Dheeraj Goud Borlla (dxb160130)

ASSIGNMENT – #3: (TEXT MINING)

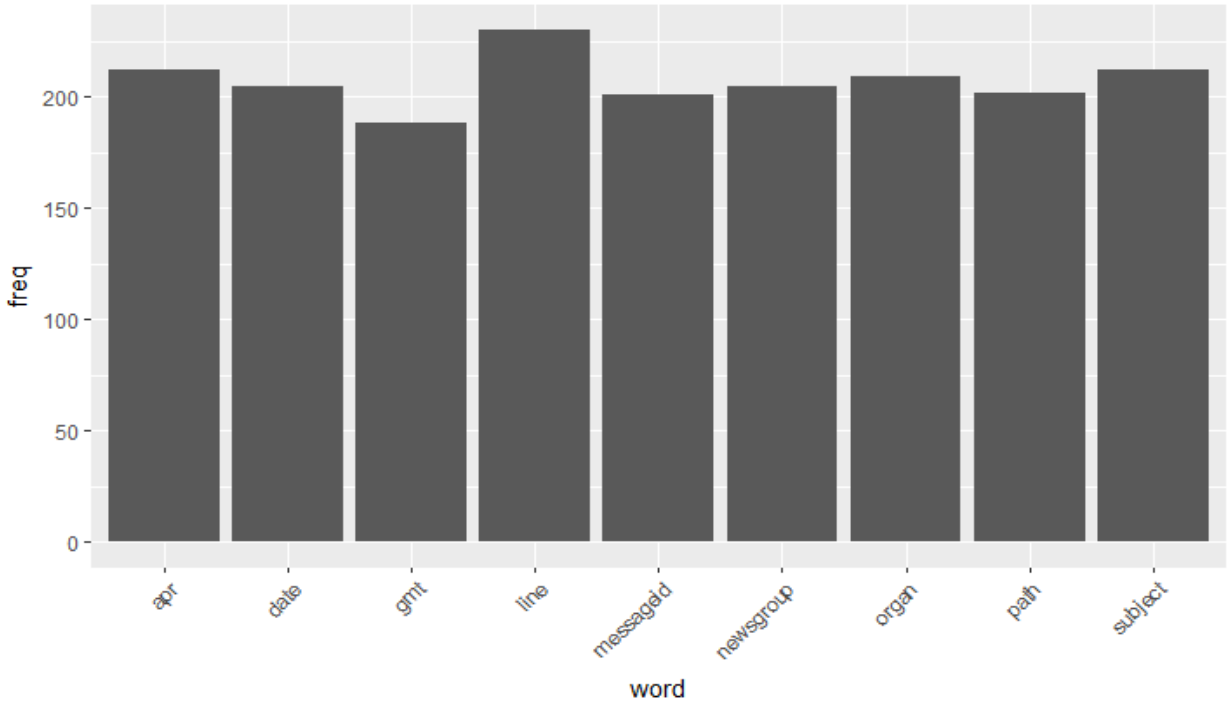
***PURPOSE*:** In this assignment, we apply the text data mining techniques to the corpus data. And we try to cluster the documents in the corpus based on the patterns that we observe from the analysis of the corpus data.

DATASETS: Assignment Corpus files of 200 text documents from the newsgroup collection.

