Zipf's Law: A Brief Analysis

Sepehr Bakhshi

Bilkent University
sepehr.bakhshi@bilkent.edu.tr

1 Abstract

In this assignment, our aim is to analyze several corpora from different authors and genres. To accomplish this goal, I first conducted a brief literature review to gain a high-level understanding of the assignment tasks and objectives. Through a thorough analysis of several corpora, I found that these works are closely allied with the Zipf's law, which states that as the rank of a word type increases, its frequency decreases. Additionally, my experiments revealed that although features mentioned in the analysis, such as the slope of a linear regression fit for the log-log curves, could be used as a possible feature for clustering various corpora, however in many cases these kind of features might not be sufficient for the clustering task, and we might need extra set of features for a more accurate clustering. Despite this fact, they may still be helpful in distinguishing certain writing styles from others. At theend of this work, we also do an anlysis to see if randomly generated text follows the Zipf's law or not.

2 Introduction

In natural language processing, both the language and its processing play a pivotal role. The analysis of the language and its various theoretical aspects are both needed for a robust analysis and meaningful discussion. Based on this fact, our first assignment has focused on both of these aspects. In terms of preprocessing, we are asked to collect diverse corpora, tokenize them, and remove the stop words to have a clean and ready-to-use corpus. In terms of theoretical analysis, Zipf's law is considered the main topic of this assignmenet. We are asked to use our preprocessed corpora to study the Zipf's law, and see if it is applicable to our copora collection. Then we are also asked to examine whether the curves obtained for type-token relations are helpful for distinguishing the authors or genres.

Zipf's law is still utilized for theoretical analysis of the natural language. Some scientists argue about the fact that Zipf's law is an oversimplification and the word frequencies distributions can not be simplified by a simple formula [5]; however, Some recent works considering Zipf's law are the evidence for the ongoing discussion and analysis of this law [4]. Other than its applicability in natural language processing, it is also considered in various other fields related to computer science [1,2].

The report is structured as follows: in **Corpus Construction and Implementation** section, I go over the assignment questions and analysis one by one (from part (a) to part (e)). Then in **Results** section, I provide the results for parts (f) to (m). Finally, in **Discussion and Conclusion** section, I conclude my work, briefly explain what I have learned in this assignment, and mention possible improvements for future work.

3 Corpus Construction and Implementation

In this section, I explain the steps that I took for preparing my work before I start my analysis on the corpora. Parts (a) to (e) of the assignment are explored in this section, and various details mentioned in these parts are addressed.

3.1 Part (a)

For this part, I choose three different authors, two of them from the Russian literature, and one from the English literature.

Table 1 shows the authors, their chosen books and the corpus size.

Table 1: Authors and their books

Author	Book name	Book Size(MB)
Charles Dickens	Bleak House	1.9
	David Caperfield	2.0
	Our Mutual Friend	1.9
Fyodor Dostoyevsky	Crime & Punishment	1.1
	The Brothers Karamazov	2.0
	The Devils	1.4
Leo Tolstoy	Anna Karenina	2.0
	Resurrection	1.0
	War & Peace	3.2

There are two main reasons for choosing these writers for analysis:

- These authors are famous for their long narrative styles. All of them have famous works with a large vocabulary size.
- Their style of writing is known to be different. Two Russian writers (Tolstoy and Dostoevsky) are very famous in their differences in the narrative style. In the case of English literature, we know that it has definitely different style compared to Russian literature. This leads to a large and at the same time diverse vocabulary usage.

3.2 Part (b)

For this part, I choose three different literature and three books from each literature with different authors. Table 2 shows the details of the chosen literature and its corresponding authors, books, and size of each book.

I tried to choose diverse writers from each category. As I was not that much familiar with this type of writings, finding books with larger than 1MB size was quite challenging. However, I tried my best to keep the size of the books relatively large.

