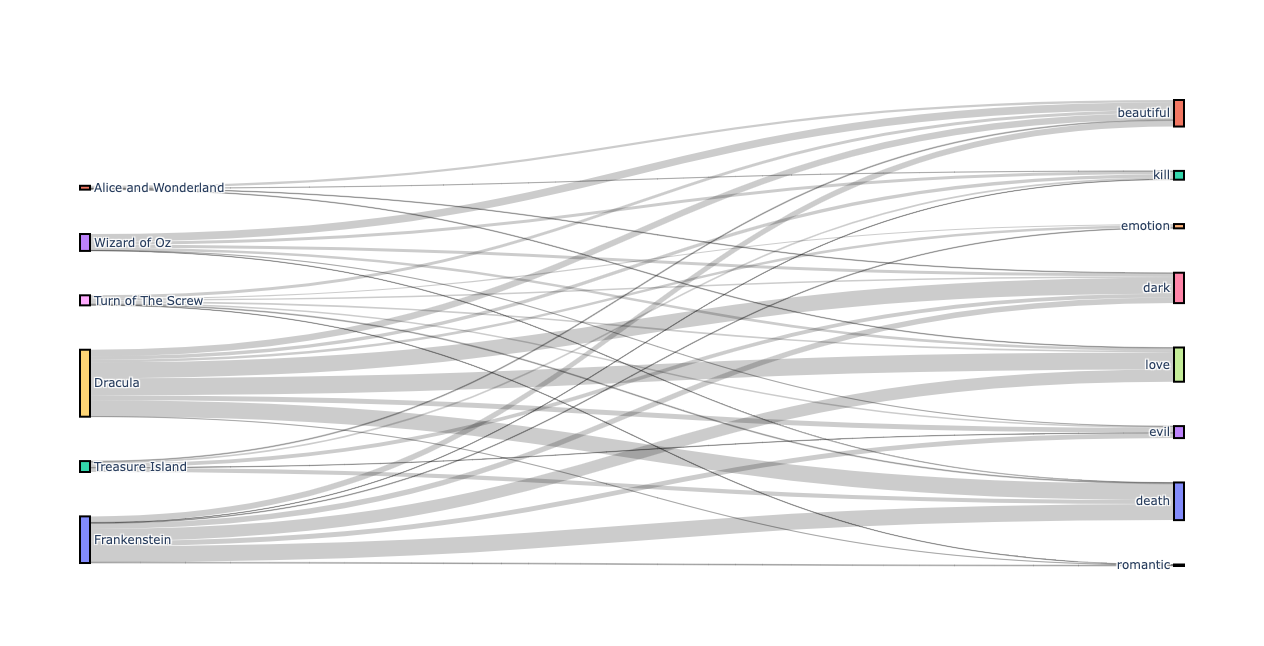
Elias Karikas, Billy Wallace, Malick Tobe, Filip Tomovski

DS 3500 - Homework 2 Report

GitHub Repository : <https://github.ccs.neu.edu/eliaskarikas/Movie-Critic-NLP>

The text files we decided to use are six novels. Within the six novels are three that are related to each other as they belong to the same or similar genres, those being horror or fantasy/adventure. The novels are: Dracula by Bram Stoker, Frankenstein by Mary Wollostonecraft, The Turn of the Screw by Henry James, Alice and Wonderland by Lewis Carroll, Treasure Island by Robert Louis Stevenson, and The Wonderful Wizard of Oz by L. Frank Braum. Although half of these novels belong to a different genre, we believed that the genres and novels have similar overarching themes, messages, and possibly sentiments, and hoped to discover some similarities or differences through text analysis. These books/text files can be found on the project gutenberg website: <https://www.gutenberg.org>.

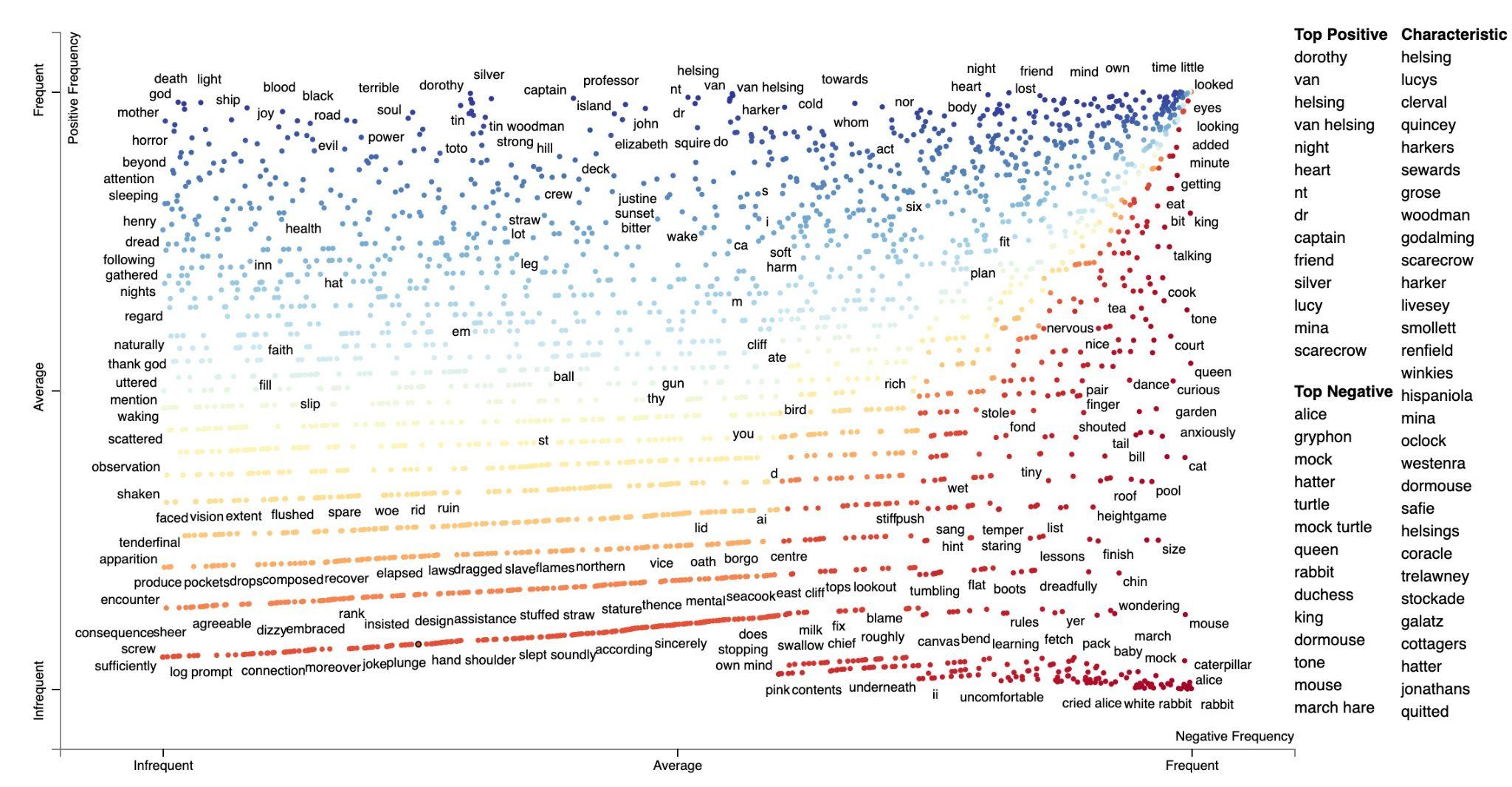
To get initial data on word occurrences we created a sankey diagram, comparing each of the text's most common occurrences. For our first Sankey Diagram we took the top 3 occurrences of words in each text, and there was not many major takeaways (appendix A). These in most cases included main characters' names, or fairly insignificant words. Nevertheless there was still an interesting overlap of words; such as the use of silver presumably in radically different contexts (Dracula, Frankenstein, Treasure Island). Because this did not give us much to interpret we decided to define our own word list to find particular word occurrences.

When defining our own wordlist our Sankey diagram was more interesting. We picked four positive words (love, beautiful, emotion, romantic) and four negative words (kill, dark, evil, death). Occurrences of positive words (beautiful, love, emotion, romantic) were pretty common in Horror novels (in Dracula the usage of the word love outweighs death by four occurrences), and the use of emotion, and romantic was only in horror novels when comparing fantasy and horror. Negative words however were used in similar volumes when comparing either genre, with the exception of ‘evil’, and ‘death’, which were more frequently used in horror. However, when summing word totals our adventure/fantasy novels used the word ‘kill’ 23: 18, with the Wizard of Oz using the word a total of 14 times, 2 less than Dracula’s 16, and 12 more than Frankenstein’s 2.

Although Sankey diagrams aggregated word totals, there wasn’t a large focus on individual texts word usage, so we decided to create WordClouds for each novel.



We can see some visible overlap between many of the texts. Many of the texts use words such as time, eyes, looked, found, little, and others. But our texts also have many of their own particular words that tie it to the novel. Particular character names or details relevant to the novel such as ‘captain’ or ‘father’ differentiate each novel from one another. We also allowed users the option to include their own folder of jpeg files to be used as masks to better personalize Word Clouds.

