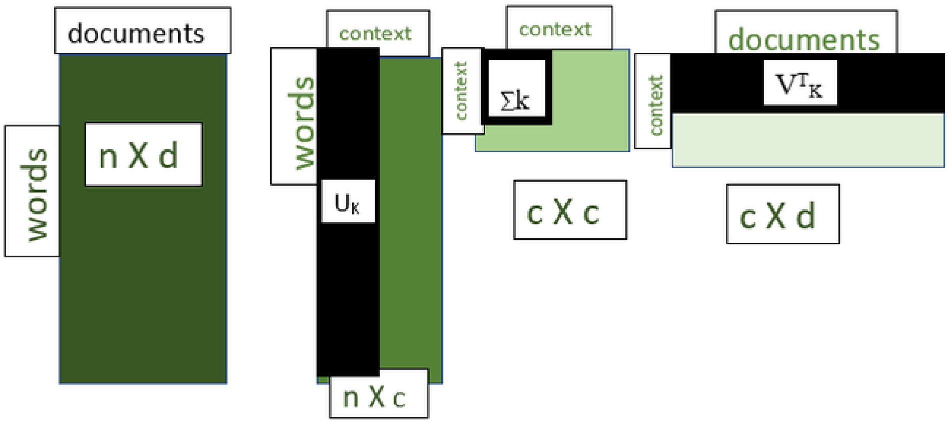


Figure 1: Diagram of SVD

values that represents the significance of the context from highest to lowest (Figure 1). These values can be used to reduce the dimensions of the matrix. Selecting the k, the largest diagonal values, within the Σ matrix we get,

$$A_k = U_k\Sigma_kV_k^T$$

A visual representation of this is well drawn out in figure 2.

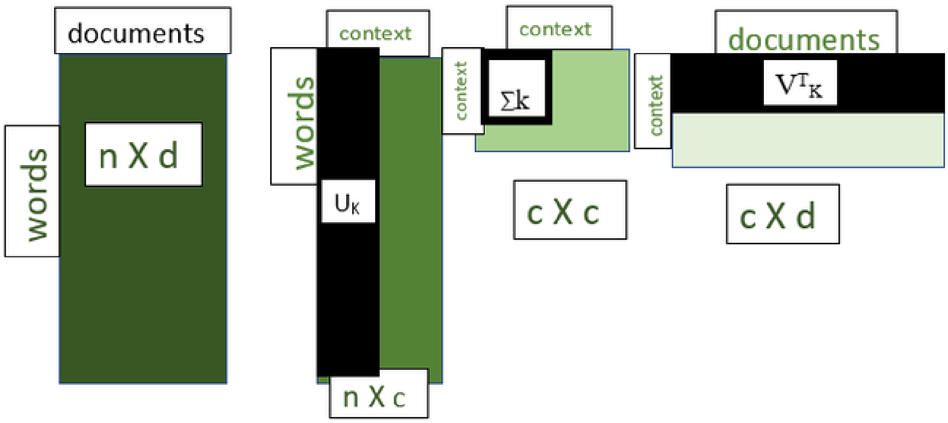


Figure 2: Truncated SVD after choosing k value

3 Is NMF possible In LSI?

Now that we saw how SVD plays a crucial role in LSI, one question that comes up is can Non-negative Factorization be used in the same way, after all its a decomposition technique?

3.1 NMF

NMF is used when A is a matrix that is composed of only non negative entries ($A_{ij} \geq 0$). It factorizes A into two non-negative matrices $W \in \mathbb{R}^{m\times k}$ and $H \in \mathbb{R}^{k\times n}$ as the following:

$$A \approx WH.$$

Here, W and H are chosen so that their product is the approximation of A, where the size, k, is smaller than both m and n. From this factorization, every element in both matrices are non-negative, or greater than 0.

The goal of NMF is to make the factorization, WH, as close as possible an approximation of A. This is achieved through minimizing the error between A and WH, which would be the difference between the matrices. This is expressed as as:

$$\min_{W,H} \|A - WH\|_F^2,$$

where $\|\cdot\|_F$ represents the magnitude of error between the original and the approximation given by NMF through using the Frobenius norm. Through this error check, we see that all entries in W and H must be nonnegative.

Although NMF and LSI share the goal of uncovering latent semantic structures in text data, their underlying assumptions differ.

NMF imposes a non-negativity constraint on the factor matrices, ensuring that the resulting factors are interpretable as parts-based representations.

This can lead to negative values in the factor matrices, which can be less intuitive to interpret.

4 Data Set

The dataset that we are using for this analyis are 10 different book that are taken from the **Project Gutenberg**[1].

Project Gutenberg is a digital library that offers free access to thousands of eBooks, particularly classic literature and out-of-copyright works. It was founded by Michael S. Hart in 1971, making it the oldest digital library in the world. Named after Johannes Gutenberg, the inventor of the printing press, the project aims to make books accessible to everyone, fostering global literacy and knowledge sharing. The dataset comprises a selection of 10 books spanning a diverse range of genres, themes, and styles. It includes fantasy, science fiction, historical works, memoirs, and satirical novels.

4.1 Books Abstract

- *Alice's Adventures in Wonderland by Lewis Carroll* is a fantasy novel that describes the titular character, Alice, who falls into a rabbit hole and enters a surrealist dream landscape that defies her logic. She encounters characters such as the Cheshire Cat and the Mad Hatter, and tries her best to navigate her new absurdist reality. The book uses surrealism to describe how a child enters adulthood, and the challenges that growing up may bring.

- *A Christmas Carol in Prose; Being a Ghost Story of Christmas by Charles Dickens* is a fantasy novel that centers around Ebeneezer Scrooge, a greedy old man with a hatred for his community and Christmas as a whole. Due to his unchanging ways, he is visited by his former business partner, Jacob Marley, on Christmas Eve. Jacob tries to warn Scrooge to change from his misery and greed. He refuses, and three spirits - the Ghosts of Christmas Past, Present, and Yet to Come, visit him over the course of the evening.

- *Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft Shelley* is a science fiction novel that tells the story of Victor Frankenstein, a "mad scientist" who feels the need to prove himself, and in that journey successfully creates a living creature from body parts. The monster he has created horrifies him, and despite being the first scientist to create life, he abandons the monster he has created. Frankenstein's creation then begins to try and track down Frankenstein for vengeance of being created as a monster. The title's reference to Prometheus notes the punishment in trying to steal from the Gods, just as Frankenstein played God in his creation of life.

- *The Time Machine by H. G. Wells* is a science fiction novel that deals with a character known as the Time Traveler who constructs a machine that lets him travel through time. He invites his friends over to demonstrate the invention, where he travels to the distant future, hundreds of thousands of years later. He finds that humans have been split into two groups - Eloi, who roam on the surface of Earth, and the Morlocks, who live underground and survive by preying on the Eloi. The Time Traveler escapes, but is disturbed by humanity for what will become of it and the societies it builds. The novel is a metaphor for heightening of class structure and how it will cause decay in the evolution of human culture.

- *The Art of War by Sun Tzu* is a historical instructional guide on warfare, which places particular emphasis on the importance of planning within warfare, and how important it is to have an acute understanding of the enemy you are facing. The book also emphasizes deception as a tactic for warfare, with guidance on deceiving your enemy. Of note in the guide is how the author describes warfare as a last resort, placing emphasis on never taking unnecessary battles and minimizing losses. This is a nonfiction book for warfare, but can be applied as guidance to various fields, such as positions of leadership.

- *A Pickle for the Knowing Ones by Timothy Dexter* is a historical autobiography written by an American colonialist businessman who made money through often illogical and nonsensical methods. In this short writing, he brags about the money he has accumulated through these ventures, and critiques society down to both local and national figures. The writing is rambling, lacks punctuation, and acts to showcase the absurd personality of Timothy Dexter. The work has been viewed as a satirical comment on early American society, but might also be a genuine work expressing the unironic views and beliefs of the author.

- *Heart of Darkness by Joseph Conrad* is an adventure novel describing the journey of Charles Marlow as he sails into further into the Congo River to retrieve an ivory trader named Kurtz - before coming on the mission, Charles hears stories of how Kurtz has created a new society within the Congo wherein he has established himself as a God to be worshipped. As Charles travels deeper into Africa, the novel progressively gets darker as he sees horrific brutality as a direct effect of European colonization on the continent. The novel is a critique of what really forms society and culture, as well as the destructive effects of power and imperialism.

- *Gulliver's Travels into Several Remote Nations of the World by Jonathan Swift* is an adventure novel that describes how the titular Lemuel Gulliver, the surgeon on a sailing ship, encounters bizarre societies in his adventures across the globe. The stories describe fantastical civilizations - most famously, a land of tiny humans who have aristotical groups that fight in wars with one another. Additionally, he encounters a land of giant people, a land of intellectuals who do not care about the world, and a land where horses reign over beasts. Gulliver thinks less of his society and the human race as a result. The novel is satirical, highlighting the worst of humanity and how it is expressed through our culture and politics.

- *Pygmalion by Bernard Shaw* is a comedic play that takes place in London during the early 1900s. Eliza Doolittle, who is a flower girl living in poverty with a heavy Cockney accent, is taken for study by linguist and professor Henry Higgins. Henry bets a colleague that he can transform Eliza into an aristocrat by only teaching her how to speak in proper English, and believes that proper manners and mindset will follow. Eliza is rapidly transformed into a proper English lady, but realizes that the life of nobility comes with restrictions and dependence. The novel satirizes early 20th century English culture and social class structure.

- *Life on the Mississippi by Mark Twain* is a comedic memoir, recounting the author's life in his youth before the Civil War, where he worked as a steamboat pilot on the Mississippi river. While the novel has elements of bittersweet longing and remembrance for youth, the author details the towns and residents of the Mississippi and riverboat worker culture with humor. The novel also compares childhood dreams with eventuality of work, and the challenges that come with the realities of work.

