Lab Assignment 2

Please use python to do this assignment. Please submit the following:

1. A report answering the questions mentioned below
2. Your code and README file in a zip/tarball archive

Deadline: 9 pm, September 6 (Thursday)

Mode of submission: We will send a google form for submitting this assignment.

Data

First download English data (Jane Austen corpus) via the piazza link below:

http://www.piazza.com/class_profile/get_resource/jk9h3tq2wr57iy/jlg7i9dm5ic1cn

Unzip the Austen corpus and write a python program to perform the following operations on this dataset and submit a report documenting your findings.

Questions

For each letter and each non-punctuation word in this corpus, do the following steps:

* Convert all words in the above datasets to lower case.

* Calculate the frequency of occurrence of each word as well as each letter after stripping off punctuation marks (Refer lecture slides for relevant commands).

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

* Assign rank 1 to the first item in each 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. Comment on these words.

* 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?

Plots

* For plots, you may use python or even something like gnuplot or excel. If you plan to use other softwares for generating plots, please print to file (Refer lecture slides for this) the required numerical values (X-axis and Y-axis data points).