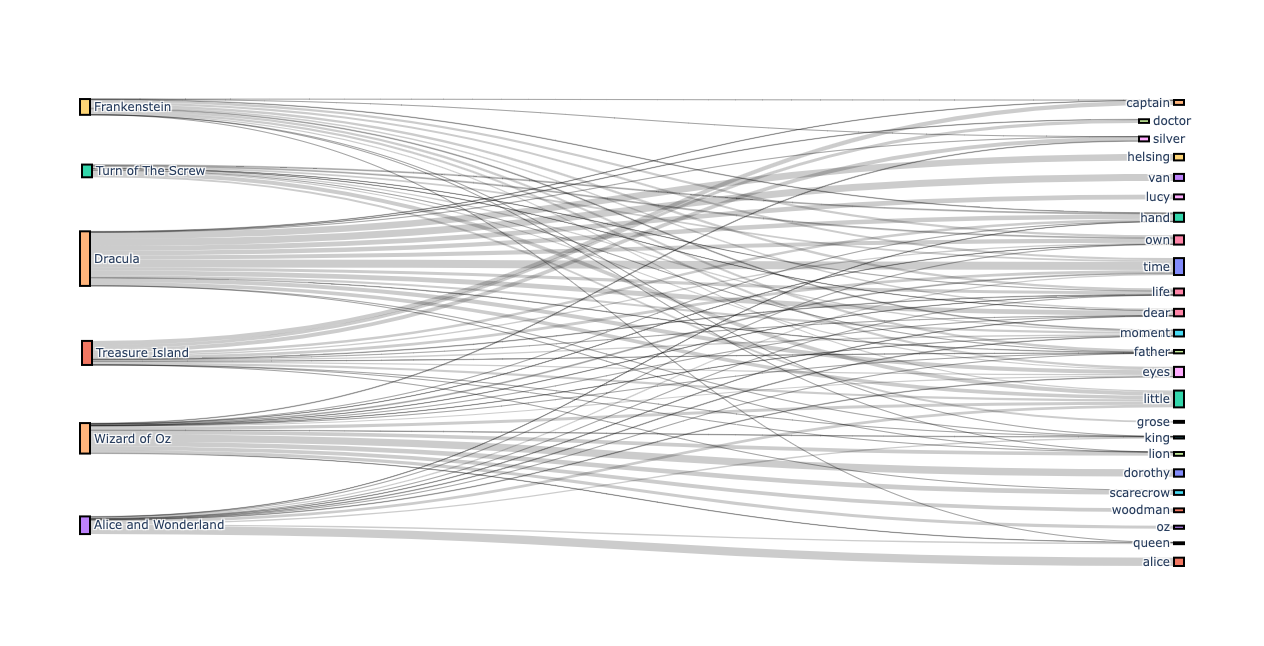
Finally, we created a sentiment analysis function to understand the overall sentiment of each novel (Appendix B). Overall sentiments were not statistically significant (not exceeding ~.07 positive and -.04 negative). However, the results were still interesting to gain some understanding of the general tone. Alice and Wonderland was the only text that had an overall negative sentiment of -.04, Dracula and Treasure Island had more of a neutral tone (.0096, .0030 respectively), and Turn of the Screw, Frankenstein, and Wizard of Oz had more of a positive tone (.034, .057, .066 respectively). Now that we had our novels categorized by general sentiment we wanted to get more of an understanding of the frequency of word use. While we had other analyses of word use, there was no analysis of word use relevant to text size. So we decided to categorize our novels by sentiment and use the scattertext library to create an interactive scatterplot to see common and uncommon word occurrences as a ratio of words used in each text.

Within our graph we categorized by positive and negative sentiment (Alice and Wonderland receiving its own bucket), in the middle diagonal of the graph is general overlap between all texts, but by varying frequencies, the top left being frequent words just for positive texts, and the bottom right being frequent words for negative texts.

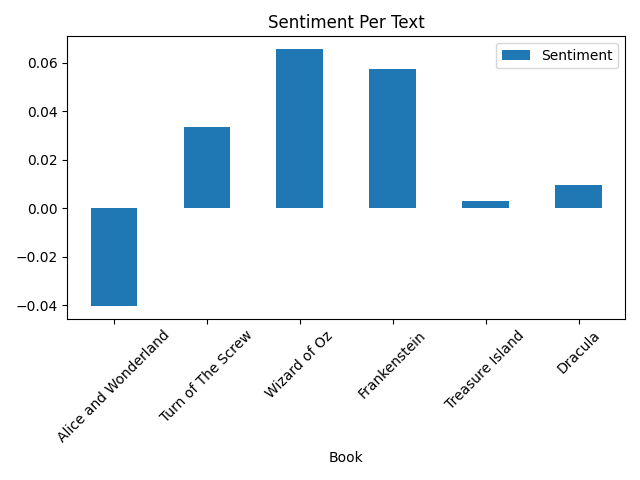
There is a lot of interesting information to be gained through an interactive text analysis but we feel that within our time restraints it was difficult to gain a deep understanding of the question we wanted to answer. Although we did not have any overwhelming outcomes in our study as we had hoped, there is great room for improvement and we believe that there is great potential in text analysis to find overarching themes, sentiment, and overlap between genres.

Additional Visualization Appendix:

A.



B.



C.

