Mariapaola Ambrosone

Mbf2fz

DS 5001 Final Project

Exploring Literary Landscapes: A Text Analytics Analysis of Novels by Woolf, Shelley, and Eliot

For my final project, I will be analyzing three classic novels written by three notable female authors: Virginia Woolf, Mary Shelley and George Eliot.

Virginia Woolf was a prominent modernist writer known for her innovative narrative techniques and exploration of characters' inner thoughts and emotions. The novel chosen for this analysis is “The Voyage Out”.

Mary Shelley’s writing often explores themes of creation, science, and the consequences of pushing ethical boundaries. The novel chosen for this analysis is “The Last Men”.

George Eliot delves into deep psychological insights and complex character development. Her works and are known for their exploration of moral and social issues, as well as their vivid portrayal of rural life. The novel chosen is "Middlemarch".

The general idea is to analyze the frequency and co-occurrence of words and phrases to reveal the prominent themes and topics addressed in each novel. Sentiment analysis can provide insights into the emotional tone of the novels. Are they generally positive, negative, or neutral?

Data acquisition and cleaning

After importing the necessary packages for my analysis, I imported the raw text files for the corpus. The three novels were obtained from the Project Gutenberg online repository. Project Gutenberg e-books contain both a header and a footer, which were removed. The text files were manipulated in the following way:

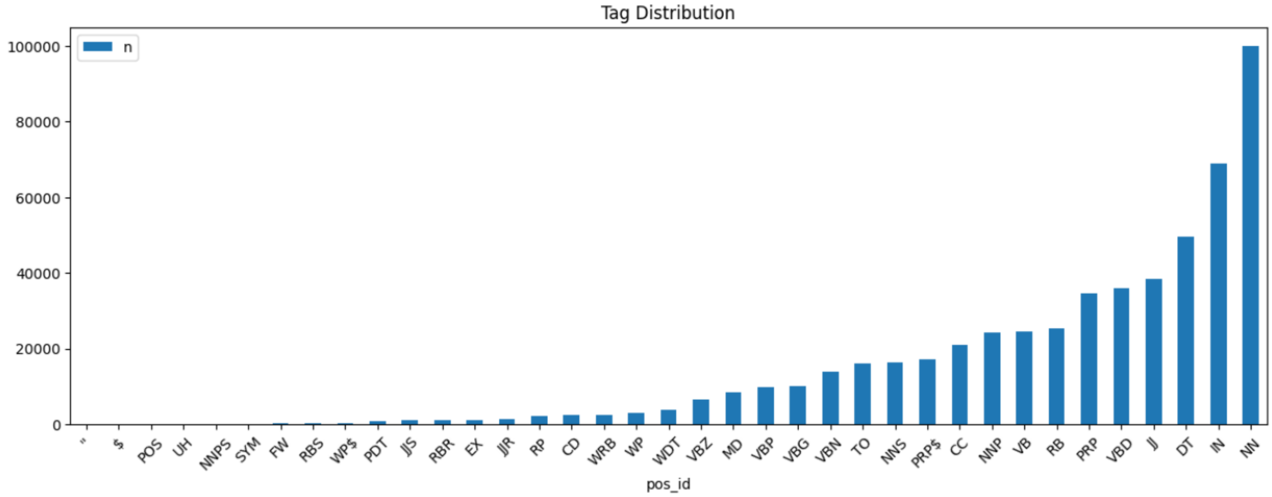
- DOC table was generated by using the acquire_epubs function, regex expressions to identify chapters. The code processes a list of e-books, chunks them by chapters, and splits them into paragraphs;
- LIB table was generated by extracting the author, book title, and file directory from the original text files;
- TOKEN table was generated using a function that turns each paragraph into a sentence, and then into tokens. The data was tokenized and annotated using NLTK. The index was set to OHCO. Each token fits under a book_id, chapter number, paragraph number, sentence number, and token number;
- VOCAB table was generated from the TOKEN table. It contains a bag of unique words across the set of novels. Each word is encoded as a row, and fields for each row includes number of times used (‘n’), and frequency of times used (‘p’). I created a Boolean value to identify the stopwords (‘stop’) in the corpus and then calculated the Stem Porter, Snowball Stemmer and Lancaster Stemmer by using NLTK. I finally obtained the POS Max for each row;
- CORPUS a list of tokens, organized by OHCO. Each token has an assigned part of speech. This table was parsed by NLTK;
- LOADINGS a list of terms that maps words and their relationship to principal components. Principal components are the components calculated to have the greatest variance for the text;

- EMBEDDINGS, the VOCAB table of terms, with associated emotional embeddings for anger, anticipation, disgust, fear, joy, sadness, surprise and trust.

POS tag distribution

In this section of the analysis, I analyzed the frequency and distribution of different parts of the speech. Part of speech tagging consists in assigning a specific grammatical category, such as nouns, verbs, adjectives, verbs, to each word in the text.

