Text Mining

Wesley Clark

iamwesleyclark@gmail.com

(301) 254-9395

Introduction

Each year the president of the United States of America gives a State of the Union(SOTU) address. The latest was given by the current president, Donald Trump on January 30th. This annual message usually covers issues such as national budget, economic report, as well as serving for an outline onto legislative agenda and national priorities. I thought this would be an interesting topic for this assignment as the tone as well as the content of a SOTU can change. It is reflective not just of the president, but also the times. I was curious to see how this would be shown through text analysis. The purpose of this analysis is to see what the common themes that are presented in different SOTU's.

For this assignment several different State of the Union Addresses were used. Each State of the Union was separated into a different transcript, and different text document. The following presidential state of the union transcripts were compiled: Trump 2018, Obama 2009, Bush 2001, Clinton 1993, Bush 1989, Regan 1982, Carter 1978, Ford 1975, Nixon 1970, Johnston 1964.

Text Preprocessing

Data preprocessing is an essential aspect of any text mining project. Before the data can be mined at all, it must be transformed. Through several transformations, we the data was tokenized. Tokenization is the task of breaking a stream of textual content up into meaningful tokens(). In this case, the tokens we generate are words. Several things occurred to make this tokenization happen. The first of which was to remove the URL. In each document, the URL where the document was taken from was present. This does not present any meaningful

information and would be erroneous in our investigation. It is not only meaningless, it would throw off our results as the URL would be counted in the subsequent analyses.

Special characters “@“, “&”, “*” were removed. Special characters are characters that may be present in a document and throw our results off. In order to eliminate this error, three special characters were removed. The @ symbol appeared in some documents relating to a link. Interestingly, the & symbol was present in some of the older documents, but not recent ones. Because “and” is a stopword, its equivalent in symbol made sense to remove. The last special character that was removed was *. This was present in some recordings of the transcript, used as a marker for a reference. It was imperative to remove all three of these special characters before analysis could be conducted.

Numbers and punctuation were also removed. Most of the numbers were used to reflect a year. There were also some numbers present to discuss a quantity or amount. Punctuation was present in every document. By removing the punctuation, we ensured a list of characters.

Next whitespace was removed. In order to fully tokenize the data, a simple bag of words was needed. While a human may intuitively be able to differentiate some common pitfalls, a computer will not. For example the difference between “end.” and “end”.

All of the characters were changed to lowercase next. The difference between “Nation” and “nation” may seem like a small difference, but they are completely different strings to a computer. By changing all the text to lowercase, this mistake is avoided.

Stopwords were removed next. Stopwords are common words that occur in many forms of text and will likely be meaningless to our results. Examples of stopwords removed are “a”, “the”, “you” “youre”. It is easy to see how all of these words could be counted many times, were they not removed. In addition to the standard stopwords, I decided to remove another list of stopwords referred to as “SMART” stop words. This is a different dictionary of words, and was chosen as it is one of the better lists of stopwords. Additionally, I removed two custom words, “america” and “american”. These words are of little importance, it was suspected that these words would appear a lot given the nature of the presidential texts.

Next the words were stemmed. Stemming is a way of keeping the core of the word, while trimming any other properties of the word that may change the word itself. An example of this could be “given”, “give”, “gives. For these three words, the stem of “give” or perhaps “giv” would be used and the words would be collectively put together. This is an important part of any text mining project, as pluralizations and different forms of the same word inevitably appear in text. The SnowballC package was used to accomplish this. In short, the following steps were taken to preprocess the data:

- Create a corpus
- Remove URL
- Remove whitespace
- Remove punctuation
- Remove numbers
- Remove special characters
- Convert text to lower case
- Ensure that the documents are Plain Text
- Remove SMART stop words as well as two custom stop words

- Stemm the document

Methodology

The next step was to create a document term matrix(DTM). DTM's are a commonly used way of visualizing preliminary results of text mining. In this method a matrix is created that displays the frequency of terms in a corpus. The column represent the terms, and the documents are displayed along the rows. The DTM gave an overview of the corpus and listed the number of documents, terms as well as Sparsity. Sparsity reflects the threshold of how often a term appears in the documents in order for it to be counted. Utilizing sparsity will change the corpus, in a way that is more inclusive of words that have been used in the the majority of the documents.

