Lab 1 Report
Zipf's law

Dominik Cedro
Complex systems Labs

16th October 2024
Group: Fridays 3:15pm-6pm

1 Introduction

Zipf's law is an empirical law stating that when a list of measured values is sorted in decreasing order, the value of the nth entry is often approximately inversely proportional to n.
In this case we focused on example of Zipf's law occurence in languages. Books downloaded via ProjectGutenberg website were the datasource. In finished followin task during this exercise:

- Analysis of Zipfs law in english literature accessed via Project Gutenberg website.
- Comparing empirical distribution from my analysis to theoritical Zipf distribution.
- Fitting the a and b: constants in Zipf-Mandelbrot law for different languages.
- Checking if LLM generated text is following Zipf-Mandelbrot equation for given language.

2 Method

To perform analysis I used Python 3.11, Link to the repository is in the last section of this report.
Firstly in "*tokenization_fileprep.py*" files (should be "*.txt*" format) are tokenized. I create DataFrame objects with columns:

- "rank" : Rank of the word. The most common one is 1.
- "word" : Word of current rank.
- "count" : Number of occurences that each word provides.
- "freq" : Frequency of the word, calculated as freq = $\frac{rank}{totalcountofwords}$

Then each task is solved individual "*.py*" file. For custom book selection user must edit contents of the lists on top of the files. Plots are stored in "*plot_output*" directory.

- Task 2 : With help of already created csv files I perform analysis of Zipf's law phenomena in each book (represented by each file). I compare it to theoritical Zipf's distribution calculated in the script. I plot the results, they will be shown in "Results" section.
- Task 3 : I decided to analyze 6 languages: Dutch, Esperanto, English, Spanish, Russian and Hebrew. I created a function that represents Zipf-Mandelbrot equation. With help of "SciPy" library I fit this function based on each language for parameters "a" and "b". Results are presented on graphs, respective parameters can be compared.
- Task 4 : I chose "Chat GPT 4.0" as my LLM of choice. I generated 1000-word long stories in 4 of previously discussed language. Then I utilize functions from previous script to analyze a and b parameters. Similary these values are then compared, but this time also with previous "natural" parameters obtained from real books in each language.

3 Results

3.1 Tasks 1,2 - Zipf comparison.

Plots present strong argument for Zipf's law occurence in english literature of choice. In linear scale empirical and theoritical functions are overlapping. Log-log scale is more clear, showing distrubution of words. Very few words represent high frequency and a big number of words represent very small frequency.

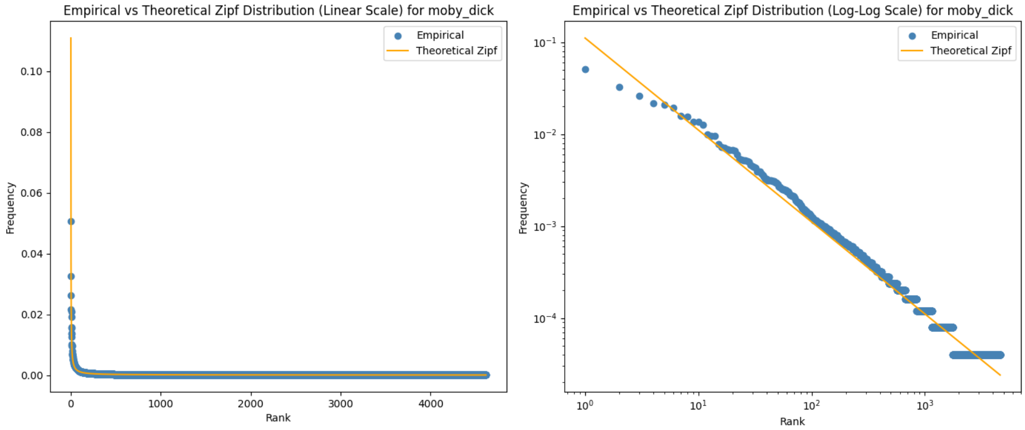

Figure 1: Results in log-log and linear scale for "Moby Dick".

