EEE 586 Statistical Foundations of Natural Language Processing Assignment 1 Report

Burak Taşdemir 22003996
Bilkent University Computer Engineering Department

***Abstract*—In this assignment, i have examined the fundamental laws of natural language processing such as Zipf's Law [1]. Also i have gained experience on working with language data, forming a corpus and performing basic preprocessing on this data. Also we have examined the effect of the writer and genre on the corpus and removing stopwords on the relationship between token size and the word type frequencies.**

I. Introduction

In the field of natural language processing, it has been noticed that there is a relationship between factors such as the frequency of use of words, the number of meanings a word carries, and the distances between words. The first studies in this sense were made by George K. Zipf. Zipf published these studies in the article and became law under his name. According to Zipf, when we sort the words in a text in descending order of frequency of use, regardless of the content and language, a list with a certain pattern is obtained. The order of usage of a word in the list multiplied by the frequency gives a fixed number.

In this report i have gathered 3 works of 3 different authors in UTF-8 format and 3 different works in 3 different genres to examine the effects of the author and the genre on the Zipf's Law. Report consists of the corpus construction details, experiment results and the discussion of the results, respectively.

II. Corpus Construction and Implementation

For constructing the corpus i have selected 3 different works of 3 different great authors. I have selected The Tale of Two Cities, Great Expectations and Oliver Twist written by Charles Dickens, The Great Gatsby, This Side of Paradise, Tales of the Jazz Age written by F. Scott Fitzerald, Frankenstein, The Last Man, Mathilda written by Mary Shelley. For different genres, i have selected the three genres as crime, horror and mythology. After that to create to whole corpus, some standardisation are made. I have deleted the replicated Gutenberg Project part from the 18 text files to prevent the disruption of the results. After that i have cast whole text to lowercase characters for same reason before using built-in string libraries of python. Also i have get rid of every punctuation mark for the same reason. Stop words are words that are filtered out during text preparation for further processing. For stop words removal i have tried two different dictionaries and decided to go with the stacy library of conda dictionary. After removal of the stop words i had two versions of the 18 texts, one with stop words and one with no stop words. I have conducted whole of my experiments on both versions of the texts. After stop words removal, i have created vocabulary files with word types and frequencies for all 36 texts.

III. Results

In this part we have experimented Zipf's Law for different authors' works. I have assembled a larger corpus for each author consisting of 3 works of them both with and without stop words. After that i have plotted the Zipf's Law plots Figure 1 in normal and log-log form. After that, for each author i have plotted a single plot with 3 different works together to see whether they are all obedient with the Zipf's Law. Figure 2. We can see that all of the plots are obedient with zipf's law yet some without stop words are relatively less obedient to it.

For the next experiment, i have analyzed the relationship between the corpus and the vocabulary size. I have kept track of the unique word types with increasing token size. With a step size, every increment in token size corresponds a increment in word types. I have plotted the relationship with normal and log-log plots for each writers works with the previous approach data Figure 3. There is seem to be a ratio between the two axis. That is called type-token ratio which represents the relationship between different unique words in a text file and the number of total words. This depends on the text file's purpose, for example in written text it is higher than normal daily conversations.

For this part of experiments i have plotted the Relationship Between Token Size and Word Type plots for each writer in their own works and all works assembled in log-log plot form with and without stop words in Figure 4. Also i have put fitting lines to each scatter plot to find the slope of the text which is type-token ratio of that text of that writer. It is seemed that with or without stop words, relationship of type-token ratio is similar for each writer. But when we have looked at the assembled plot, we can see that each author has his/her own type-token ratio range. Which is understandable since they are from different years and different backgrounds.

For now, we have put genre on authorship. We are experimenting the effect of genre in the previous experiment. I have plotted the relationship Between Token Size and Word

Type plots for each genre in their own works and all works assembled in log-log plot form with and without stop words at Figure 5. We can see that except for a text in horror genre, most of these fiction texts are resembling similar type-token ratio.

A. *Discussions & Conclusions*

Through these experiments, i have seen that, clustering can be done in genre sense or author sense by using type-token ratio. With five different authors, it would be maybe harder since our writers are coming from different times yet if they were closer to each other in year sense, there can be some problems aroused due to similarity of their writing styles and word types.

Removing stop words would affect the zipf behavior and the type-token ratio since we are removing common words. These effects are yet different. Without stopwords, corpora are getting further away from zipf's law behavior but corpora is behaving better and closer to a line without stopwords in type-token ratio experiments.

In last experiment, i have looked to the paper [2] to examine the effect of randomness on the corpus. I have created a random corpus as the paper suggested and experimented with zipf's law and the type-token ratio. I have seen that, zipfian behavior is not there as the paper suggested yet there was a type-token ratio which is high due to randomness. Yet i have figured out to lower type-token ratio by feeding the corpus with some shorter random words. Figure 6.

At Figure **??** one can see

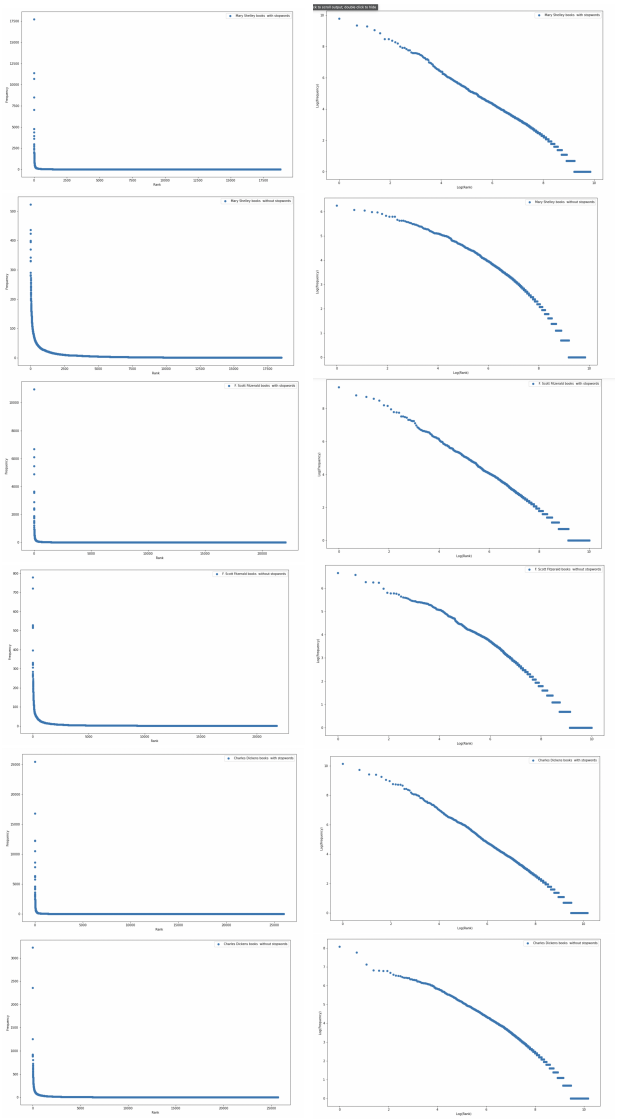

Fig. 1: For each author corpus, normal and log-log plots with and without stop words

IV. Conclusion

References

[1] Zipf, G. K. (2013). Selected studies of the principle of relative frequency in language. Harvard university press.

[2] Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. IEEE Transactions on information theory, 38(6), 1842-1845.

