Zipf's Law

Complex Systems Laboratory Project 1

Michał Skrzypczyński

October 2024

1 Introduction

1.1 Zipf's law formula

$$f(r) = \frac{1}{(r+b)^a} \tag{1}$$

For English language: $a \approx 1.0$ and $b \approx 2.7$.

1.2 Zipf's law in language

A hierarchy of words' frequency can be noticed in human languages. It is described by upper formula. Following analysis was done in order to check whether the **Zipf's distribution is in fact true**. For this manner 4 books were chosen written by Jane Austen in 19^{th} century.

Later a following hypothesis was put to a test: **Languages can be distinguished by peculiar** a **and** b **parameters**. In order to prove it, books from different languages were analysed: 'Emma' by Jane Austen and 'Pan Tadeusz' by Adam Mickiewicz. Both come from the same literary era. After optimization of parameters one could compare them and state a prove or negation of stated hypothesis.

2 Method

2.1 Preparing data and plotting

Extracting each word from text file of a book. *read()* method was used from *typing.IO* class in python then *re* library in order to split and tokenize words. Frequency values were normalized in a following way, for each frequency value only one rank value was assigned (so if frequency values of different words were the same, only one rank value was defined to these words).

Following equation was used to calculate model values (theoretical values of Zipf's distribution).

$$f(r) = \frac{0.1}{r} \tag{2}$$

Calculating it with following formula yield not accurate data:

$$f(r) = \frac{\left(\sum_{r=1}^{r_{max}} 1/r\right)^{-1}}{r} \tag{3}$$

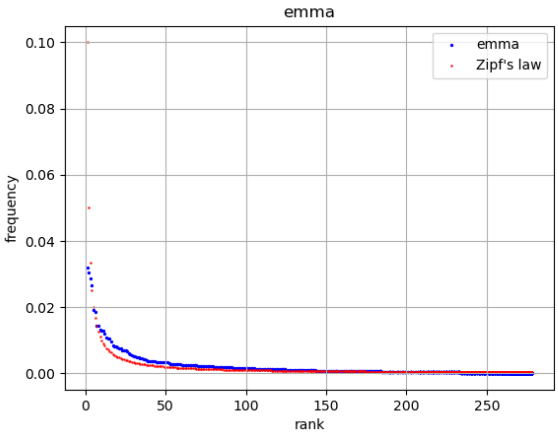
Next the data was put to *pandas.DataFrame* structure and plotted with *matplotlib* library.

2.2 Fitting the function

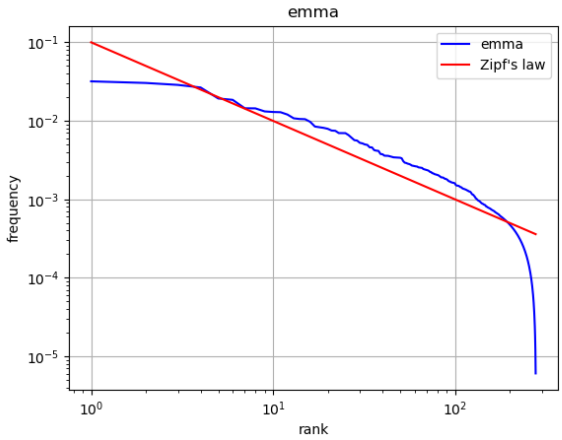
In order to find **optimal** a and b **parameters** (from *formula 2*) *scipy.optimize* library was used and following function: *curve_fit*. Later these values were used to prove the hypothesis.

3 Results

3.1 Plots

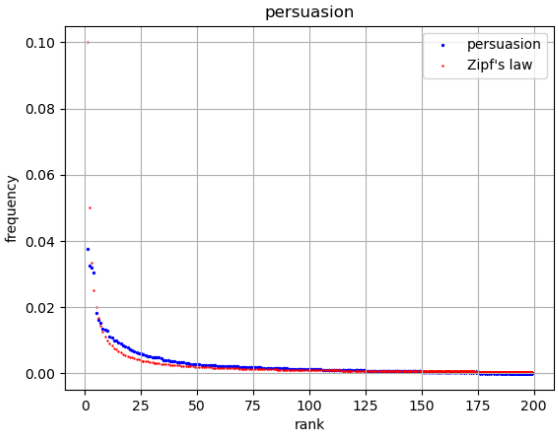


(a) Linear plot for 'Emma'