4.2 Dataset Description

Book Title	Author	Genre	Page Count (Approx.)	Word Count (Approx.)	Copyright Laws
Alice's Adventures in Wonderland	Lewis Carroll	Fiction	96	29,610	Published in 1865, this work is in the public domain.
A Christmas Carol in Prose; Being a Ghost Story of Christmas	Charles Dickens	Fiction	104	31,650	Published in 1843, this work is in the public domain.
Frankenstein; Or, The Modern Prometheus	Mary Wollstonecraft Shelley	Science Fiction	280	78,100	Published in 1818, this work is in the public domain.
The Time Machine	H. G. Wells	Science Fiction	118	35,530	Published in 1895, this work is in the public domain.
The Art of War	Sun Tzu	History	68	15,000	Written in the 5th century B.C., this work is in the public domain.
A Pickle for the Knowing Ones	Timothy Dexter	History	24	17,101	Published in 1802, this work is in the public domain.
Heart of Darkness	Joseph Conrad	Adventure	72	41,597	Published in 1899, this work is in the public domain.
Gulliver's Travels into Several Remote Nations of the World	Jonathan Swift	Adventure	306	107,293	Published in 1726, this work is in the public domain.
Pygmalion	Bernard Shaw	Comedy	80	36,718	Published in 1913, this work is in the public domain.
Life on the Mississippi	Mark Twain	History	624	174,000	Published in 1883, this work is in the public domain.

Table 1: Dataset Description: Summary of the 10 books used for analysis.

5 Methodology

5.1 Data Processing

For the 10 books that we used in this project, the first step we took to perform LSI in Python was to download each book as a PDF file. After we downloaded each book as a PDF file, we used a Python script that read each book and saved it into a variable called text data. However, before we could start our LSI analysis, further steps had to be taken to prepare the data for LSI analysis.

The first that we had to take is called tokenization. In the Natural Language Process (NLP) tokenization is breaking down large amounts of text into individual words or tokens. Once the text data had been tokenized, we then had to stop word removal which is the process of removing commonly used words in the English language. These words include but are not limited to “the”, “a”, “an”, or “in”. These words are removed to save up space in our data but also removing them tends to lead to a more accurate model.

The last step we had to take was to lemmatize words in the text data. In NLP, lemmatization is the process of reducing a word into its base form. For instance, it will turn the word "running" into "run", "better" into "good" and "cats" into "cat". Lemmatization allows the model to ensure that words with different forms like in tenses or their plural forms are treated as its base words. The process of lemmatization helps the model to have better accuracy(IBM).

5.2 Latent Semantic Indexing

- **Brief explanation of LSI**:
  Latent Semantic Indexing (LSI) is a technique used in text analysis to identify hidden relationships between terms and documents. It works by applying SVD to a term-document matrix, reducing its dimensions while retaining key semantic information. This process captures patterns in the data, addressing issues like:
  - Synonymy: Different words that have the same meaning.
  - Polysemy: Words or phrases that have multiple meanings based on their context.
- **Description of Term-Document Matrix creation.**:
  A Term-Document Matrix (TDM) represents the frequency of terms across documents in a corpus. Rows correspond to unique terms, columns correspond to documents, and cell values demonstrate term frequency or weighted values like TF-IDF. To create it, you must:
  - Preprocess the text.
  - Define the vocabulary.
  - Populate the matrix with counts or weights.

  The TDM is essential for processing methods like LSI and helps analyze term relationships, document similarities, and patterns in text data.

- **Use of weighting schemes (e.g., TF-IDF).**:
  Weighting schemes like TF-IDF assign importance to terms in a document corresponding to a body of text. They highlight significant terms while reducing the impact of common ones. TF-IDF combines Term Frequency (TF), or how often a term appears in a document, with Inverse Document Frequency (IDF), how rare the term is across the body of text. To understand this, you must:

5.3 Tools and Libraries

As mentioned above, we leveraged Python to perform the LSI. We choose Python over because it is widely used in industry and there are many different libraries made for NLP. We leveraged the following libraries in our analysis.

Library
scikit learn
NLTK
matplotlib
Gensim
seaborn

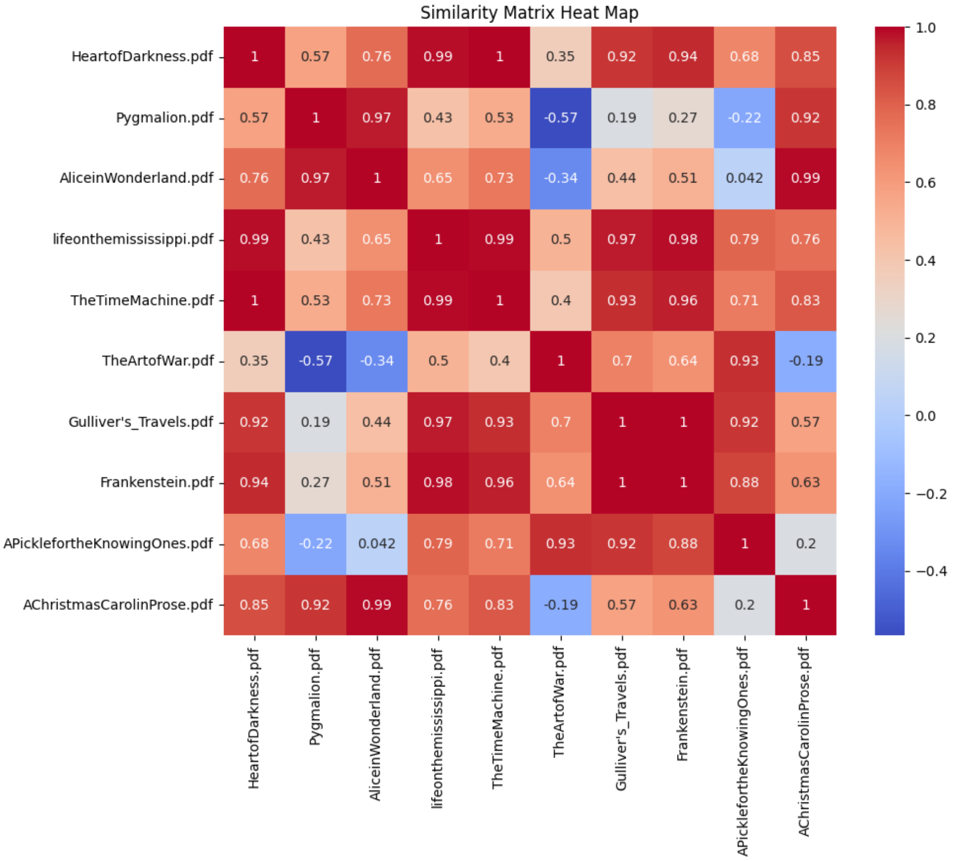
6 Analysis

6.1 Dimensionality Reduction

Like any other data science project, we often have data in higher dimensions. To make data amenable for machine learning models, it is desirable to reduce the dimensions of the data while retaining as much information as possible (Benjamin Fayyazuddin Ljungberg). In LSA dimension reduction is achieved by the singular values to reduce the number of features in the term-document matrix while retaining most of the semantic information. Having a reduced-dimensional data set allows LSA to focus on the latent structure within the data, leading to better insights.

6.2 Exploratory Data Analysis (EDA)

Our preliminary results showed that despite differences in genres, semantic relationships were able to link certain books with an element of similarity. For instance, the earlier(Figure 3) provided heat map that A Pickle for the Knowing Ones was closely linked *to The Art of War* - while both these books acted as our 2 examples of historical works, this similarity is more likely due to the political conflict that centers around conflict within both novels. *A Pickle for the Knowing Ones* is a work that describes conflict that a member of a community has with politicians - while not entirely true in thematic content for *The Art of War*, the language used is similar enough that it provokes the analysis to detect similarity. We can attest for the differences in works through differences in genre - in another example, *Pygmalion* was vastly different from *The Art of War* due to containing little of the same language. Looking at the genre of the two books, we see that this is because *The Art of War* is a historical political book that details conflict, while *Pygmalion* is a modern comedy that deals with class relations. These two works have little to do with one another in both thematic content and in language used - this is why the similarity comparison found these to be the two books that are the least alike.



	HeartofDarkness.pdf	Pygmalion.pdf	AliceinWonderland.pdf	lifeonthemississippi.pdf	TheTimeMachine.pdf	TheArtofWar.pdf	Gulliver's_Travels.pdf	Frankenstein.pdf	APicklefortheKnowingOnes.pdf	AChristmasCarolinProse.pdf
HeartofDarkness.pdf	1	0.57	0.76	0.99	1	0.35	0.92	0.94	0.68	0.85
Pygmalion.pdf	0.57	1	0.97	0.43	0.53	-0.57	0.19	0.27	-0.22	0.92
AliceinWonderland.pdf	0.76	0.97	1	0.65	0.73	-0.34	0.44	0.51	0.042	0.99
lifeonthemississippi.pdf	0.99	0.43	0.65	1	0.99	0.5	0.97	0.98	0.79	0.76
TheTimeMachine.pdf	1	0.53	0.73	0.99	1	0.4	0.93	0.96	0.71	0.83
TheArtofWar.pdf	0.35	-0.57	-0.34	0.5	0.4	1	0.7	0.64	0.93	-0.19
Gulliver's_Travels.pdf	0.92	0.19	0.44	0.97	0.93	0.7	1	1	0.92	0.57
Frankenstein.pdf	0.94	0.27	0.51	0.98	0.96	0.64	1	1	0.88	0.63
APicklefortheKnowingOnes.pdf	0.68	-0.22	0.042	0.79	0.71	0.93	0.92	0.88	1	0.2
AChristmasCarolinProse.pdf	0.85	0.92	0.99	0.76	0.83	-0.19	0.57	0.63	0.2	1

Figure 3: Cosine Siimilariy between the books

Across genres, we see a large element of separation in similarity. The genres that are the most dissimilar are comedy and historical, as well as fantasy and historical. Science fiction appears to have a strong correlation with adventure - having as close to a 1 similarity ranking as is possible within our analysis. Thematically, these connections make sense - the science fiction books we used have the same element of civil unrest as the adventure books we used. *Heart of Darkness* and *Gulliver's Travels* are both the adventure novels we used and both use their context of adventure to highlight fundamental critiques that the author feels toward human civilization and the societies they wrote the novels. Similarly, both of the science fiction novels we used, *The Time Machine* and *Frankenstein*, use the science fiction genre as a critique of the society in which the authors wrote. As both of these genres used their respective framework to thematically note the issues in the world around them, it is clearer why these are the two most correlated genres in the semantic index similarity processing.

