Rumeer Keshwani

Professor Zhi Li

MIS3640

3/29/2020

Project Overview

For this assignment I utilized the gutenburg library of books in order to run a text analysis on Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft Shelley. The analysis I ran was a word frequency analysis of the top 20 words displayed in a graph. I was hoping to gain a better understanding of key words that are used in the book, outside of stop words.

Implementation/Results

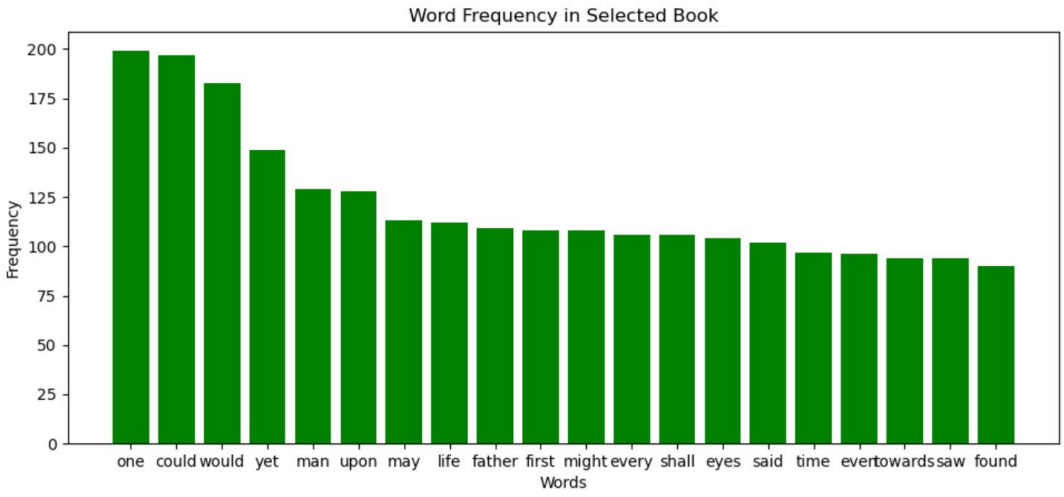
The structure I decided to follow was composed of three primary functions and the analysis of the cleaned data. First I imported all the relevant libraries in order to run the frequency analysis. The major components were cleaning the data in order for it to be analyzed, removing the undesirable stop words that yield zero insight and formatting the data so that it could be understood easily.

Importing the NLTK library in order to remove the stop words was critical to this process and fit together very well with the visualization portion of the code. Without this package, the stop words would have to be listed out manually, or I would have to display the top 100 words and reason out the stop words myself and then manually remove them. In terms of formatting the

data, I decided to display the words in a table format of Words: Frequency of Words because there were only two components to my analysis.

```
Words : Frequency of words
--------------
one  : 199
could  : 197
would  : 183
yet  : 149
man  : 129
upon  : 128
may  : 113
life  : 112
father  : 109
first  : 108
might  : 108
every  : 106
shall  : 106
eyes  : 104
said  : 102
time  : 97
even  : 96
towards  : 94
saw  : 94
found  : 90
```

Additionally, a graph in descending order also lends itself easily to the human eye for immediate understanding.



As is made clear, outside of major stop words, the number one word in Frankenstein is "One." One reason why this could be is because the creature in Frankenstein has no distinct gender and therefore is assigned a gender neutral pronoun. Beyond could and would, which are second and third because of the Creature and other characters referring to potentiality in general, man is the fifth most used word. Man is a key concept in the book as the creature spends his time observing a family in order to gain a deeper understanding of human nature.

Reflection

Ultimately, I think that the process was fairly clean given the scope of the analysis. Some other potential ideas I could have explored was a sentiment analysis of the book or running a test of similarities to another book by Mary Shelley. I learned how to display data as a visual within python, which is something I had not previously explored on my own. Beyond visualization, exploring the NLTK package is also quite useful. I had no real insight into natural language processes (outside of in-class activities) and NLP in python is extremely relevant today. I worked on this project by myself and did not have to split up the work.