Zipf's Law

Zipf's Law is a power-law distribution that states that the frequency of any element is inversely proportional to its rank.

Mathematically it can be expressed as:

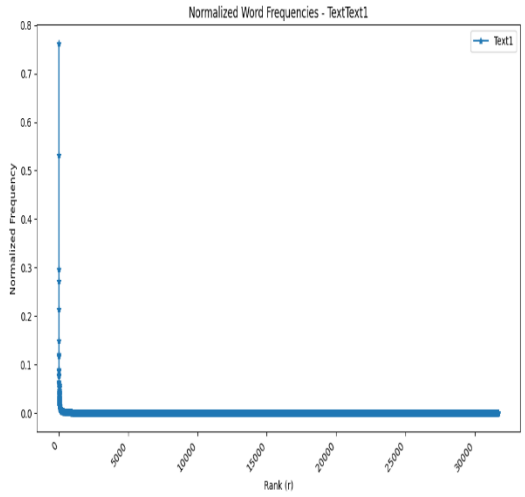
$$\text{frequency } \alpha \frac{1}{(rank+b)^a}$$

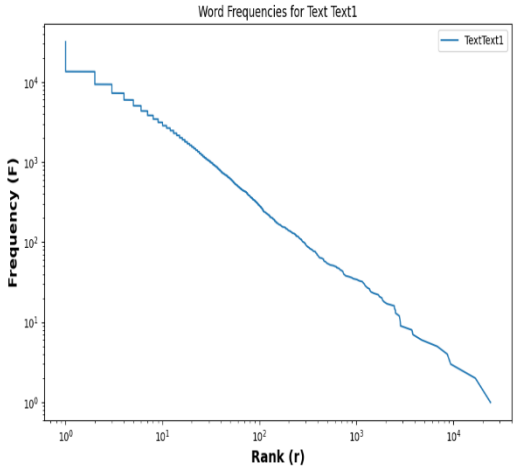
where a and b are the fitted parameters.

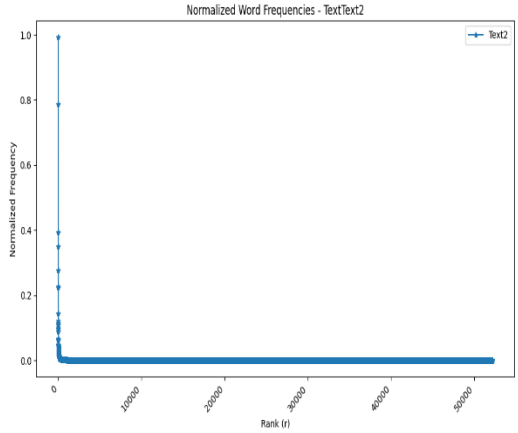
In this report I apply the Zipf law on four different books using Python and calculate the frequency and rank of the word and plot the graph between the frequency and rank of the words in the text.

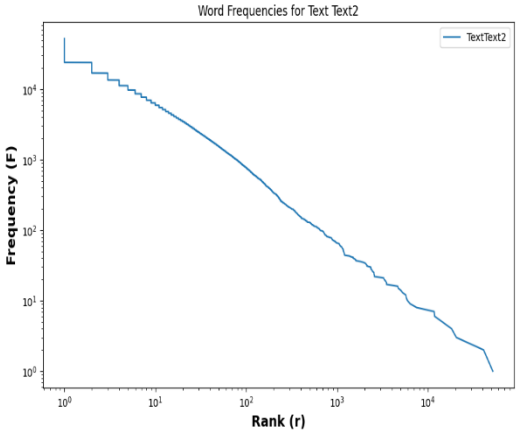
As we can see top five word with high frequency is Counter({'the': 24154, 'of': 16877, 'and': 9392, 'to': 8658, 'in': 6812}), total no of word in each text is 31679, 52188, 40029, 39750.

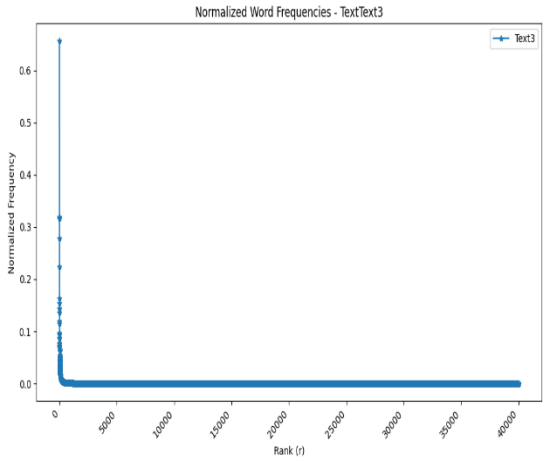
The graph between the frequency and rank is given below for all four texts.

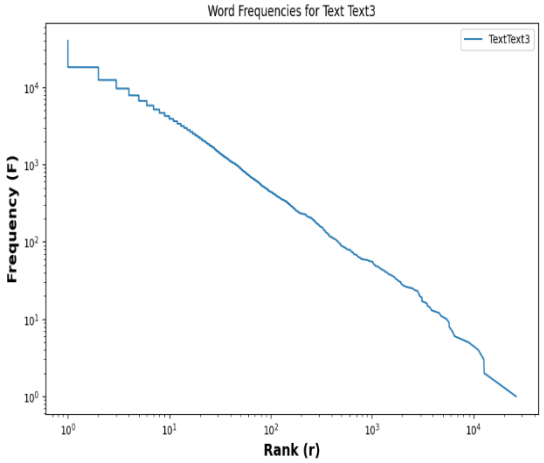


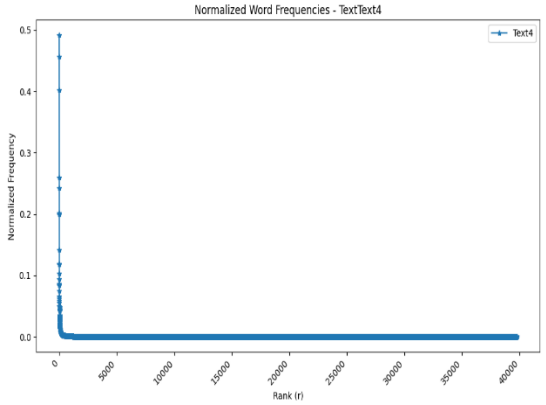


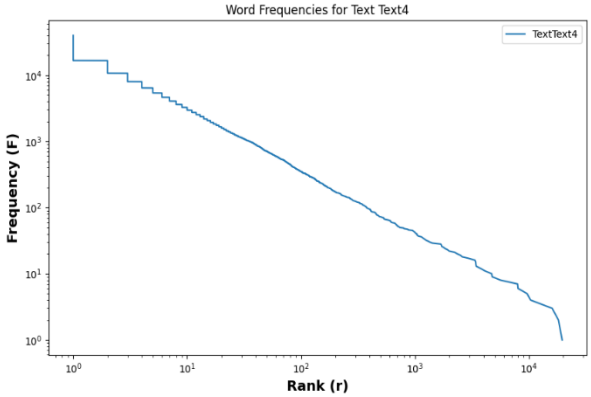




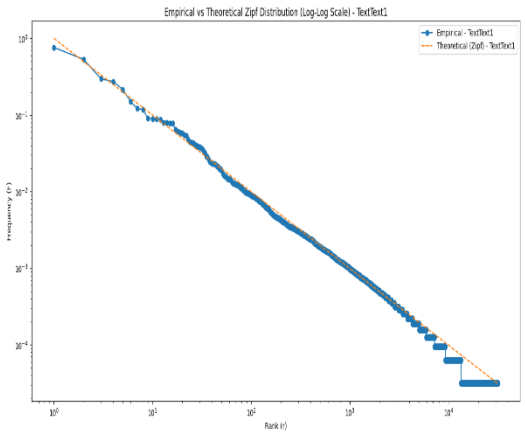


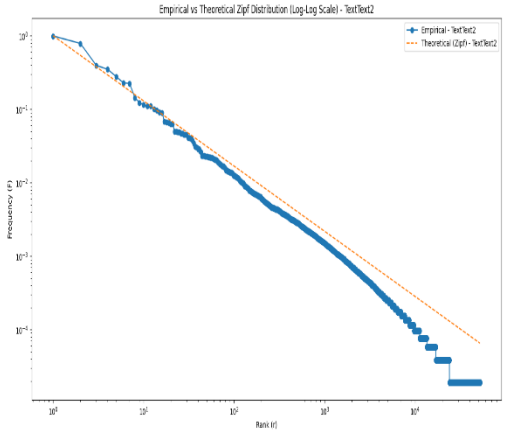


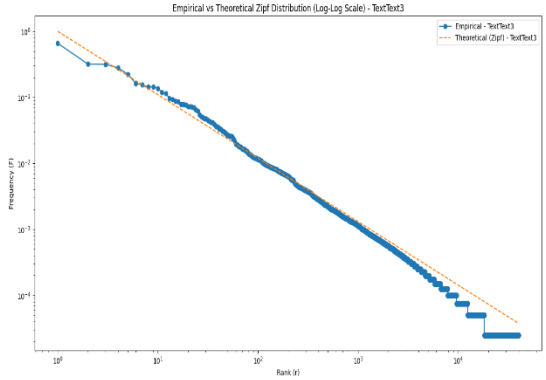


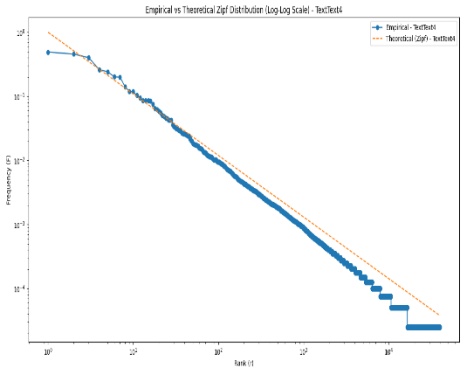


The Comparison between Empirical and Theoretical distribution is given below:







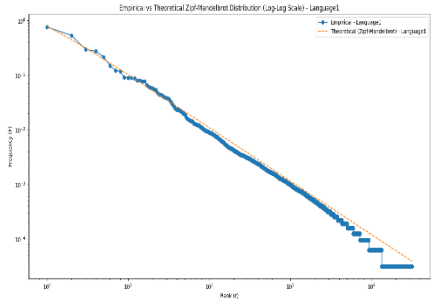


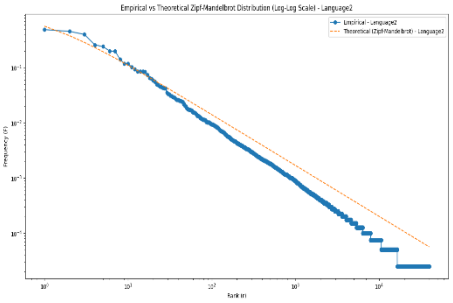
To fit ‘a’ and ‘b’ constants in Zipf-Mandelbrot law freq. ∝ 1/(rank + b) a for two different languages

Language1 - Fitted parameters (a, b): [0.97968575 0.27351501]

Language2 - Fitted parameters (a, b): [0.9233879 0.84764076]

And the graph are below:





Sometime it is not possible to classify language based on a, b only. Because these value are not fixed.

For Checking LLM generate text I used ChatGPT and generate random text in Polish language the graph of fit is given below: