Assignment 2

You may use any programming language or tool(s) 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, March 25, 2016

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 letter and each non-punctuation word in these datasets, calculate the following:

* Calculate the frequency of occurrence of each word

* Rank the terms in descending order of frequency (ties do not matter)

* Assign rank 1 to the first term in the list (highest frequency)

* Assign ranks in ascending order to the rest of the list so that highest rank is lowest frequency

* Write down the five most frequent words. Comment on these words.

* Write down the five most frequent letters.

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

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

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

*Write a short note on the Zipf's law.

Data Processing

Converting all text to lower case can be achieved by unix tools:

tr '[:upper:]' '[:lower:]' < input.txt > output.txt

Even creating word frequency lists can be achieved using unix tools like “uniq” and “sort”. For tips, please refer to :

http://www.cs.upc.edu/~padro/Unixforpoets.pdf

For plots, you need to use something like gnuplot or even excel will work.