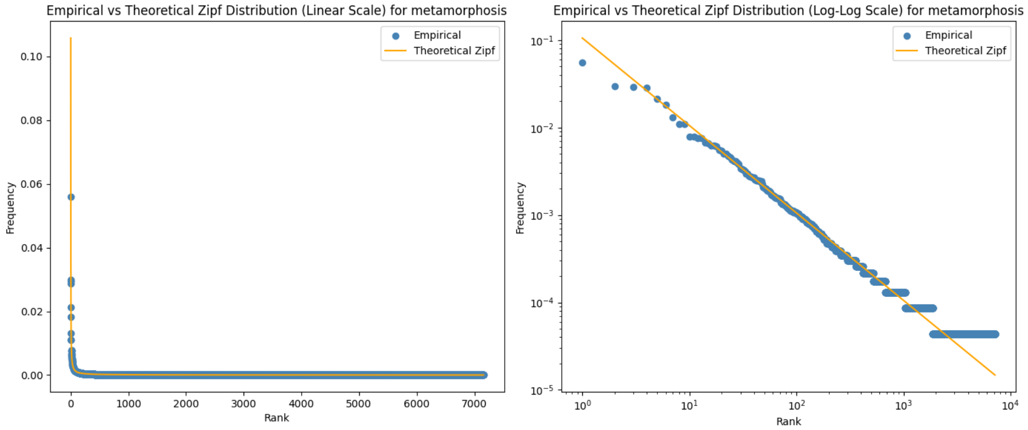

Figure 2: Results in log-log and linear scale for "Metamorphosis".

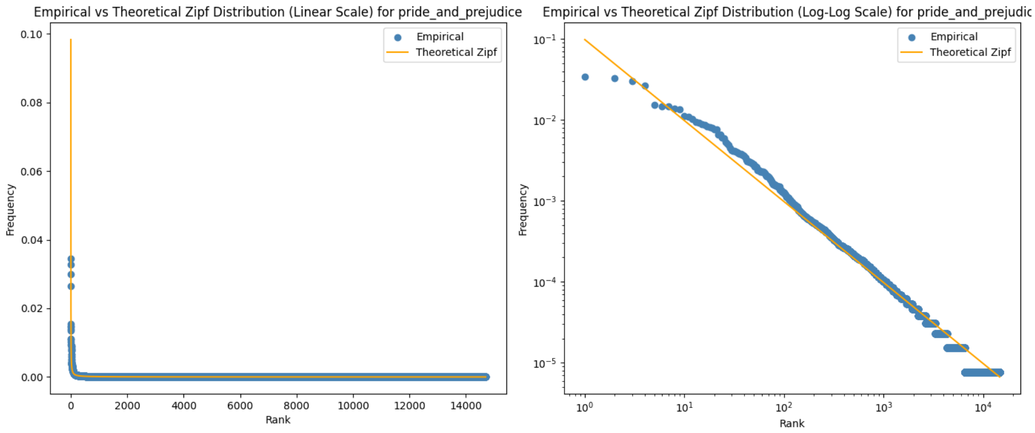

Figure 3: Results in log-log and linear scale for ”Pride and prejudice”.

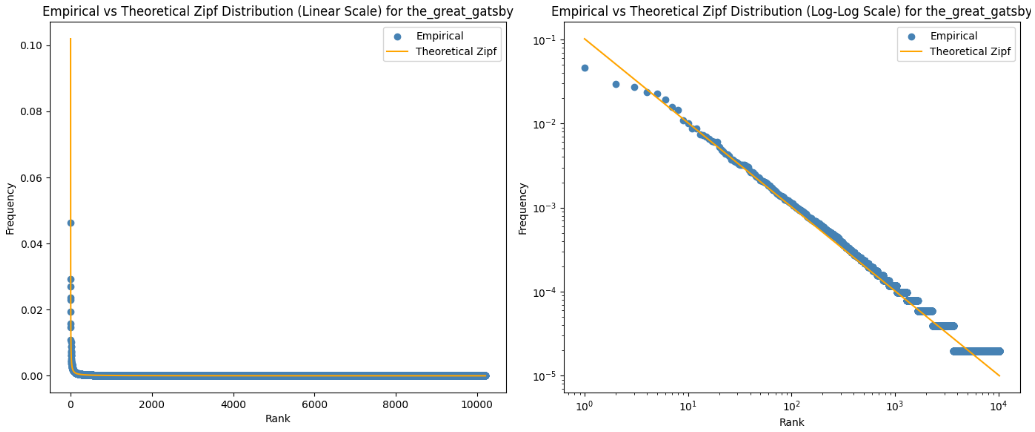

Figure 4: Results in log-log and linear scale for ”The Great Gatsby”

3.2 Task 3 - Different languagues.

Plot shows different values of parameter pairs for different languages. An interesting observation can be done comparing esperanto and danish, which have relatively close parameters a and b. English is much more distinct.