Table 2: Literature, Author and Books'details

Literature	Author	Book	Book Size(MB)
Horror	Bram Stoker	Dracula	0.8
	Bram Stoker	The Lady of Shroud	0.7
	William Hope Hodgson	The Night Land	1.0
Romance	R. D. Bleakmore	Lorna Doone	1.1
	Egerton Castle	The Light of Scarthey	0.9
	Alexandre Dumas	The Man in The Iron Mask	1.0
Scifi	Mark Twain	A Connecticut Yankee in King Arthur's Court	0.7
	Jules Verne	The Mysterious Island	1.1
	Jules Verne	Twenty Thousand Leagues Under the Seas	0.9

3.3 Part (c)

My work is implemented in python. Except some general libraries like Numpy, Pandas, Pickle, String, Matplotlib, etc. I do not use any NLP library. Based on the instruction in the assignment, I removed the text related to Gutenberg Project from the books. Then I preprocessed the text by separating the tokens. For removing the apostrophe from the text, I concatenated the string before and after the apostrophe. For more details about the implementation please check the *"HW1.py"* file.

3.4 Part (d)

I found a comprehensive set of English stopwords in Gitgub (link). I generated two corpus for each book; one that contains stopwords, and the other without stopwords.

The details of implementation is in "*HW1.py*" file.

3.5 Part (e)

I implemented the code needed for finding the word types and their frequencies for each book. Please check *"HW1.py"* for the code.

Table 3 for the detailed output of my implementation. The table demonstrates interesting results. In most of the books, the most frequent word is the novels is the protagonist's name. In other cases, it is totally relevant to the title of the book. For instance, in *"Twenty Thousand Leagues Under the Seas"*, the most frequent word is captain, which is quite relative to the title of the book that gives us hint about having a submarine and a captain as the main content of the book.

In some books we can see that surprisingly the word time is the most frequent word (*The Devils*, *Dracula*, and *The Lady of Shroud*). I have read one of these novels (The Devils) and I totally remember that the chronological order of the events were really important. I think it could be a possible reason that makes the word "time" the most frequent word in this novel.

4 Results

After the implementation and preprocessing of the texts, I provide experimental results for parts (f) to (m) in this section.

Table 3: Word type frequency details. For each column after the name of the book, there are two numbers or words in the parentheses. The first and second ones are retrieved before and after stopword removal, respectively

Book	No. Word Types	Most Freq. Word	No. Occurrences
Bleak House	(17,727, 17,070)	(the, sir)	(14912, 1,018)
David Caperfield	(16,936, 16,284)	(the, micawber)	(13,642, 769)
Our Mutual Friend	(17,271, 16,605)	(the, boffin)	(14,658, 1,118)
Crime & Punishment	(10,330, 9,716)	(the, raskolnikov)	(7,807, 722)
The Brothers Karamazov	(13,545, 12,901)	(the, alyosha)	(15,189, 1,176)
The Devils	(12,603, 11,964)	(the, time)	(10,847, 580)
Anna Karenina	(14,347, 13,703)	(the, levin)	(17,507, 1,512)
Resurrection	(10,222, 9,614)	(the, nekhludoff)	(11,568, 1,349)
War & Peace	(20,258, 19,556)	(the, prince)	(34,388, 1,886)
Dracula	(10,541, 9,925)	(the, time)	(7,858, 369)
The Lady of Shroud	(10,056, 9,455)	(the, time)	(7,749, 293)
The Night Land	(6,061, 5,606)	(the, great)	(13,866, 1,088)
Lorna Doone	(14,497, 13,892)	(the, john)	(12,412, 741)
The Light of Scarthey	(13,879, 13,262)	(the, adrian)	(9,055, 522)
The Man in The Iron Mask	(11,429, 10,852)	(the, king)	(10,991, 870)
A Connecticut Yankee in King Arthur's Court	(11,565, 10,945)	(the, sir)	(6,225, 355)
The Mysterious Island	(10,426, 9,860)	(the, pencroft)	(17,154, 1,047)
Twenty Thousand Leagues Under the Seas	(13,230, 12,596)	(the, captain)	(9,364, 735)

4.1 Part (f)

Zipf's Law states that for most of the data that we study in social science, the rank-frequency relation is inverse [6]. Meaning that by increasing the frequency, rank decreases and vice versa.

