Assignment 1: Text Mining Analysis Using R

Daanish Ahmed

DATA 650 9041

Spring 2018

dsahmed2334@yahoo.com

Professor Elena Gortcheva

UMUC

February 7, 2018

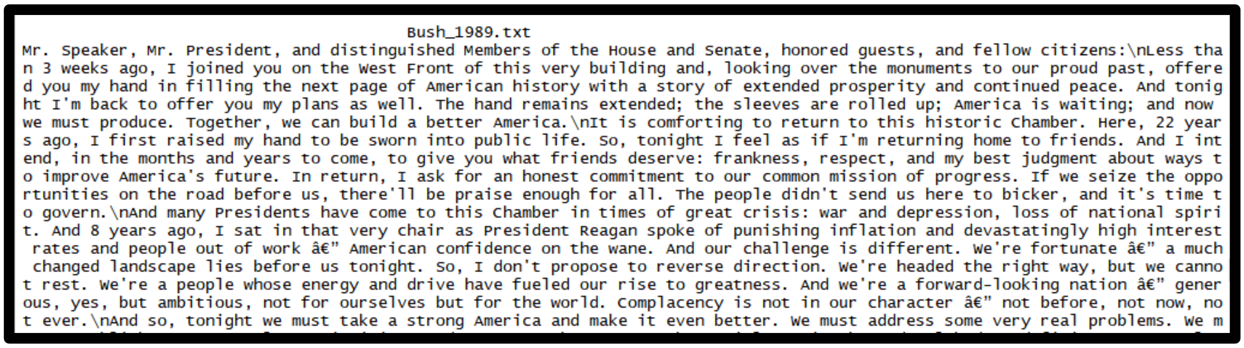
Introduction

Text mining is an increasingly important process which allows data scientists to extract useful information from text documents. Although text is a form of unstructured data, there are techniques that allow us to gain insight regarding the usage of each word. One such technique is the bag-of-tokens approach, which involves examining documents as a set of words that appear at least once in the text (Bramer, 2016). Key words are identified in a process known as tokenization, and unnecessary terms are removed using preprocessing steps such as stemming and removing stop words (Han, Kamber, & Pei, 2011). The texts can then be analyzed by creating a document-term matrix (DTM), which shows the number of times each word appears in every text ("Basic Text Mining," n.d.). From here, several methods can be used to analyze the texts, including word clouds, word correlation plots, and clustering models. However, the bag-of-tokens approach has several shortcomings—including a loss of word order information, a limited ability to determine word context, and an inability to recognize a word's synonyms (Han et al., 2011). Still, this method gives data scientists the ability to make educated inferences regarding a word's importance, which would be far more difficult and time-consuming without text mining techniques.

My goal is to use the bag-of-tokens approach in R to analyze the State of the Union addresses delivered by the last five U.S. presidents every year between 1989 and 2017. The documents consist of the past 29 State of the Union speeches in .txt format (Tatman, 2017). These speeches were given by the five most recent U.S. presidents: George H.W. Bush, Bill Clinton, George W. Bush, Barack Obama, and Donald Trump. The reason why I am including all of these speeches is to understand the most important issues facing the U.S. in the modern day, rather than looking only at individual presidents and their policies. My analysis will have three components: word frequency analysis, word relationship mining, and clustering analysis. The first component will involve creating word clouds to study some of the most commonly-used terms. The second component will use both word associations and correlation plots to study the relationships between important words. The last component will use k-means clustering to group terms into clusters based on their usage together. I expect that my techniques will provide insights regarding the common issues facing the nation, regardless of the change in political landscape over time.

Data Preparation

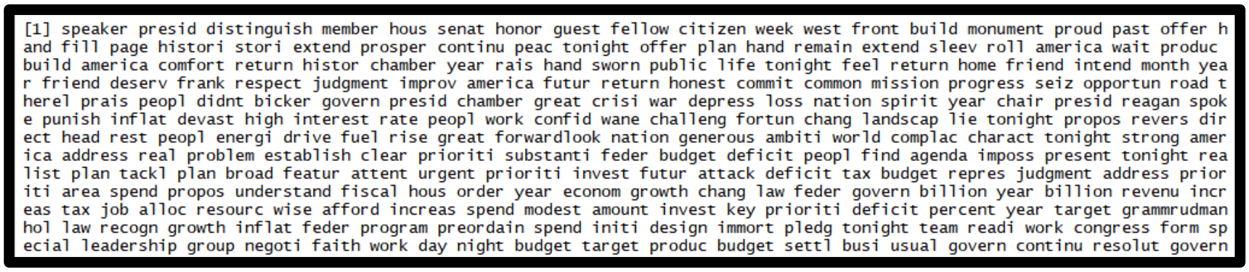
There are several preprocessing steps that I performed before beginning the text analysis. I first initialized all of the required packages, namely “tm,” “SnowballC,” “wordcloud,” and “cluster.” These packages are required for features such as stemming, stop lists, word clouds, and k-means clustering. Next, I created a corpus that includes all 29 State of the Union Addresses. This allows us to compile the text documents together to perform text mining operations. By examining the first speech in the corpus (see Figure 1), we see that the texts need to be cleaned before beginning the analysis. This can be achieved by using the “tm_map” method in R. This function requires the “tm” package, and it allows us to transform the text in our corpus (Feinerer & Hornik, 2017). I began by removing all numbers, punctuation, and URLs. These characters are not useful for understanding a word’s context, and they can result in duplicate terms if they are not removed. Likewise, I also set all letters to lowercase to ensure that identical words with different cases are recognized as the same word.



```
                                   Bush_1989.txt
Mr. Speaker, Mr. President, and distinguished Members of the House and Senate, honored guests, and fellow citizens:\nLess tha
n 3 weeks ago, I joined you on the West Front of this very building and, looking over the monuments to our proud past, offere
d you my hand in filling the next page of American history with a story of extended prosperity and continued peace. And tonig
ht I'm back to offer you my plans as well. The hand remains extended; the sleeves are rolled up; America is waiting; and now
we must produce. Together, we can build a better America.\nIt is comforting to return to this historic Chamber. Here, 22 year
s ago, I first raised my hand to be sworn into public life. So, tonight I feel as if I'm returning home to friends. And I int
end, in the months and years to come, to give you what friends deserve: frankness, respect, and my best judgment about ways t
o improve America's future. In return, I ask for an honest commitment to our common mission of progress. If we seize the oppo
rtunities on the road before us, there'll be praise enough for all. The people didn't send us here to bicker, and it's time t
o govern.\nAnd many Presidents have come to this Chamber in times of great crisis: war and depression, loss of national spiri
t. And 8 years ago, I sat in that very chair as President Reagan spoke of punishing inflation and devastatingly high interest
 rates and people out of work â€" American confidence on the wane. And our challenge is different. We're fortunate â€" a much
 changed landscape lies before us tonight. So, I don't propose to reverse direction. We're headed the right way, but we canno
t rest. We're a people whose energy and drive have fueled our rise to greatness. And we're a forward-looking nation â€" gener
ous, yes, but ambitious, not for ourselves but for the world. Complacency is not in our character â€" not before, not now, no
t ever.\nAnd so, tonight we must take a strong America and make it even better. We must address some very real problems. We m
```

Figure 1. Excerpt of George H.W. Bush's 1989 SOTU Speech Before Preprocessing.