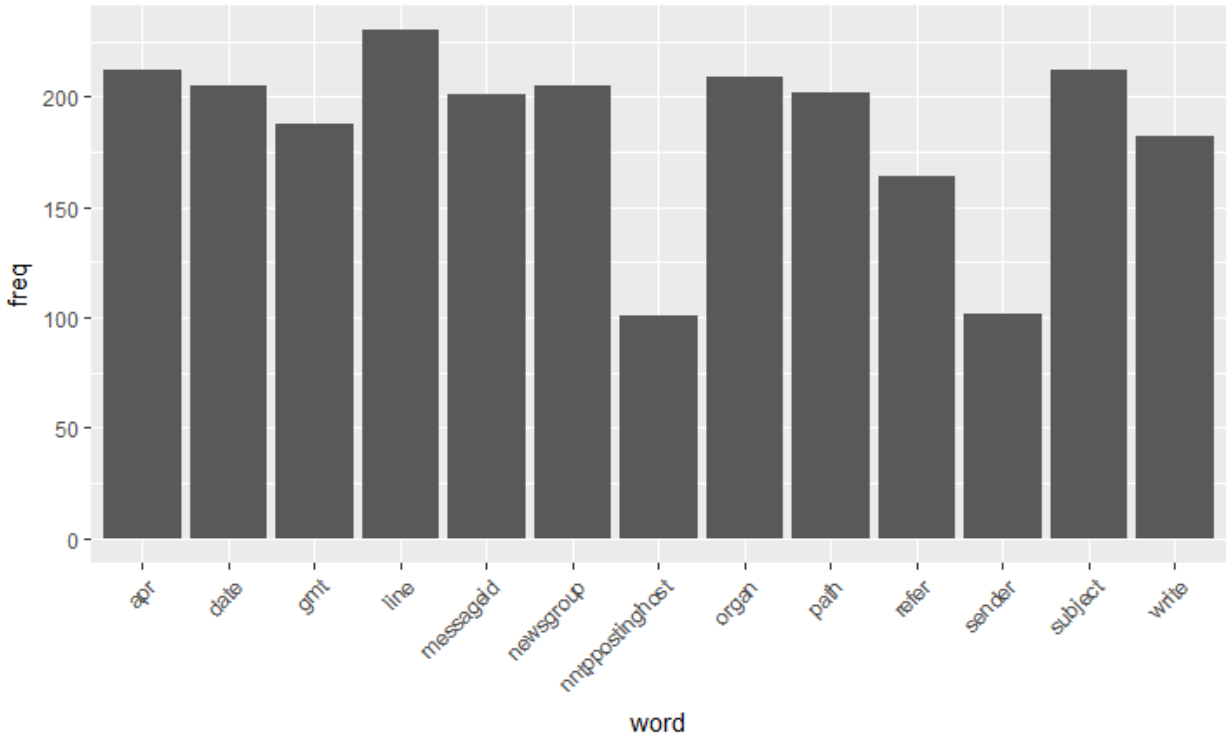
APPROACH:

- First, we created a volatile corpus using "tm" package and loaded the data into RStudio.
- Next, we go on to pre-process the loaded text files. In this process, we remove the punctuations, numbers, capitalization, common words and if any special characters from the text documents. We stem the documents that means removing common word endings (ing, es, s) and stripping of the whitespace after completion of the pre-processing. This makes the text data ready for analysis.
- Now we create a Document Term Matrix (DTM), where each column represents words and each row represent document name. We can also create Term Document Matrix where words take rows and documents names take columns.
- Now with the DTM, we can explore the data. We can organize terms by their frequency. We see that that the DTM has 98% sparsity with 6266 terms and 200 documents. The word 'line' is the most frequent occurred word with 230 times and there are 2931 words that occur exactly once in all the 200 documents.
- We will reduce the matrix by removing sparse words with the maximum of 10% empty space. This creates a DTM of 9 terms in 200 documents with the sparsity of 1%. The ggplot of the removed sparse DTM is shown in graph 1. Now we will reduce the matrix by removing sparse words with the maximum of 50% empty space. The created DTM has 13 terms with the sparsity of 13%. The ggplot is shown in graph 2.
- We applied hierarchal clustering to the terms and found four clusters seems reasonable for the text files. When the K-means clustering is applied with k=4, on the reduced sparse DTM, the ratio between_ss / total_ss is 74.8% which seem four clusters are reasonable to divide 200 documents into four different clusters.
- Term frequency-inverse document frequency (tf-idf) is a numerical statistic which can be used to show how much the word is important to a document of the corpus. Since words like 'the' are common in every document, we can't differentiate documents. So, the uncommon words carry more weight than the common words.

GRAPHS:



(Graph 1)



(Graph 2)

SUMMARY:

- Text mining is a text analytical learning where we retrieve information from the corpus data (collection of text files).
- In this process, first we pre-process text files means the files are cleaned before observing the pattern in the data. The cleaning process includes removing symbols, numbers, common words and so on which imply no important information while observing the data.
- Next the matrix is formed with the words in columns and document names in rows. This matrix has more sparsity because each word is checked against each document. By reducing the sparse terms to some extent and we can try to find the clustering of documents using hierarchal clustering / K-means clustering techniques.
- Tf-idf is applied to the words which are uncommon. Because if the word is common to all the documents then the idf value will be zero, that is tf * idf results zero which means the word is not very informative. So, the uncommon word carries more weight giving more information about the documents than the common words.