DATA MNING APPLICATIONS  
  
  
ALY6040, WINTER 2020  
MODULE 5 PROJECT ASSIGNMENT

TEXT MINING

SUBMITTED BY: SHRADDHA GOPALAKRISHNAN

NUID: 001376444

SUBMITTED TO: PROF. JUSTIN GROSZ

­GRADED: 100/100

DATE: 03/022/2020

**Introduction**

In this module we learn about one of the important concepts in machine learning i.e., Text Mining. It is a process of transforming unstructured data into meaningful material. Text mining is a mighty tool in the world of data mining which helps in faster computations of text analytics and help businesses infer essential elements from the tons of data generated every day. Manually this procedure can induce many errors and hence there is a need of automating this entire thing in a nutshell which is termed as “Text Mining”.

**Analysis**

In this homework, we perform text mining on the text files containing Donald Trump’s Tweets/Speeches that originally has huge amounts of raw data. Our job here is to use that raw information, create models and run the predictive models which help negate the redundant messages in order to lessen the cognitive load on the reader’s mind. Let us now dive into the analytics of the same step by step. The initial process is tokenization, converting long sentences into words. We then remove unwanted punctuations, frequent words which don’t have semantics like “es”, “ing”. We put codes to stem/stop few words by dropping characters which are not needed for example: “elect” for “election”, “histori” for “historical”; the stemming algorithm removes suffix from the word. Using tm\_map a text mining function in R, we remove numbers, convert to lowercase, map to plain texts, stop/stem words, clear away the unused white space after which we move on to stage the data. Staging comprehends a matrix of words in the document with that of their corresponding frequencies of occurrence. We even create data frames using the word count, plotting visuals and finding associations between different words. Word clouds also known as a collection of tags are visual tools of data which focus on important words in a story and communicate ideas in a direct way. Lastly, we implement clustering algorithms on word similarity.

**Objectives of Donald Trump:**

From the text analytics we draw a summary based on Trump’s viewpoints. Trump seems to be happy about the change in the mindset about African Americans and speaks positively of them in the Black History Month. He praises the contributions of the Black Americans and throws a light on the impact they have made in the American History. He says he’s proud that USA has a museum in the National Mall to let people know more about the African American antiquity. To the CIA, he pays obeisance and he stresses to get rid of ISIS and motivates CIA to work on eradicating Radical Islamic Terrorism. In the Congressional Address, he highlights the greatness of the American Leaders in the past. His motive here is leaving imprints for the future generations in terms of education, society welfare, employments, economy and prosperity of the country.

In all of his consequent speeches [ CPAC, Florida Rally, Immigration, Inauguration, National Prayer Breakfast, Nomination ], Trump lays emphasis on securing the country borders, voicing your opinions, protecting freedom of religion, asking the citizens of the country to be bold and dream big for themselves and their children, sustaining the idea of bonds of loyalty amongst the nation. To all the great sheriffs and cops, Trump thanks them for their glorious service on behalf of the people. In response to the Healthcare bill, which was a failure, he says a better bill will be proposed in which bipartisan might occur but both parties will be focused in the betterment of the healthcare facilities of the nation. He aims to gain the trust of Democrats and bring out a substance that is superior to the Obamacare. The ultimate common goal behind all the views of Trump is to make AMERICA Proud, Safe, Great and Strong again; as he believes ANCIENT AMERICA was a SUPER POWER and it was after the migration of the global clan to USA, which decreased it’s materiality and snatched away the vigor of the nation. He now is on the verge of attaining a BIG DREAM for AMERICA in all countenance.