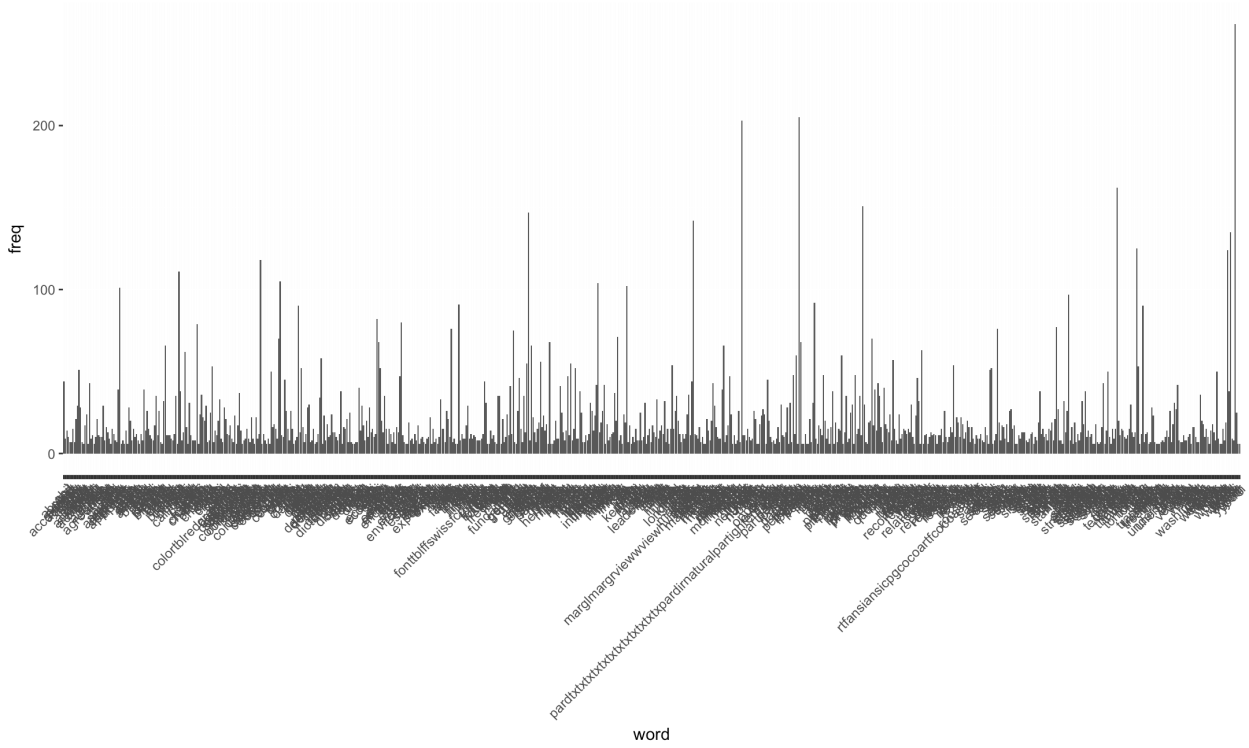
```
> dtm
<<DocumentTermMatrix (documents: 10, terms: 3325)>>
Non-/sparse entries: 8873/24377
Sparsity           : 73%
Maximal term length: 57
Weighting          : term frequency (tf)
```

Once the DTM was created, term frequencies were counted. It was interesting to see that there were many terms listed at least 10 times. In order to provide more insight into the corpus, the terms were narrowed to include only terms with a frequency of 40.

```
> findFreqTerms(dtm, lowfreq=40)
 [1] "act"       "administr" "american"  "billion"   "budget"
 [6] "busi"      "care"      "children"  "congress"  "continu"
[11] "cost"      "countri"   "creat"     "cut"       "day"
[16] "deficit"   "dollar"    "econom"    "economi"   "educ"
[21] "end"       "energi"    "famili"    "feder"     "forc"
[26] "fund"      "futur"     "give"      "good"      "govern"
[31] "great"     "growth"    "health"    "high"      "histori"
[36] "home"      "hope"      "incom"     "increas"   "inflat"
[41] "invest"    "job"       "live"      "major"     "make"
[46] "meet"      "million"   "money"     "nation"    "opportun"
[51] "pay"       "peac"      "peopl"     "percent"   "plan"
[56] "polici"    "presid"    "problem"   "program"   "propos"
[61] "provid"    "put"       "rate"      "reduc"     "reform"
[66] "respons"   "save"      "school"    "secur"     "spend"
[71] "state"     "support"   "system"    "tax"       "time"
[76] "today"     "tonight"   "unit"      "weve"      "work"
[81] "world"     "year"      "_"
```