I proceeded to remove stop words, which are common but meaningless terms such as "and," "the," and "to." Failure to remove these terms will result in the analysis producing trivial and uninteresting results. The "tm" package contains built-in stop lists in the "English" and "Smart" lists (Feinerer & Hornik, 2017), which I used to remove stop words from my corpus. I also created an additional stop list containing words that were not found in these two lists (such as "just," "open," and "bring"). After handling stop words, I removed special characters such as "@" and "â." Next, I performed stemming to reduce words with common stems to the same root word. For instance, the words "educate" and "education" are reduced to the root term "educ." The stemming command requires the "SnowballC" package in R to be installed. The problem with stemming is that it often forms terms that are not English language words. For example, it transforms the word "president" into the term "presid." It can be difficult to tell whether this term refers to "president" or "preside." But the benefit of stemming is that it reduces the number of redundant terms by preventing variations of the same word from appearing as separate words. After stemming, I removed stop words again—since some stop words may have been formed by the stemming process. Finally, I removed the extra whitespace between words, since having excess whitespace can cause problems when identifying terms. By examining the first speech again (see Figure 2), we see that the text has been cleaned.



```
[1] speaker presid distinguish member hous senat honor guest fellow citizen week west front build monument proud past offer h
and fill page histori stori extend prosper continu peac tonight offer plan hand remain extend sleev roll america wait produc 
build america comfort return histor chamber year rais hand sworn public life tonight feel return home friend intend month yea
r friend deserv frank respect judgment improv america futur return honest commit common mission progress seiz opportun road t
herel prais peopl didnt bicker govern presid chamber great crisi war depress loss nation spirit year chair presid reagan spok
e punish inflat devast high interest rate peopl work confid wane challeng fortun chang landscap lie tonight propos revers dir
ect head rest peopl energi drive fuel rise great forwardlook nation generous ambiti world complac charact tonight strong amer
ica address real problem establish clear prioriti substanti feder budget deficit peopl find agenda imposs present tonight rea
list plan tackl plan broad featur attent urgent prioriti invest futur attack deficit tax budget repres judgment address prior
iti area spend propos understand fiscal hous order year econom growth chang law feder govern billion year billion revenu incr
eas tax job alloc resourc wise afford increas spend modest amount invest key prioriti deficit percent year target grammrudman
hol law recogn growth inflat feder program preordain spend initi design immort pledg tonight team readi work congress form sp
ecial leadership group negoti faith work day night budget target produc budget settl busi usual govern continu resolut govern
```

Figure 2. Excerpt of George H.W. Bush's 1989 SOTU Speech After Preprocessing.

After the initial preprocessing steps, I built the document-term matrix (DTM) using my corpus as the input. The DTM allows us to find the number of times each term appears. Each row of the DTM represents a document, while each column represents a word that appears in any of the documents ("Basic Text Mining," n.d.). My initial DTM contains 29 documents and 5601 words. However, it has a sparsity of 82%—meaning that the majority of words rarely appear throughout most of the documents. I solved this issue by removing sparse terms that appear in less than 50% of the documents. This results in 589 terms remaining, with a sparsity of only 25%. Finally, I created the variable "freq" which contains all unique words and their frequency counts. It will be needed to visualize and analyze terms based on their frequencies.

Methods

With data preparation complete, I can now describe the methods I will use in my analysis. First, I will study how often certain words appear in the State of the Union speeches. This is achieved by using word clouds, which can help viewers easily identify the most common words. To use this method in R, we need to install and activate the "wordcloud" package. I will create two distinct word clouds in my analysis. The first word cloud will include terms that appear at least 150 times while allowing a maximum of 40 words. It will use a color scheme that colors terms according to their frequency, which makes it easier for us to determine which words appear more often than others. The purpose of this word cloud is to highlight the general themes mentioned in the State of the Union addresses, regardless of the year or political climate. My second word cloud will use a term frequency-inverse document frequency (TF-IDF) matrix, which measures words based on their importance within the document rather than their frequency ("Basic Text Mining," n.d.). This cloud will be limited to 30 words, and its purpose is to include terms that are more relevant to the political landscape and current events during this period.

The next part of my analysis involves studying the relationships between important terms. This will be done using two methods: word association mining and correlation plots. Word associations are one of the most powerful methods used in my analysis, since they allow us to see how often certain words appear together. By finding terms that are used together, it becomes easier to make inferences regarding a word's context. This method is performed by using the "findAssocs" command with the following inputs: the document-term matrix, the terms(s) to analyze, and the minimum percentage of cases where the related terms appear together (Maceli, 2016). I will use this method to find words related to the following terms: jobs, schools, health, economy, taxes, debt, war, terror, and immigration. For each term, I will use a percentage between 0.4 and 0.65 to return useful word associations while minimizing the size of the output. I expect that my results will provide insights regarding U.S. presidents' attitudes toward major issues such as terrorism, the economy, education, health care, debt, and immigration.

After studying word associations, I will build word correlation plots that show the correlations between common terms. These plots involve using lines to connect pairs of correlated terms, where the strength of their correlation is greater than or equal to the minimum correlation threshold ("Text Mining Analysis," 2018). Using this method requires installing the "Rgraphviz" package, which is obtained from bioconductor.org. In my analysis, I will create two correlation plots. The first plot will contain a set of 10 words that appear at least 150 times while using a correlation threshold of 0.4. The second plot will consist of 6 terms that appear at least 150 times with a correlation threshold of 0.2. However, the second plot will be weighted—for which the width of each line represents the strength of that correlation. These plots are helpful for finding not only the relationships between terms, but also the strengths of those relationships. Unlike the word associations, this method will only focus on correlations between frequent terms.

The final component of my analysis involves building a clustering model to group words based on their appearance together. I will use k-means clustering, which is an unsupervised learning method that groups terms into k clusters such that entries in the same cluster will have strong similarities with each other and few similarities to those in different clusters (Han et al., 2011). This method will require installing the "cluster" package in R. The value of k will be selected using two distinct methods. The first approach will use the formula $k \approx \sqrt(n/2)$, where n is the number of cases (Bati, 2015). The second approach is to use the elbow method, which involves building a plot of the within clusters sum-of-squares over k and choosing the approximate k-value where the plot bends to form the shape of an "elbow" (Bati, 2015). I will implement this model using only the terms that appear in every single document—eliminating all spare terms. By focusing on words that are found in all documents, I will be able to form clusters that are more representative of the common themes addressed in the State of the Union speeches.

