# ALY6040: DATA MINING APPLICATIONS

## MODULE 5: TECHNICAL PRACTICE

Submitted by:- Jeseeka Shah

## Introduction

We live in a world where availability of data is not scare, but the structure and usable of data are some of the biggest problems for the new age data engineers. An estimated 80-90% of the data in an enterprise is textual data or unstructured data and only 10-20% of data is structured. Hence making corporate decisions based on the small 10-20% of structured data does not make sense. Therefore, text data mining and natural language processing (NLP) have risen to popularity. Text mining is simply the process of transforming the raw unstructured text data into a structured data for analysis. The aim of text mining is to uncover hidden relationships within the unstructured data and facilitate further analysis such as summarization, text categorization, and sentiment analysis using NLP techniques.

In this project, I will be using one of the most memorable speeches of all times, "I have a Dream" by Martin Luther King Jr. It is a call for equality and freedom, it became one of the defining moments of the civil rights movement and one of the most iconic speeches in American history. This historic masterpiece provides the perfect platform to explore NLP methods for analysing the speech for frequent words and to create a visual representation of text data such as word cloud, find word associations to identify the hidden meaning from the speech and understand the emotions related to the words using sentiment analysis.

## Data Cleansing

The analysis is done in R software. Text mining packages in R like "tm" for text mining, "SnowballC" for text stemming that is reducing the inflected or derived words from the dataset to their word stem, base or root form, "wordcloud" for visualizing the frequent words from the speech, "RColorBrewer" for colour palettes, "tidyr" for creating tidy data and extracting values from string objects, "tidytext" for conversion of text to and from tidy formats, and "qdap" package to bridge the gap between quantitative and qualitative analysis. Along with the above-mentioned text mining packages, "ggplot2" is used for bar graph visualizations, and "dplyr" for working with data frame objects.

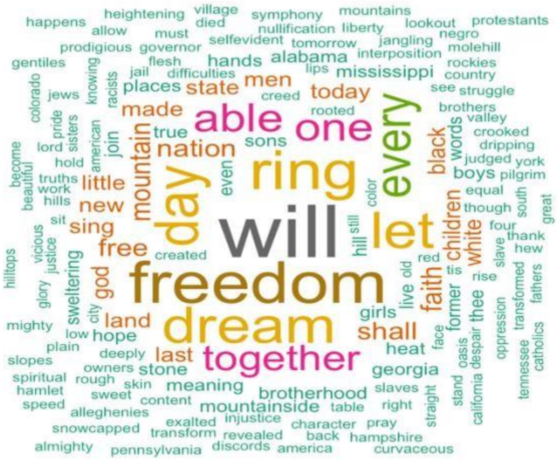
The speech is pulled from a website with the whole text and read into a variable using "readLines". The text is loaded as a Corpus into an object variable. Corpus is a text processing package that helps reading data for normalizing, tokenizing text, searching for term occurrences, and computing term frequencies. Once inspecting that it is the right text needed for analysis, I did transformation on the text.

Transformation includes using "content_transformer" to convert all text into lower case for easy analysis. I then removed of special characters including front/back slashes, white space, special signs like '@' or punctuation, stop words in English, common words like 'the'/'we', and numbers using "tm_map" function. Text stemming was also done to normalize words into their word root by reducing common prefixes or suffixes or plurals or verd/noun forms into the root word. Further,

I created a document term matrix using "TermDocumentMatrix" from the text mining package. The document term matrix has information on frequency of the words which is sorted to get the most frequent words from the speech. Using this matrix a word cloud showing the most frequent words can be generated and analysed word associations and finally sentiment analysis on the most frequent words.

## Data Analysis and Interpretation

**Word Cloud**: Using thedocument term matrix created, I created a visualization of frequency of words as a ***word cloud*** for the 'I have a Dream' speech. Word cloud is a cluster of words from the text source with varying size and font based on its frequency of occurrence in the text of at least one and a maximum words of 200 is plotted. The more a specific word appears in a source of textual data, the bigger and bolder it appears in the word cloud. Bigger words appear more cantered and grab attention. Figure 1 below shows the word cloud generated. It can be noticed that the words "will", "freedom", "dream", "day", "ring", "together" are some of the words that pop out. These are the words with highest frequency of occurrence in the speech as seen in the bar plot of Figure 2. This resonates with the theme of the speech which signifies a dream that MLK had that one day freedom will be the basic right of every citizen of the country and everyone can stay together harmoniously.