For the first part of this question which asks for a linear plot that shows the rankings vs. word frequencies of all the books of the authors compared with each other, I manually concatenated the books of each author and made a collection of three books for each author. As we can see in Figure 2, the plot is in harmony with the Zipf's law. Our observation for all of the authors show that as the rank of a word increases, its frequency decreases.

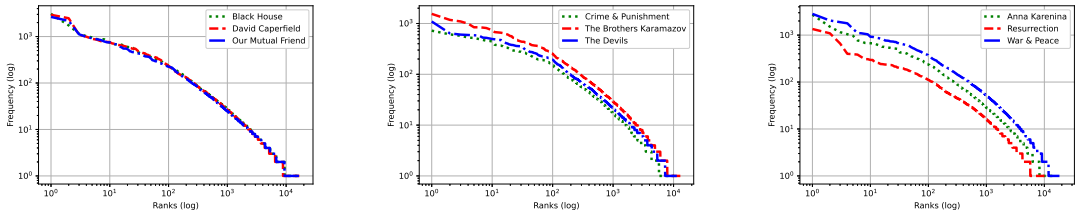
In Figure 1, we have the log-log relation between word type frequency and ranks of the words. The results are again compatible with our expectations and Zipf's law.

4.2 Part (g)

Figure 3 shows the relationship between the tokens and word types. As we can see in Figure 3, it is following the **Heap's law**. This law indicates the number of occurrences of unique words in a corpus decreases as we increase the size of corpus in terms of number of tokens. We can observe this relationship in our plot 3. Besides,it is obvious that that two authors from Russian literature, have closer curves. However, the slop of the type-token curve of the English writer (Dickens) is a little bit different from that of the Russian ones.

The parameters for the Heap's law are the followings:

- V is the unique vocabulary size
- N is the total number of tokens



(a) (b) (c)

Fig. 1: Log-log plot of freq. vs. rank for the authors and their books. a) Dickens, b) Dostoyevsky, c) Tolstoy

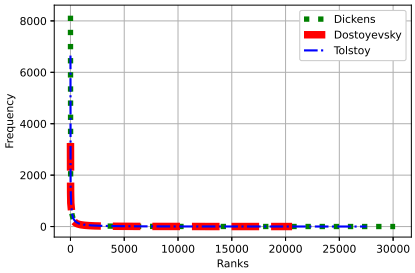


Fig. 2: Frequency-rank plot for three authors

– β is the scaling component
– k is a constant

The following formula is utilized to find the relationship between tokens and number of unique words:

$$V = k \times N^{\beta} \tag{1}$$

4.3 Part (h)

The loglog plot for Heap's law should be linear. Our plot in Figure 4 also follows this rule. As we observe in these figure, the linear lines in the figures are closer to eachother. The log scale usually does not provide a visually comprehensible difference, however, in later stage, we will have a deeper analysis on the loglog plots.

In one of the log-log plots in Figure 4.a, we see a change in the slop of the line. This line belongs to the **War & Peace** book of Tolstoy. This change in the slope is also called a "knee" in the plot, may be due to several reasons:

– The writer uses more diverse words from that point on. Meaning that the frequency of the unique words increases.
– Another reason could be due to the complex nature of novel. There might be new events happening in the novel, and it may oblige the writer to add more explanations to the novel. This necessitates using novel words and results in the so called "knee"in the log-log plot.

