NLP ANALYSIS OF POPULAR NOVELS (PROJECT GUTENBERG)

Authors: Justin Guthrie, Ryan Huang, Julia Luo

Author Affiliations:

Justin Guthrie:

Assisted in parsing, gathering text files and the Sankey visualization. Contributor to analysis paper and powerpoint poster.

Ryan Huang:

Lead contributor for parsing, library creation, and visualizations of text files.

Julia Luo:

Lead contributor to analysis paper and powerpoint poster.

Abstract:

In this project, we built a Natural Language Processing library able to perform comparative text analysis on a variety of datasets. We demonstrated our reusable framework's usability by comparing 10 text files from Project Gutenberg, a library that contains over 60,000 eBooks.

To do so, we implemented a preprocessor to clean the simple unstructured text files by removing unnecessary whitespace, punctuation and capitalization. In the meantime, the library also stores details such as a word counter and a list of all words. We also implemented a default parser that reads .txt files from Project Gutenberg and a generic parser for other types of files.

The library also supports three visualizations: a Text-To-Word Sankey diagram, which displays the word count of the most common words in each text; a polarity versus subjectivity scatter plot for the texts; and a word cloud that visualizes the prevalence of words in each text.

The results of the visualizations convey interesting and valuable information about the texts. Both the Sankey and word cloud diagrams show that some of the most common words are function words such as “like”, “as”, “the”, and pronouns. The word cloud diagram, however, includes some irrelevant punctuation and letters. Finally, the polarity and subjectivity plot follow a clear trend that polarity increases (both negatively and positively) with subjectivity. While the comparative text analysis was not perfect, it was still extremely effective and efficient overall.

Introduction:

Natural Language Processing (NLP) is a subdivision of artificial intelligence referring to the ability to process human language in the form of text or voice data and “understand” it as humans would. To do so, computer programs combine statistics, machine learning models, and deep learning. In day-to-day life, NLP can be found in the form of smart voice assistants, speech-to-text dictation software, customer service chatbots, and more, all helping make simple tasks easier and more streamlined for humans.

One particular task of NLP is sentiment analysis, which classifies text into different subjective categories concerning overall emotions, most commonly as either positive or negative.

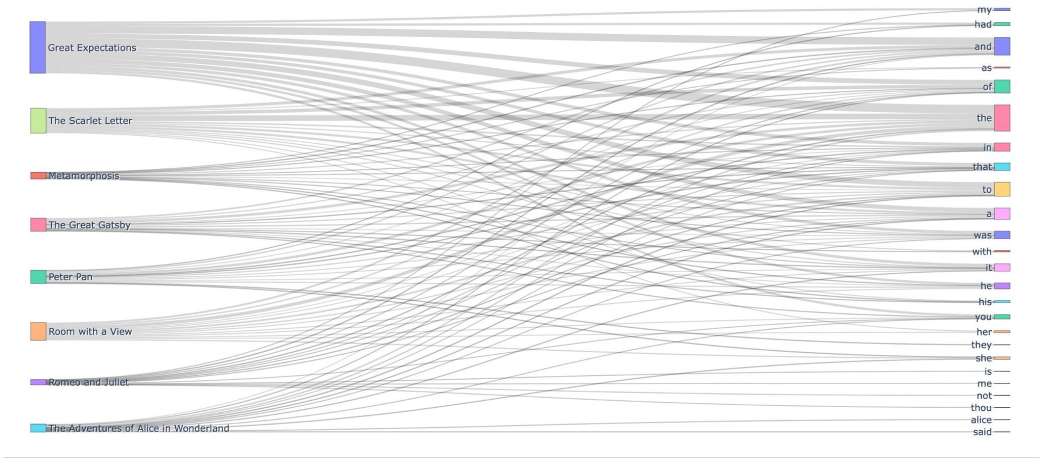
Sentiment analysis allows data processing at a huge scale and can be very helpful in social media and feedback from product reviews. Though the process is not always perfect, it is slowly being improved and built upon, allowing it to become more accurate.

Data Sources:

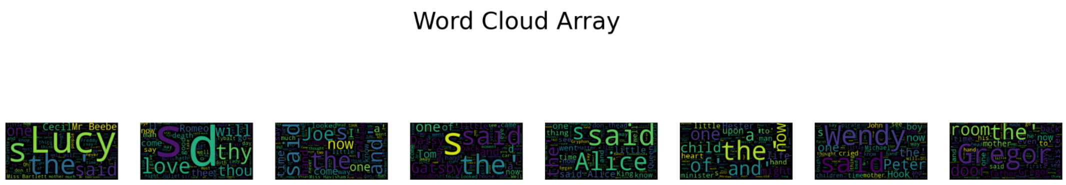
In this project, the following books were used as text files: “Great Expectations” by Charles Dickens, “Moby Dick; Or, The Whale” by Herman Melville, “Romeo and Juliet” by William Shakespeare, “A Room With A View” by E. M. Forster, “Alice’s Adventures in Wonderland” by Lewis Caroll, “Metamorphosis” by Franz Kafka, “Peter Pan” by J. M. Barrie, “The Great Gatsby” by F. Scott Fitzgerald, and “The Scarlet Letter” by Nathaniel Hawthorne. These files all are .txt files that contained punctuation, proper nouns, and other miscellaneous characters that had to be parsed in order to perform the proper NLP analysis.

Framework Capabilities:

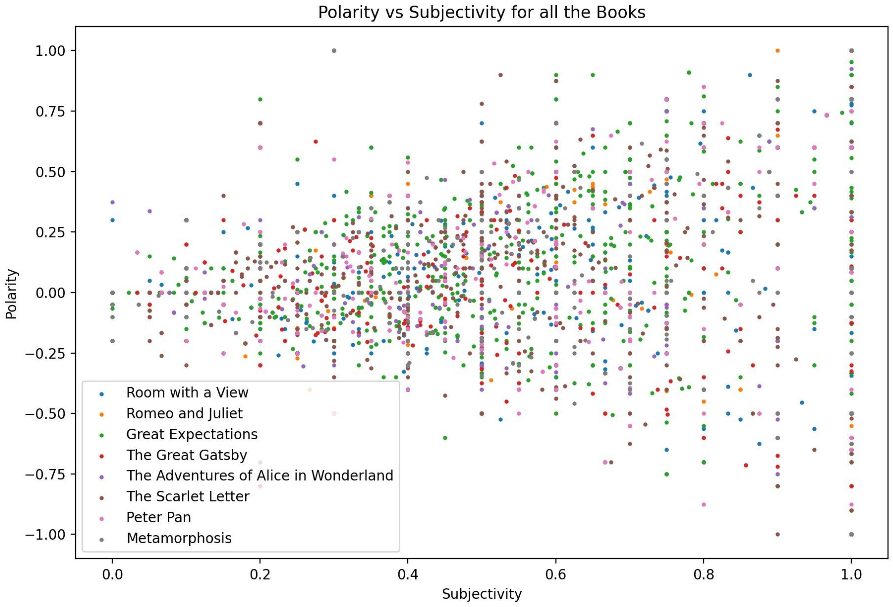
Our project is capable of taking multiple text files, parsing them, and visualizing the word counts of each and the statistics regarding them for each text. In this project, our functions take in ten text files, but this can be used to parse any number of text files. The library supports importing/writing/developing alternate parsers to parse any file. Our parser function cleans the data by removing proper nouns, punctuation, and any text in brackets to provide the cleanest word count for each of the novels possible. After cleaning the data, we create the visualizations. The text to word Sankey diagram displays the novels on the left and connects the word count to each word on the right side.



In addition to this, we also create separate word clouds for each of the texts and display those subplots as a visualization, which examines the frequency of each word in the text, where the bigger the word is in the subplot, the more frequently it occurs in that text.



Lastly, we create a scatter plot plotting polarity versus subjectivity for each of the novels, overlaying them onto the same plot. This measure of sentiment analysis examines how positive or negative the text is as the subjectivity of the text increases.



Conclusions:

For this project, a few common trends were derived from the visualizations. In the Sankey diagram, the most common words were “the”, “and”, and “to”, while the novel with the most usable words was Charles Dickens’ Great Expectations. In the polarity vs subjectivity plot, a conical pattern is created where as the subjectivity of the novels increased, the polarity (what is seen as “positive” vs “negative”) varied more and more drastically. Our findings have ramifications for NLP processing and text parsing for many different types of files, and can derive interesting insights like the ones shown in this project.

Link to code repository:

https://github.khoury.northeastern.edu/jmguthrie/ds3500_booklib

References

"Alice's Adventures in Wonderland by Lewis Caroll." *Project Gutenberg*, 14 November 2022, https://www.gutenberg.org/ebooks/11.

"A Room With A View by E. M. Forster." *Project Gutenberg*, 1 May 2001, https://www.gutenberg.org/ebooks/2641.

"Great Expectations by Charles Dickens." *Project Gutenberg*, 1 July 1998, https://www.gutenberg.org/ebooks/1400.

"Metamorphosis by Franz Kafka." *Project Gutenberg*, 17 August 2005, https://www.gutenberg.org/ebooks/5200.

"Moby Dick; Or, The Whale by Herman Melville." *Project Gutenberg*, 1 July 2001, https://www.gutenberg.org/ebooks/2701.

"Peter Pan by J. M. Barrie." *Project Gutenberg*, 25 June 2008, https://www.gutenberg.org/ebooks/16.

"Romeo and Juliet by William Shakespeare." *Project Gutenberg*, 11 May 2022, https://www.gutenberg.org/ebooks/1513.

"The Great Gatsby by F. Scott Fitzgerald." *Project Gutenberg*, 30 October 2021, https://www.gutenberg.org/ebooks/64317.

"The Scarlet Letter by Nathaniel Hawthorne." *Project Gutenberg*, 19 October 2022, https://www.gutenberg.org/ebooks/25344.

"What Is Natural Language Processing?" *IBM*, https://www.ibm.com/topics/natural-language-processing#:~:text=Natural%20language%20processing%20(NLP)%20refers,same%20way%20human%20beings%20can.