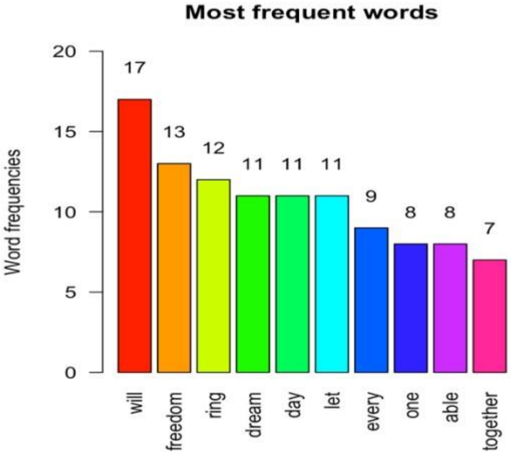


Figure 1: Word cloud of MLK speech

Figure 2: Bar plot of Most Frequent Words

**Word Associations:** Further analysis included finding ***word associations*** from the text speech to analyze the speech further. I used “findAssocs” function, from the text mining package, on the document term matrix to find word associations for one of the most frequent words from the above bar plot in Figure 2, “freedom”. I chose the word “freedom” irrespective of “will” being the top most as the word freedom resonates more with the MLK speech where he stresses on a dream of freedom for racial equality where everyone is treated based on their characteristics and not by their colour. This exercise helps in understanding the terms that correlate with “freedom”.

The function results in a score which ranges from 0 to 1. A score of 1 means that two words always appear together in documents, while a score approaching 0 means the terms seldom appear in the same document. It is important to remember that the word association is done at the MLK speech document level and would not be the same for any other document. Using a minimum correlation unit of 0.3, associations with the word “freedom” is found.

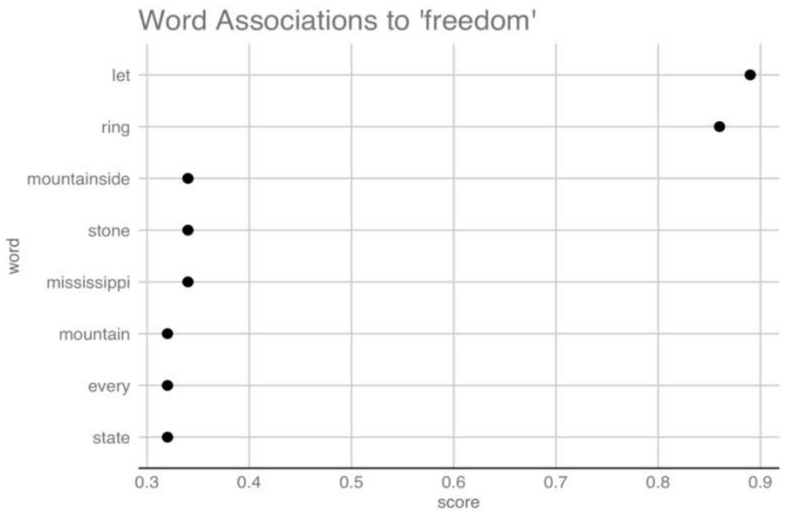


Figure 3: Shows scatter plot visualization of word association with the target word ‘freedom’.

| Word Associations with Freedom | | | | | | | |
|---|---|---|---|---|---|---|---|
| Let | Ring | Mountainside | Stone | Mississippi | Mountain | Every | State |
| 0.89 | 0.86 | 0.34 | 0.34 | 0.34 | 0.32 | 0.32 | 0.32 |

Table 1: Word Associations with the word “Freedom” from the speech.

The word ‘let’ has the highest association of 0.89 with ‘freedom’, i.e., 89% correlation. This emphasises the theme of the speech which is one of the biggest demonstration in the history of USA showcasing the right to freedom. It is interesting to see the word ‘ring’ as second most highly correlated word with freedom from the speech. It has a correlation of 86%. The reason being, MLK using ring in the context of “Let Freedom Ring” many times throughout the speech. It is a statement that emphasizes ideals of life, liberty, and the pursuit of happiness should be spread across the Earth and allowed to flourish. Other words have lower associations, but one that stands out is “Mississippi”. It has a low correlation of 34% but it signifies a major point from MLK’s speech, that is showcasing the very essence of Southern Segregation at that time. He references this state to even say words of inspiration which states that he dreams of a day when freedom will ring even from the states like Mississippi. The words ‘mountainside’, and ‘stone’ have equal correlation of

34%. The least correlated words are ‘mountain’, ‘every’ and ‘state’ with equal correlations of 32% with the word ‘freedom’. All of these words are mentioned in the context of giving hope to people that one day every man will be treated equal everywhere and given opportunities equally everywhere.

**Sentiment Analysis**: This is a process of detecting positive or negative sentiment from the speech. It is extensively used in the industry understand customer satisfaction levels, and gauge brand reputation. This analysis focusses on polarity of text (positive, negative, neutral). I used tidytext package function called “get_sentiments” to get specific lexicons in a one word per row format. I used “bing” as my lexicon which includes sentiments data frame used for sentiment analysis of the speech text.

