Complex Systems — Problems for lab

Fall 2024/2025

---

2 Zipf law

The topic of this project is Zipf's phenomenological law applied to linguistics. **The tasks should be finished till 07.11.2024**.

Visit the Project Gutenberg website (`https://www.gutenberg.org/`), which is a library of over 70 000 free eBooks. Download four eBooks of your choice and conduct the Zipf analysis, and then:

1. Save the result of the analysis in the text file. The name of the file should include the title of the book and the number of words it contains. The file should contain four columns: (1) rank, (2) word, (3) number of a given word in the text, (4) frequency of a given word in the text. Why do I need both columns (3) and (4)? Think about it.

2. Compare the empirical distribution from your analysis with the theoretical Zipf distribution by plotting both on the same graph in log-log scale (one graph for each book). Try to do the same in linear scale. What do you observe?

3. Try to fit a and b constants in Zipf-Mandelbrot law freq. $\propto 1/(\text{rank}+b)^a$ for two different languages. Is it possible to distinguish languages using a and b parameters only?

4. (extra) Check if LLM (large language models) generate text following Zipf-Mandelbrot law for a given language.