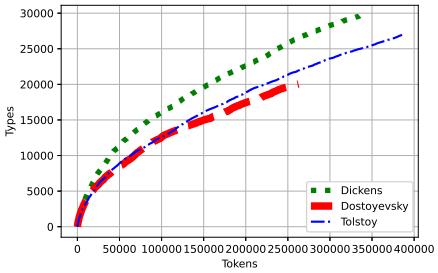
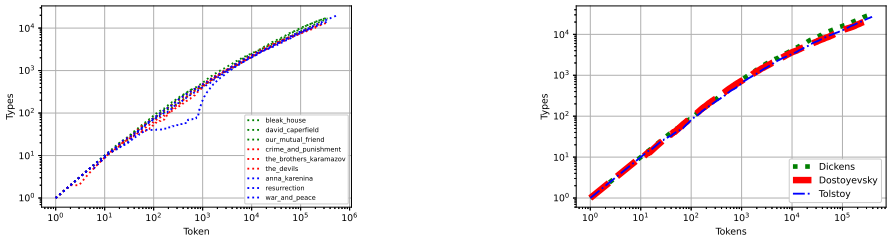


Fig. 3: Frequency-rank plot for three authors



(a) Log-log plot of each book (b) Loglog plot of all books of each authors concatenated

Fig. 4: log-log plots for token vs. type. In (a), green, red, and blue colors are for Dickens, Dostoyevsky, and Tolstoy, respectively

4.4 Part (i)

In this part, I calculated the slope of the log-log plots obtained in the previous sections. I have also plotted the slope of the best fitting line for one book from each author in Figure 5. The results of my calculations are also presented in Table 4. The implementation details of my work is in *"HW1_find_slope.py"* file.

The results of my calculations shows some interesting results. First of all, the slopes of the works of Dicken's are pretty close to each other (min: 0.632, max: 0.643). The same also applies for Dostoyevsky's books (min: 0.604, max: 0.621). However, for Tolstoy, the slopes are more diverse and quite deviated form the norm of his books (min: 0.618, max: 0.657). We will go back to analyzing the results in the upcoming sections in more details.

4.5 Part (j)

I have repeated the experiments in Part (h) and Part (i) for the literary types. The results are presented in Figure 6. The results are also similar to that of previous sections where we analyzed the books of three authors. However, the fluctuations in the log-log ("knee" in the plot) file that we observed in previous sections (Figure 4.a), is softer in these set of books.

The results obtained in Table 5 are much diverse compared to the results in Table 4. There could be several reasons for this observation. One reason could be diversity of authors, and the

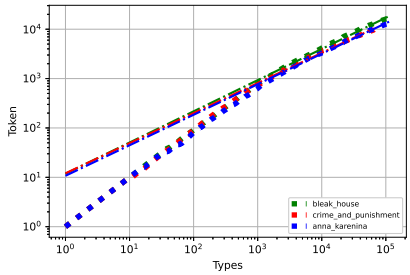


Fig. 5: Slop (-.), line (–)

Table 4: Authors and their books

Author	Book name	Slop	Intercept
Charles Dickens	Bleak House	0.632	1.072
	David Caperfield	0.642	0.993
	Our Mutual Friend	0.643	1.025
Fyodor Dostoyevsky	Crime & Punishment	0.610	1.080
	The Brothers Karamazov	0.604	1.095
	The Devils	0.621	1.056
Leo Tolstoy	Anna Karenina	0.618	1.031
	Resurrection	0.637	0.967
	War & Peace	0.657	0.830

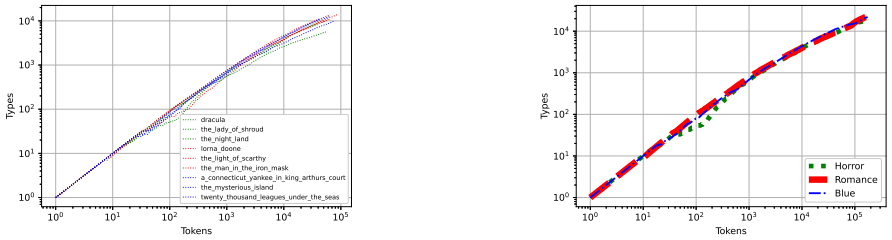
dominance of their style over the literature style. For instance, in horror genre, the slopes found for the same books that have the same author (Bram Stoker as the author) are pretty close to each other (0.700 and 0.736). However, the slope found for the book written with another author in the same genre, is far away from the other too books (0.565). The same is also applicable to scifi genre. The works in the romance genre however, have a pretty closer slopes compared to other genres.

Figure 7 shows the slope of three chosen books (one from each genre). The results are similar to the authors. In the next section, we will have a thorough analysis on this matter.

4.6 Part (k)

In this section, we are doing the analysis on the findings in the previous sections. We are asked to check if we can derive a method for clustering the authors' and genres' books. For this purpose, I utilize simple k-means algorithm on the slopes that I found on the previous section.

I used the k-means library of scikit-learn for this purpose (Check python file *HW1_kmeans.py*). The results are shown in Figure 9. The two figures 9.b and 9.d are the real clusters. The clustering results of k-means are shown in figures 9.a and 9.c. As I mentioned in the previous section. We have an easier path to cluster the books based on their authors compared to that of various literature. This shows that the style and the vocabulary used by authors has more dominance in forming the type-token relationship, and it also plays an important role in finding the style of writing a certain writer rather than identifying its genre.



(a) Log-log plot of each book (b) Loglog plot of all books of each authors concatenated

Fig. 6: log-log plots for token vs. type. In (a), green, red, and blue colors are for Dickens, Dostoyevsky, and Tolstoy respectively

Table 5: Literature, Author and Books'details

Literature	Author	Book	Slope	Intercept
Horror	Bram Stoker	Dracula	0.700	0.762
	Bram Stoker	The Lady of Shroud	0.736	0.604
	William Hope Hodgson	The Night Land	0.565	1.112
Romance	R. D. Bleakmore	Lorna Doone	0.635	1.027
	Egerton Castle	The Light of Scarthey	0.680	0.912
	Alexandre Dumas	The Man in The Iron Mask	0.684	0.797
Scifi	Mark Twain	A Connecticut Yankee in King Arthur's Court	0.751	0.609
	Jules Verne	The Mysterious Island	0.619	1.000
	Jules Verne	Twenty Thousand Leagues Under the Seas	0.675	0.900

As we can see in Table 6, I added two new authors to the list. I choose one woman author and one philosophical author and the results of my clustering based on the slopes earned for five authors is in Figure 9. The results show that adding more authors makes the clustering method more confused and the results are worse than that of three authors. The reason is that the features that we use for clustering in this part, are not enough for determining the cluster of authors. However, we can still use it as a feature in our clustering approaches alongside with some other features.

Table 6: Authors and their books

Author	Book name	Slop	Intercept
Hegel	Hegel's Philosophy of Mind	0.701	0.746
	Philosophy of Fine Arts Vol.1	0.659	0.828
	Science of Logic	0.663	0.793
Charlotte Bronte	Jane Eyre	0.681	0.898
	Shirley	0.668	0.984
	Vilette	0.716	0.775

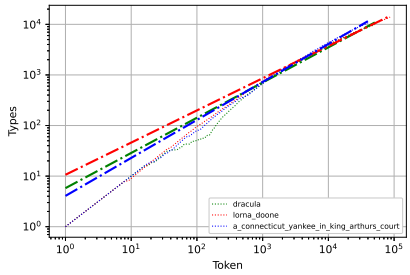


Fig. 7: Slop (-.), line (–) of three literatures

Table 7: Authors and their books

Author	Book name	Slop	Intercept
Charles Dickens	Bleak House	0.566	1.134
	David Caperfield	0.585	0.991
	Our Mutual Friend	0.573	1.108
Fyodor Dostoyevsky	Crime & Punishment	0.541	1.163
	The Brothers Karamazov	0.531	1.211
	The Devils	0.552	1.139
Leo Tolstoy	Anna Karenina	0.563	1.050
	Resurrection	0.564	1.074
	War & Peace	0.580	0.987