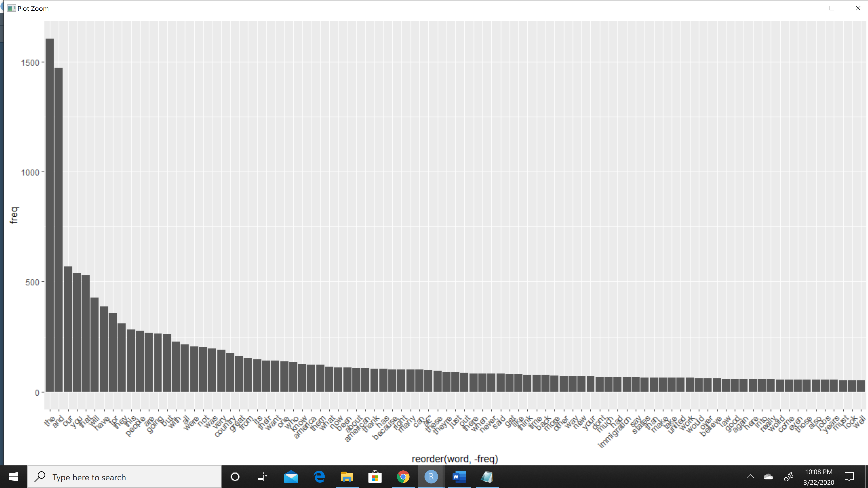
**Text Analytics:**

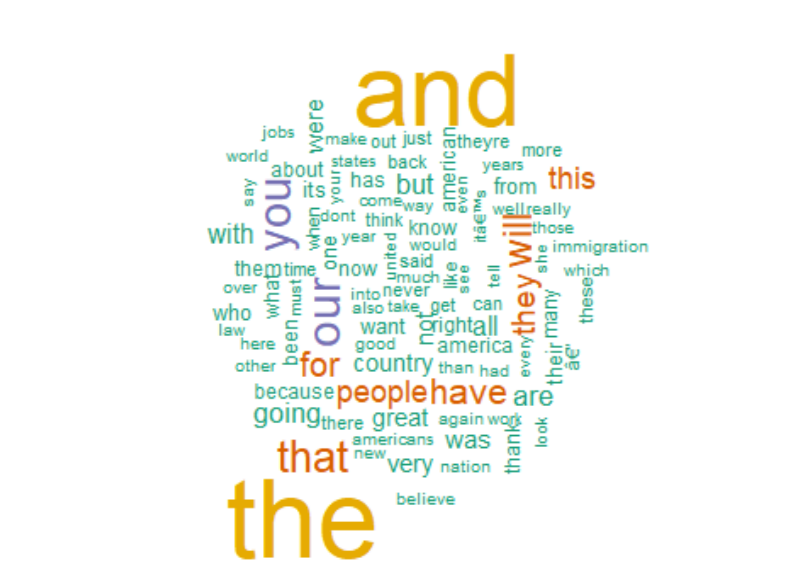
Figure : Words that appear at least 50 times

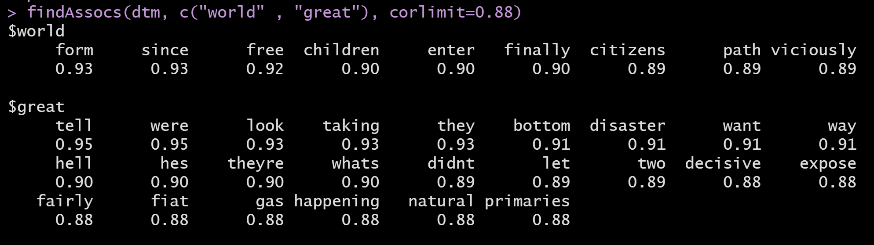
We plot a graph of word frequency which shows “the” word to be used more than 1500 times being highest in all of the scripts followed by “and”, “our”, “you”, “that”, and “will”.

Relation of “world” and “great” with other words in the script having correlation limit 0.88. The term “world” has 9 associations and great

has 24 relations in all the 11 text files.

Figure 2: Word Cloud

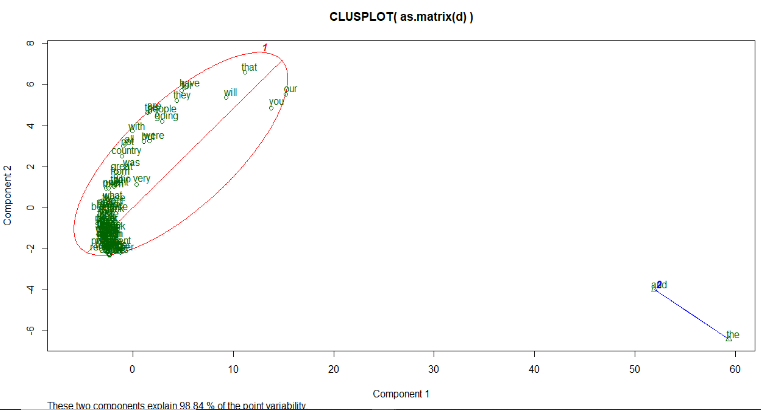




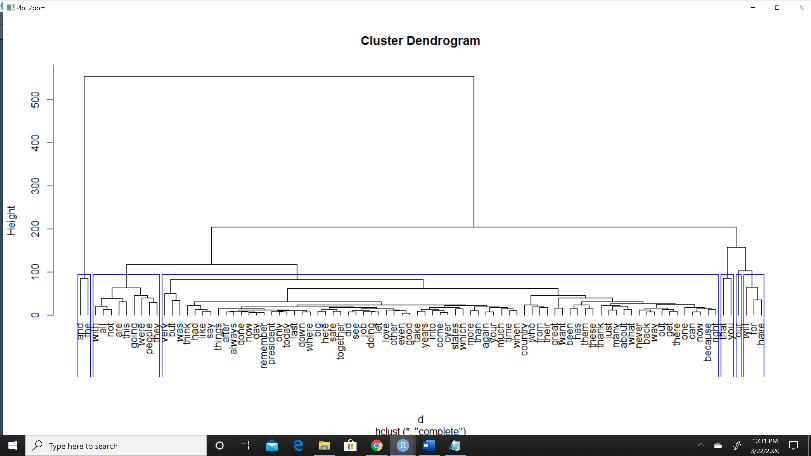
Word clouds are the best part of data visualization as it engages the reader. The tag cloud besides shows the 100 most frequently occurring words.

**Clustering-**

We perform just two clustering algorithms *Hierarchical Clustering* and *K-means Clustering Technique*, after the removal of sparse words. The output of both obtained are as follows.



K-means on the other hand aims at clustering words into an itemized number of group. It is an unsupervised machine learning algorithm.

**Conclusion**

Text mining helped in obtaining valuable insights and take decisions based on the data given. We identified essential abstracts and provide qualitative results with respect to the speeches of Trump. We illustrated Text analytics through charts and diagrams. From this fragment of the course, we define Text Mining as the combination of following-

Clustering Dendrogram shows 6 clusters as the cutree fit is arbitrarily chosen as 6. This process groups similar terms in one cluster.

* Text Pre-processing
* Staging of textual data
* Exploratory Data Analysis
* Feature Extraction
* Applying ML Algorithms

**References**

[1] Text Mining: The Beginner's Guide. (2019, May 8). Retrieved from <https://monkeylearn.com/text-mining/>

[2] Nabi, J. (2019, June 24). Machine Learning - Text Processing. Retrieved from <https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958>

Note: The **Text Analytics** part is an optional portion and it was my choice to keep it in the paper.