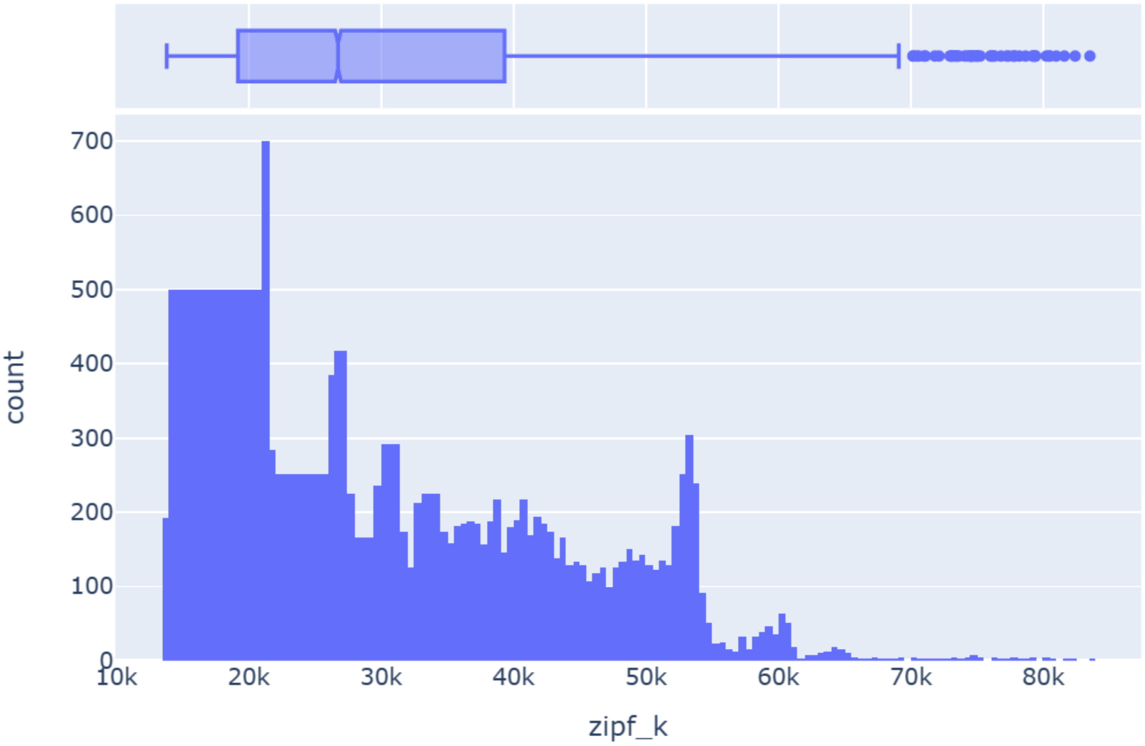
The distribution of POS counts in the corpus is shown below. The x axis represents the different POS tags and the y-axis represents their frequencies. The top 5 most frequent parts of speech in the corpus are nouns, prepositions, determiners, adjectives, and past tense verbs.



The Zipf’s score was calculated and added to the VOCAB table. In the context of Zipf's law, the statement "the frequency of an item is inversely proportional to its rank" implies that the more frequently an item (such as a word) is in a dataset, the lower its rank will be. In other words, the most frequent items are assigned lower ranks, while less frequent items are assigned higher ranks.

From an interpretive perspective:

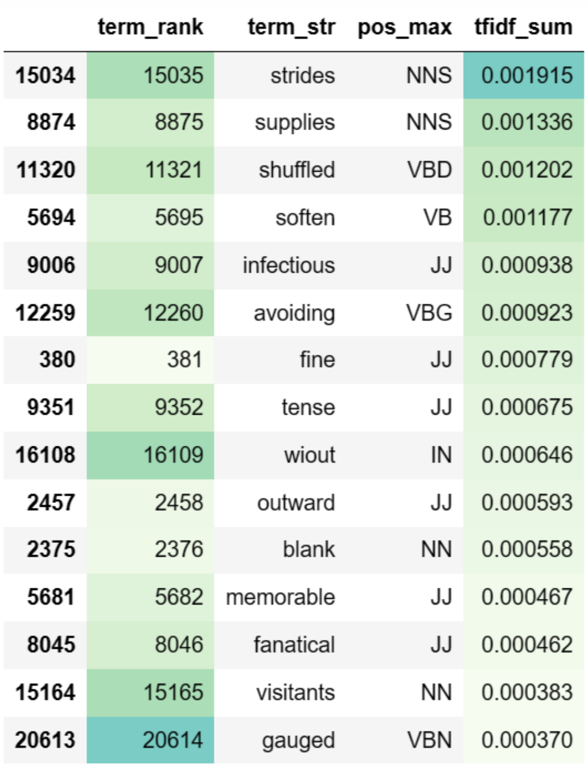
- The most frequent words, often referred to as "stop words," tend to be less informative because they occur frequently in many contexts and don't contribute much to the uniqueness or understanding of a text.
- Less frequent words, also known as "content words," are typically more meaningful and carry more specific information about the content of the text.



