**Q3 Report**

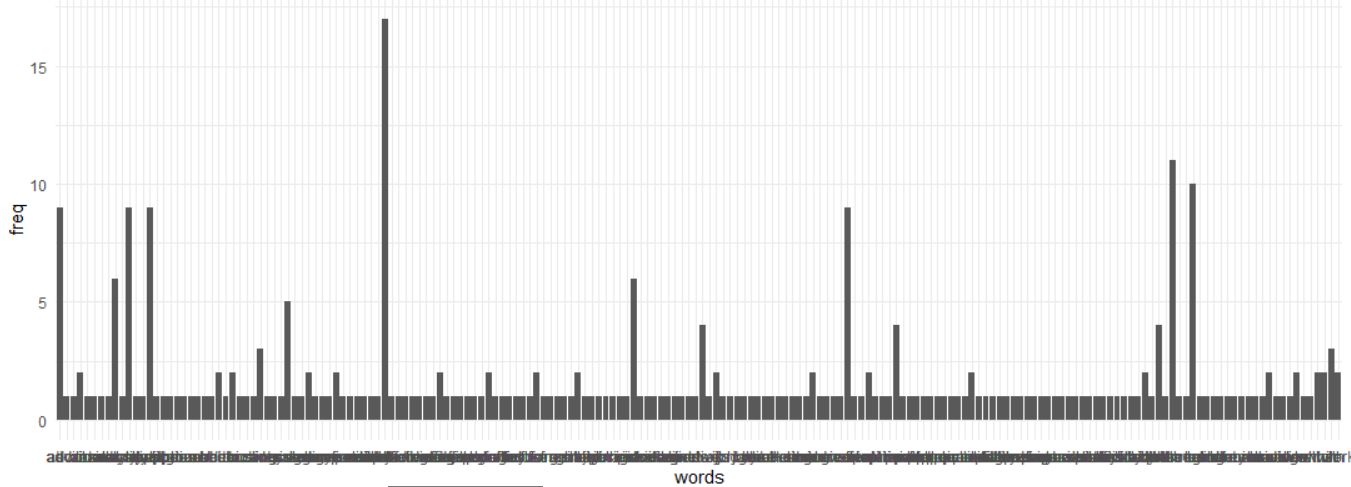
1. Firstly, upon converting to the txt file text pre processing was performed on the text file to remove number, convert to lower case and etc. Libraries such as tm,SnowballC in R was used to perform text pre-processing. Pre-processing steps were:

* Convert to lower case
* Remove punctuation
* Remove numbers
* Remove white spaces

Stopwords were not remove as we want to preserve the words of the text file for later task.

Furthermore, I computed the frequency of words in the document and converted it into a dataframe for easy viewing. It can be seen that occurrence of word data is total of 17 times in a text file. Lastly, after computing the probability it can be seen that probability of word data occurring in each line is 5.7%

1. My initial data frame contains the frequency of distinct words in the text file. Thus, to check the distribution of words across all the lines histogram was used to visualize the frequency. Ggplot2 library in R was used to visualize as it provided multiple functionality for data visualization .However, due to large amount of distinct words the graph was not clear. As a solution, we can further tokenise by removing stop words which does not bring context or extra information to the text.



1. Since my dataset contains frequency of each word in a text file I could simply calculate the probability dividing frequency of word analytics by frequency of word data and multiplying it by 100. The probability I obtained was 52.94%.