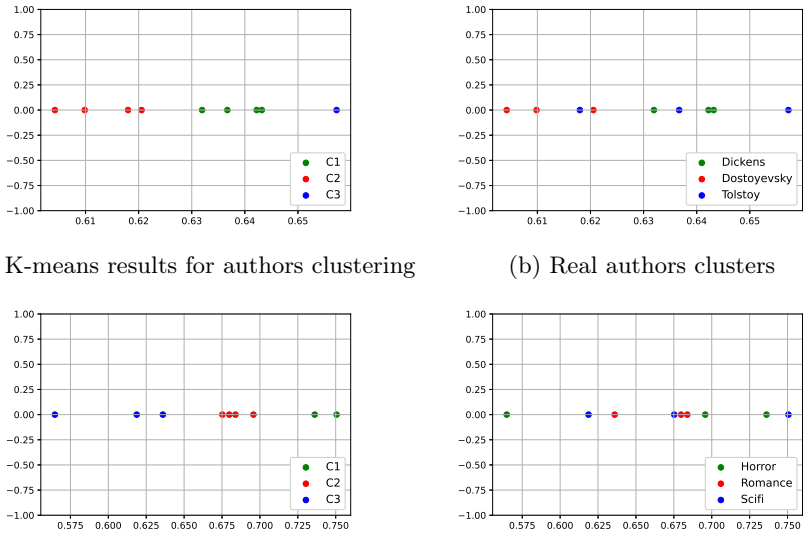
4.7 Part (l)

I used the corpora of three authors (Dickens, Dotoyevsky, Tolstoy) for analysis of this part. I used the corpus with all stopword included. Then I generated the same plots like the previous sections. I also calculated the slope and applied the k-means clustering one more time on the new version which includes the stopwords.

As we can see in Figure 10, compared to Figure 2, the frequency in y-axis is almost multiplied by seven. There is also a slight change in the x-axis. It is obvious that the number of ranked elements should increase by including the stopwords. We can see that for instance in Dickens' plot there is a slight increase in ranks, compared to the ranks of words with removing stop words.

Also the log-log results for each author separately is shown in Figure 11. Again the values in x and y axis are slightly higher compared to Figure 1. Also the log-log results before the stopword removal has less steep points and is much like a straight line. This is exactly in harmony with my expectations. Since removal of stopwords and their frequencies may change the shape of the loglog file in exactly the same way as our observations.

In Figure 12, we can see that number of word types is slightly increased. Also the number of tokens is increased by a large margin. Since we have removed lots of stopwords in Figure 3, and it is the obvious consequence of this stopword removal. Also the slope of the plot in Figure 12 is much less than that of Figure 3.



(a) K-means results for authors clustering

(b) Real authors clusters

(c) K-means results for genre clustering

(d) Real genre clusters

Fig. 8: Clustering results for three authors and three literature type based on the slope of their loglog plot.

I also calculated the slope of the corpora related to different authors. The results are demonstrated in Table 7. By comparing the results in this table with that of Table 4, we can see that the slope of the log-log plot is drastically decreased.

The clustering results based on the slopes of the log-log plots before removal is demonstrated in Figure 13. Comparing the results in Figure 9 and 13, results in interesting outcomes. If we score the clustering results based on the $\frac{\text{No. of correctly clustered books}}{\text{Total No. of books}}$, then we have the following results for three authors *Dickens, Dostoyevsky, Tolstoy*.

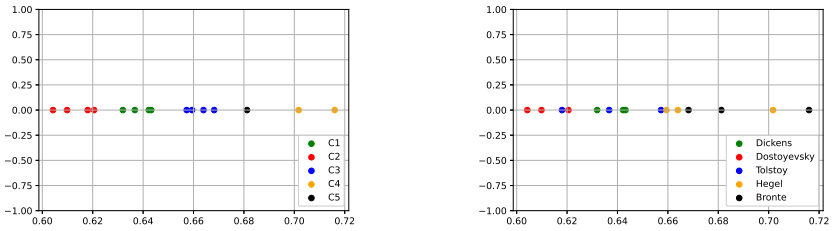
$$\textit{Correctly clustered}_{before} = \frac{4}{9} \tag{2}$$

$$\textit{Correctly clustered}_{after} = \frac{7}{9} \tag{3}$$

The results in Eq. 2 and 3 shows that the number of correctly clustered books before removing stopwords is less than that of after removing the stopwords. This results stresses that removing stopwords is an important part of analyzing any textual data.

4.8 Part (m)

In [3], the authors do some experiments to see if the Zipf's law is the result of the styles of the texts generated with natural language, or it has something to do with statistical process in complicated systems. The results in this paper strongly confirms that the Zipf's law that is observed in many corpora is not the property of natural language and has something to do with the statistical process of complex systems.



(a) K-means results for five authors clustering (b) Real clusters of five author

Fig. 9: Clustering results for three five authors based on the slope of their loglog plot.

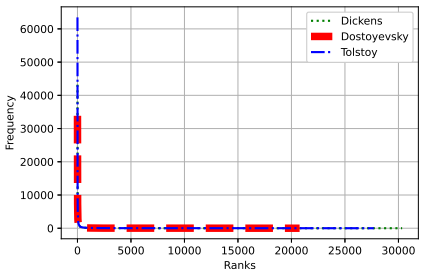
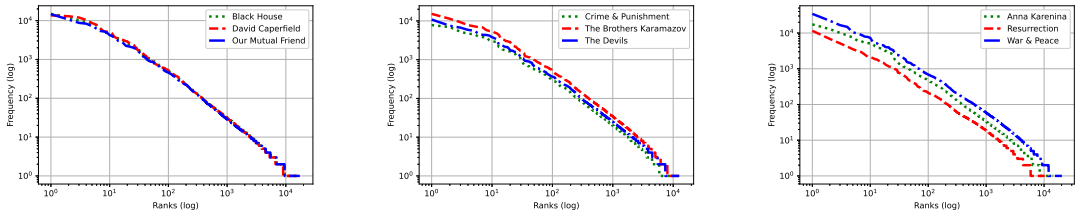


Fig. 10: Frequency-rank plot for three authors before stopword removal

To do further study on this important law, we do a similar experiment. For this purpose, I generated a random text. First, I concatenated the text of all of the corpora of three authors discussed in previous sections (Dickens, Dostoyevsky, Tolstoy). Then using the Numpy's random number generator, I randomly selected words from this large corpora, and made a new random corpora with **218,246** tokens (excluding the stopwords). We can see the plots for frequency-rank in Figure14. Our results is in harmony with the results observed in [3], and confirms that the Zipf's law is also present in a randomly generated corpora and it is not necessarily a property of natural language (Check *HW1_random_text_generator.py* for the code).

5 Discussions and Conclusions

The detailed interpretation of my results is provided separately under each subsection. Most of the results that I observed in these set of experiments were the results that I was expecting. From Zipf's law to the effect of stopword removal in analyzing the text were proved to be true in my experiments. The parts that I might consider it as a failure are mostly related to (part k) were we are asked to derive a clustering methodology. I was expecting that the slopes that we found in the previous parts, would be a determining factor in separating the books to various genres. However, this wasn't the case for the genre separation. However the results in this part were helpful in separating the the books of various authors, and with only single feature, we were able to have an accuracy of 0.77% in this task.



