CSCI-722 - Data Analytics Cognitive Comp(Assignment-1)

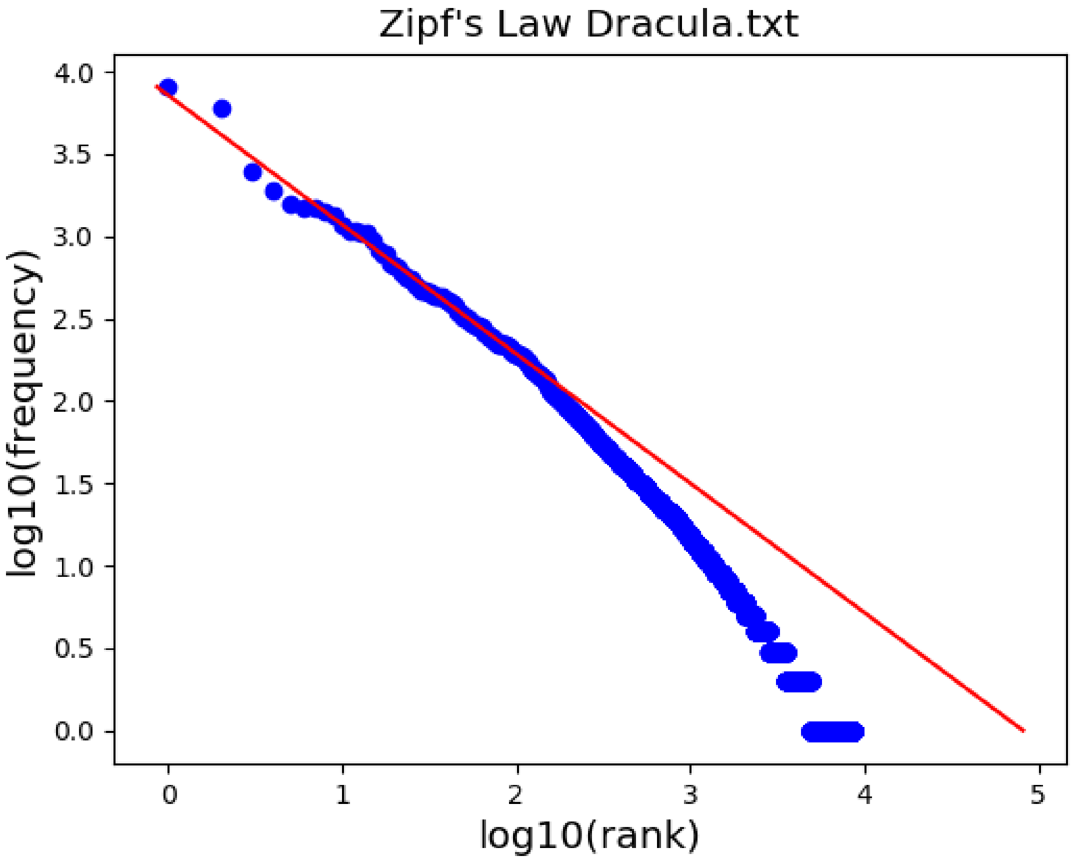
Name: Rajkumar Pillai

Q.1)

