Word Length Analysis

Instructions

You may use any programming language to do this assignment using the dataset provided below. Please submit the following:

1. A report answering the questions mentioned below.
2. Your code in a tarball archive.

Datasets

First, download the English data from the links provided:

- `http://www.gutenberg.org/cache/epub/10/pg10.txt`
- `http://www.gutenberg.org/cache/epub/35997/pg35997.txt`

Task

Combine these two datasets and convert all words in them into lowercase. For each non-punctuation word in these datasets, calculate the following:

1. Measure word length in terms of number of letters (2 marks).
2. Calculate the number of words at different word lengths (5 marks).
3. What are the shortest words in your dataset? Comment on these words (5 marks).
4. Plot a graph with length on the X-axis and frequency on the Y-axis (3 marks).
5. Plot a graph with $\log_{10}$(word length) on the X-axis and $\log_{10}$(frequency) on the Y-axis (3 marks).
6. Calculate Pearson's coefficient of correlation between length and frequency (2 marks).
7. Write a short note on: "Are word lengths optimized for efficient communication?" (5 marks). Please connect your answer to the following research paper:

   Word lengths are optimized for efficient communication by Steven T. Piantadosi, Harry Tily, and Edward Gibson.
   `https://www.pnas.org/doi/10.1073/pnas.1012551108`