To build the clustering model, I will first create a dissimilarity matrix (DSM) that computes the distances between and within clusters. The next step is to create the model using the "kmeans" method in R, which takes the DSM and the number of clusters as input. I can then display the cluster plot, which shows the clusters and the terms contained in each cluster. Implementing the elbow method requires plotting the between clusters sum-of-squares (BSS) and the within clusters sum-of-squares (WSS) on the same graph. The "elbow" can be found on the WSS, which will appear in blue. After selecting the ideal number of clusters, I will build the clustering model again using the same steps but with the new k-value. I will then evaluate the results to determine which terms frequently appear together and why. In the following sections of this paper, I will analyze and discuss the results of every method used in my analysis.

Word Clouds

In this section of the paper, I will describe and evaluate the output from my two word clouds. The first word cloud contains terms that appear at least 150 times with a maximum of 40 words (see Figure 3). It also uses a color scheme that places words into categories based on their frequency. According to this image, the five most frequent terms are "America," "year," "work", "people," and "nation." These findings are expected, since they reflect general terms used by any U.S. president in a speech towards the nation. Some of the more useful terms include "tax," "secure," "job," "economy," "health," and "reform." Such terms are more indicative of actual policies that are important to the American people. A president is expected to make progress in fields such as health care, tax reform, national security, job creation, and overall economic growth. Although the context of these words may be difficult to obtain, we can still confirm their importance towards shaping a president's successful career.

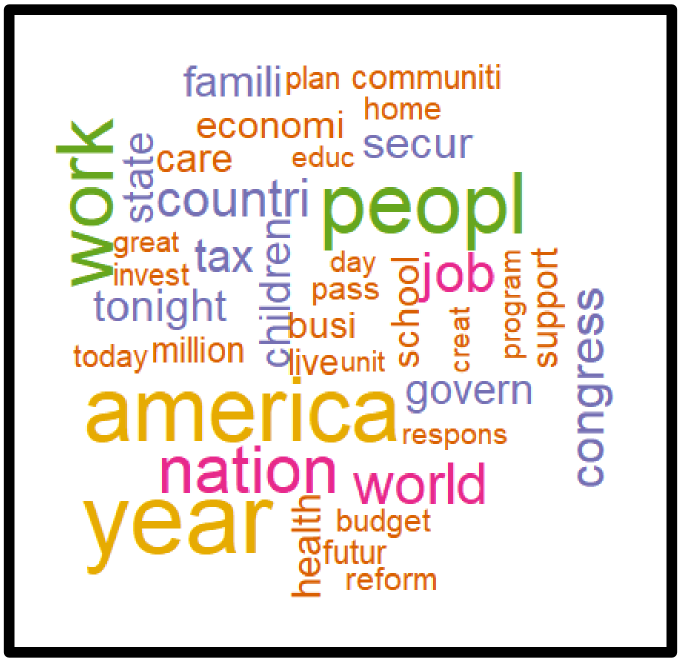


Figure 3. Color Word Cloud of the 40 Most Frequent Words.

To gain further insights, we can examine the coloring used for some of these terms. Words using the same color have similar frequencies with one another, while words of different colors have a greater difference in frequency. With this in mind, we notice that the word "job" (labeled in pink) has a visibly higher frequency than the words "tax" and "secure" (both labeled in blue). One possible implication is that job creation is more important than taxes or security with regards to a president's achievements. However, the fact that a term is mentioned more often than others does not guarantee that it is more important. Furthermore, the word "job" may not necessarily refer to job creation at all. The term might refer to the president's job, or even the jobs that other people are performing. This highlights the difficulty of understanding context when it comes to text mining. One possible way to gain greater contextual insight is to study word associations. Although this method does not guarantee finding the actual context, it can be helpful since it shows how often certain terms are found together. In the next section of this paper, I will use word association mining to gain a deeper understanding of words and their importance.

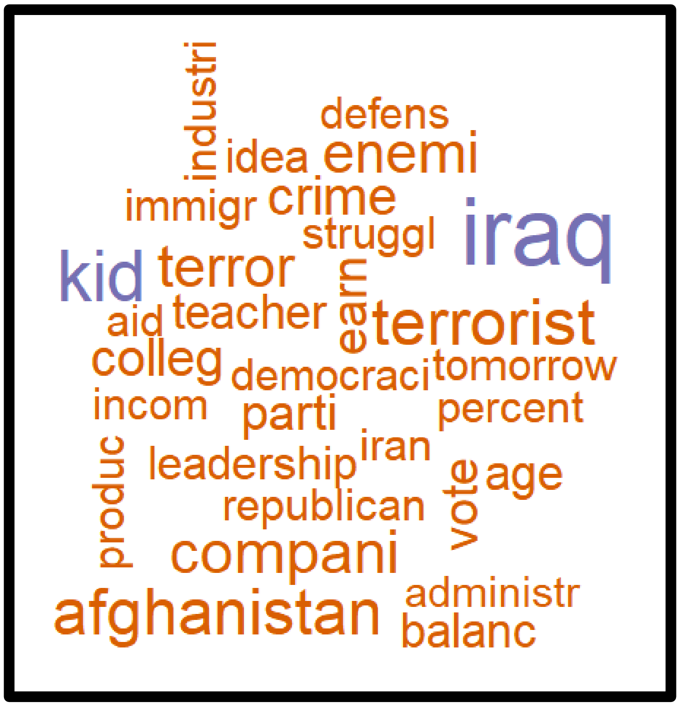


Figure 4. Word Cloud Using TF-IDF Matrix with Maximum of 30 Words.