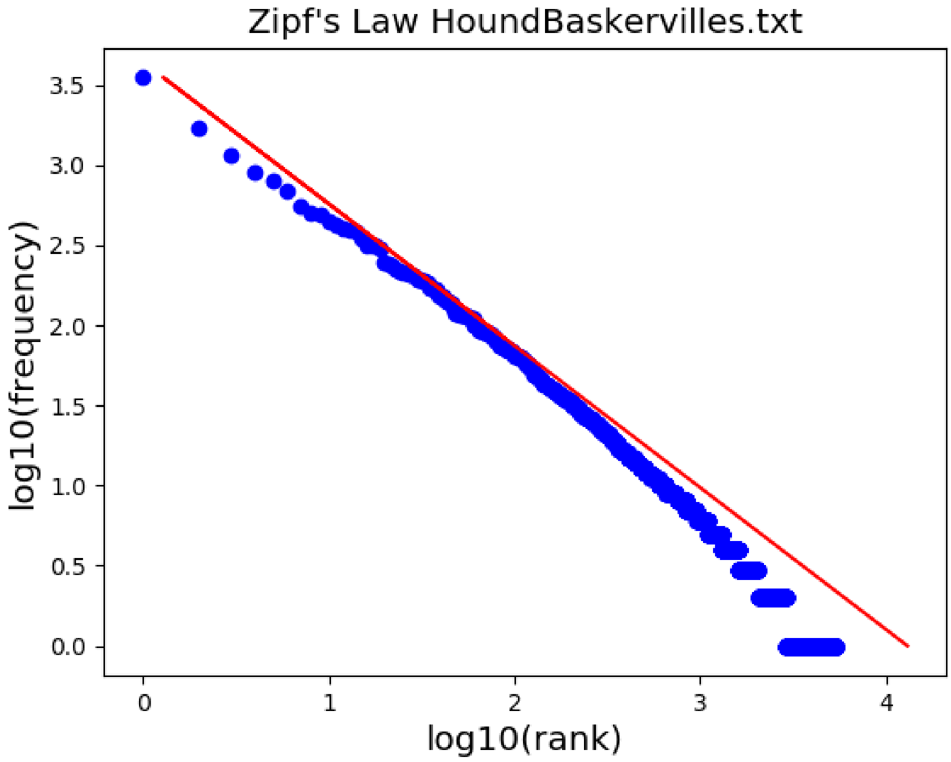
Solution:

Plots for Zipf's Law:

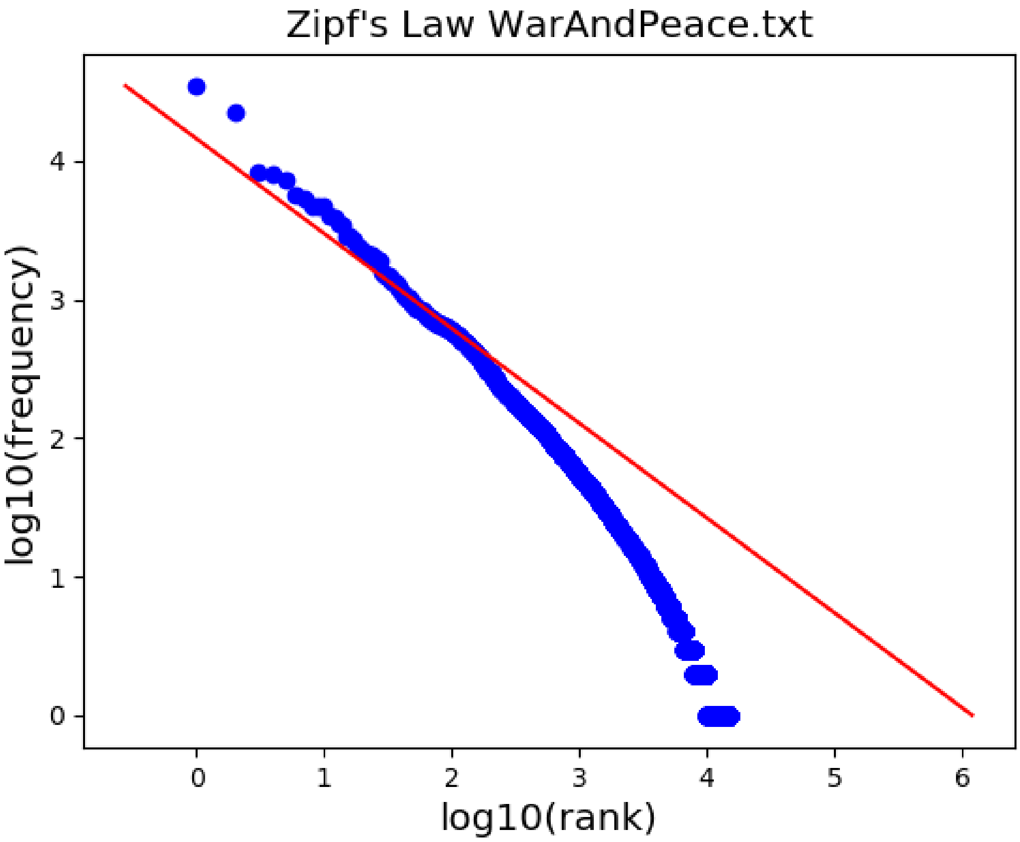
The red line indicates the ideal case for Zip's law and the blue dots represent the actual data.



From the above plot we can observe that, the Zipf's law does not hold true for low frequency words but holds true for most of the high frequency words.



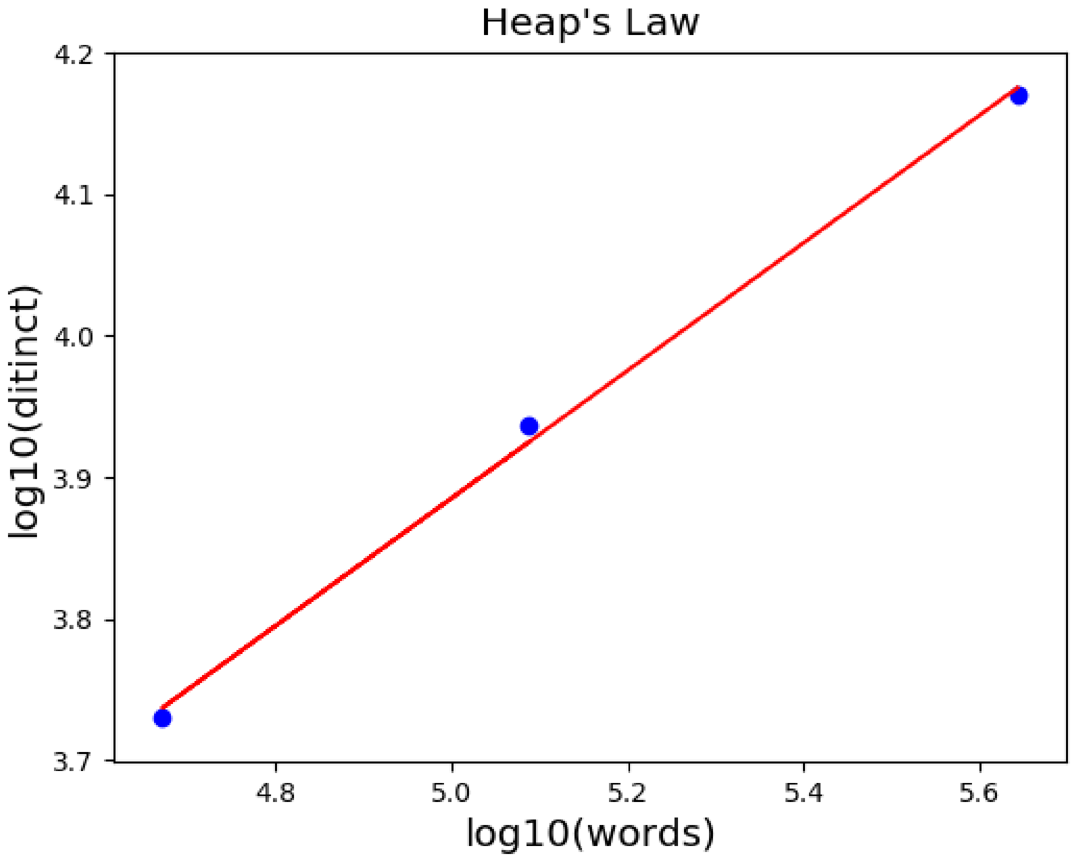
From the above plot we can observe that, the Zipf’s law does not hold true for very low frequency words but holds true for high frequency words and so the frequency of words in this text file are inversely proportional to their rank in frequency table.



From the above plot we can observe that, the Zipf's law does not hold true for most of the words except those with frequencies between 2.5 and 3.

Plot for Heaps' Law:

The red line indicates the ideal case for Heaps' law and the blue dots represent the actual data.



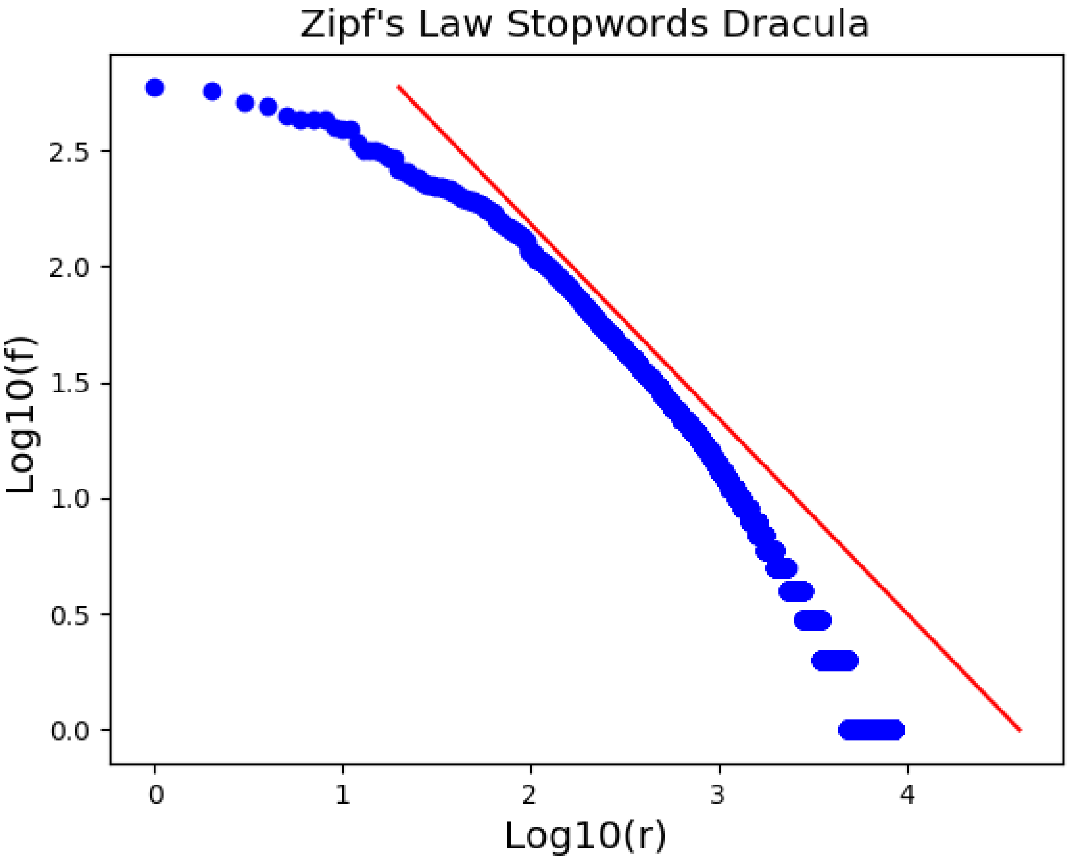
From the above plot we can observe that, the Heaps' law holds true so the dictionary size keeps increasing if the number of documents keeps on increasing. So, the size of dictionary will be large if the number of tokens keep increasing.

Q.2)

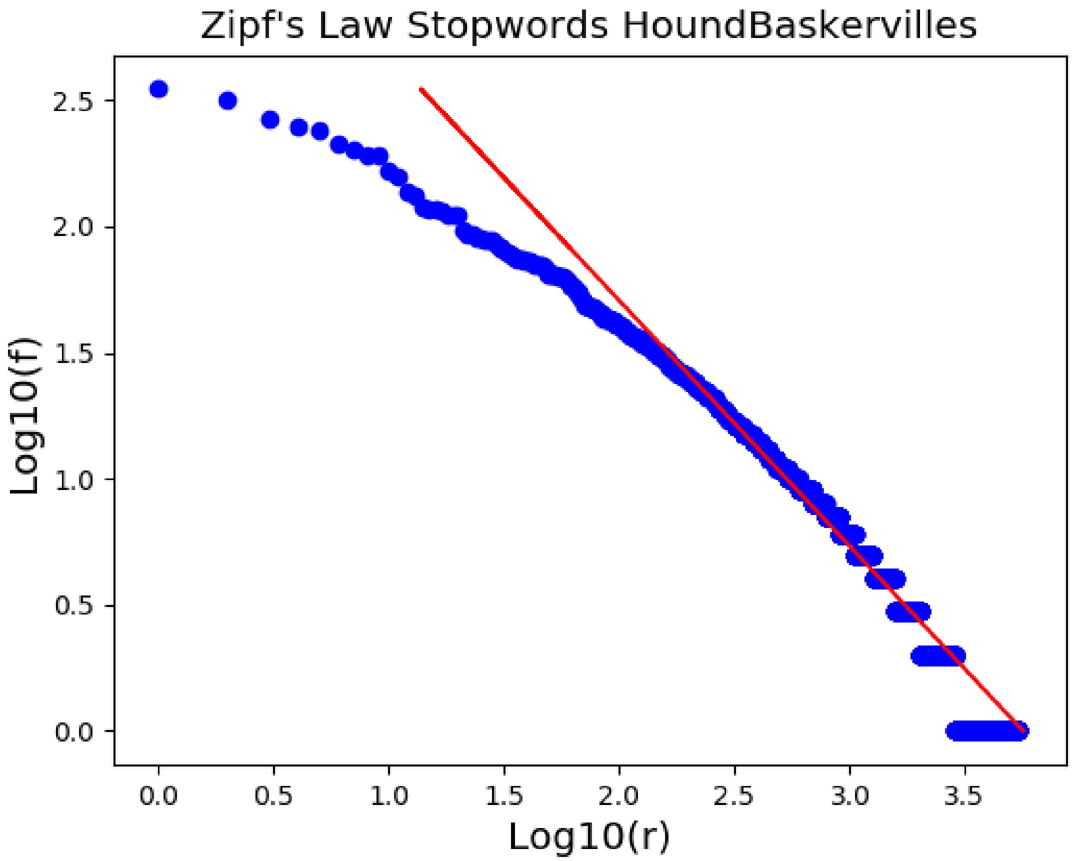
Solution:

Plots for Zipf’s Law using Stop Words:

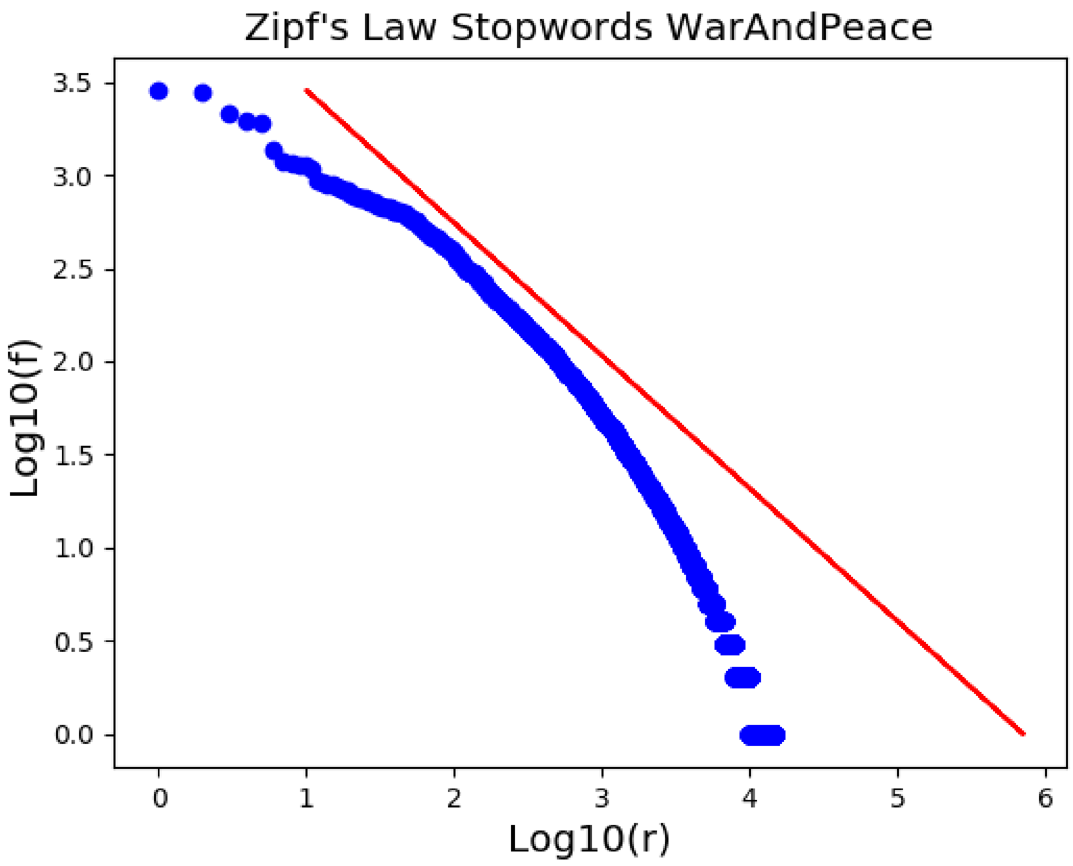
The red line indicates the ideal case for Zip’s law and the blue dots represent the actual data. In these plots, the stop words which are equivalent to useless data and reduces the efficiency of the database system are not considered.



From the above plot we can observe that , the Zipf’s law does not hold true for any words in this text file if the stop words are removed from consideration.



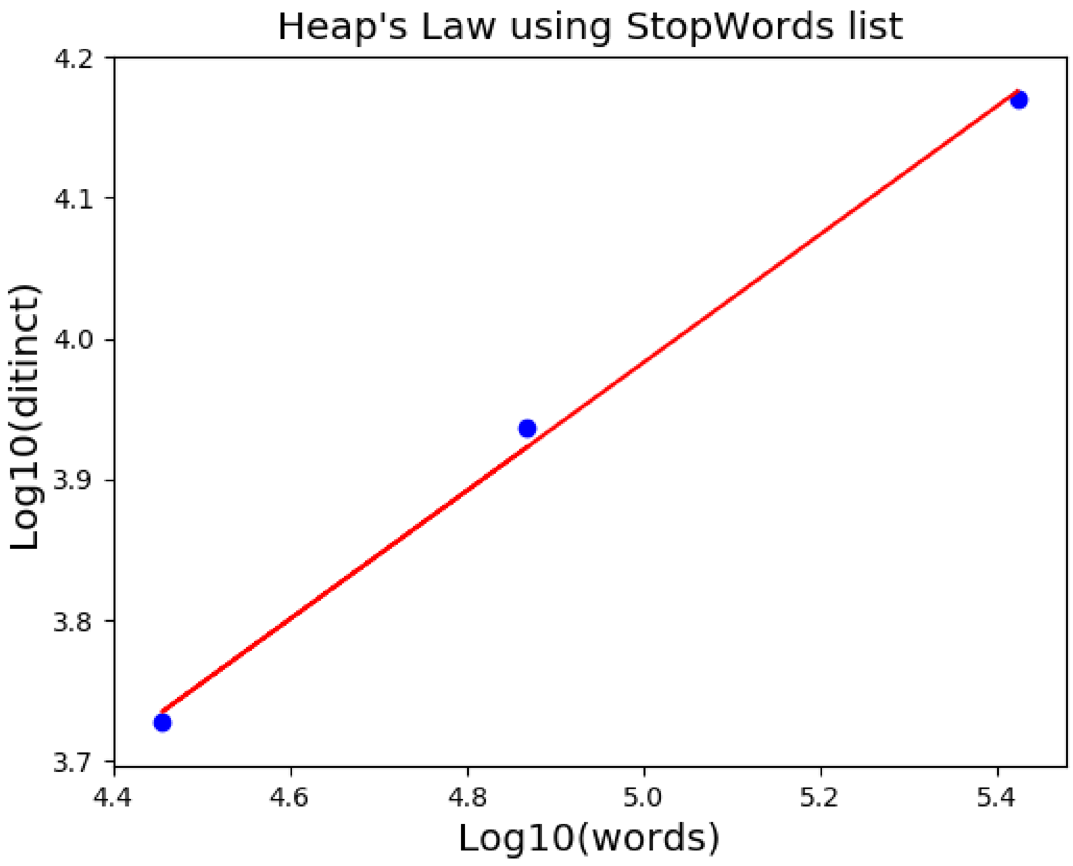
From the above plot we can observe that, the Zipf's law holds true for low frequency words but does not holds true for high frequency words which have frequency above 1.5 . So in future as the size of the text increases the law may not hold true for the entire corpus. This was not the same result when stop words were considered where the Zipf's law was holding true for the entire corpus which proves that not removing the stop words from corpus may follow the Zip's law but the system will be very less efficient.



From the above plot we can observe that , the Zipf's law does not hold true for any word in the text file after removing the stop words from consideration.

Plot for Heaps' Law using Stop words:

The red line indicating the ideal case for Heaps' law and the blue dots representing the actual data.



From the above plot we can observe that, the Heaps' law holds true so the dictionary size keeps increasing if the no of documents keeps increasing. So , the size of dictionary will be large if the number of tokens keeps increasing even if we don't consider the stop words. So compressing the dictionary size will be a challenge if the number of documents are more.

Note: To run the program, use the python command and file name(no inputs required). q1.py will plot results of Heaps' law and Zipf's law for three .txt files. q2.py will plot results of Heaps' law and Zipf's law for three .txt files using the StopWords List.