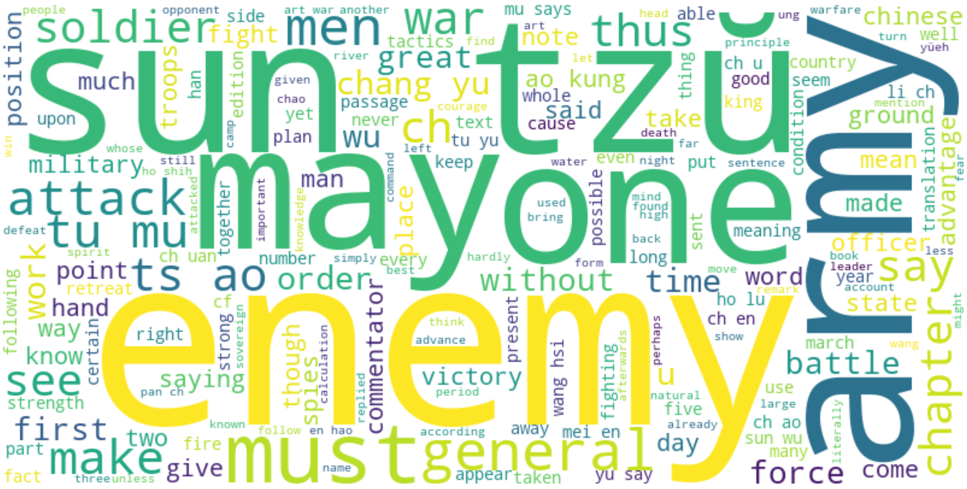


Figure 4: Word Cloud for the book 'The Art of War'

The word cloud for The Art of War by Sun Tzu highlights the most frequently used terms in the text, providing an overview of its central themes and concepts. Prominent words such as "enemy," "army," "soldier," "general," and "attack" dominate the visualization, reflecting the text's focus on military strategy, tactics, and warfare principles. Other key terms like "victory," "force," "order," and "ground" emphasize the practical and philosophical aspects of achieving success in battle. The presence of words like "must," "may," and "thus" indicates the prescriptive and instructional tone of the work. This word cloud serves as a visual summary of the text, capturing its thematic essence and highlighting the importance of strategy, leadership, and conflict resolution in its content.

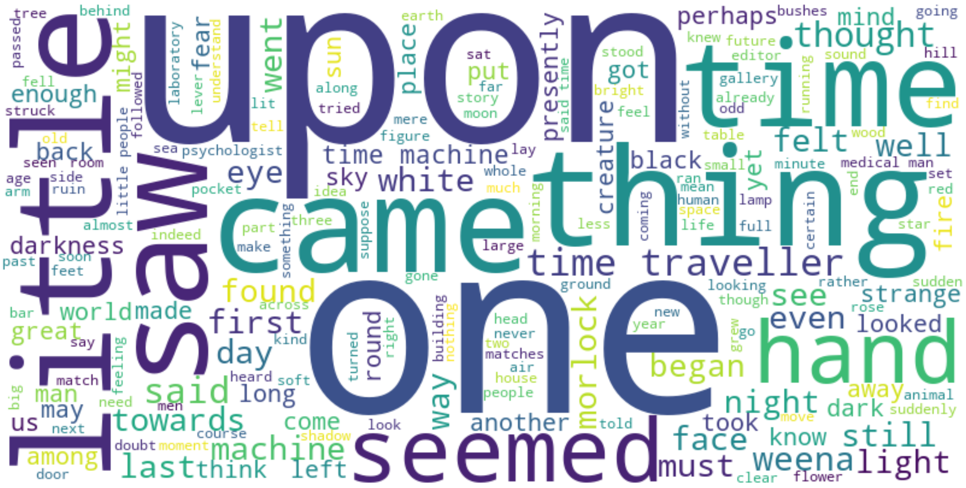


Figure 5: Word Cloud for the book 'The Time Machine'

The word cloud for The Time Machine by H.G. Wells provides a visual representation of the most frequently used words in the text, shedding light on its narrative focus and themes. Prominent words such as "time," "one," "upon," "saw," and "came" suggest a descriptive and observational style, reflecting the protagonist's experiences and discoveries during their journey through time. Other significant terms like "machine," "traveller," "morlock," and "darkness" highlight key elements of the story, including the central concept of time travel and the portrayal of futuristic civilizations. Words like "seemed," "thought," and "felt" emphasize the introspective and speculative nature of the narrative. Overall, the word cloud captures the imaginative and exploratory essence of the novel, showcasing its blend of adventure, science fiction, and philosophical inquiry.

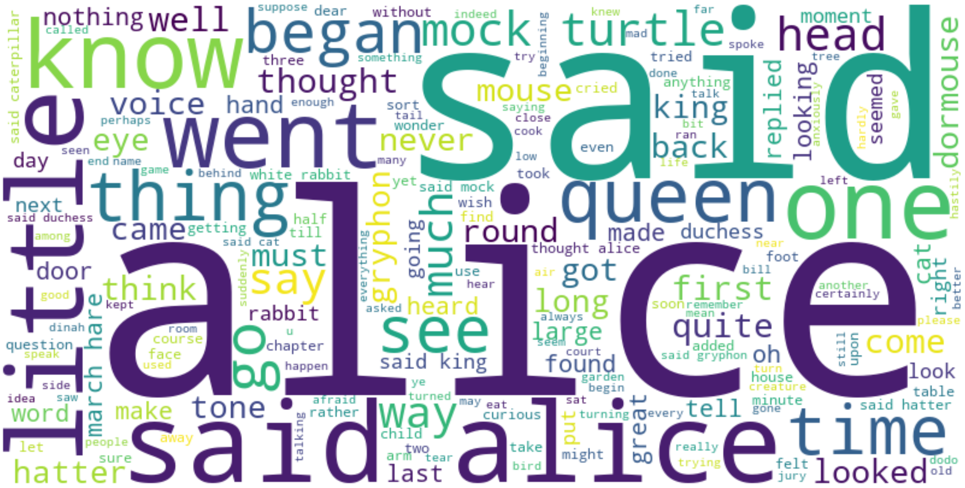


Figure 6: Word Cloud for the book 'Alice's Adventures in Wonderland'

The word cloud for Alice's Adventures in Wonderland by Lewis Carroll highlights the whimsical and dialogue-driven nature of the text. The central prominence of the words "Alice" and "said" reflects the protagonist's role as the focal point of the narrative and the story's reliance on character interactions. Terms like "queen," "rabbit," "hatter," and "turtle" emphasize the fantastical cast of characters that populate Wonderland, each contributing to the surreal and imaginative atmosphere. Words such as "thought," "know," and "time" underscore the introspective and curious journey Alice embarks upon, often questioning the logic and absurdity of her surroundings. The recurrence of action-oriented words like "went," "came," and "began" conveys the dynamic and eventful progression of the plot. Overall, the word cloud captures the playful, nonsensical, and richly descriptive elements of Carroll's iconic tale.

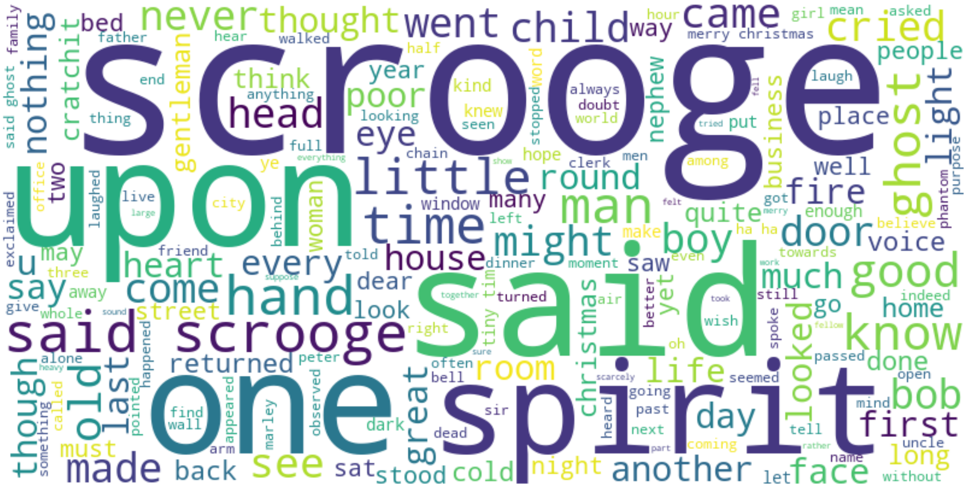


Figure 7: Word Cloud for the book 'A Christmas Carol in Prose; Being a Ghost Story of Christmas'.

The word cloud for "A Christmas Carol" by Charles Dickens showcases the novella's thematic emphasis on transformation and redemption. The dominant presence of the words "Scrooge," "said," and "one" highlights the central character's journey and the narrative's focus on his inner world. Terms like "spirit," "ghost," and "Christmas" underscore the supernatural elements driving Scrooge's change, while words such as "child," "boy," and "man" trace his development from a miserly old man to a kinder, gentler soul. The recurrence of words like "time," "night," and "day" conveys the story's condensed timeline and the swift nature of Scrooge's transformation. Overall, the word cloud encapsulates the novella's exploration of personal growth, morality, and the redemptive power of kindness.

