NLP ASSIGNMENT

Arushi Dogra
201302084

Programming Language : Python

Tool : Regex (for extracting the tokens) ,matplotlib (for plotting graphs)

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called *tokens* , perhaps at the same time throwing away certain characters, such as punctuation.

Count is also kept for the tokens and it is notices that there are certain words which have a large frequency and then there is a sudden decrease in frequency of words.

The frequency and rank of the tokens follow the Zipfs Law.

$$\text{Frequency} = 1/(\text{rank})$$

All the plots are in the respective folders.

The top words by frequency occuring in the english dataset are the function words which are commonly used in everday language and their meaning is not restricted to a real world object.

The words with median frequency , all have their frequency at one. This suggests that the frequency of unique words goes down rapidly initially . It is illustrated by the graph which follows Zipf's law.

English has more tokens because of more punctions and also due to a more varied dataset.

The median words appear to be carrying specific meaning . They are the content word.