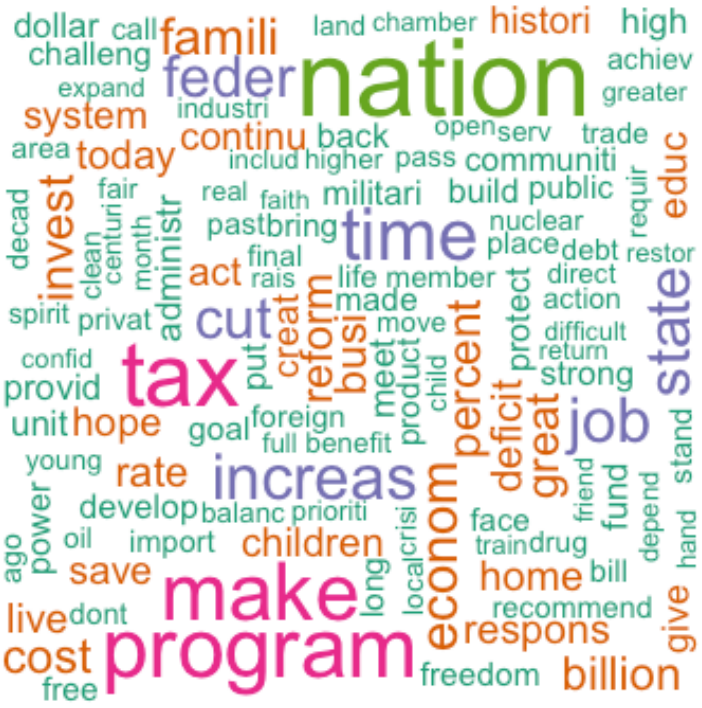
The next step was to remove the sparse terms. This step is important to change the corpus in a way that benefits the overall corpus. A closer look into the document names and terms revealed 836 terms. Word associations were derived, and it was observed that the most associated word was net.

The data has been observed in matrix form. Moving forward the data was visualized in other ways. A word frequency plot was generated to display the different terms that were used.



Although there were too many words used to gather much meaningful information from this plot, it was interesting to see the overall frequencies of all the words. Clearly there were many words that were used with a small frequency, and only a few words that were used with a large frequency. Future visualizations provided more insight into the words that were used the most.

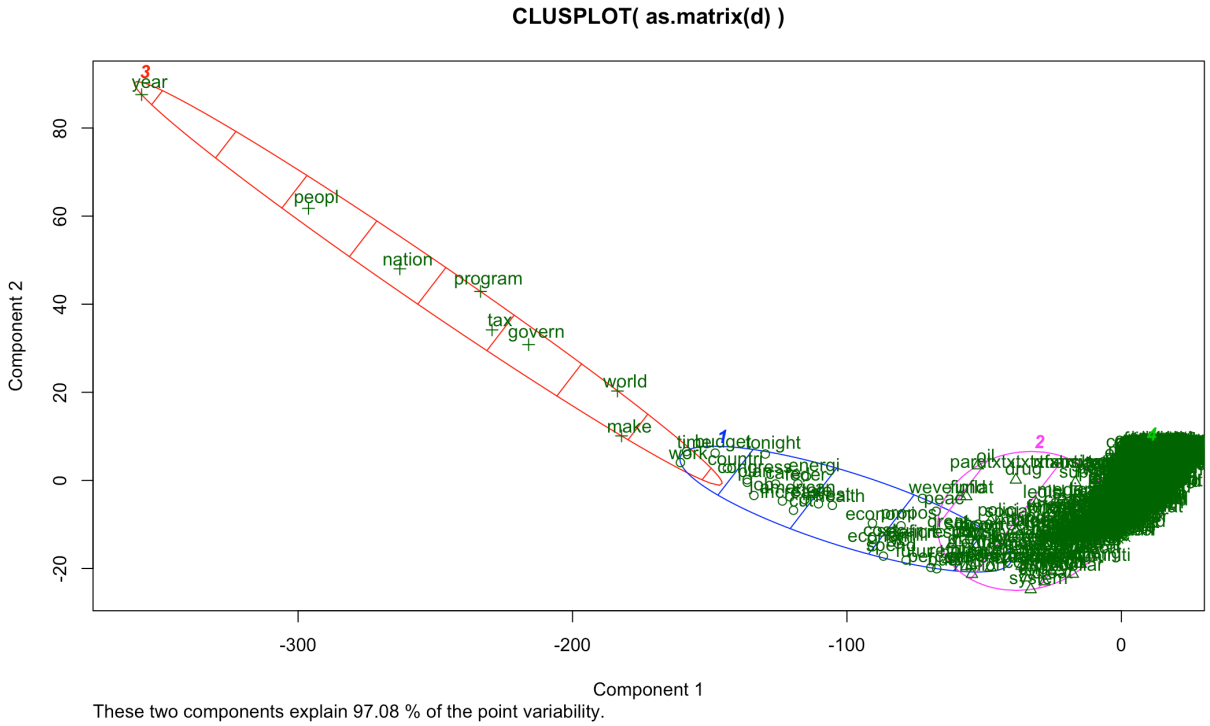
The best method for understanding the most frequent words at a glance is a word cloud. As such, a word cloud was created to accomplish this.



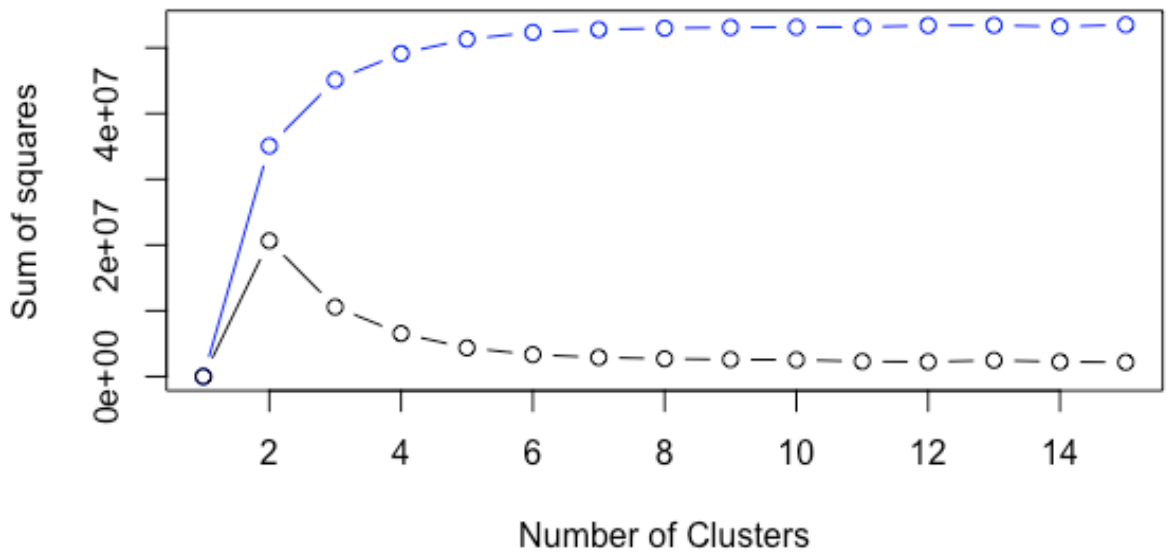
It was observed in this color word cloud what some of the most frequent words were. This form of visualization is especially useful for presentation, when the results are needed quickly at a glance. In this example, it can quickly be observed that words such as “nation”, “make”, “tax” and “program” were the most common words. The next step was to cluster the words together as a means of understanding how the words were related to each other. Through this analysis it can be seen that some words were easily distinguishable, such as year+people.

While this k-means cluster is useful, it was not known if it was the most useful cluster or not. In order to decipher how many clusters was optimal for this corpus the elbow method was employed.

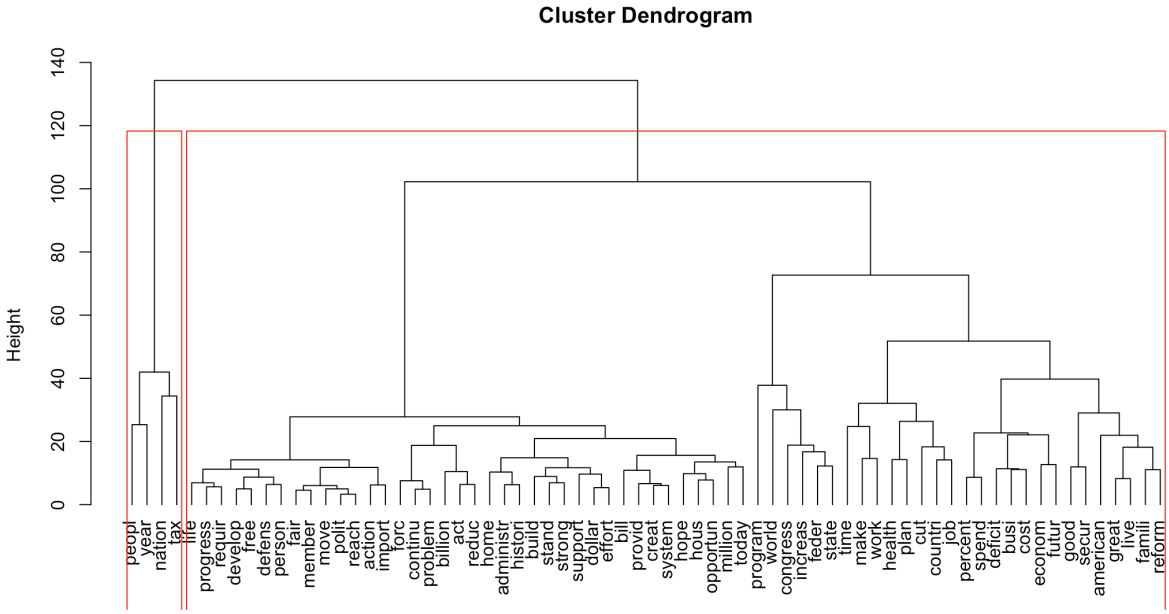
It was observed that the optimal number of clusters using this method was 2. The code was appropriately changed to reflect this, and a new plot was derived.



Interestingly the overwhelming majority of the words were lumped together into one cluster. The other cluster contains the words “people”, “nation”, “program”, “government”, “world” and “year” suggested strong correlation between these words.



Lastly a dendrogram was constructed. Initially the dendrogram included too many words to be of use, as they were not legible. To change this, the sparse term removal was changed to 0.01.



As can be seen from this dendogram, the clusters and their likeliness to be related are significant.

Conclusion

It was noteworthy that the majority of the work in this analysis was done in the preprocessing section. Assembling the data, and analyzing the results took a fraction of the time comparatively. This text analysis of 7 different SOTU transcripts revealed insight into the the speech itself. It was very interesting to see which words were related, and likely to occur together in the same document. Perhaps most interestingly was the word cloud. This method of visualizing data does have a way of departing the results onto a user at a glance. Previously, my experience with word clouds was leaving something to be desired, however after this analysis, the appeal of word clouds is evident, especially when presenting results in a short amount of time. Two things could be done to further this study. First, include more SOTU's from presidents past as well as multiple SOTU's from a single president. By increasing the sample size the results are likely to yield more credibility and probably change. Second, it would be interesting to see how these results change for each president. Generation of a top 5 word list for each SOTU would accomplish this.

Bibliography

Thomas, R. (n.d.). IBM Big Data Success Stories.

Brillinger, D. R., Hannan, E., Kanal, L., Krishnaiah, P. R., Rao, C. R., Rao, M., . . . Raghavan, V. V. (1980). Handbook of statistics. Big data analytics. Amsterdam: North-Holland/Elsevier.

Introduction to the tm Package. (2017, December 6). Retrieved from https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf