Jasmine Malik (jm5xx@virginia.edu) DS 5001 Spring 2023

Introduction

The corpus used for this project is comprised of a subset of 100 english novels from the github repository, "Computational Stylistics Group." The age range for the novels in the corpus range from the 19th century to the beginning of the 20th century. I am a huge fan of novels, and for this project I wanted to conduct exploratory analysis on a corpus of novels. The major question posed around this project is specifically to compare male versus female authors of English novels during this time period. I conducted the standard analytical procedure to analyze the corpus of novels, and then went one step further to visualize and separate the corpus by gender of author to see if there was any significant findings between the two groups.

Source Data

The corpus of novels used for this project is a subset of a corpus of 100 English novels. (https://github.com/computationalstylistics/100_english_novels) This repository is a part of a larger initiative started from a group the "Computational Stylistics Group." This group is a research team that specializes in text analytics and has many similar corpora comprising of novels from a variety of different countries available for public use. From the 100 novels provided in the original corpus, I selected five female and five male authors and two books from each other, resulting in a corpus of 20 books total, available in my UVA Box at the following link. (https://virginia.box.com/s/j9m1j7v4jojqaek7mmovtho8lp0ynyu9). The twenty books' selected initial file format was in text file format. To create a combined corpus of all 20 books with OHCO formatting, some manual coding had to be done by finding the correct regex for chapter identification for each book and creating and combining the separate book dataframes into one corpus. A LIB table was made to hold the corpus metadata that has information about the books such as the number of chapters and characters. In this corpus, the average length of books is about 94323 lines, and the average number of chapters is 26.

Data Model

Describe the analytical tables you generated in the process of tokenization, annotation, and analysis of your corpus. You provide a list of tables with field names and their definition, along with URLs to each associated CSV file.

The first table created in the analytical process was the CORPUS table. This is a compilation of all of the text files of the individual books, completely tokenized, and indexed using OHCO. This had to be done manually both because each book had a different chapter formatting, and the TextParser function was not working properly. The LIB dataframe was created simultaneously with the metadata for the files, such as regex, author name and gender, and book title. In the process of creating the rest of the analytical tables, I would use the LIB table to merge, join and visualize certain features by categorizing using the information in the LIB table. The next table is the VOCAB table. This is a table about information at the term, or token, level. This includes amount of times a term appears in the whole corpus, number of characters it has, part of speech and more. Information that was later found out such as TDIDF, stopwords, and stemwords information was merged as it was found. The next tables created was a POS and