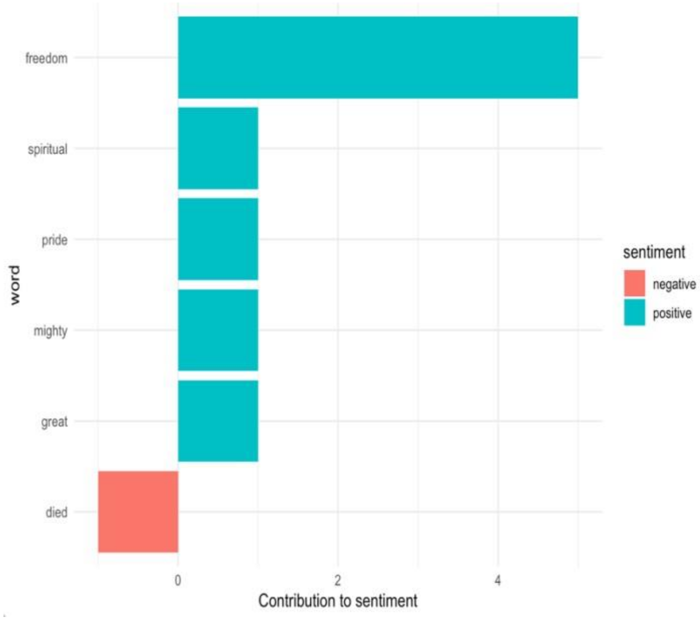


Figure 4: Shows bar graph of frequency of words grouped by the positive or negative sentiments.

I identified positive and negative sentiments and sorted them according to their frequencies. One advantage of having a data frame with both sentiment and word is that we can analyze word count that contribute to each sentiment. We can see that majority of the most frequent words show positive sentiment making MLK’s speech an inspirational speech. Words like “freedom”, “spiritual”, “pride”, “mightly”, “great” all signify the great hope for Unites States to be void of segregation and a call for equality and freedom. The negative sentiment word mentioned once is “died”. MLK refers to this word in the sentence “Land where my fathers died”. It signifies patriotic appeal to the audience along with shedding light on the struggle his forefathers had to endure.

## Conclusion and Recommendations

The above report shows analysis done on the historic Martin Luther King Jr's speech, "I have a Dream" on August 28th 1963. He stressed on civil and economic rights of all people in the country of United States and sent a strong message on end to racism. This historic proclamation provided a ray of hope to millions of enslaved African-Americans who had been engulfed in the raging fires of injustice. It was a joyful sunrise that brought an end to their captivity.

The analysis was done using various natural language processing packages in R software to decipher the hidden patterns of word occurrences, word associations and sentiment analysis of the words from the speech. I found words like 'will', 'freedom', 'ring', 'dream', 'day', 'let', 'every', 'one', 'able', and 'together' to be the most used words from the speech. My interpretation of these words from the speech is that MLK wanted to stress on a hope of complete freedom for the people who were oppressed and shed light on what the future can hold for the country. Finding the most frequent words helped in analysing the theme of the speech. Visualization of the frequency of words using word cloud and bar graph were done to showcase the understanding clearly. My recommendation for any such speech is to use NLP techniques like shown above to understand the main theme of the speech by the words used.

Further analysis on word associations for the word 'freedom' as it is one of the most frequent words and it can also be considered as the one word representation for the theme of the speech. I found 'let' and 'ring' to be highly correlated with the word freedom which signifies "Let Freedom Ring" famous phase used multiple times in the speech to encourage the public to move towards the goal of equal rights for everyone and everyone being judged by the colour of their skin but by the content of their character. Another major part of text analysis is sentiment analysis which shows that majority of the words are positive which signifies MLK's speech was encouraging and this is what makes the speech more memorable. He intended to use positive reinforcement to take a big step towards the civil rights movement in 1960s and it proved so powerful that it even after 50 years it has been regarded as the greatest speech of all times. My recommendation is to use sentiment analysis techniques to contextualize the words used which can help understand the feel of the textual data along with understanding the topic trends from the speech. Further analysis can be done on the speech to categorize the words into different emotions like anticipation, excitement, hope etc, and can also be compared with other famous human civil rights speeches like Eleanor Roosevelt, The Struggle for Human Rights, 1948, Harold Macmillan, The Wind of Change, 1960, Nelson Mandela, I Am Prepared To Die, 1964, and many more.

# References

IBM Cloud Education. (2020, November 16). What is Text Mining? IBM - Deutschland | IBM. https://www.ibm.com/cloud/learn/text-mining

Padmaperuma, D. (2019, March 11). Representing and Text Mining. Representing and Text Mining. https://rstudio-pubs-static.s3.amazonaws.com/547454_c9cbbbd180484c9f9accf76c349a17a3.html

Surles, W. (2017, September 25). RPubs - Text Mining Bag of words. RPubs. https://rpubs.com/williamsurles/316682