6.3 Latent Concepts

- **Identification of latent semantic dimensions.**:
  Identification of latent semantic dimensions involves uncovering hidden patterns and relationships in text data that aren't immediately obvious. Through techniques like LSI, SVD, and other dimensionality reduction methods, we reduce the original high-dimensional term-document matrix into a lower-dimensional space. The retained dimensions capture the most important semantic structures, such as the underlying themes or topics, by grouping related terms and documents. This helps improve tasks like document retrieval, clustering, and topic modeling by focusing on the deeper meaning rather than surface-level words.

- **Examples of terms that are grouped in the same concepts.**: Examples of terms grouped in the same concepts through latent semantic analysis would be:
  - **Synonyms**: Words with similar meanings, like "car" and "automobile" or "happy" and "joyful".
  - **Related Terms**: Words that are related in context, such as "doctor", "physician", and "nurse" in the medical field, or "apple", "banana", and "fruit" in food-related contexts.

- **Contextual Associations**: Terms used in similar contexts but not exact synonyms, such as "bank" (financial institution) and "bank" (riverbank), which might be distinguished by surrounding words.
- **Topic-Related Terms**: Words like "climate", "weather", "temperature", and "environment" are grouped under environmental science concepts.

These term groups allow for the identification of latent semantic dimensions and provide a better understanding of relationships between terms in a document.

6.4 Book Similarities

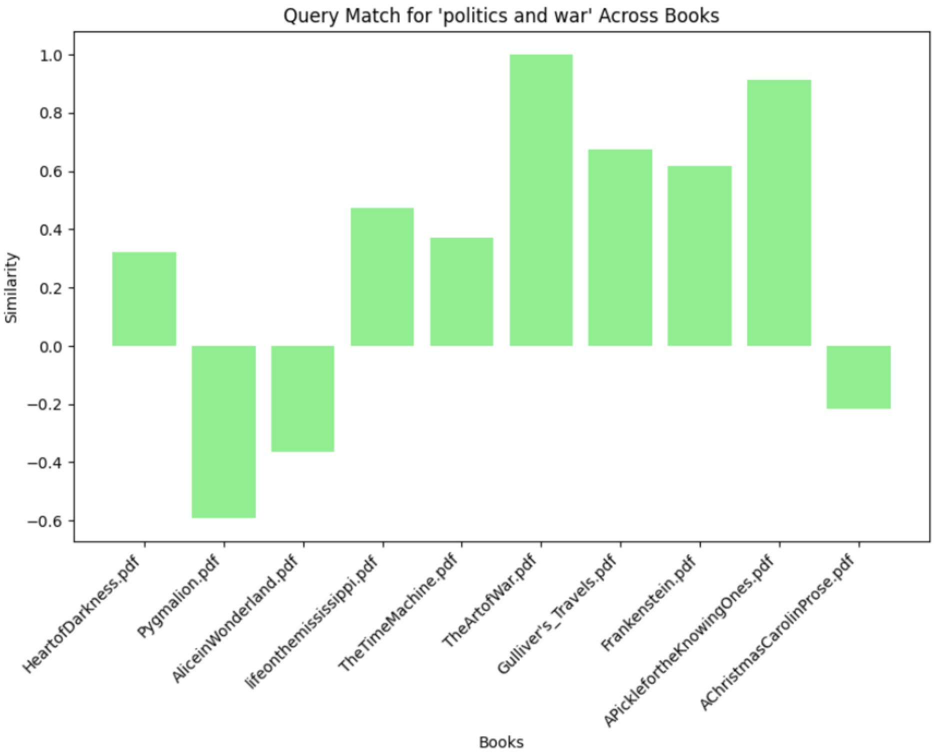


Figure 8: Query Match for **Politics and War** across different books

The bar graph illustrates the query match for 'politics and war' across various books, with the y-axis representing similarity scores ranging from -0.6 to 1.0. The x-axis lists 14 books, including "HeartofDarkness.pdf" and "AChristmasCarolInProse.pdf". Notably, "TheArtOfWar.pdf" and "AChristmasCarolInProse.pdf" exhibit the highest similarity scores, while "AliceinWonderland.pdf" and "AChristmasCarolInProse.pdf" display the lowest. The graph is set against a white background, providing a clear visual representation of the query match results.

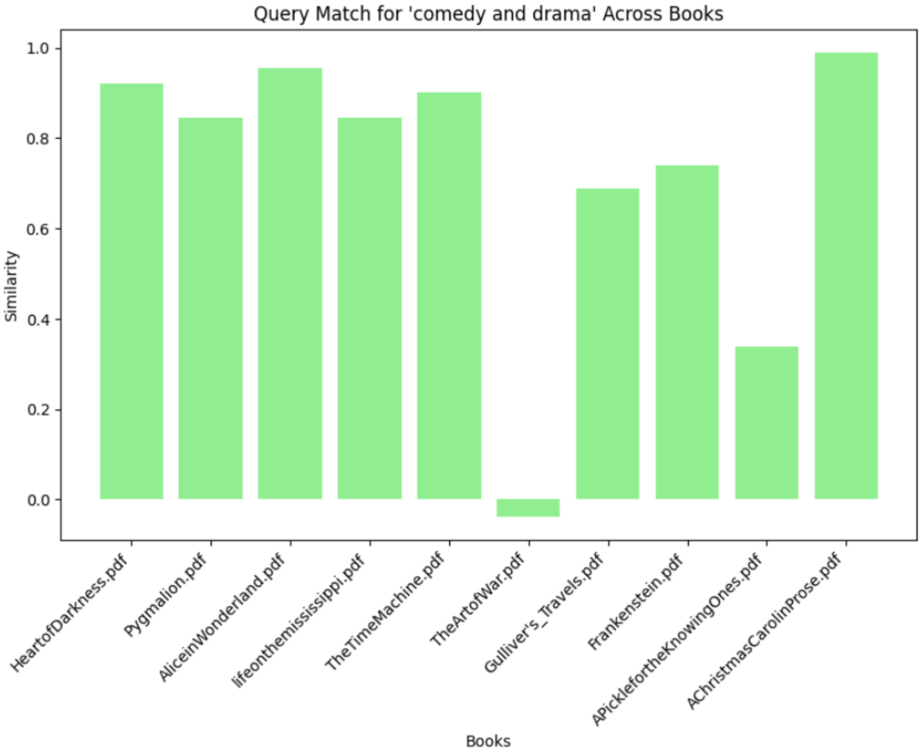
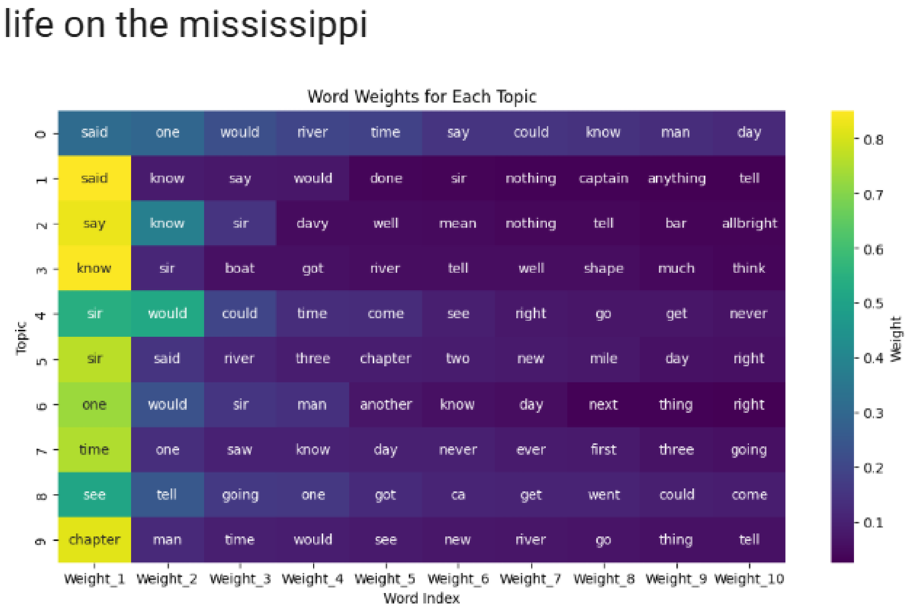


Figure 9: Query Match for **Comedy and Drama** across different books

The query match plot for comedy and drama books reveals a fascinating pattern, with "Heart of Darkness.pdf" and "Pygmalion.pdf" emerging as the top matches, boasting a similarity score of 0.9. In contrast, "Gulliver's Travels.pdf" stands out as the least similar, with a score of 0.1. The remaining books cluster around a similarity score of 0.8, indicating a moderate level of similarity. This plot suggests that certain themes or elements are more prevalent in comedy and drama books, while others are less common. Further analysis could uncover the underlying factors driving these similarities and differences.

6.5 Topic Analysis

The heat plot(Figure 5) consists of the 10 most common words associated with each topic extracted using Latent Semantic Analysis (LSA). Each row represents a distinct topic, and the columns correspond to word weights, indicating the importance of specific words within that topic. The color intensity reflects the weight of each word, with brighter colors signifying higher contributions. This visualization helps in interpreting the thematic structure of "Life on the Mississippi," showcasing how key terms cluster together to form meaningful topics across the text.

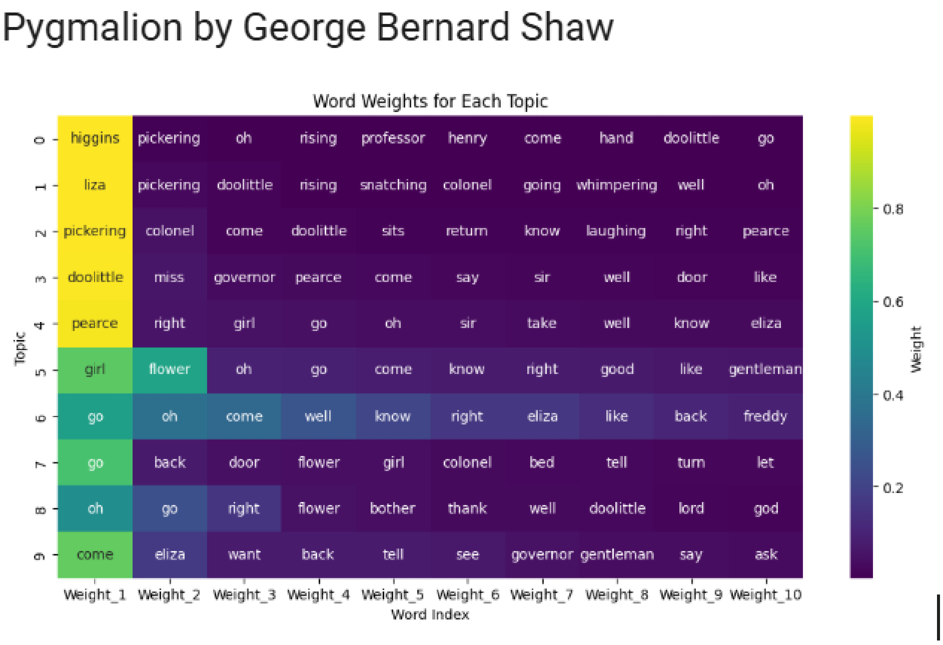


Topic	Weight_1	Weight_2	Weight_3	Weight_4	Weight_5	Weight_6	Weight_7	Weight_8	Weight_9	Weight_10
0	said	one	would	river	time	say	could	know	man	day
1	said	know	say	would	done	sir	nothing	captain	anything	tell
2	say	know	sir	davy	well	mean	nothing	tell	bar	allbright
3	know	sir	boat	got	river	tell	well	shape	much	think
4	sir	would	could	time	come	see	right	go	get	never
5	sir	said	river	three	chapter	two	new	mile	day	right
6	one	would	sir	man	another	know	day	next	thing	right
7	time	one	saw	know	day	never	ever	first	three	going
8	see	tell	going	one	got	ca	get	went	could	come
9	chapter	man	time	would	see	new	river	go	thing	tell

Figure 10: Life on the Mississippi

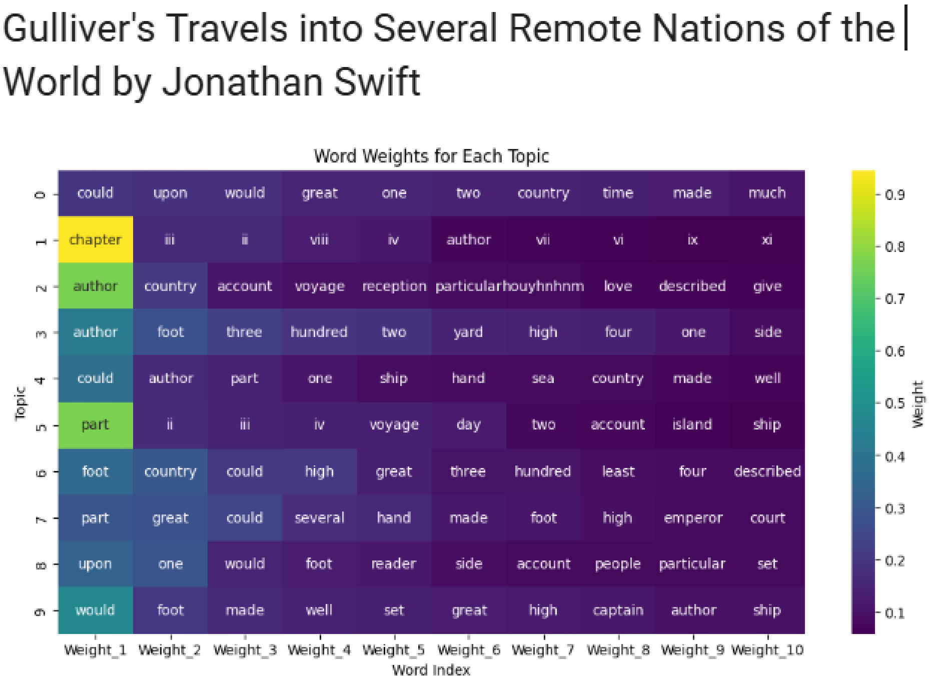
The same type of heat plot analysis applies to the other books in the dataset by extracting and visualizing the top 10 most significant words for each latent topic identified through Latent Semantic Analysis (LSA). For each book, the heat plot reveals the thematic structure and highlights key terms that are most strongly associated with each topic. This allows for a comparative exploration of the textual and thematic characteristics of works across different genres, authors, and styles, ranging from fiction and science fiction to history and adventure. The consistent approach ensures that the visualizations provide insight into the unique linguistic and thematic patterns within each book.

Below are all the other heat maps for the other books in the dataset.



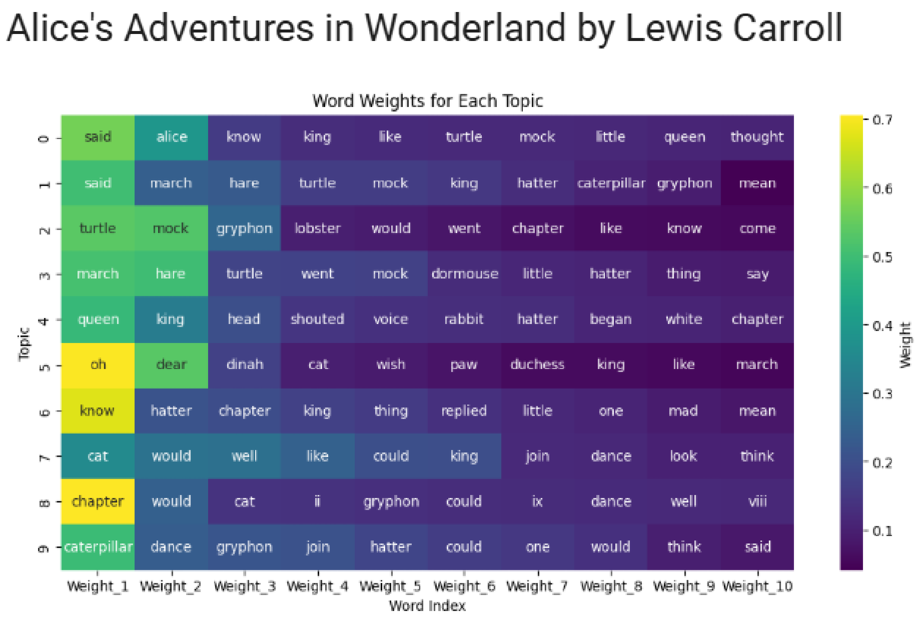
Topic	Weight_1	Weight_2	Weight_3	Weight_4	Weight_5	Weight_6	Weight_7	Weight_8	Weight_9	Weight_10
0	higgins	pickering	oh	rising	professor	henry	come	hand	doolittle	go
1	liza	pickering	doolittle	rising	snatching	colonel	going	whimpering	well	oh
2	pickering	colonel	come	doolittle	sits	return	know	laughing	right	pearce
3	doolittle	miss	governor	pearce	come	say	sir	well	door	like
4	pearce	right	girl	go	oh	sir	take	well	know	eliza
5	girl	flower	oh	go	come	know	right	good	like	gentleman
6	go	oh	come	well	know	right	eliza	like	back	freddy
7	go	back	door	flower	girl	colonel	bed	tell	turn	let
8	oh	go	right	flower	bother	thank	well	doolittle	lord	god
9	come	eliza	want	back	tell	see	governor	gentleman	say	ask

Figure 11: Pygmalion by George Bernard Shaw



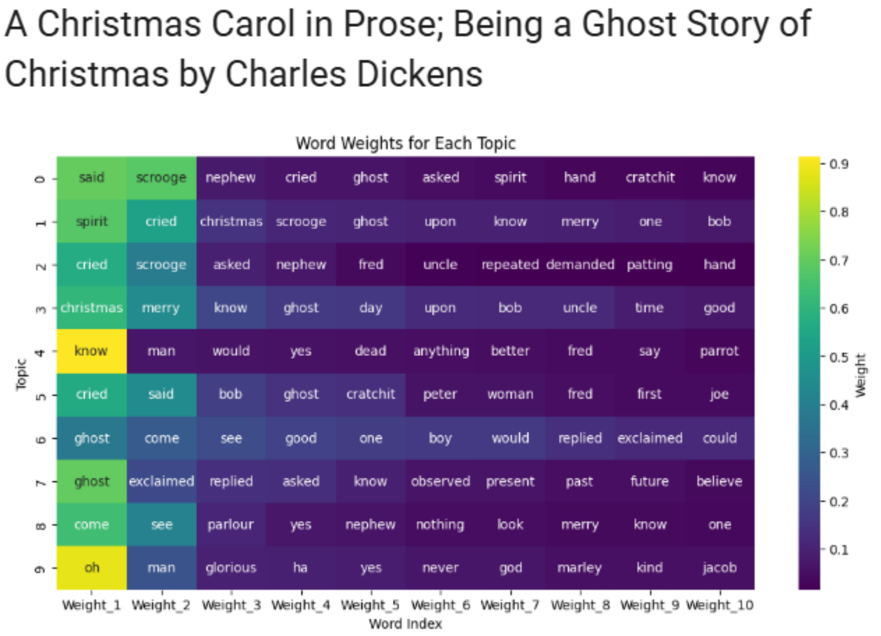
Topic	Weight_1	Weight_2	Weight_3	Weight_4	Weight_5	Weight_6	Weight_7	Weight_8	Weight_9	Weight_10
0	could	upon	would	great	one	two	country	time	made	much
1	chapter	iii	ii	viii	iv	author	vii	vi	ix	xi
2	author	country	account	voyage	reception	particular	houyhnhnm	love	described	give
3	author	foot	three	hundred	two	yard	high	four	one	side
4	could	author	part	one	ship	hand	sea	country	made	well
5	part	ii	iii	iv	voyage	day	two	account	island	ship
6	foot	country	could	high	great	three	hundred	least	four	described
7	part	great	could	several	hand	made	foot	high	emperor	court
8	upon	one	would	foot	reader	side	account	people	particular	set
9	would	foot	made	well	set	great	high	captain	author	ship