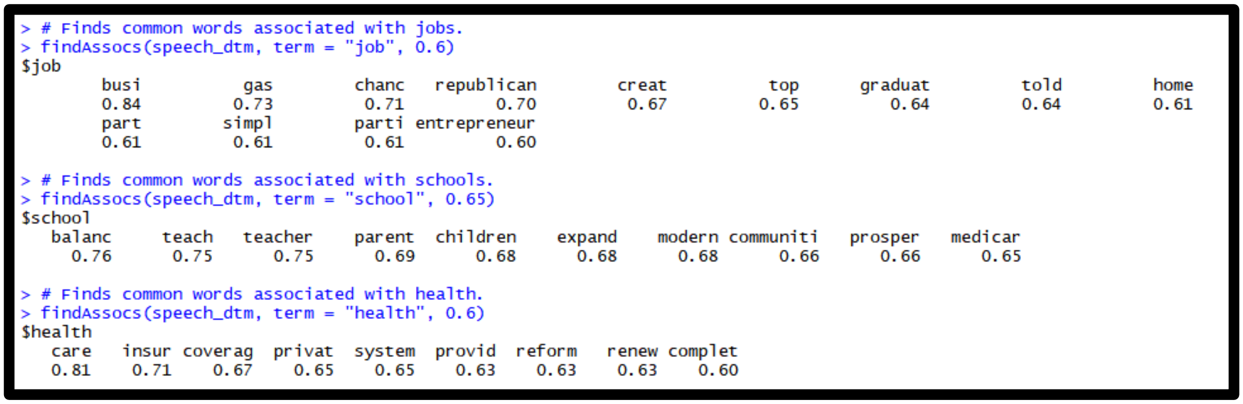
My first word cloud focuses on common themes such as the economy, taxation, security, and health care. Such topics are important policies that will be mentioned in every State of the Union address, regardless of political events at the time. But to discover words that are more relevant to current events, we need to examine my second word cloud. As mentioned, this cloud uses term frequency-inverse document frequency (TF-IDF) to evaluate words based on their importance rather than their frequency (see Figure 4). This word cloud is visibly different from my previous cloud, since it excludes many common words such as “America,” “people,” “job,” and “economy.” Instead, prominent terms include “Iraq,” “terrorist,” “Afghanistan,” “crime,” “defense,” and “immigr” (i.e. immigration). Many of these terms relate to major crises or wars that occurred during this period, and they reflect serious concerns that were addressed during several State of the Union speeches. Furthermore, many of these words imply a negative sentiment (such as “terrorist” or “crime”). Although these words appear less often than those used in the previous word cloud, they nevertheless offer more specific information about the important political concerns mentioned in State of the Union addresses.

But one limitation of this word cloud is that it groups all of these terms together, regardless of the president or the year. Some of these words—such as "crime" and "immigration"—are relevant regardless of the year or political landscape. However, terms such as "Iraq" and "Afghanistan" relate to the wars in those two countries, meaning that these words have less relevance during times of peace (such as during Bill Clinton's presidency). This is not a major issue for this analysis, but it would be more problematic if analyzing speeches covering a longer period of history. One possible solution would be to conduct separate text mining studies for each individual president and compare the important terms used in their speeches. This approach would also provide useful findings about the issues that were most important to each president. Overall, this word cloud helps to shed light on specific political issues during this period, while my first cloud highlights policies that are important to U.S. politics in general.

Word Associations

With the word frequency analysis complete, I can now focus on the association mining results. Word associations show how often certain words appear together, and they are useful for determining the relationships between important terms. I will first examine the associations for the terms "job," "school," and "health" (see Figure 5). For the term "job," some of the associated terms include "create" (67%) and "top" (65%). Such findings suggest that the word "job" may indeed refer to job creation in most cases. This helps to address my earlier concern with regards to the meaning and context of the word "job." Additionally, the word "Republican" is included in

70% of cases where the term “job” is used. One possible explanation is that Republican politicians place a larger emphasis on job creation when compared to their Democratic counterparts. However, these word relationships are based on whether the terms appear in the same document—not necessarily in the same sentence or phrase. Since each document consists of an entire speech, we have no way of knowing the ordering of these words or their exact context.

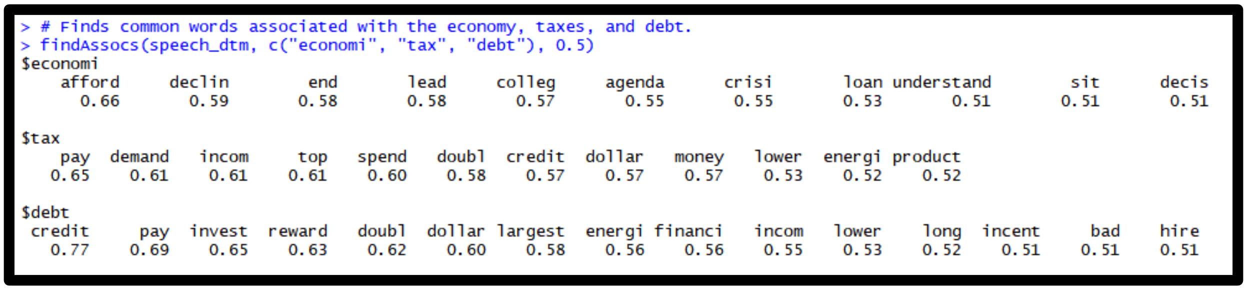


```
> # Finds common words associated with jobs.
> findAssocs(speech_dtm, term = "job", 0.6)
$job
        busi          gas        chanc   republican        creat          top      graduat         told         home
        0.84         0.73         0.71         0.70         0.67         0.65         0.64         0.64         0.61
        part        simpl        parti entrepreneur
        0.61         0.61         0.61         0.60

> # Finds common words associated with schools.
> findAssocs(speech_dtm, term = "school", 0.65)
$school
   balanc     teach   teacher    parent  children    expand    modern communiti   prosper   medicar
     0.76      0.75      0.75      0.69      0.68      0.68      0.68      0.66      0.66      0.65

> # Finds common words associated with health.
> findAssocs(speech_dtm, term = "health", 0.6)
$health
   care   insur coverag  privat  system  provid  reform   renew complet
   0.81    0.71    0.67    0.65    0.65    0.63    0.63    0.63    0.60
```

Figure 5. Common Words Associated with “Jobs,” “Schools,” and “Health.”

When looking at the term “school,” we find associated words such as “expand,” (68%), “modern” (68%), and “prosper” (66%). And although we are lacking the exact context for each term, we can infer that these words are reflective of the U.S.’s goal to improve and modernize its education system. This is largely because the U.S. is underperforming in academic achievement compared to other industrialized nations—especially in math and science (DeSilver, 2017). When analyzing the word “health,” some of the related terms include “coverage” (67%), “provide” (63%), and “reform” (63%). These terms may suggest that presidents wish to improve the health care system to provide coverage to as many Americans as possible. Additionally, the term “private” (65%) may imply that most of the past five presidents prefer a private health care system as opposed to a national approach such as the Affordable Care Act.