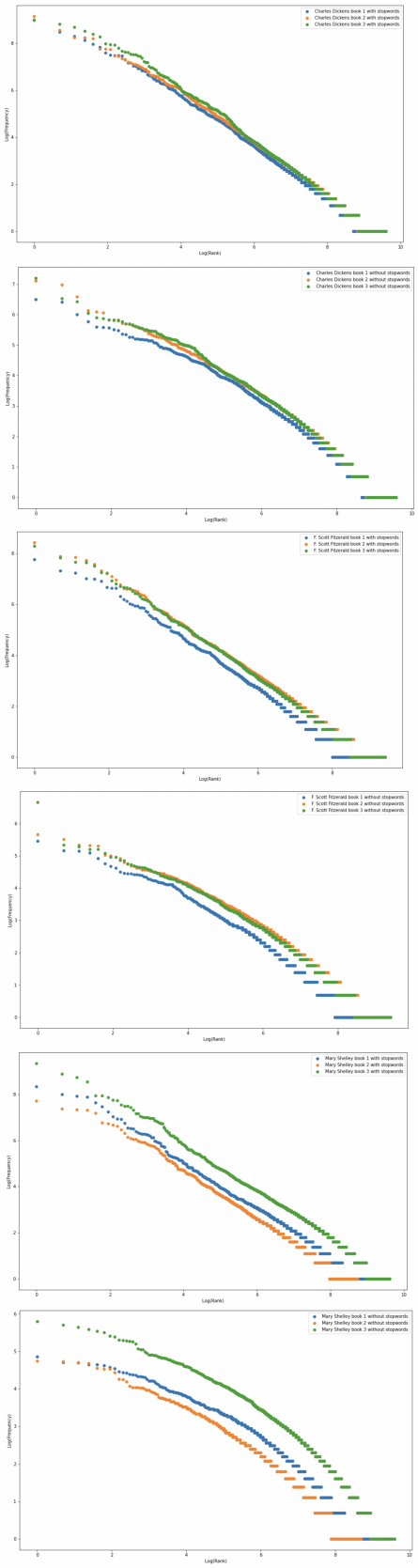

Fig. 2: For each author 3 different works together plotted, log-log plots with and without stop words

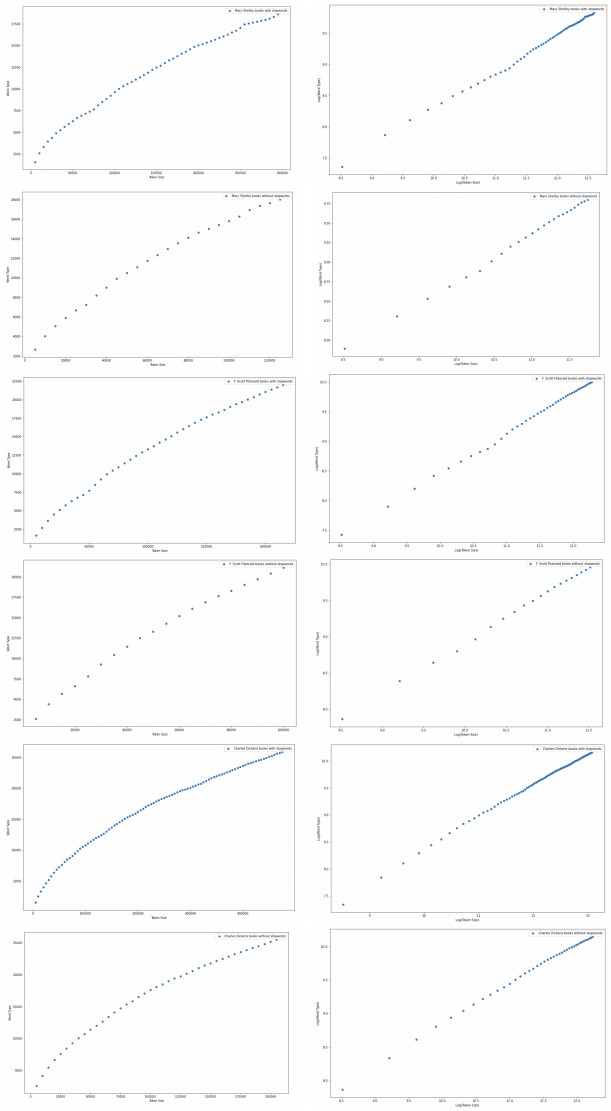

Fig. 3: Relationship Between Token Size and Word Type plots in normal and log-log form with and without stopwords

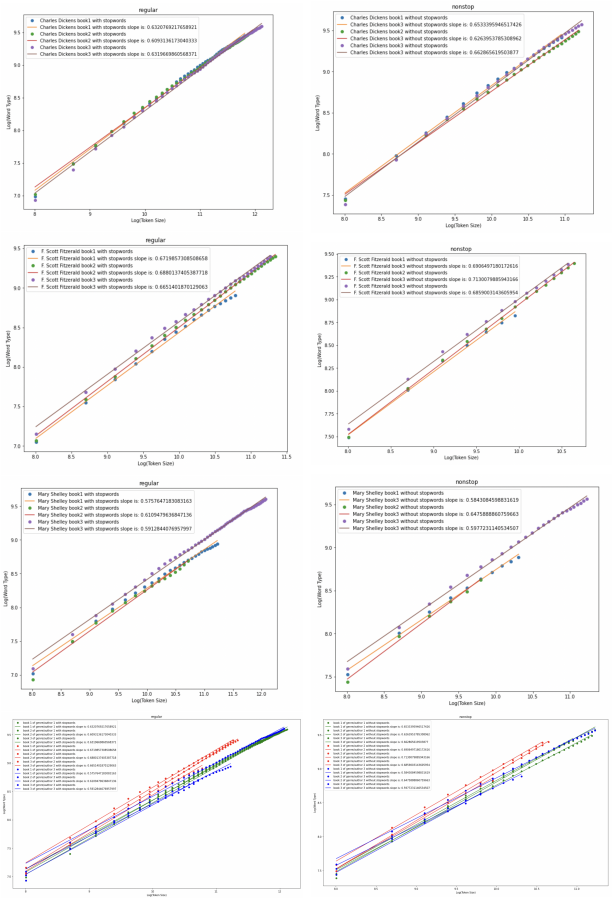

Fig. 4: Relationship Between Token Size and Word Type plots for each writer in their own works and all works assembled in log-log plot form with and without stop words

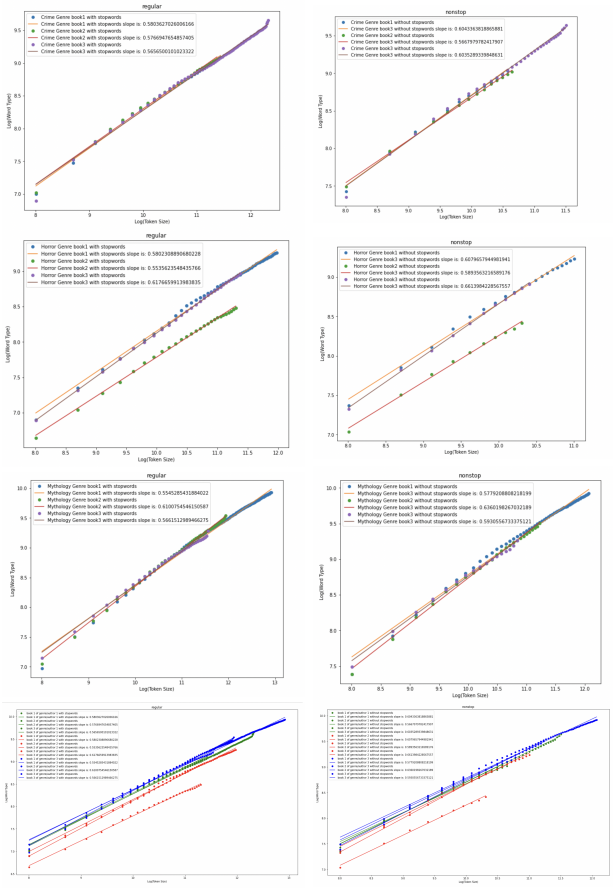

Fig. 5: Relationship Between Token Size and Word Type plots for each genre in their own works and all works assembled in log-log plot form with and without stop words

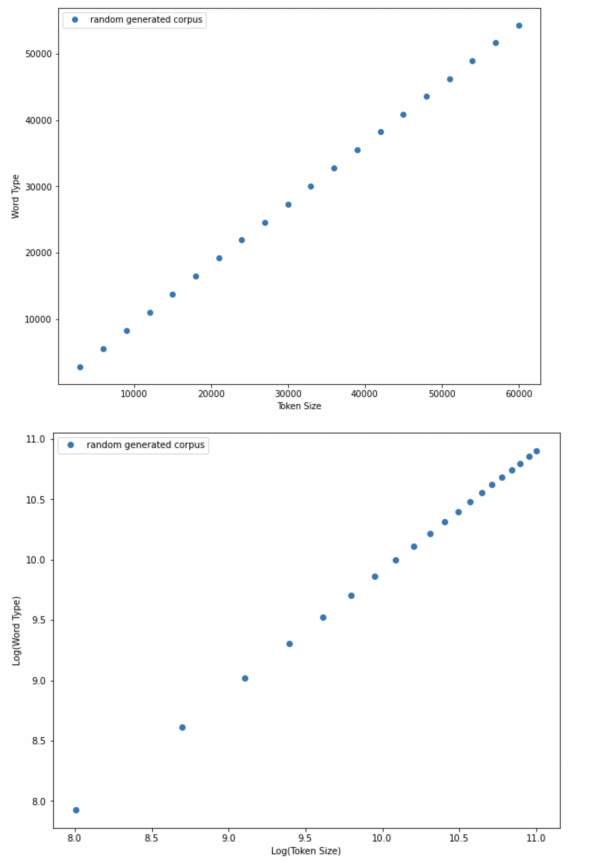

Fig. 6: Relationship Between Token Size and Word Type plots for random corpus in normal and log-log plot form