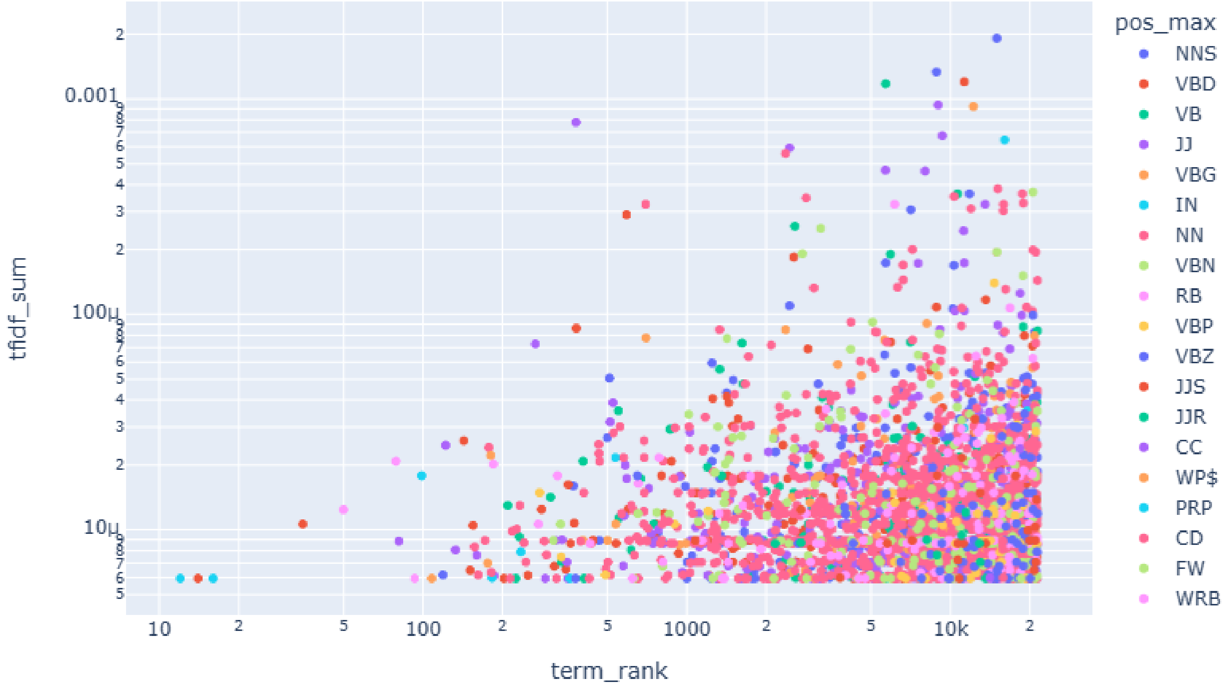
Generating the Term Frequency Inverse Document Frequency (TFIDF) matrix

To generate the TFIDF I used a function showed during the semester, tfidf_matrix, and used the results of the TFIDF analysis to enrich the VOCAB table adding the columns 'df' representing the document frequency for each term, 'idf' representing the inverse document frequency for each term, and the 'tfidf_sum' representing the sum of TFIDF values for each terms across all documents.

In the table below, I visualized the top 15 terms. The terms with higher significance in the text corpus are the one with the highest TFIDF score. As we can see from the table below, these terms are mostly adjectives and nouns. The prevalence of adjectives could indicate that all three authors might utilize rich descriptive language to create vivid and detailed imagery in their writing.



	term_rank	term_str	pos_max	tfidf_sum
15034	15035	strides	NNS	0.001915
8874	8875	supplies	NNS	0.001336
11320	11321	shuffled	VBD	0.001202
5694	5695	soften	VB	0.001177
9006	9007	infectious	JJ	0.000938
12259	12260	avoiding	VBG	0.000923
380	381	fine	JJ	0.000779
9351	9352	tense	JJ	0.000675
16108	16109	wiout	IN	0.000646
2457	2458	outward	JJ	0.000593
2375	2376	blank	NN	0.000558
5681	5682	memorable	JJ	0.000467
8045	8046	fanatical	JJ	0.000462
15164	15165	visitants	NN	0.000383
20613	20614	gauged	VBN	0.000370



The above scatterplot is another visualization of the TFIDF.

For the analysis of the text, I used Principal Components Analysis (PCA), Topic Modeling, Word Embedding and Sentiment Analysis. These different methods were used to extract meaningful insights and patterns from the data.