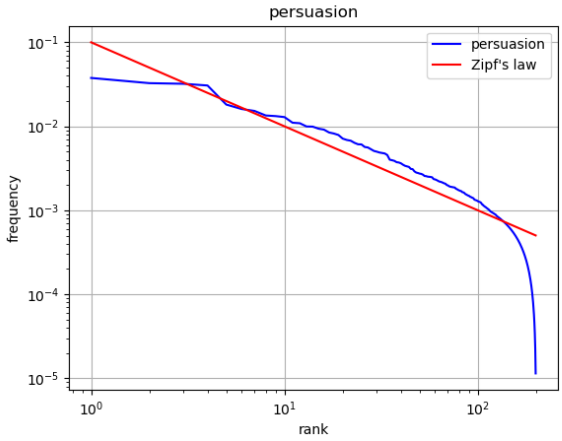


(b) Logarithmic plot for 'Emma'

Figure 1: 'Emma' plots

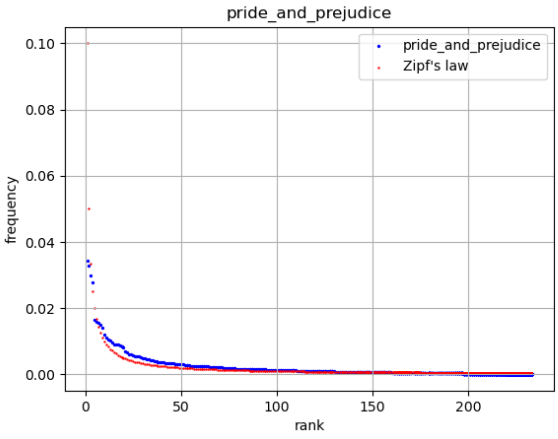


(a) Linear plot for 'Persuasion'

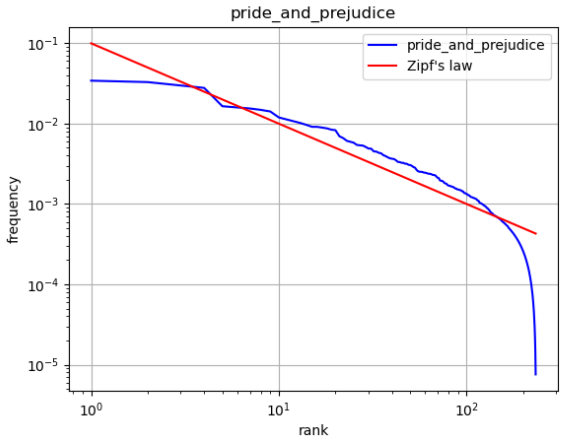


(b) Logarithmic plot for 'Persuasion'

Figure 2: 'Persuasion' plots

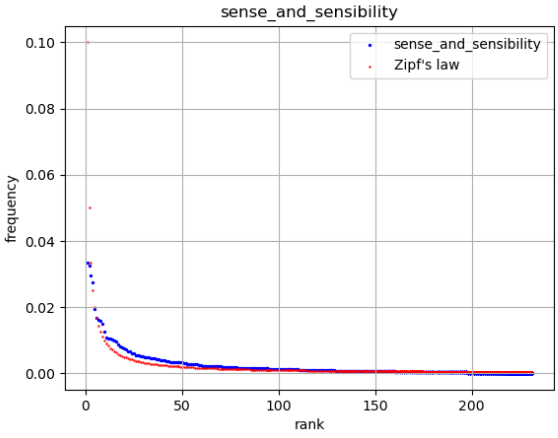


(a) Linear plot for 'Pride and Prejudice'

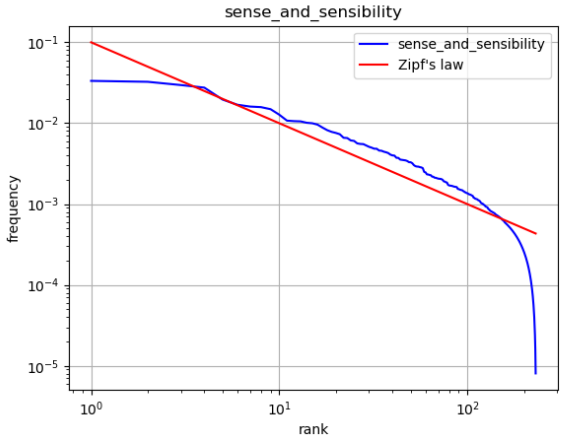


(b) Logarithmic plot for 'Pride and Prejudice'

Figure 3: 'Pride and Prejudice' plots



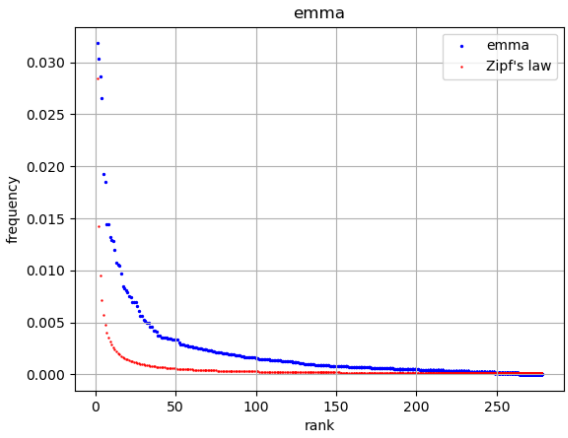
(a) Linear plot for 'Sense and Sensibility'



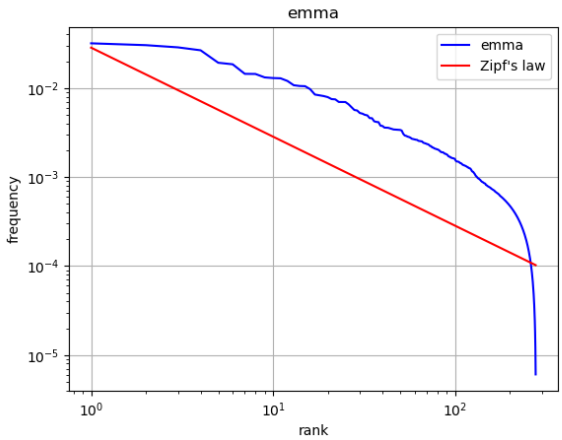
(b) Logarithmic plot for 'Sense and Sensibility'

Figure 4: 'Sense and Sensibility' plots

Exemplary plots obtained with use of *formula 3*:



(a) Linear plot for 'Emma' with *formula 3* used



(b) Logarithmic plot for 'Emma' with *formula 3* used

Figure 5: 'Emma' plots with *formula 3* used

3.2 Optimal parameters for two languages

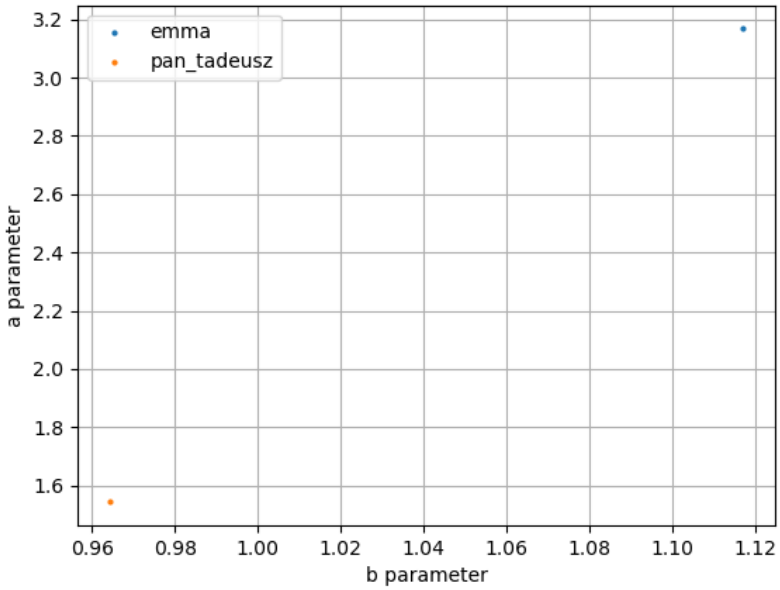


Figure 6: a and b parameters for Polish and English

4 Conclusions

4.1 Zipf's distribution

Looking at diagrams one could notice that middle parts of functions overlay with theoretical values (calculated according to *formula 2*). Thus it can be stated that **Zipf's law is truly relevant**.

4.2 Language peculiar parameters

We can see that a and b parameters differ greatly for **Polish** and **English**. For Polish: $a \approx 1.54$, $b \approx 0.96$, for English: $a \approx 3.16$, $b \approx 1.11$. In Wikipedia following values for English are presented: $a \approx 1.0$ and $b \approx 2.7$. One cannot say that they are very close to ones obtained by me, yet the ratios of $\frac{a}{b}$ are as follows: (English) my results — 2.84 and Wikipedia's — 2.7. For Polish language ratio is: 1.6. Thus, a great difference in ratio and values of parameters of two different languages can be noticed and **one language be can distinguished from another** by peculiar a and b parameters. *Quod erat demonstrandum.*

5 Repository

6 Bibliography

https://www.pythontutorial.net/python-basics/python-read-text-file/
https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curv_fit.html