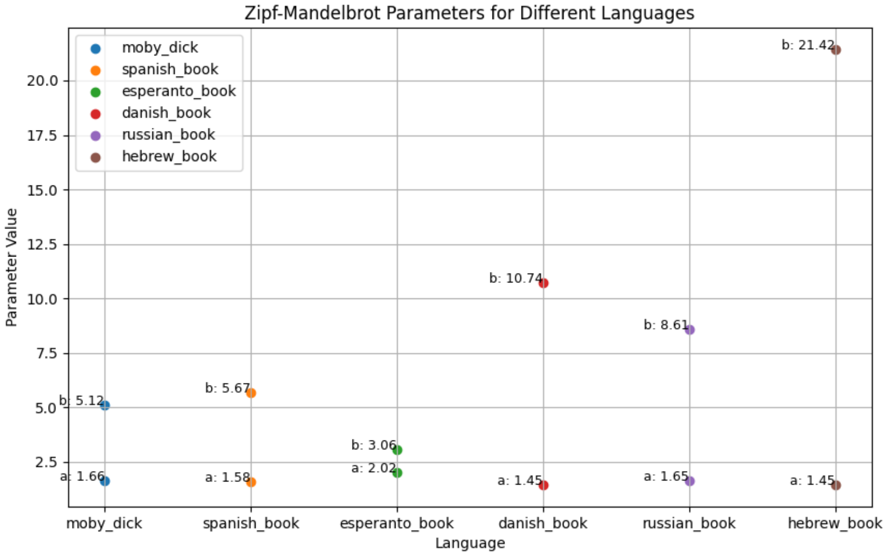

Figure 5: Results for different languages

3.3 Task 4 - LLM generated language.

Plot shows that LLM provided text show similar results for Zipf-Mandelbrot equation for Danish. Other languages are not very well represented compared to the real book results.

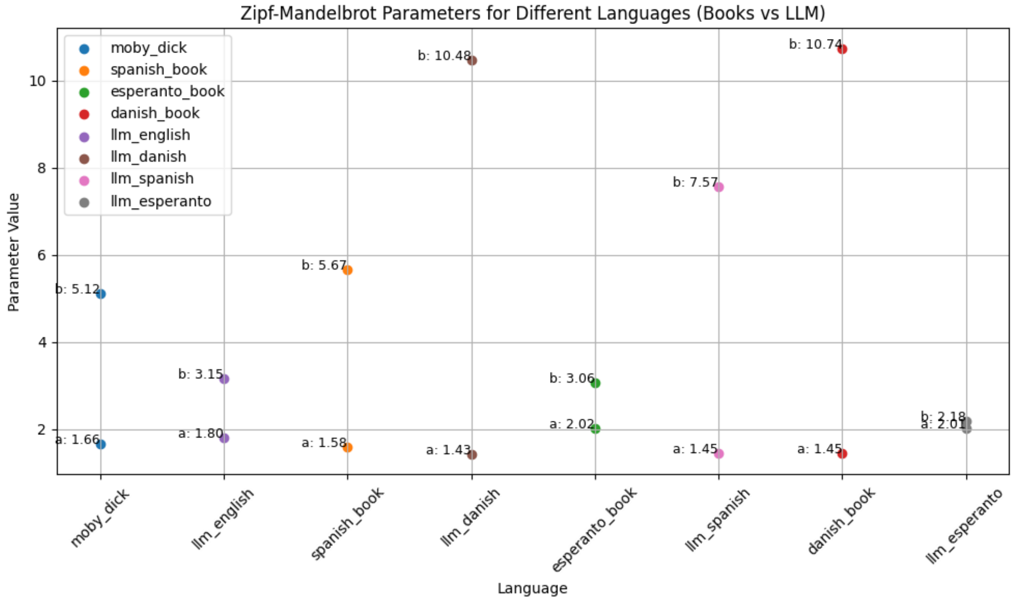

Figure 6: Results for text in different languages generated by LLM.

4 Conclusion

4.1 Tasks 1,2

Zipf's distribution is clearly visible, therefore the occurence of this phenomena is proven. All english books showed similar results.

4.2 Task 3

Parameters differ among languages therefore it is possible to distinguish them based on a and b solely. An interesting observation can be done comparing germanic languages to hebrew or russian.

4.3 Task 4

Plot shows, that "Chat GPT 4.0" generated danish text that follows a and b parameters. In case of other languages resemblence is not present. Therefore I conclude, that LLMs have great potential in generating text that follows Zipf-Mandelbrot law, but only in their respective specialization. In my case prompts were provided in danish, maybe that would be the reason for good performance in this language.

Repository and resources

- Article on Zipf's law
- My github repository for this laboratory. Lab1 solutions are in ”lab1” folder.