DATA MINING APPLICATIONS

Northeastern University

ALY6040, WINTER 2020

INCORPORATING UNSTRUCTURED DATA INTO MODELS

SUBMITTED BY: SHIVANI ADSAR

NUID: 001399374

SUBMITTED TO: PROF. JUSTIN GROSZ

DATE: 03/22/2020

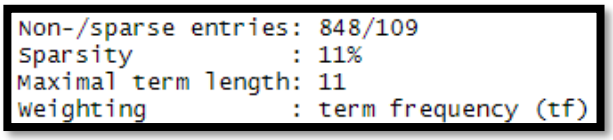
Introduction

In this assignment, we have used the "Trump Tweets" dataset for performing text analysis. On the basis of text mining, decision trees and clustering algorithms, we have analyzed Trump's motives and politics. This has given us understanding of the his incentives from the speeches he has delivered. (1)

Analysis

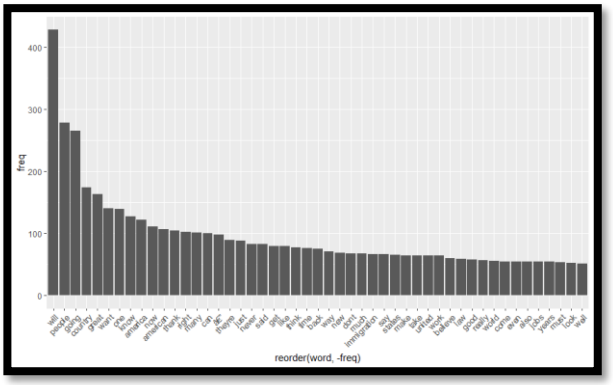
We have imported the text files containing Donald Trump's speeches and created a corpus of the files. Further, the data is processed to remove the stop words for performing better analysis.

- In order to perform Data Staging, we have implemented Document Term Matrix which tracks the term frequency in the documents and this is Transposed to get the Term Document Matrix where each row represents a document vector.



On interpreting that the sparsity is 79%, we have tried to remove the sparsity after performing the term document matrix, the sparsity was reduced to 11%.

Fig.1: Term Document Frequency Matrix

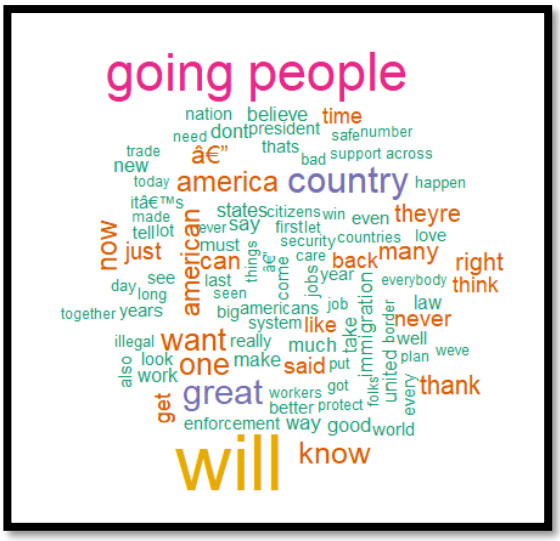


Observation: The frequency plot has been plotted for the highest frequency words occurring in the document. We can observe that the word "will" with 428 was used most frequently, while "well" was the least used word.

Fig.2: Frequency Bar Graph

- On the basis of the highest frequency words, we can interpret that, Donald Trump wants his country, america and americans to do great and he will ensure that his country performs well. (1)

- Moreover, based on the associations between the words in the TDF, we have interpreted the correlations between certain words like, “nation”, “people” and “going”. (2)

- 

As we can note from the Word Cloud, that the highest occurring word is “will”, this shows that Trump is determined to do better developments for America and its citizens.

We can observe that Trump intends to improve jobs in the United States, keep up with the promises of Americans and care for the military veterans.

Fig.3: Word Cloud for highest frequency of words

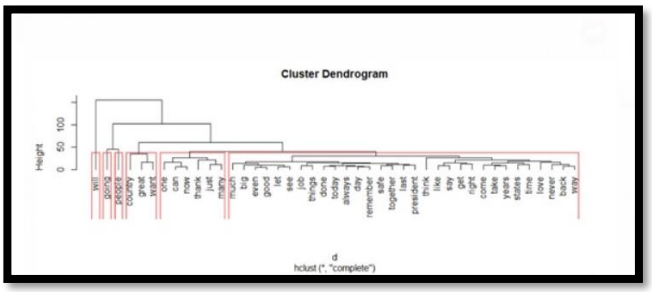
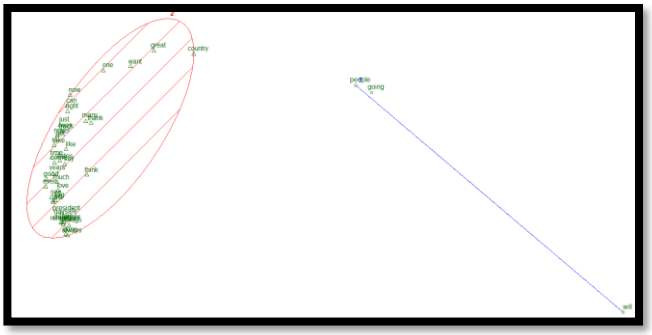
- 

Fig.4: Cluster Dendogram

As we can observe from the Cluster Dendogram, the words are grouped on the basis of their frequencies with respect to the heights. There can be 6 clusters that can be formed on the basis of data, “will”,”going” and “people” being the highest frequent have been grouped individually based on their height while remaining words are arranged in 3 clusters based on the similarities in their occurrences.

- On performing the above analysis, we can note that Donald Trump would work upon enforcement of immigration laws for the border security. Moreover, he noticed that he was endorsed as the Presential candidate, he intends to stop the illegal immigration and save dollars over taxes for improvement of cities in America and thanks his supporters in his tweets.



Observations: As can be seen from the K-means clustering that clusters the data on the basis of shortest Euclidean distance, there are 2 clusters. This clustering is done on the basis of most frequent words. The clusters are based on their occurrences of words in the documents.

Fig.5: K-means Clustering

Conclusion

- As per the text analysis, it was observed that, Donald Trump intends to invest more finances for the industries, increase jobs and improve employment amongst workers.
- Donald Trump has shown that he has been more effective in improving the nation's economy than the pre-president.
- The clustering algorithms, associations and word cloud has helped in analyzing the data through text mining and deriving Trump's policies. The analysis shows that Trump is a very determined president, who would definitely work for the nation and improve employment for the Americans.

References

- Building a simple sentiment classifier in R using Trump's tweets. (n.d.). Retrieved from https://www.datacareer.co.uk/blog/building-a-simple-sentiment-classifier-in-r-using-trump-s-tweets/
- Fitzgerald, J. D. (2018, January 25). Sentiment analysis of (you guessed it!) Donald Trump's tweets. Retrieved from https://www.storybench.org/sentiment-analysis-of-you-guessed-it-donald-trumps-tweets/