Principal Component Analysis (PCA)

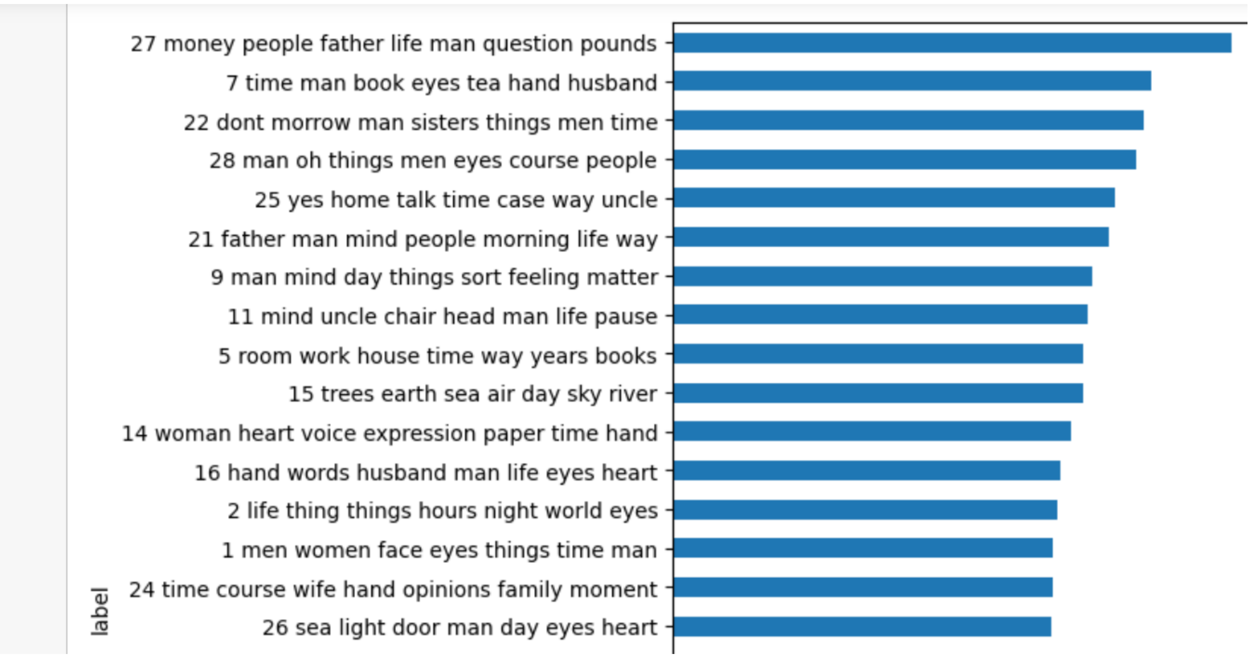
Principal component analysis was used with each novel to see if there would be any similarities. I obtained the following loadings for PC0, PC1 and PC2.

```
Books PC0+ strides supplies shuffled soften infectious avoiding wiout outward blank memorable
Books PC0- fine tense visitants hev suck amazement trembles voyager beneficently milk
Books PC1+ shuffled soften wiout outward blank fanatical gauged satisfactions navigation idols
Books PC1- strides supplies infectious mournings landholders lengthy anachronism law finely amanuensi
s
Books PC2+ batti cared forsaking retreats stimulant preyed womens undeceived nervous tick
Books PC2- epigrams attend emigration paws orthodox plumage concert lurched fragments organ
```

It is interesting to note that some of the words like shuffled, soften, blank are positive for PC0 and PC1 but also negative for PC0. If the same word has a positive loading for one principal component (PC0) and a negative loading for another principal component (PC1), it could indicate that the word contributes differently to these two principal components in terms of the patterns it captures in the data.

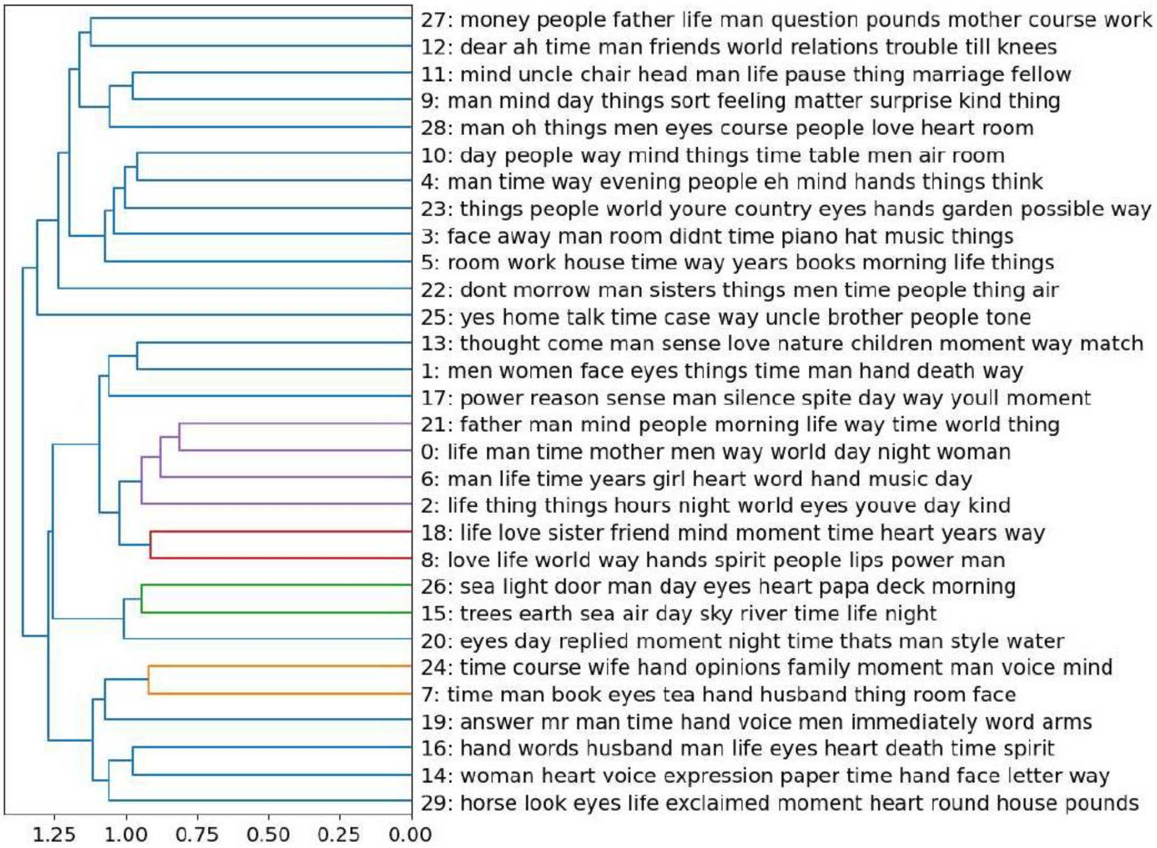
LDA

The main goal of this analysis is to understand the similarities between the three writers, Latent Dirichlet Allocation was used to derive 5 topics from the texts and determine which words were most relevant to each topic.



We notice some recurrent words throughout the analysis, words like father, man, husband, sisters, home, house, woman, heart all seems to convey towards subjects related to domestic life, familial relationships and emotions. It seems that importance is given to feelings and emotions in shaping the everyday life of the characters in the novels. The presence of these recurring keywords may indicate that the three authors, Woolf, Shelley, and Eliot, were all exploring these themes within their works.

The dendrogram was generated as a means of evaluating how closely associated the novels were with each other. Figure below.



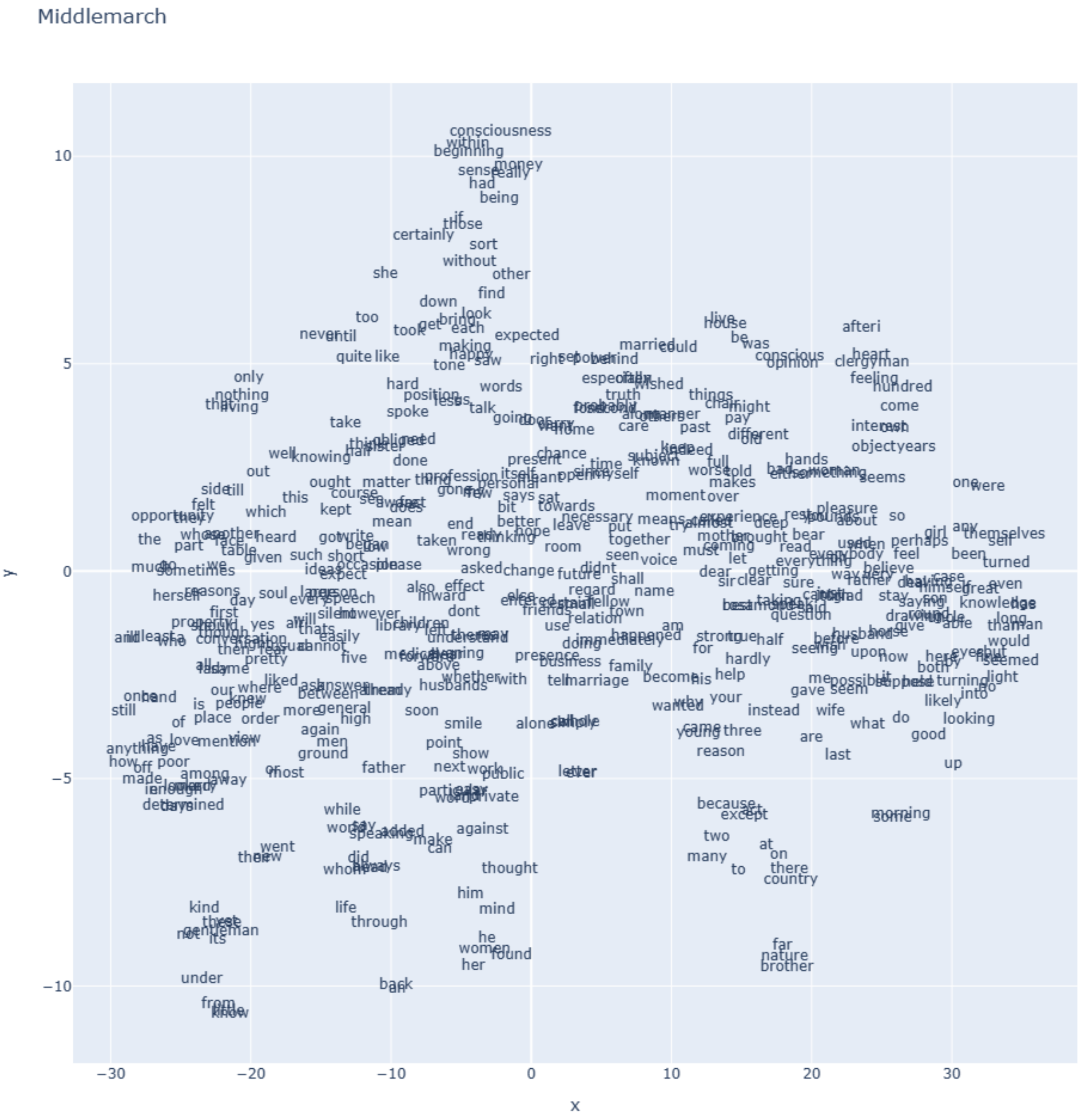
The words within each cluster seem to share semantic similarities or contextual relationships.