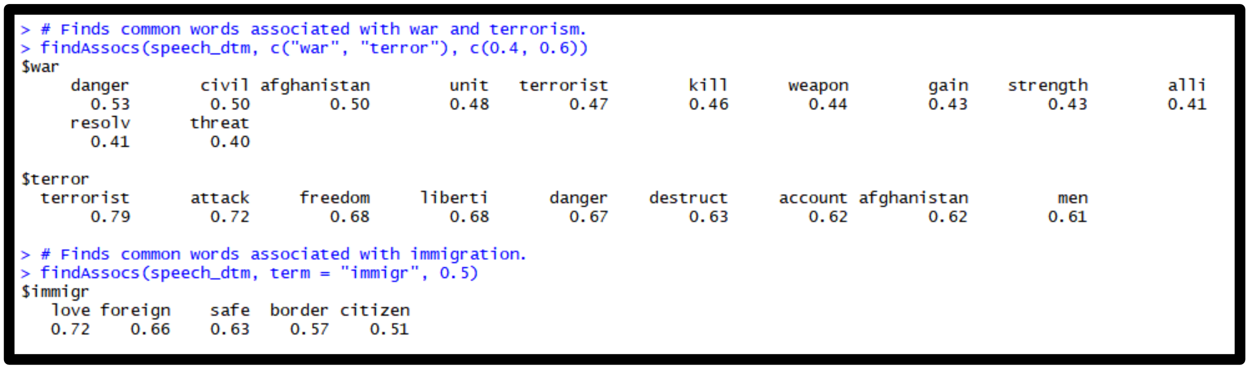
Next, I will look at word associations for the terms “economy,” “tax,” and “debt” (see Figure 6). Interestingly, some of the words associated with the economy are “decline” (59%), “end” (58%), and “crisis” (55%). These negative terms seem to suggest a struggling or underperforming economy. This does not mean that the economy was struggling during the entire period from 1989 to 2017. However, it is likely that the economy will be mentioned more often during times of economic hardship, since Americans are more likely to be concerned about the economy when it is not performing well. For the word “tax,” some of the related words include “pay” (65%), “income” (61%), and “credit” (57%). Such terms are commonly mentioned when discussing taxes, but they do not seem to refer to any specific policies or tax reforms. Likewise, the word “debt” is commonly associated with “credit” (77%), “pay” (69%), and “reward” (63%). These terms seemingly refer to peoples’ individual debt, and they do not seem indicative of any policies regarding the U.S.’s national debt.



```
> # Finds common words associated with the economy, taxes, and debt.
> findAssocs(speech_dtm, c("economi", "tax", "debt"), 0.5)
$economi
    afford     declin        end       lead     colleg     agenda      crisi       loan understand        sit      decis 
      0.66       0.59       0.58       0.58       0.57       0.55       0.55       0.53       0.51       0.51       0.51 

$tax
    pay  demand   incom     top   spend   doubl  credit  dollar   money   lower  energi product 
   0.65    0.61    0.61    0.61    0.60    0.58    0.57    0.57    0.57    0.53    0.52    0.52 

$debt
 credit     pay  invest  reward   doubl  dollar largest  energi financi   incom   lower    long  incent     bad    hire 
   0.77    0.69    0.65    0.63    0.62    0.60    0.58    0.56    0.56    0.55    0.53    0.52    0.51    0.51    0.51 
```

Figure 6. Common Words Associated with “Economy,” “Taxes,” and “Debt.”

Finally, I will analyze the terms related to the words “war,” “terror,” and “immigration” (see Figure 7). Some of the word associations for “war” are trivial, such as “danger” (53%), “kill” (46%), and “threat” (40%). Likewise, many words related to “terror” are obvious as well, such as “freedom” (68%), “danger” (67%), and “destruction” (63%). However, the word “terrorist” is mentioned in 47% of documents mentioning war. This reflects the fact that many of the wars involving the U.S. over the past 30 years have been related to terrorism. Additionally, the word “Afghanistan” is related to both “war” (50%) and “terror” (62%). This key phrase links together U.S. policy on both war and terrorism, and it highlights the importance of the War in Afghanistan as one of the defining events of modern U.S. history.



```
> # Finds common words associated with war and terrorism.
> findAssocs(speech_dtm, c("war", "terror"), c(0.4, 0.6))
$war
     danger       civil afghanistan        unit   terrorist        kill      weapon        gain    strength        alli 
       0.53        0.50        0.50        0.48        0.47        0.46        0.44        0.43        0.43        0.41 
     resolv      threat 
       0.41        0.40 

$terror
  terrorist      attack     freedom     liberti      danger    destruct     account afghanistan         men 
       0.79        0.72        0.68        0.68        0.67        0.63        0.62        0.62        0.61 

> # Finds common words associated with immigration.
> findAssocs(speech_dtm, term = "immigr", 0.5)
$immigr
   love foreign    safe  border citizen 
   0.72    0.66    0.63    0.57    0.51 
```

Figure 7. Common Words Associated with “War,” “Terror,” and “Immigration.”

And when looking at the word immigration, we will notice that some of the related words include “love” (72%), “safe” (63%), and “citizen” (51%). Based on these findings, it may seem that most of the recent presidents have a compassionate attitude towards immigrants. But due to the lack of context, we cannot confirm whether this is the case. For instance, the term “safe” may refer to protecting the rights of immigrants—or it may refer to keeping American citizens safe from crime caused by illegal immigrants. In addition, the prevalence of the term “border” (57%) suggests a need for the U.S. to secure its borders. Furthermore, the term “illegal” does not show up on this list since it only appears in 34% of documents relating to immigration. Therefore, it is difficult to tell whether these word associations refer to illegal immigration. Overall, the issue of determining context is one of the greatest challenges facing the field of natural language processing (NLP). However, studying these word associations has allowed us to gain a greater understanding of the text when compared to looking at word frequencies alone.

Correlation Plots

From here, I will describe the results of my correlation plots. As I stated earlier, these plots show the correlations between words by using lines to connect them. One difference between my correlation plots and word associations is that these plots only show the relationships between frequent terms. They are also more visually-appealing due to their layout. However, the main limitation is that they provide less information regarding the actual numeric strength of each correlation. I created two correlation plots in my analysis—the first of which contains 10 words with a minimum frequency of 150 and a correlation threshold of 0.4 (see Figure 8).