(a) (b) (c)

Fig. 11: Log-log plot of freq. vs. rank for the authors and their books before removing stopwords. a) Dickens, b) Dostoyevsky, c) Tolstoy

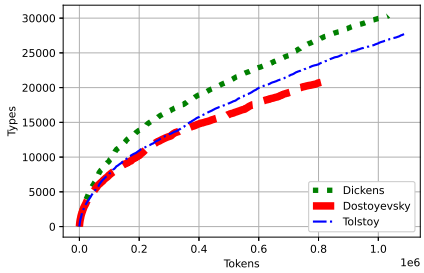


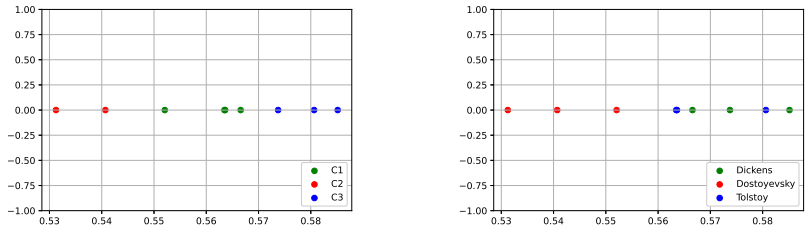
Fig. 12: Frequency-rank plot for three authors before stopwords removal

During the assignment, I learned a lot from preprocessing the text to analyzing various aspects of a text collection based on the things that we learned in class. Implementing these ideas from scratch was another important thing that I learned. In term of the concepts that I learned from this assignment, I can mention the following ones:

- Any text corpora, even a randomly generated one, follows the Zipf's law.
- The type-token relation is a really important factor for any corpora, and while used alongside with other features of a corpora, could play a pivotal role for natural language analysis.
- Removal of stopwords for any corpora is not just a routine task without any reason. Including them as part of our analyzed corpus can adversely affect our analysis.

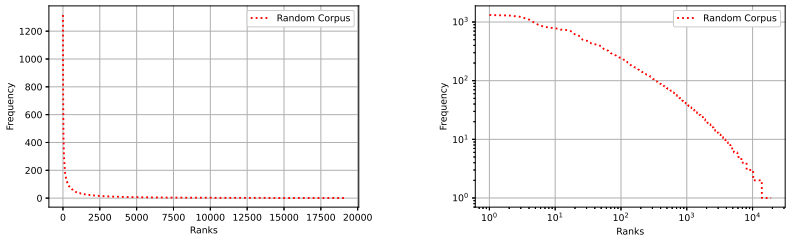
References

1. Cheng, Z., Liu, Y.: A novel method for studying 3d-manet capacity based on zipf's law. In: 2022 3rd Information Communication Technologies Conference (ICTC). pp. 14–18. IEEE (2022)
2. Hou, Z., Wang, D.: New observations on zipf's law in passwords. IEEE Transactions on Information Forensics and Security **18**, 517–532 (2022)
3. Li, W.: Random texts exhibit zipf's-law-like word frequency distribution. IEEE Transactions on information theory **38**(6), 1842–1845 (1992)
4. Linders, G.M., Louwerse, M.M.: Zipf's law revisited: Spoken dialog, linguistic units, parameters, and the principle of least effort. Psychonomic Bulletin & Review pp. 1–25 (2022)



(a) K-means results for authors clustering (b) Real authors clusters

Fig. 13: Clustering results for three authors before stopword removal based on the slope of their log-log plot.



(a) Normal (b) Log-log

Fig. 14: Frequency-rank normal and log-log plot for the randomly generated corpus.

5. Piantadosi, S.T.: Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic Bulletin & Review **21**(5), 1112–1130 (Oct 2014). https://doi.org/10.3758/s13423-014-0585-6, `https://doi.org/10.3758/s13423-014-0585-6`
6. Zipf, G.K.: Selected studies of the principle of relative frequency in language (1932)