# Zipf's Law Assignment

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 and README file in a tarball archive

A) First, download Mark Twain's novel "The Adventures of Tom Sawyer" from Project Gutenberg:

https://www.gutenberg.org/files/74/74-0.txt

First calculate the token-type ratio of the above text [2 marks].

B) After converting the text to lower case and excluding punctuation marks from the dataset, calculate the following for **each** of the *2 classes of words*: **Words, Letters [2x5 =10 marks]**

- ❑ Calculate the frequency of occurrence of each item. Rank them 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 the lowest frequency
- ❑ Write down the five most frequent items and comment on them.
- ❑ 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 Y-axis.
- ❑ What is the Pearson's coefficient of correlation between rank and frequency?

C) Write a short note on Zipf's law and discuss any patterns which emerged from the above data by using your best judgment and going beyond the numerical results. For example, since you have studied both words and letters, are there any interesting similarities (or differences) which you observe? Comments should be along these lines. **[5 marks]**

**For HSS600 students only [Extra work]:**

D) Quantify and demonstrate a relationship between word frequency and word length [5 marks].

E) Download a text of your own native language or Hindi corpus provided with this assignment. Repeat the entire exercise **only for words** and report your results.

**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 commands like "uniq" and "sort". For tips, please refer: http://www.cs.upc.edu/~padro/Unixforpoets.pdf

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