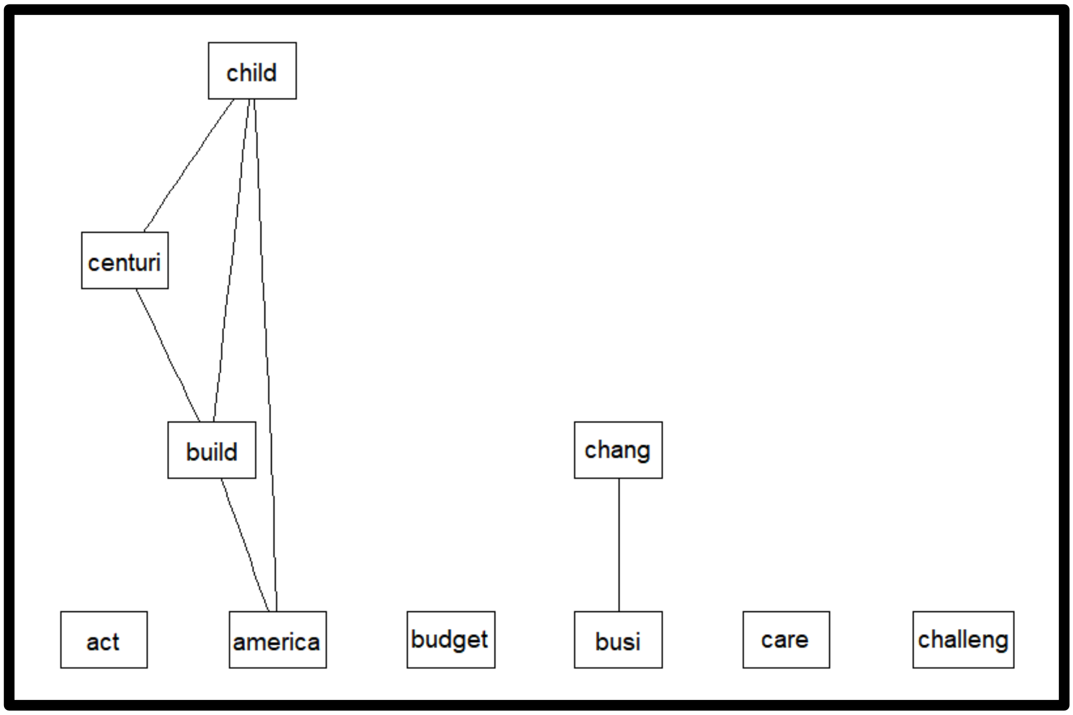


Figure 8. Correlation Plot of Common Words Using Correlation Threshold of 0.4.

Based on this figure, we find that correlations exist between word pairings such as "America" and "build," "America" and "child," "build" and "century," and "business" and "change." Since the correlation threshold is 0.4, each pair of correlated terms has a correlation strength of at least 40%. The correlation between "America" and "build" may imply a strong incentive to build up America and make it stronger. Likewise, the correlation between "build" and "century" may suggest a suggest a desire to promote growth (in infrastructure, the economy, or other areas) within the next century. And the correlation between "business" and "change" may imply a need to propose reforms or changes to help businesses in the U.S. But since correlation does not imply causation, it is still unclear whether these terms are truly related.

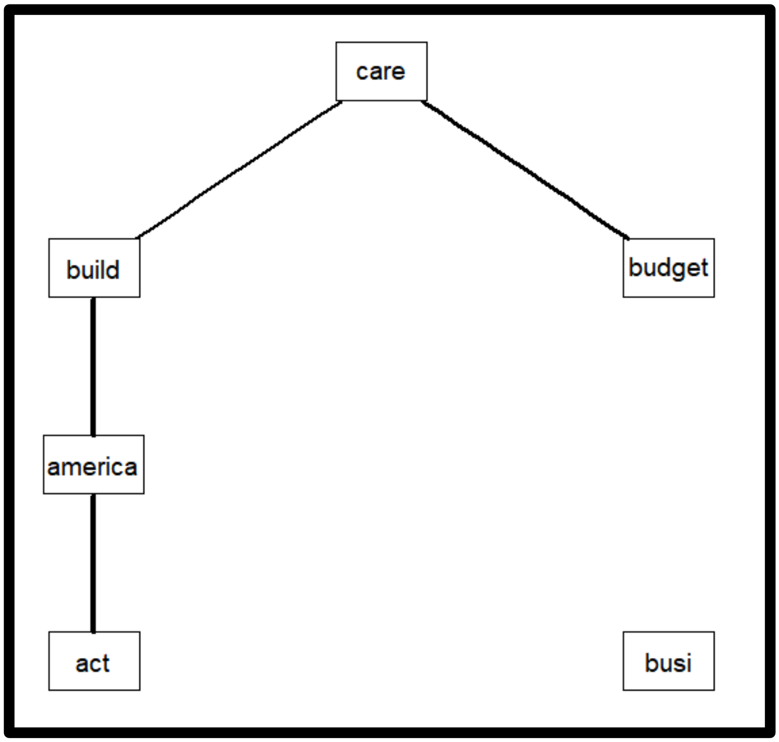


Figure 9. Weighted Correlation Plot Using Correlation Threshold of 0.2.

I will now discuss the outcome of my second correlation plot. This plot is weighted, and it contains 6 words with a minimum frequency of 150 and a correlation threshold of 0.2 (see Figure 9). The weights are represented by the width of each line, for which a thicker line indicates a stronger correlation. This plot shows that "act" is correlated with "America," which is correlated with "build," which is furthermore correlated with "care," which finally correlates with "budget."

But within this sequence of correlations, the terms “build” and “care” have a visibly thinner line than the other pairs of terms. This suggests that these terms have the weakest correlation out of the included words. Furthermore, many correlations in this plot—such as “budget” and “care”—do not appear in the previous correlation plot. This is because the first plot uses a correlation threshold of 0.4 while the second plot only has a threshold of 0.2. Thus, any correlated terms in the second plot that are uncorrelated in the first plot have a correlation strength below 40%.

Another observation is that the words “America” and “build” are the only terms that are correlated in both the first and second correlation plots. This means that these two words have the strongest correlation out of any terms included in the second plot. Likewise, they are the only pair of terms from the second plot that have a correlation strength greater than or equal to 40%. Ultimately, the main issue of using this method is that correlations are determined by words appearing in the same document. Since we have little information regarding the word order and context of each term, it is hard to determine whether the correlations between words have any real implications. But the benefit of this approach is that it identifies potential relationships between some of the most frequent terms. We may be able to find the true meaning of these relationships in a future analysis, once natural language processing technology evolves further.

K-Means Clustering

Finally, I will discuss the results from implementing the k-means clustering method. As mentioned earlier, this model only includes terms that appear in every single document. I first selected the number of clusters by using the equation $k \approx \sqrt{(n/2)}$, where n represents the number of instances. In the context of this assignment, each instance (or row) represents a document.

Since there are 29 documents, using this equation yields an approximate k-value of 4. Thus, I generated the first k-means clustering plot using 4 clusters (see Figure 10).