For example, for topics 21, 0, 6, 2 (purple color in the dendrogram) the clustering of words like "father," "man," "mother," "woman," "girl," and "heart" might suggest a group related to gender and family relationships. This could indicate discussions about familial roles and emotions. At the same time the presence of words like “night”, “day” and “hours” could indicate discussions about different times of the day and the passage of time. Words like “music”, “heart”, “hand” and “eyes” could refer to sensory aspects of life and emotional expression.

Word Embeddings

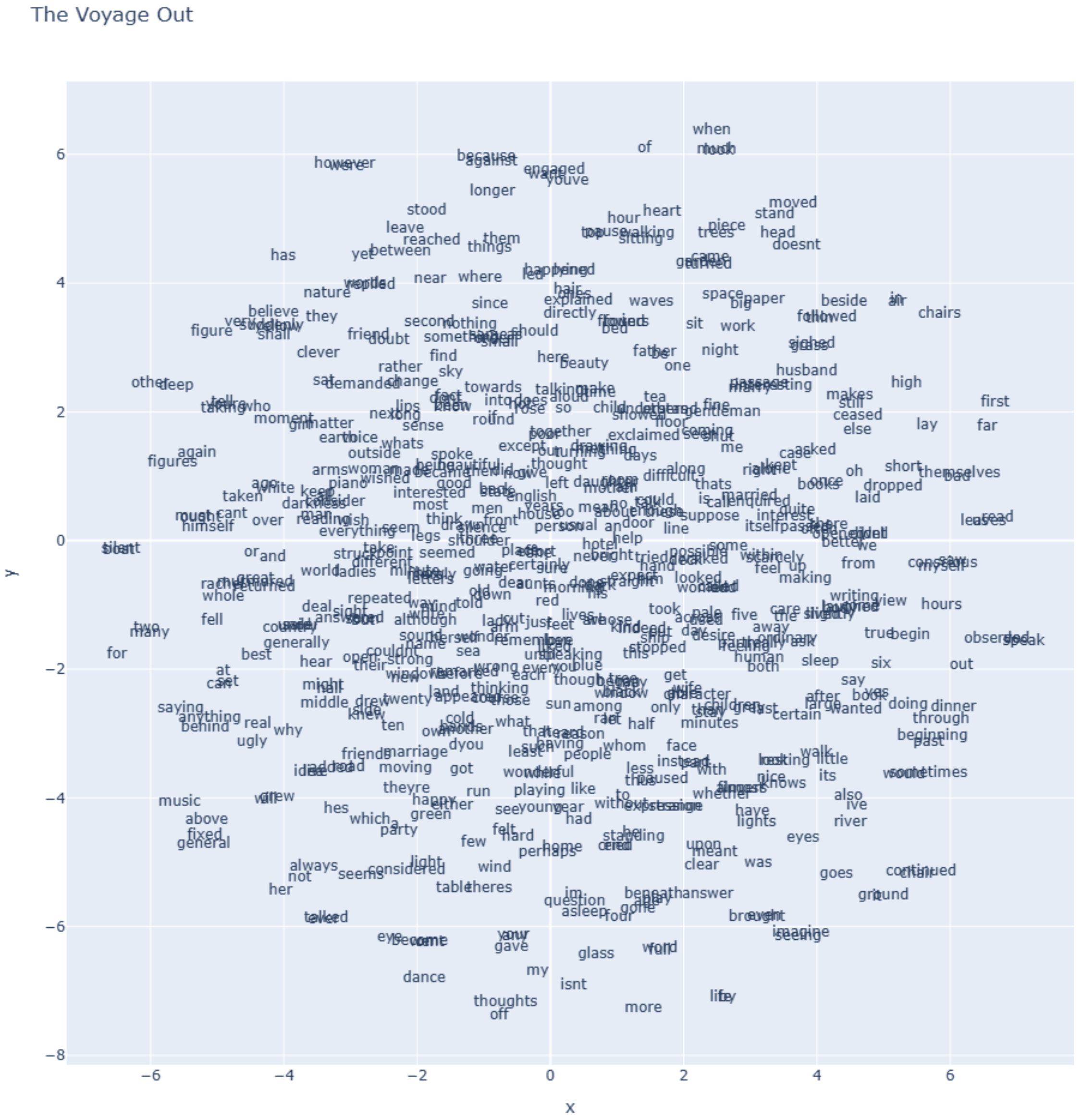
Middlemarch – George Eliot

In the top part of the word embedding plot, we see a relatively sparse cluster of words - “consciousness”, “within’, “beginning”, “money”, “sense”, “being”, “making”, “happy”, “married”, “live”, “house”, “power” – could suggest a theme related to personal introspection, self-awareness and individual agency within the context of money, relationships, and happiness. At the same time, words like "money," "making," and "power" might point towards financial and social dynamics. This could relate to characters' ambitions, their pursuit of success, or the influence of money and power on their lives.



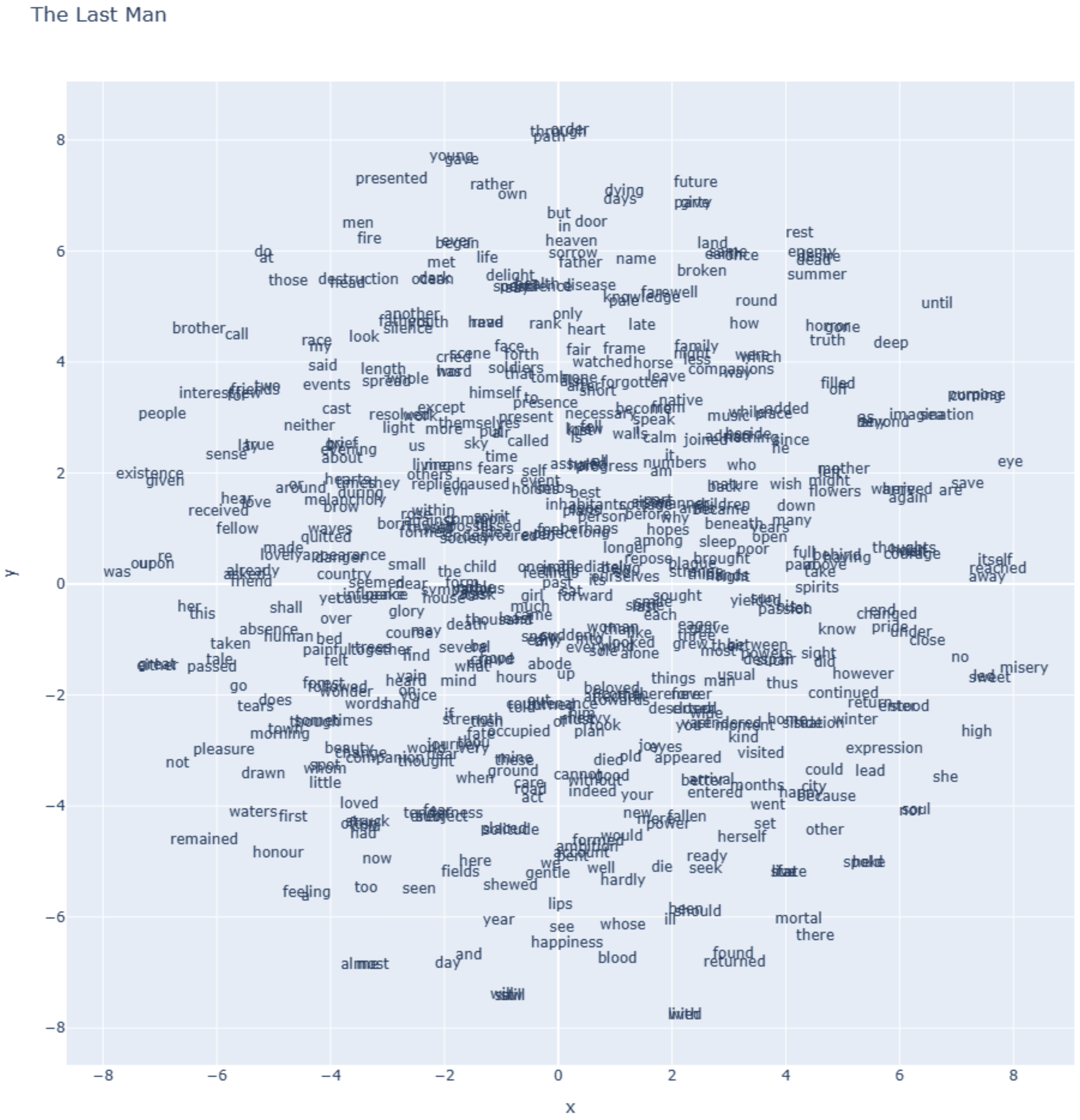
The Voyage Out – Virginia Woolf

In the word embedding plot for the Voyage Out, we find words such as "person," "year," "lives," "love," "marriage," and "friend" point to social interactions and connections. These terms could imply exploration of human relationships, romantic interests, friendships, and the passage of time. At the same time the word "water", “sea”, “land”, and “sun” could indicate a connection to nature and the environment or some sort of voyage. It might suggest scenes involving natural landscapes as part of the life of the characters. This could make sense considering the novel follows the journey of Rachel Vinrace, on a sea voyage from London to South America.



