Assignment 4: Word Length

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

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

Submission deadline: **11:55pm, 24 April**

First download English data via the links provided below:

http://www.gutenberg.org/cache/epub/10/pg10.txt

http://www.gutenberg.org/cache/epub/35997/pg35997.txt

Combine these 2 datasets and convert all words in the above datasets into lower case. For each non-punctuation word in these datasets, calculate the following:

* Measure word length in terms of number of letters.

* Calculate the number of words at different word lengths

* What are the shortest words in your dataset? Comment on these words.

* Plot a graph with length on the X-axis and Frequency on the Y-axis.

* Plot a graph with log10(word length) on the X-axis and log10(Frequency) on the Y-axis.

* What is the Pearson's coefficient of correlation between length and frequency?

* Write a short note on the question: “Are word lengths optmized for efficient communication?”

Please connect your answer to the paper “Word lengths are optimized for efficient communication” by Steven T. Piantadosi1, Harry Tily, and Edward Gibson