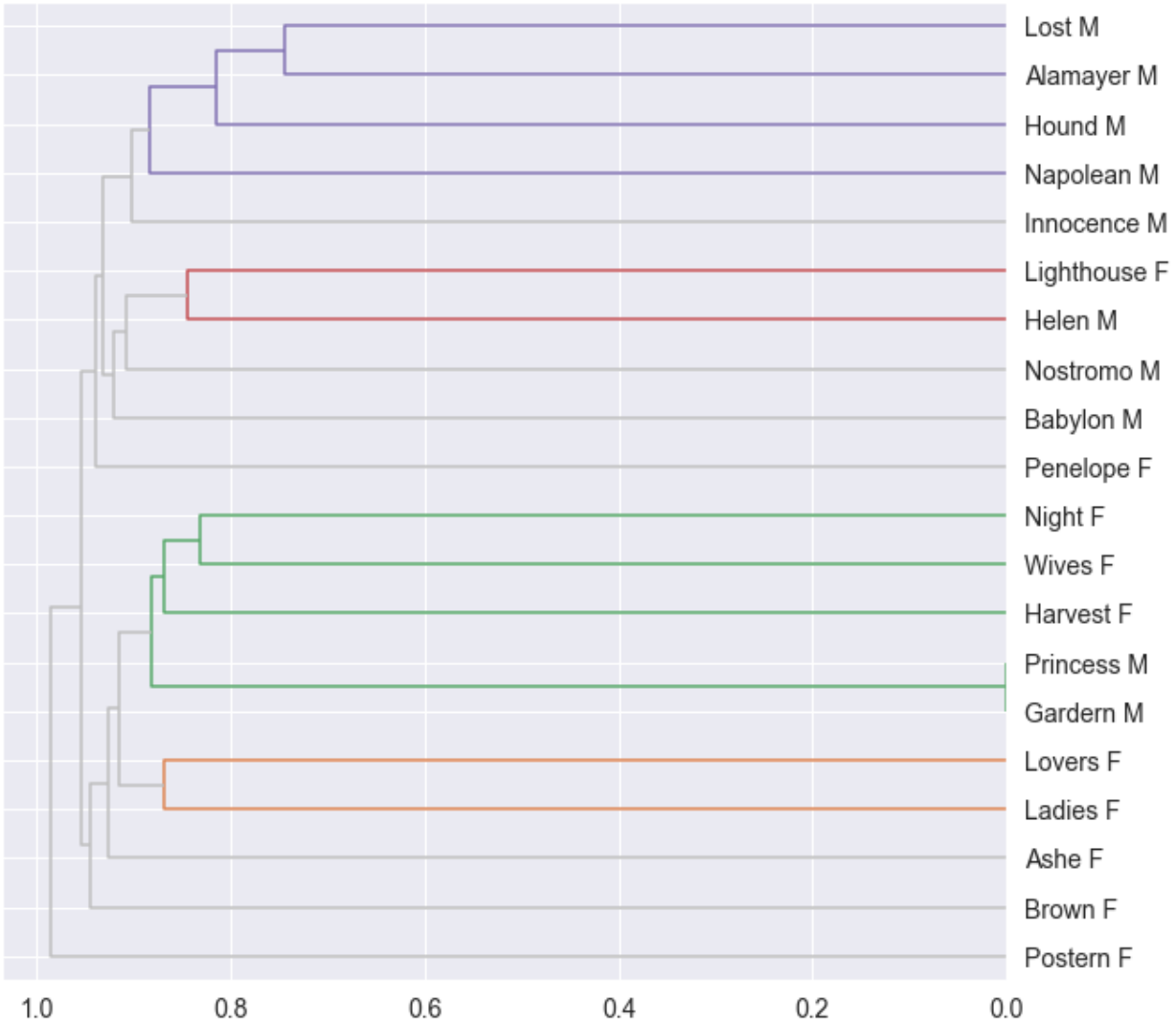
POS group table. This is a table that contains information about the parts of speech of the words in the corpus. Next I made a BOW table, or a Bag of Words table, by grouping the CORPUS table by chapter and term only. Using the BOW table, I made a Document Term Matrix table that orgainzes the terms variations' by book. Next a TDIDF table was made, which stands for Term Frequency Inverse Document Frequency. This table gives each word in the corpus a score that represents its signficance in the corpus. The values of this table were merged onto the vocab table. A PAIRS matrix was made to understand the distances between the terms in the corpus. This was useful to conduct analysis and create visualizations including hierarchal clusters. A CORR matrix was made to understand the correlations between the 20 books in the corpus. To condcut PCA analysis, a table was made for the eigenvalues, and a LOADINGS table was made and used to understand the contribution of each term in its components. To conduct LDA, a THETA and PHI table were made to make the TOPICS table that shows the top words for the the topics that were found. A word2vec table was created to save the results from the word2vec analysis that was conducted as well. When conducting Sentiment Analysis a SALEX table was made that holds sentiments of words and this was combined into the VOCAB table and the BOW table get sentiments for the words in the corpus.

Table	Desc.	Link to File
CORPUS	Table of terms categorized by OHCO (book, chapter, paragraph, sentence and term. Has part of speech information as well.	https://github.com/jasminemalik/ETA-Final-Project/blob/main/Created%20Tables/Compressed_CORP_DTCM.zip
LIB	Metadata on each book - source file path, title, regex, author, author gender, number of chapters, length of book, and the kendall sum.	https://github.com/jasminemalik/ETA-Final-Project/blob/main/Created%20Tables/LIB.csv
VOCAB	Table for each term, number of occurences, length of term, TFIDF values, part of speech, stop word (yes or no) and stem word information as well.	https://github.com/jasminemalik/ETA-Final-Project/blob/main/Created%20Tables/VOCAB_.csv
POS	Number of occurrences of each part of speech in the corpus.	https://github.com/jasminemalik/ETA-Final-Project/blob/main/Created%20Tables/POS.csv
BOW	Terms with the tfidf values.	https://github.com/jasminemalik/ETA-Final-Project/blob/main/Created%20Tables/Compressed_TFIDF_BOW.zip
POS_GROUP	Definition of each part of speech and	https://github.com/jasminemalik/ETA-Final-Project/blob/main/Created%20Tables/POS_GROUP.csv

