Project Overview

Overview: from the Gutenberg Data Base, I got a book from Ernest Hemingway, processed it using the Urllib library and created a function to calculate word frequency within the entire book (using the requests library as an additional tool to facilitate), after that, I used regular data analysis resources, including Pandas, NumPy, Seaborn and MatPlotLib to extract some data about the frequency of the words and made some interesting discoveries.

Implementation: I decided to use tools that I had previous experience before, since it would be easier to manipulate data using them, so Pandas, Numpy and Seaborn were the main tools employed in data analysis. For Data Mining, I decided to use the UrlLib given as an example, and due to no API issues, there was no need to Pickle it.

The Data gets processed using an "If" statement, which leads to a list combining the word and the number of times it appears in the book. Due to the difficulty in analyzing data inside of a list format, this dataset gets into a dataframe format. From this point on, I used the Pandas Library basic tools to treat the data, so I could perform the analysis. Pretty much, the only thing necessary here was to remove the duplicated values, as it was done and then re-order in a descending order. Instead of almost 6900 words, there were "only" 1996 unique words in the book.

On the data analysis aspect I did some descriptive statistics, including, 25% and 75% percentiles, mean, standard deviation, etc. Also, performed a normal distribution of the data, which presented an interesting bell curve plot. Finally, I included a histogram and a box plot to find outliers.

Results: this analysis led to some interesting discoveries. The first being the amount of pronouns, prepositions and interjections used in our texts, as this table shows:

```
t.head(10)
```

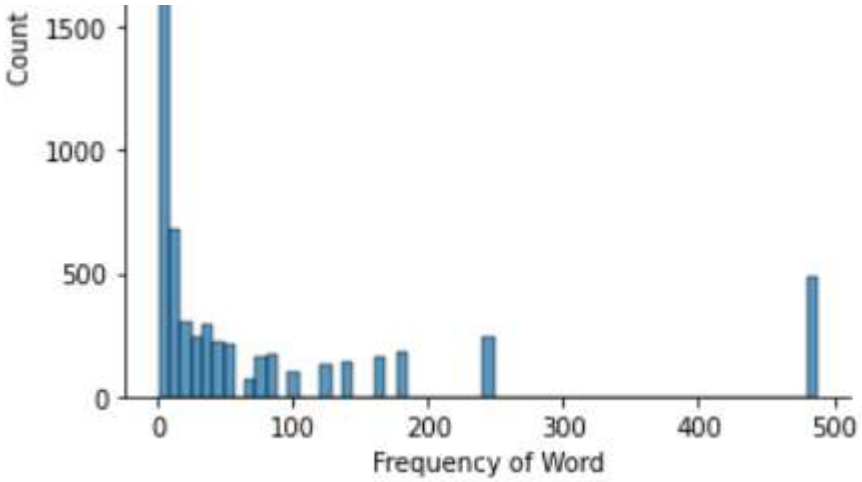
1]:

	word	frequency
15	the	488
24	and	246
4	of	179
77	to	162
5	in	142
147	a	126
388	he	100
1	project	85
40	you	85
47	or	80

Besides the word "project", all of the other words are either prepositions, pronouns or interjections, with a special mention to the word "the". I always think about this when I'm writing papers, the amount

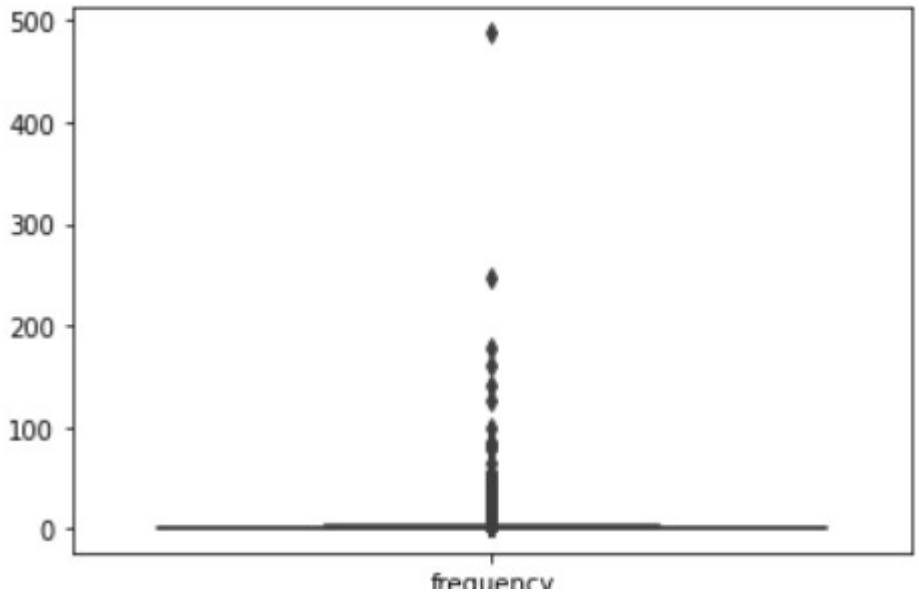
of “the” we use, as this table shows, “the” is by far the most common word in this text, with more than twice the amount of appearances of the second word, “and”. This can be an interesting sign for me when I write future papers to find some synonyms for those words and thus have a more cohesive and pleasurable text to read.

Another aspect, which complements of the one I talked above, is that most of the words in this text appear less than 50 times as this frequency plot shows:



As such, roughly 75% of the words appear in the text 50 times or less, which is pretty interesting. This presents a hypothesis that most texts use words that have meaning a lot less frequently that I thought. From a quantitative point of view, a text is a bunch of “the’s” “and’s” and “of’s” linked by meaningful words in context and not the opposite.

To prove the point above, here is a boxplot showing the exact same scenario:



Reflection: the process of defining the variables, functions, datasets, consolidating into the data frame and performing simple analysis went well, with minor issues, which were easily fixed, I believe I could definitely improve in building a more complex model, with some machine learning and AI aspects, as well as using more advanced techniques and functions that the ones used in this project. The scope of the project was simpler than what I wanted because of time restrictions as well as a lack of technical knowledge of my part. As I worked alone, there was no team issues. What I wish I could know better are better ways to gather data as well as a more depth knowledge and intensive practice in functions and iterations, something I have to work on.