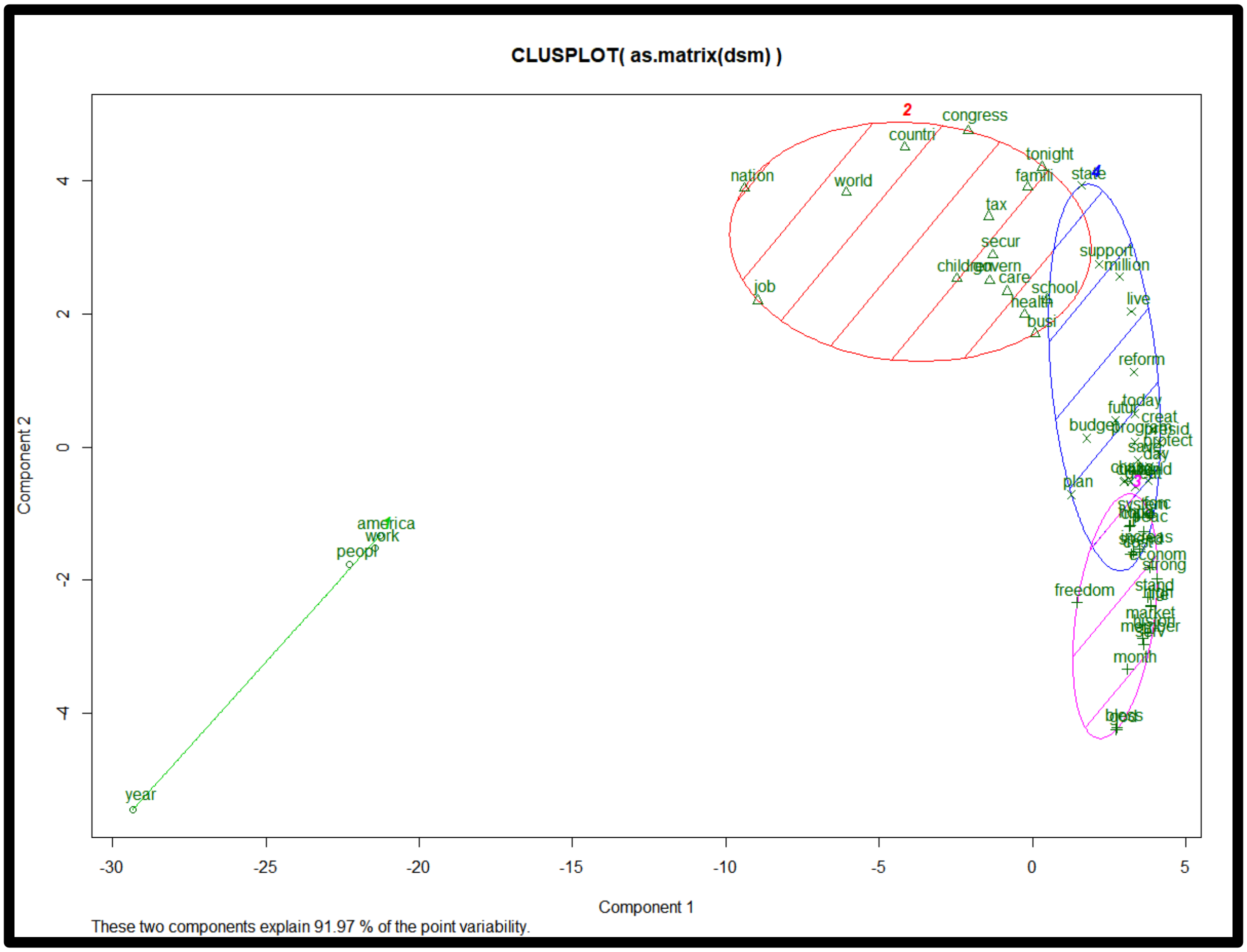


Figure 10. K-Means Clustering Model Using 4 Clusters.

Based on this image, the first cluster contains 4 terms: America, work, people, and year. The second cluster contains 15 terms, such as nation, job, world, congress, and tax. The third cluster has 21 terms, including freedom, market, economy, strong, and bless. The fourth cluster has 20 terms, which include state, support, reform, budget, and create. By analyzing these results, we notice that the first cluster contains the four most frequently-used terms across all speeches. This cluster is much farther from the other three clusters, which suggests that these four terms are unlikely to appear frequently near the words from any other cluster. The other three clusters are quite close together, which implies a higher likelihood of their terms appearing near each other. However, one issue with this model is the significant overlap between the third and fourth clusters. K-means clustering should not involve overlap between clusters, and every instance must belong to a single cluster (James, Witten, Hastie, & Tibshirani, 2013). Because of this overlap, it is likely that the current k-value of 4 is not the ideal number of clusters to use.

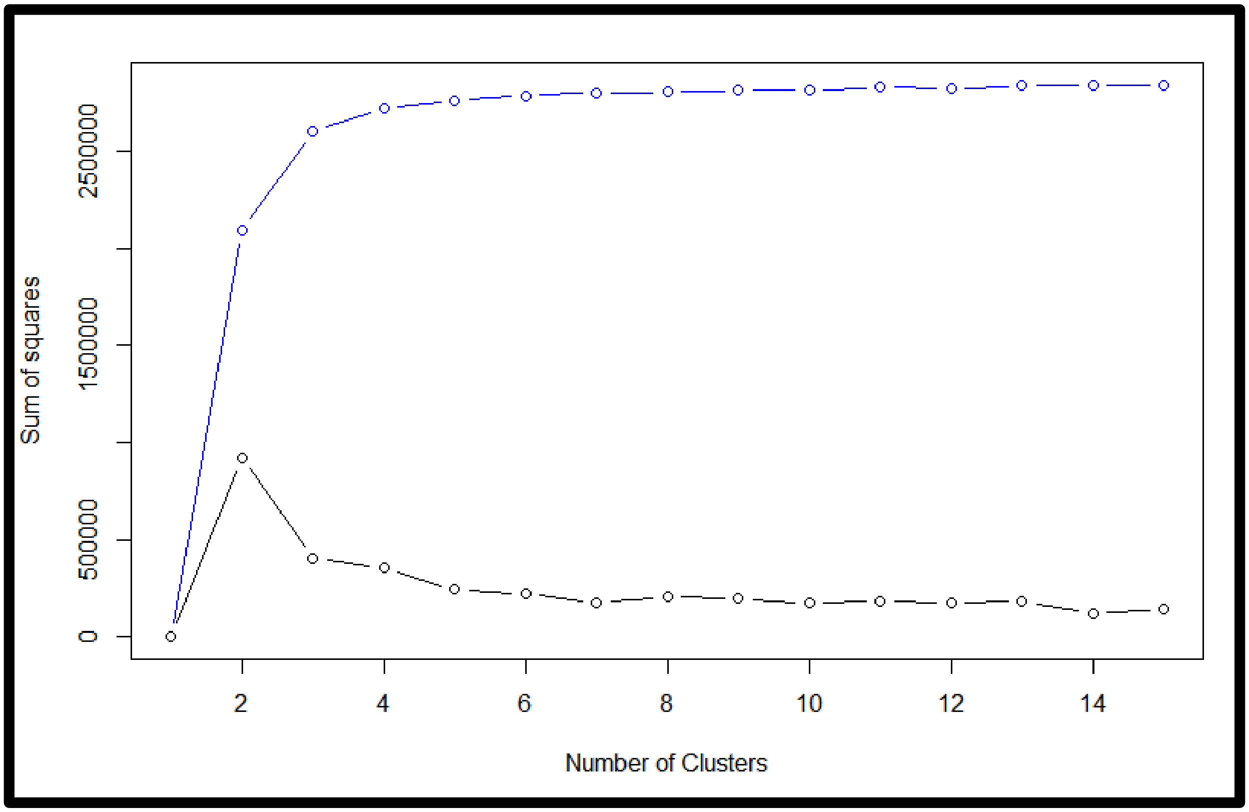


Figure 11. “Elbow Method” Visualization on State of the Union DSM.