Figure 12: Gulliver's Travels into Several Remote Nations of the World by Jonathan Swift



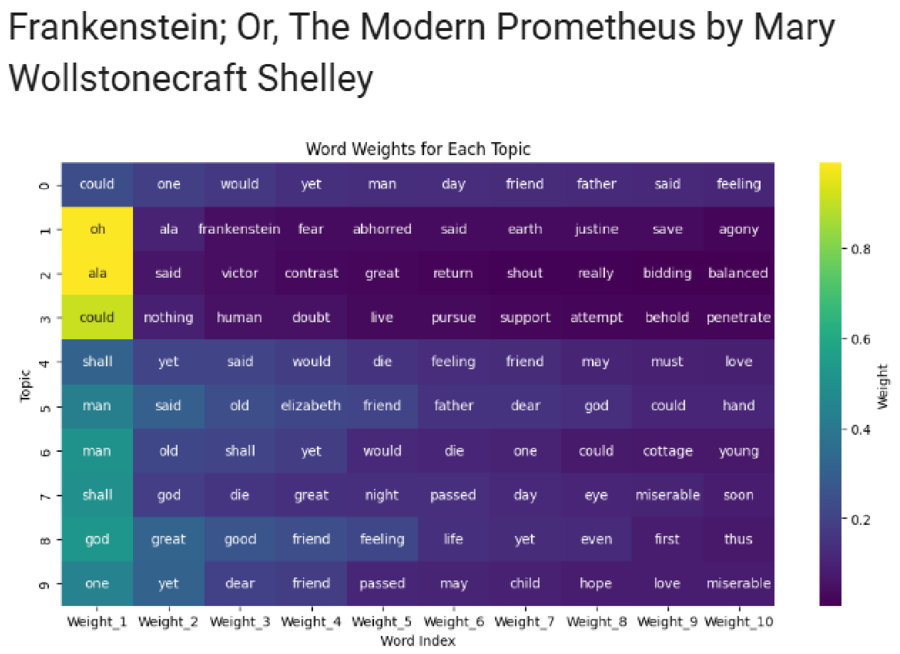
Topic	Weight_1	Weight_2	Weight_3	Weight_4	Weight_5	Weight_6	Weight_7	Weight_8	Weight_9	Weight_10
0	said	alice	know	king	like	turtle	mock	little	queen	thought
1	said	march	hare	turtle	mock	king	hatter	caterpillar	gryphon	mean
2	turtle	mock	gryphon	lobster	would	went	chapter	like	know	come
3	march	hare	turtle	went	mock	dormouse	little	hatter	thing	say
4	queen	king	head	shouted	voice	rabbit	hatter	began	white	chapter
5	oh	dear	dinah	cat	wish	paw	duchess	king	like	march
6	know	hatter	chapter	king	thing	replied	little	one	mad	mean
7	cat	would	well	like	could	king	join	dance	look	think
8	chapter	would	cat	ii	gryphon	could	ix	dance	well	viii
9	caterpillar	dance	gryphon	join	hatter	could	one	would	think	said

Figure 13: Alice's Adventures in Wonderland by Lewis Carroll



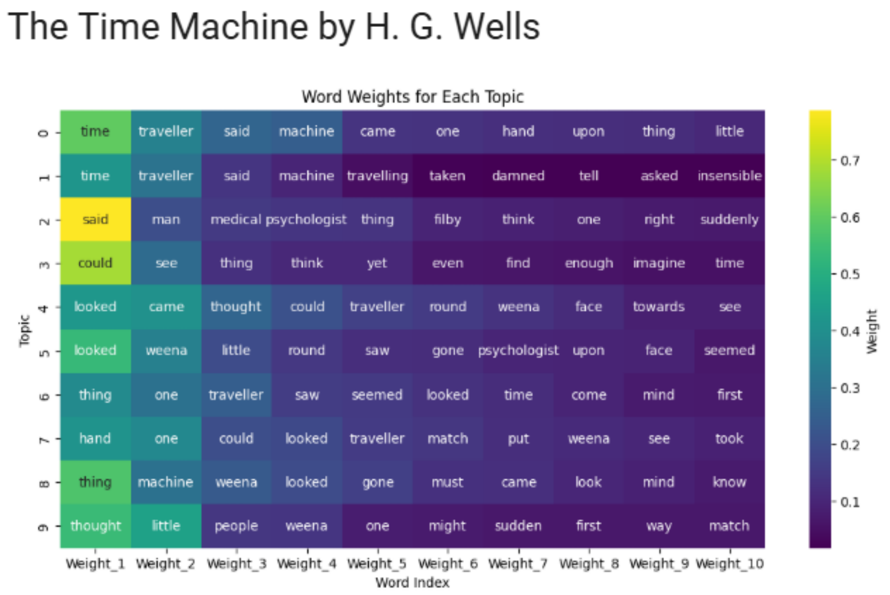
Topic	Weight_1	Weight_2	Weight_3	Weight_4	Weight_5	Weight_6	Weight_7	Weight_8	Weight_9	Weight_10
0	said	scrooge	nephew	cried	ghost	asked	spirit	hand	cratchit	know
1	spirit	cried	christmas	scrooge	ghost	upon	know	merry	one	bob
2	cried	scrooge	asked	nephew	fred	uncle	repeated	demanded	patting	hand
3	christmas	merry	know	ghost	day	upon	bob	uncle	time	good
4	know	man	would	yes	dead	anything	better	fred	say	parrot
5	cried	said	bob	ghost	cratchit	peter	woman	fred	first	joe
6	ghost	come	see	good	one	boy	would	replied	exclaimed	could
7	ghost	exclaimed	replied	asked	know	observed	present	past	future	believe
8	come	see	parlour	yes	nephew	nothing	look	merry	know	one
9	oh	man	glorious	ha	yes	never	god	marley	kind	jacob

Figure 14: A Christmas Carol in Prose; Being a Ghost Story of Christmas by Charles Dickens



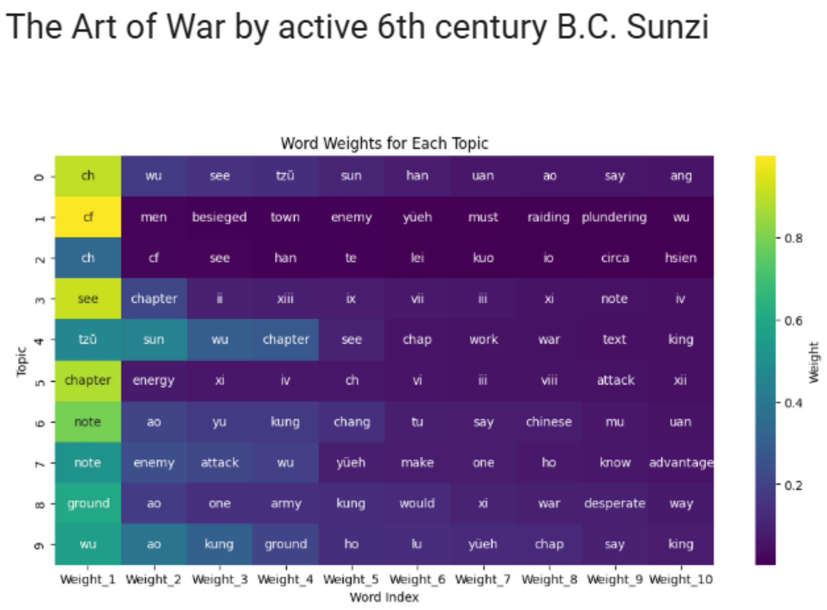
	Weight_1	Weight_2	Weight_3	Weight_4	Weight_5	Weight_6	Weight_7	Weight_8	Weight_9	Weight_10
0	could	one	would	yet	man	day	friend	father	said	feeling
1	oh	ala	frankenstein	fear	abhorred	said	earth	justine	save	agony
2	ala	said	victor	contrast	great	return	shout	really	bidding	balanced
3	could	nothing	human	doubt	live	pursue	support	attempt	behold	penetrate
4	shall	yet	said	would	die	feeling	friend	may	must	love
5	man	said	old	elizabeth	friend	father	dear	god	could	hand
6	man	old	shall	yet	would	die	one	could	cottage	young
7	shall	god	die	great	night	passed	day	eye	miserable	soon
8	god	great	good	friend	feeling	life	yet	even	first	thus
9	one	yet	dear	friend	passed	may	child	hope	love	miserable

Figure 15: Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft Shelley



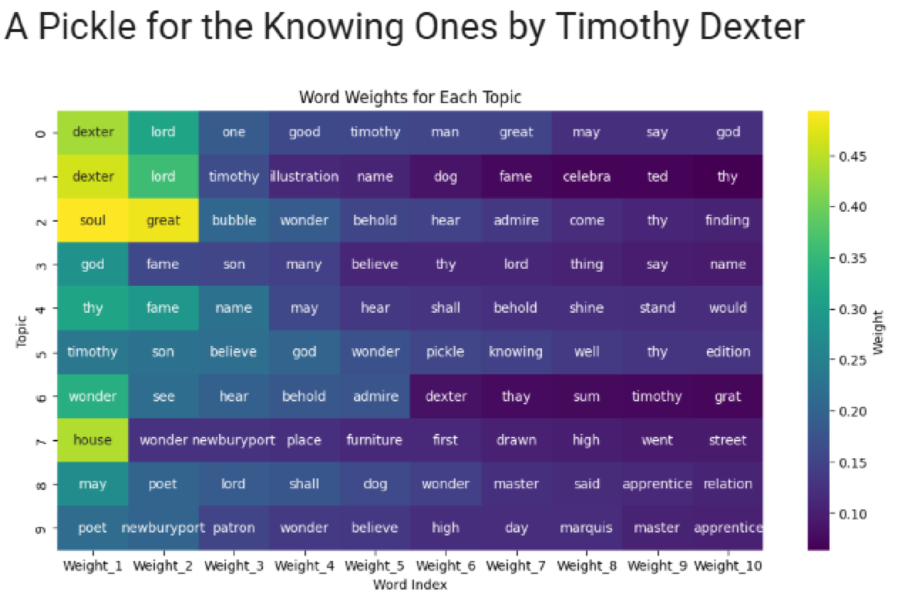
	Weight_1	Weight_2	Weight_3	Weight_4	Weight_5	Weight_6	Weight_7	Weight_8	Weight_9	Weight_10
0	time	traveller	said	machine	came	one	hand	upon	thing	little
1	time	traveller	said	machine	travelling	taken	damned	tell	asked	insensible
2	said	man	medical	psychologist	thing	filby	think	one	right	suddenly
3	could	see	thing	think	yet	even	find	enough	imagine	time
4	looked	came	thought	could	traveller	round	weena	face	towards	see
5	looked	weena	little	round	saw	gone	psychologist	upon	face	seemed
6	thing	one	traveller	saw	seemed	looked	time	come	mind	first
7	hand	one	could	looked	traveller	match	put	weena	see	took
8	thing	machine	weena	looked	gone	must	came	look	mind	know
9	thought	little	people	weena	one	might	sudden	first	way	match

Figure 16: The Time Machine by H. G. Wells



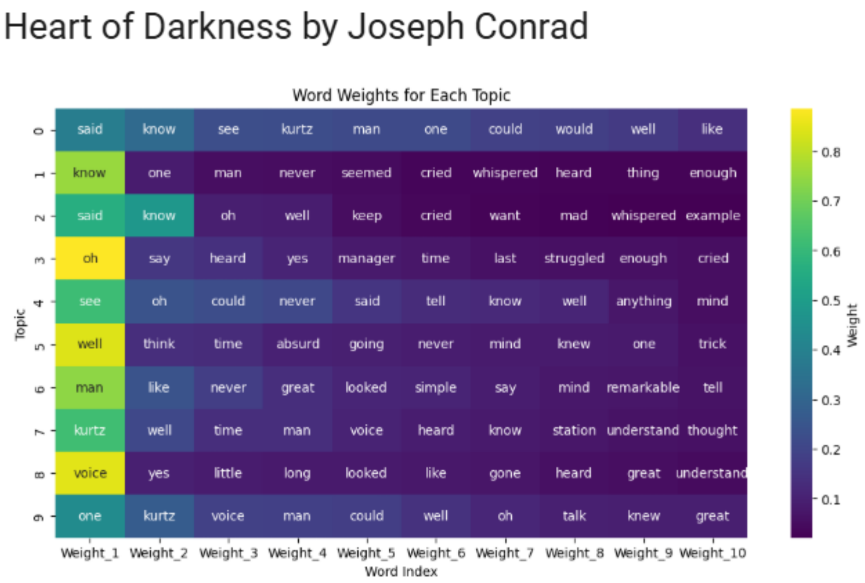
Topic	Weight_1	Weight_2	Weight_3	Weight_4	Weight_5	Weight_6	Weight_7	Weight_8	Weight_9	Weight_10
0	ch	wu	see	tzŭ	sun	han	uan	ao	say	ang
1	cf	men	besieged	town	enemy	yüeh	must	raiding	plundering	wu
2	ch	cf	see	han	te	lei	kuo	io	circa	hsien
3	see	chapter	ii	xiii	ix	vii	iii	xi	note	iv
4	tzŭ	sun	wu	chapter	see	chap	work	war	text	king
5	chapter	energy	xi	iv	ch	vi	iii	viii	attack	xii
6	note	ao	yu	kung	chang	tu	say	chinese	mu	uan
7	note	enemy	attack	wu	yüeh	make	one	ho	know	advantage
8	ground	ao	one	army	kung	would	xi	war	desperate	way
9	wu	ao	kung	ground	ho	lu	yüeh	chap	say	king

Figure 17: The Art of War by active 6th century B.C. Sunzi



Topic	Weight_1	Weight_2	Weight_3	Weight_4	Weight_5	Weight_6	Weight_7	Weight_8	Weight_9	Weight_10
0	dexter	lord	one	good	timothy	man	great	may	say	god
1	dexter	lord	timothy	illustration	name	dog	fame	celebra	ted	thy
2	soul	great	bubble	wonder	behold	hear	admire	come	thy	finding
3	god	fame	son	many	believe	thy	lord	thing	say	name
4	thy	fame	name	may	hear	shall	behold	shine	stand	would
5	timothy	son	believe	god	wonder	pickle	knowing	well	thy	edition
6	wonder	see	hear	behold	admire	dexter	thay	sum	timothy	grat
7	house	wonder	newburyport	place	furniture	first	drawn	high	went	street
8	may	poet	lord	shall	dog	wonder	master	said	apprentice	relation
9	poet	newburyport	patron	wonder	believe	high	day	marquis	master	apprentice

Figure 18: A Pickle for the Knowing Ones by Timothy Dexter



Topic	Weight_1	Weight_2	Weight_3	Weight_4	Weight_5	Weight_6	Weight_7	Weight_8	Weight_9	Weight_10
0	said	know	see	kurtz	man	one	could	would	well	like
1	know	one	man	never	seemed	cried	whispered	heard	thing	enough
2	said	know	oh	well	keep	cried	want	mad	whispered	example
3	oh	say	heard	yes	manager	time	last	struggled	enough	cried
4	see	oh	could	never	said	tell	know	well	anything	mind
5	well	think	time	absurd	going	never	mind	knew	one	trick
6	man	like	never	great	looked	simple	say	mind	remarkable	tell
7	kurtz	well	time	man	voice	heard	know	station	understand	thought
8	voice	yes	little	long	looked	like	gone	heard	great	understand
9	one	kurtz	voice	man	could	well	oh	talk	knew	great

Figure 19: Heart of Darkness by Joseph Conrad

7 Results

The application of Latent Semantic Indexing (LSI) to the dataset of 10 books effectively reduced the dimensionality of the term-document matrix, uncovering hidden semantic relationships and grouping synonyms, related terms, and contextual associations under latent semantic dimensions. This process enabled the analysis to capture meaningful semantic patterns in diverse genres. Books with similar genres demonstrated strong semantic similarities, as highlighted by the cosine similarity matrices before and after preprocessing. For example, "Alice's Adventures in Wonderland" and "A Christmas Carol" exhibited thematic connections rooted in fantastical narratives, while "Frankenstein" and "The Time Machine" displayed closer proximity in the semantic space, underscoring shared science fiction themes. The preprocessing steps, including tokenization, stop-word removal, and lemmatization, played a crucial role in improving the coherence of latent concepts, as evidenced by the clarity of the word clouds generated for each book. Additionally, topic-specific queries, such as "Politics and War" and "Comedy and Drama," revealed the distribution of these themes across the dataset. Notably, "The Art of War" was strongly aligned with "Politics and War," while "Life on the Mississippi" and "Pygmalion" exhibited significant associations with "Comedy and Drama," further demonstrating the effectiveness of LSI in thematic exploration.

8 Conclusion

In conclusion, this study has successfully unearthed latent semantic structures, providing a nuanced understanding of the thematic relationships that transcend genre boundaries across the 10 analyzed books. The application of Latent Semantic Indexing (LSI) has proven to be a robust methodology for dimensionality reduction, effectively preserving the intricate semantic relationships that underlie the textual data. Ultimately, this research demonstrates the potential of LSI to reveal novel insights into the thematic landscape of literary corpora, paving the way for further exploration and discovery.

9 Limitations and Future work

A key limitation of this study is the relatively small dataset, comprising only 10 books, which may not be entirely representative of the broader literary landscape. Consequently, the findings may not be fully generalizable across diverse genres, authors, and literary periods. Future research could seek to address this limitation by substantially expanding the dataset to include a more comprehensive and heterogeneous collection of literary works. Additionally, incorporating more advanced natural language processing techniques, such as topic modeling or deep learning-based methods, could further enhance the analysis and provide more nuanced insights into the thematic structures of literature.
reference

10 Appendix

10.1 Additional Visualizations

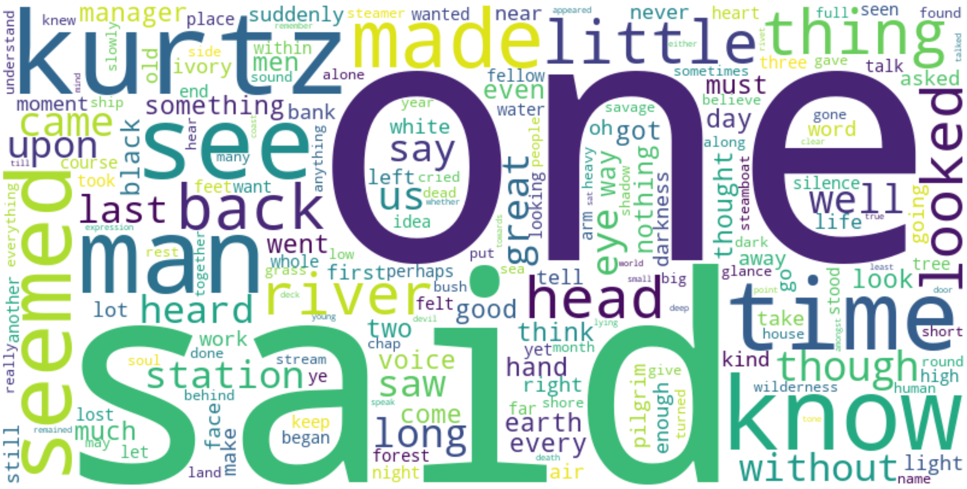


Figure 20: Word Cloud for the book 'Heart of Darkness'