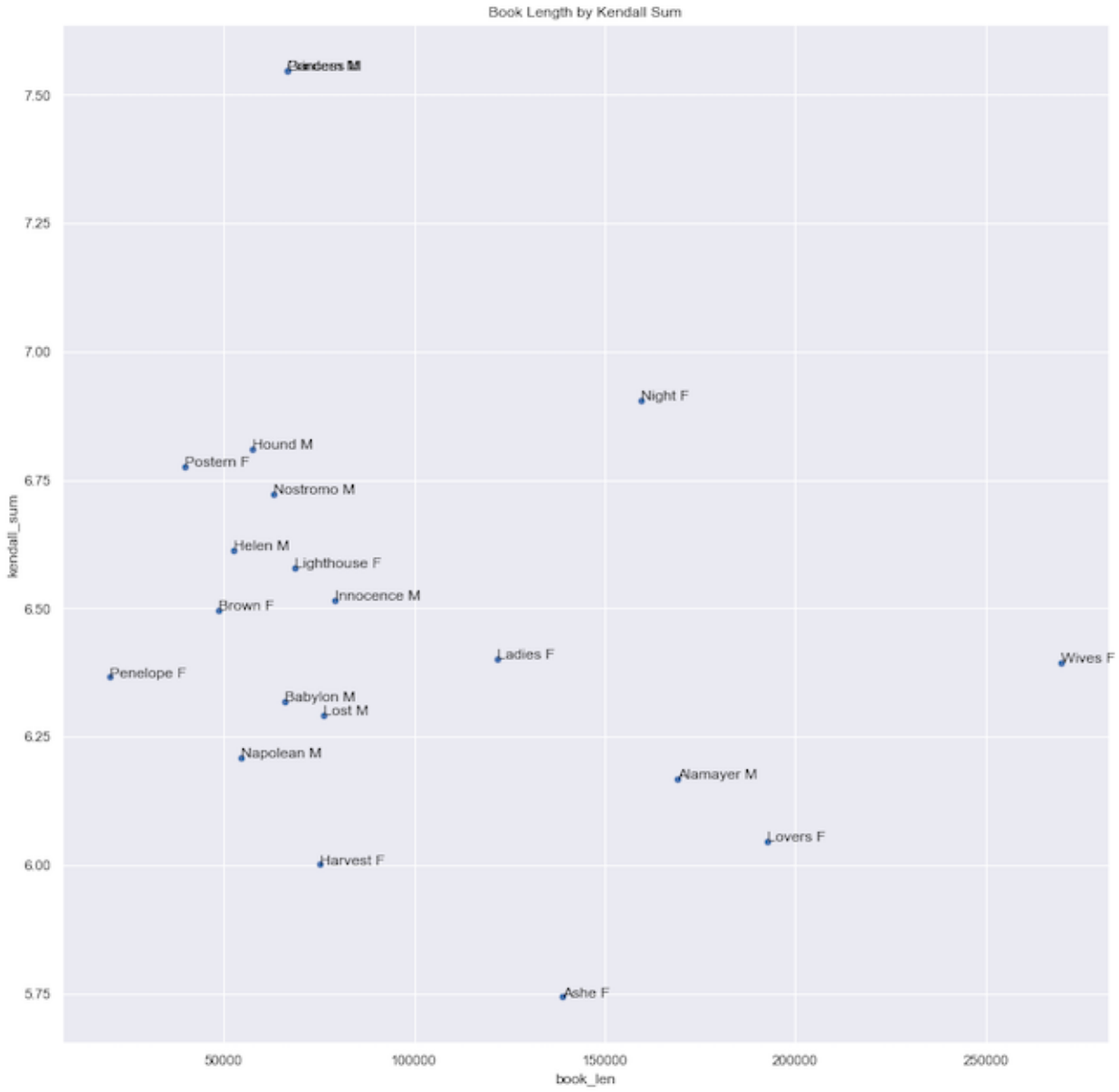
Table	Desc.	Link to File
	number of occurrences.	
DTCM	Document term matrix, categorized by the book chapters.	https://github.com/jasminemalik/ETA-Final-Project/blob/main/Created%20Tables/Compressed_CORP_DTCM.zip
TFIDF	TFIDF values of terms in the corpus.	https://github.com/jasminemalik/ETA-Final-Project/blob/main/Created%20Tables/Compressed_TFIDF_BOW.zip

Exploration

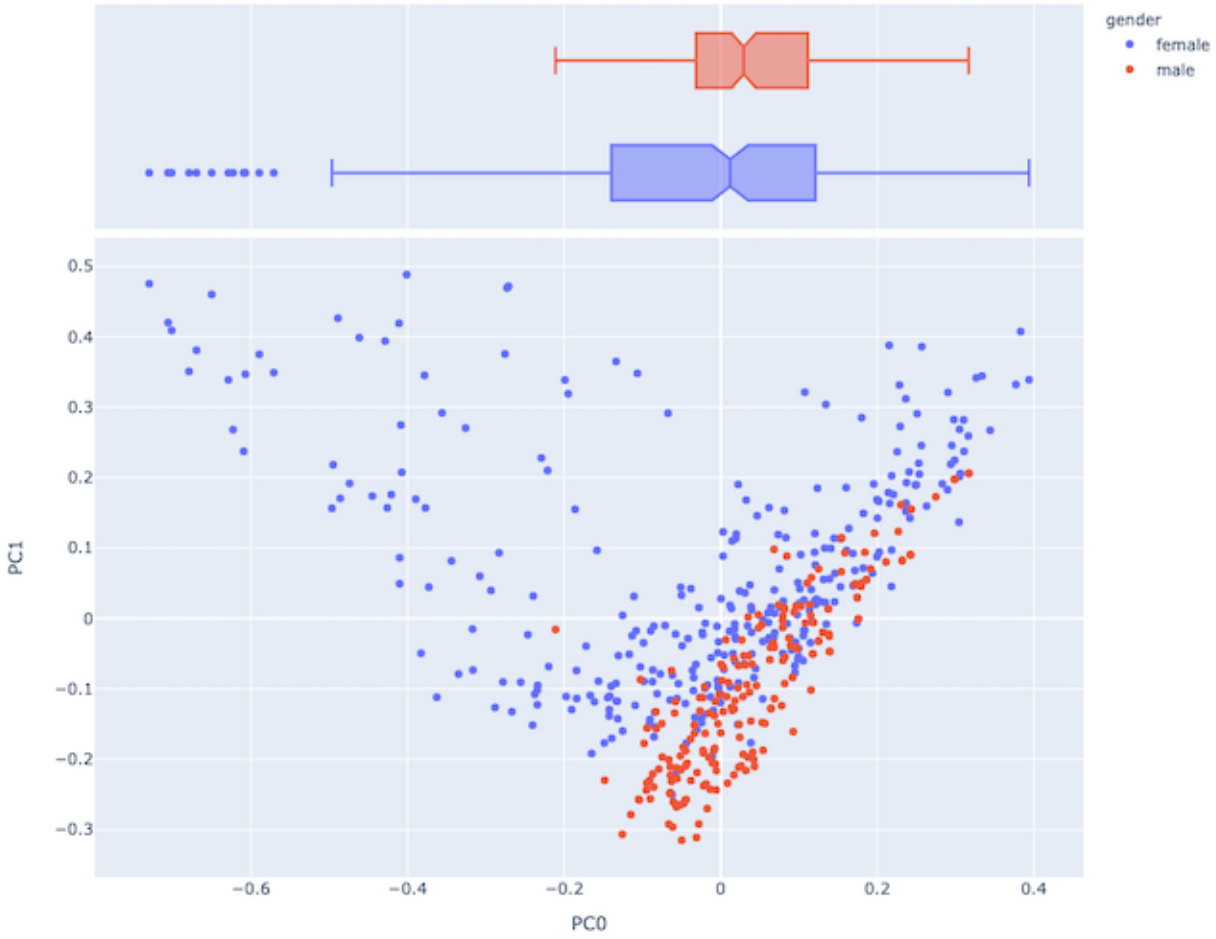
To start off a key point in my exploration, I will begin with hierarchical clustering that I created using the PAIRS and LIB tables. I used the function that was created in class called hca(). The parameters that I inputed to create this hierarchical cluster were: PAIRS.cosine, linkage_method='ward', color_thresh=1. Because I changed the titles in the LIB table to include M or F to signify the gender of the author, I was able to see both the book and the author gender clustered. As you can see below, the function was able to cluster male and female authors pretty successfully except for a couple that were separated, but altogether I was very pleased with the results.



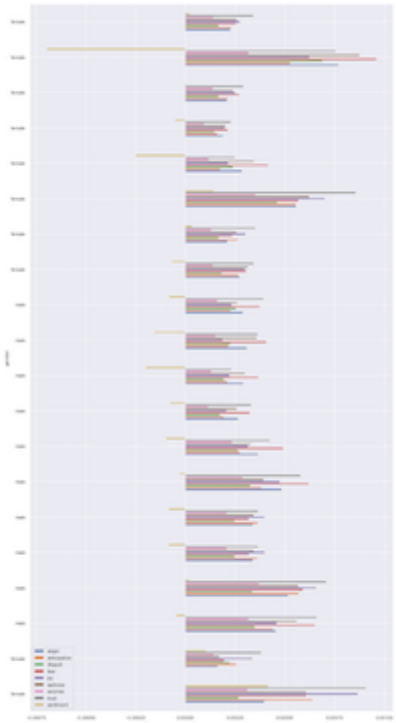
Below is a scatterplot made plotting the kendall sums of the books in the corpus, with the book lengths. There was no clear similarity or difference between the male and female authors here in terms of both the kendall sum and length of books, which is still in itself an interesting observation.



After conducting PCA, to visualize the results I used a function used in class made to visualize the loadings and the principal components using the DCM table. The parameters used to create the scatterplot was signifying PC 0 and PC 1 and coloring the points by author gender. The scatterplot shown below shows that there is a notable difference between the two groups of authors. Female authors have much more variation in the PC 0 and PC 1 loadings compared to male authors. Male author books have mainly positive principal component values, which means that their books, separated by their chapters contribute positively to the overall PC value of the corpus.



The last visualization that I want to highlight in this report out of the ones that I created during the analysis of this corpus is the one made after the sentiment analysis portion of the project. Here you can see through a visulization the gender of the author of a book, and the level of sentiment of the different emotions. Upon further inspection, male authors seem to have higher sentiments of fear and anger in their books. Male authors also had higher levels of trust and joy sentiment in their books than female authors did.



Interpretation

Through the steps of analysis that was conducted on this corpus, there are many observations that can be found both about the corpus as a whole, and about the initial research question that was posed. One, is that the hierarchal clustering method was able to successfully cluster the male and female authors together. The method used in the function was the cosine method on the PAIRS dataframe. This signifies that there is a level of correlation between the books written by female authors versus male authors in the late 20th century, however, the level to which there is a correlation is not very high as was displayed in the correlation matrix made in the same analysis. All the books in the corpus except for two that were written by the same author, all had the same level of correlations. When looking at the LDA portion of the analysis, it was very interesting to see the top terms in the topics for this corpus, seeing words such as "sir, evening, possession, evening, tea" attest to the early century writing style and topics of novels in England. In conclusion, although no strong conclusions can be made about the difference in female and male novel authors in early 20th century England, there are definitely significant correlations and similarities between the two groups that was seen and visualized through a variety of analytical methods in this project.