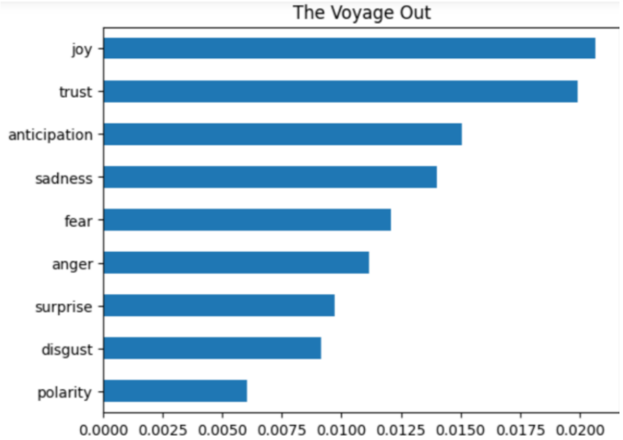
The Last Man – Mary Shelley

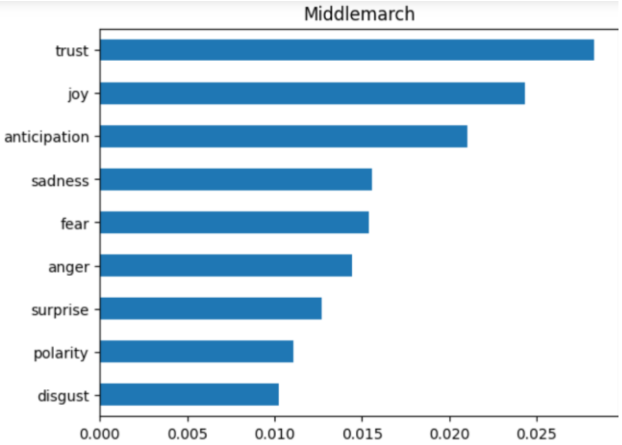
The word embedding scatterplot is highlighting words that convey themes of sorrow, loss, and melancholy. Words such “dying”, “heaven”, “sorrow”, “forgotten”, “silence”, “grief”, are strongly associated with emotions of sadness, loss, and sorrow. They suggest that the novel might explore themes of tragedy, mourning, and the emotional impact of difficult circumstances. We also find the words “companions”, “soldiers”, “horse”, “brother”, and “father” which could indicate the presence of a male character being in the army and a possible war or conflict. In addition, words such as “watched” and “forgotten” evoke a sense of introspection and solitude.



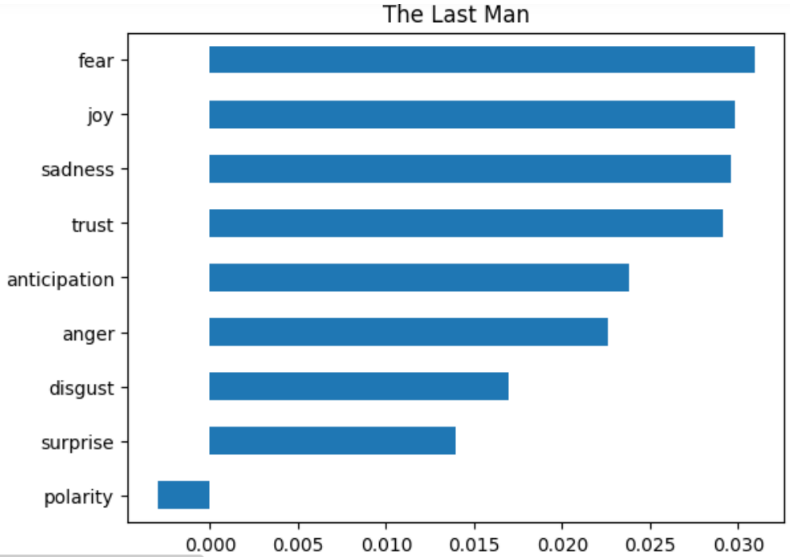
Sentiment analysis

The results of the sentiment analysis were interesting because they reflected the findings from the word embedding and the sharp distinction between the themes in the novels from Woolf and Eliot and the novel by Shelley. This distinction can be inferred from the bar plots below. Middlemarch and The Voyage Out are characterized by positive sentiments such joy, trust and anticipation.





On the other hand, in the Last Man we mostly find negative sentiment categories.


The Last Man
fear
joy
sadness
trust
anticipation
anger
disgust
surprise
polarity