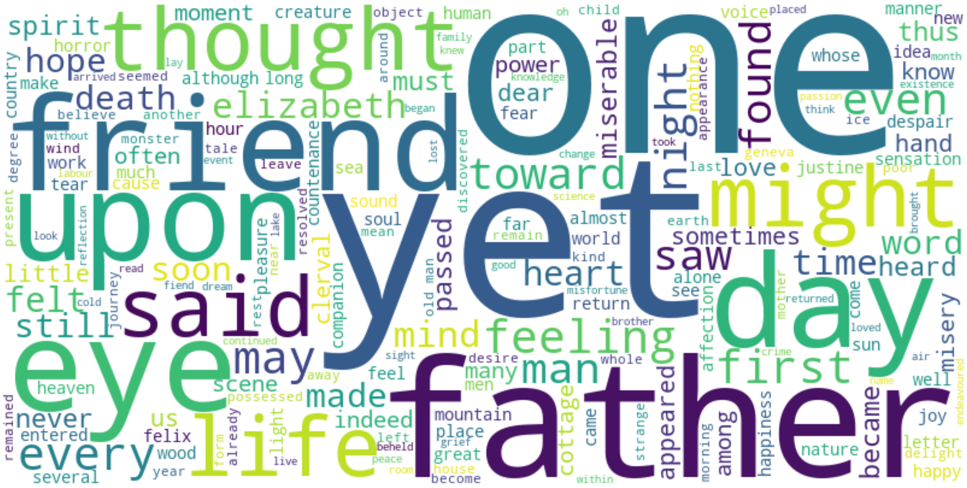


Figure 21: Word Cloud for the book 'Frankenstein'

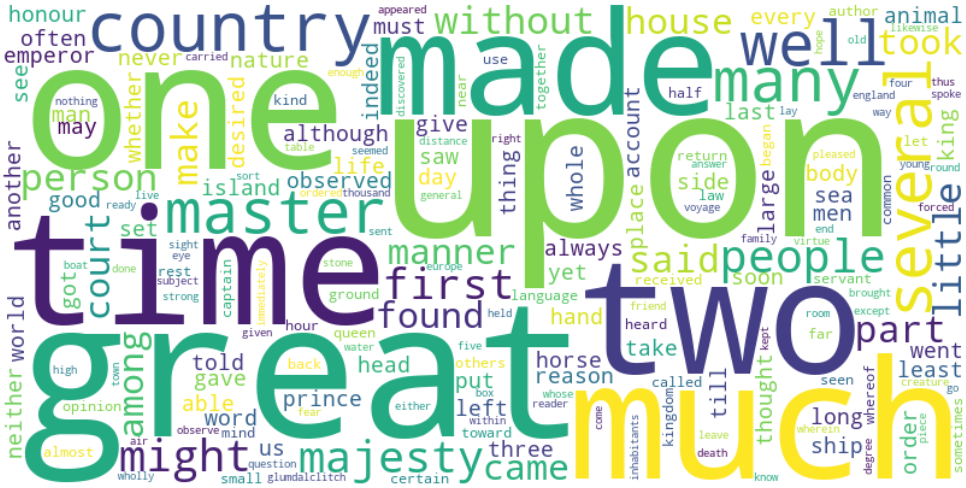


Figure 22: Word Cloud for the book 'The Time Machine'

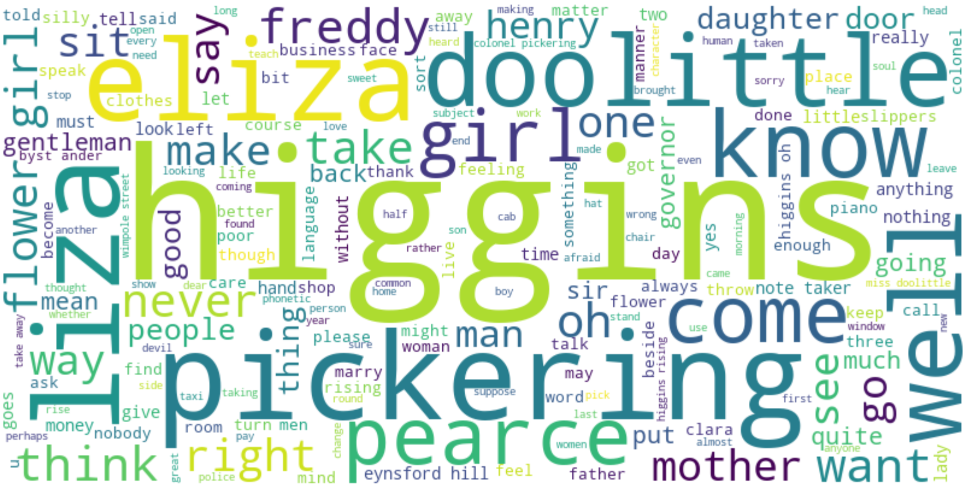


Figure 23: Word Cloud for the book 'Pygmalion'

10.2 Sample Code

```
def preprocess_text(text):
    """
    Tokenizes, removes stopwords, and lemmatizes the text.
    """
    tokens = word_tokenize(text.lower())
    tokens = [lemmatizer.lemmatize(token) for token in tokens if
       token.isalpha() and token not in stop_words]
    return tokens

def read_pdf_and_extract_text(file_path):
    """
    Reads text from a PDF file and returns it as a string.
    """
    reader = PdfReader(file_path)
    text = ""
    for page in reader.pages:
        text += page.extract_text() + " "
    return text.strip()

```

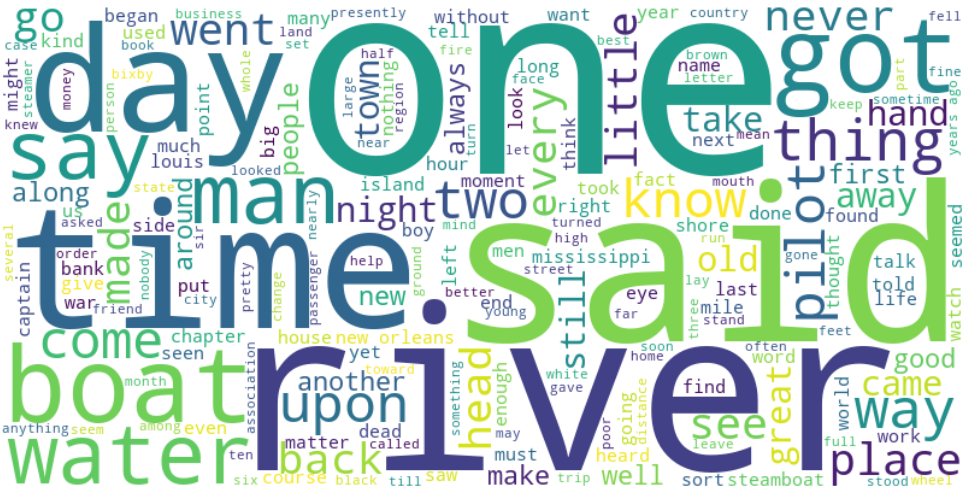


Figure 24: Word Cloud for the book 'life on the mississippi'

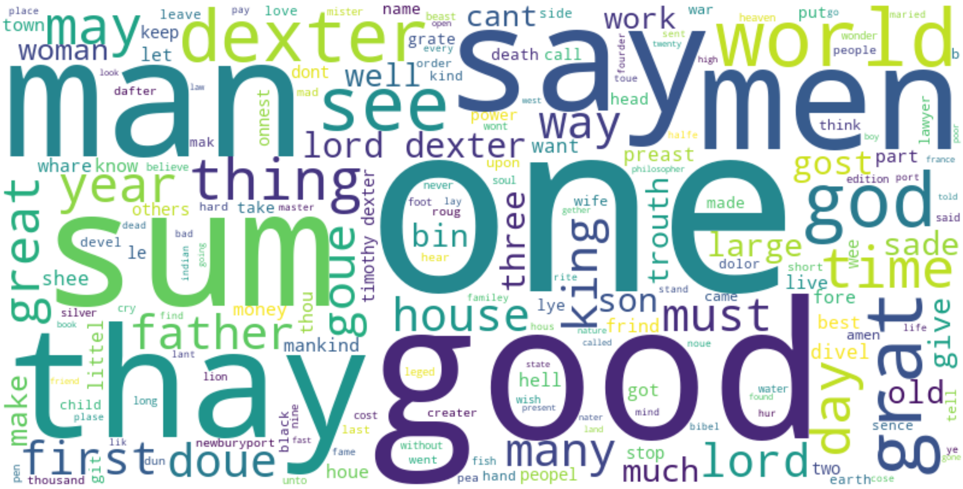


Figure 25: Word Cloud for the book 'A Pickle for the Knowing Ones'

```python
def preprocess_pdf_text(file_path):
    """
    Reads and preprocesses text from a PDF.
    """
    raw_text = read_pdf_and_extract_text(file_path)
    sentences = sent_tokenize(raw_text)
    preprocessed_sentences = [' '.join(preprocess_text(sentence))
        for sentence in sentences]
    return preprocessed_sentences
```

11 Biography

- **Ainsley Braunscheidel** is a Statistics and Data Science undergraduate student at University of Colorado, Boulder. She has experience in coding languages such as R and Python, as well as various techniques for data analysis.
- **Divya Nallawar** is a Data Science graduate student at the University of Colorado, Boulder. With a strong background in software development and data analysis, she has worked on projects involving machine learning, predictive modeling, and data-driven insights. Her recent research focuses on network traffic analysis and time series forecasting, where she explores innovative approaches. She is passionate about applying data science to solve real-world problems and continuously expanding her knowledge in the field.
- **Luke Petet** is a Data Science undergraduate student at the University of Colorado, Boulder. He has experience in data analysis with an emphasis in the application of data analysis on audio engineering.
- **Reza Naiman** was born in Kabul, Afghanistan, and moved to Englewood, Colorado, in 2016. After graduating from Englewood High School, he continued his education at the University of Colorado Boulder. He is a senior in the Statistics and Data Science program with a minor in Computer Science. He has completed two internships with Frontier Technologies Defense as a data analyst and software developer. His interest is to continue working in the defense industry or government agencies by applying the skills he learned from his education at the University of Colorado Boulder.