To obtain a better value for k, I implemented the elbow method on my data (see Figure 11). This visualization shows the within clusters sum-of-squares (blue) and the between clusters sum-of-squares (black) over k. As stated earlier, the elbow method involves plotting the within clusters sum-of squares as a function of k, and choosing k to be the point where the plot bends the most.

Based on this image, we see that the "elbow" is located at approximately k=3. Thus, I will use this k-value to build a new k-means clustering model. After plotting the new model, it is evident that the overlap between clusters has been largely eliminated (see Figure 12).

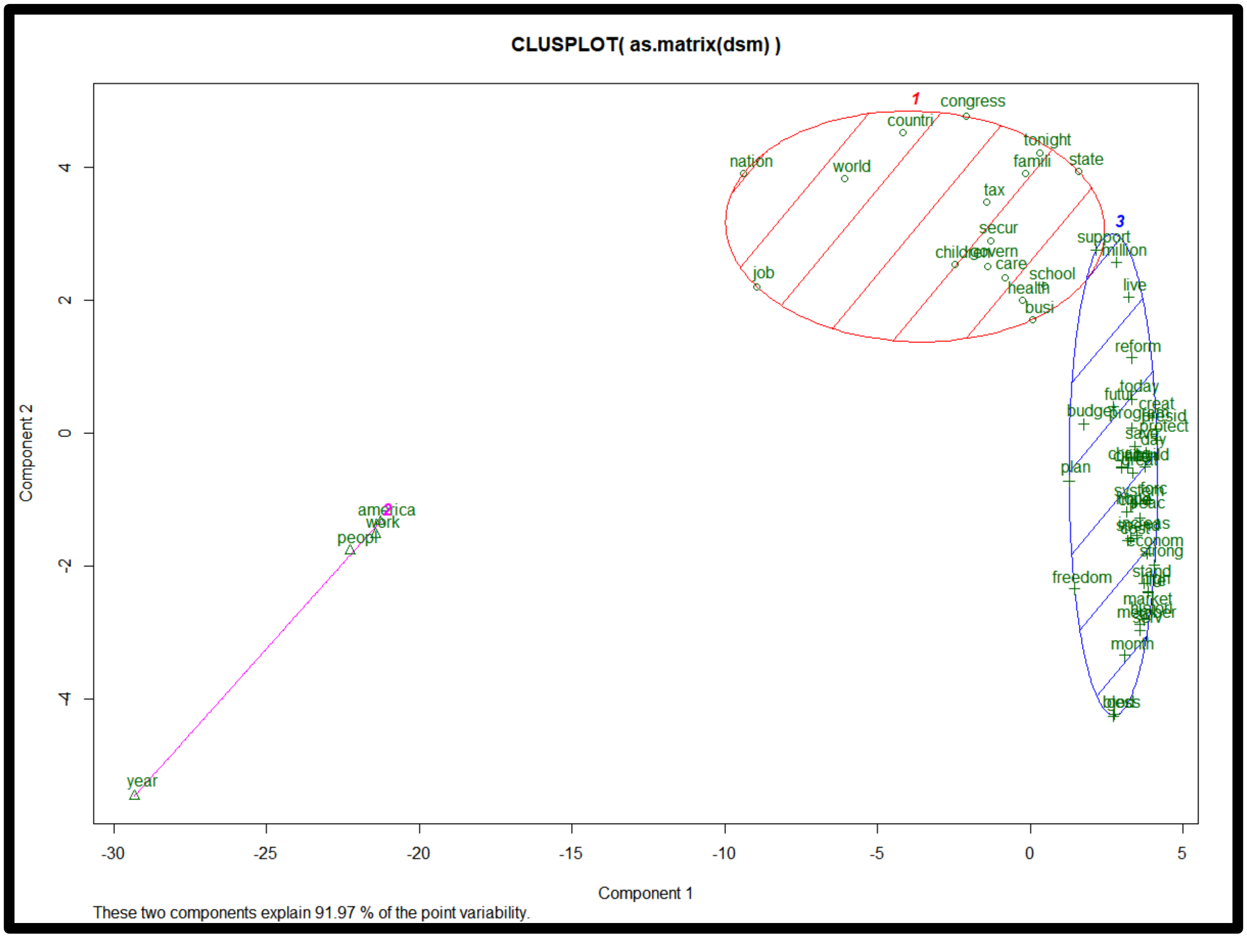


Figure 12. K-Means Clustering Model Using 3 Clusters.

The first cluster of this model contains 16 terms, including nation, job, world, congress, and tax. This cluster is mostly identical to cluster 2 from the original model—but the difference is that it has gained a new term, "state." The second cluster consists of 4 terms: America, work, people, and year. It is exactly identical to cluster 1 from the original model. The third cluster contains 40 terms, which include reform, budget, freedom, market, create, and economy. This cluster largely combines clusters 3 and 4 from the original diagram. One interesting finding is that the clusters seem to group terms based on their frequency. By reviewing my word cloud showing the 40 most common terms (see Figure 3), we find that the words in cluster 2 make up the four most frequent terms, while the words in cluster 1 are still relatively common. The words in cluster 3 are the least common out of the included terms—in fact, several of these terms are not even listed in the word cloud. Of course, the words included in my clustering model have each been included in every document at least once. Thus, the distance between words is likely based on the overall term frequency—rather than whether a term appears in a document. Overall, this method serves as an effective tool for categorizing terms based on their frequencies. Its main shortcoming is the lack of exact frequency ranges for the terms in each cluster. However, this information can be estimated using other methods such as word frequency tables or bar plots.

Conclusion

In this analysis, I incorporated several methods using the bag-of-tokens approach to understand common terms used in the State of the Union addresses between 1989 and 2017. The purpose of my analysis was to understand some of the most important issues facing America in the present day. My first method consisted of a word frequency analysis using two distinct word clouds. The first word cloud displayed the 40 most common terms, and it highlighted the usage of general policy-related terms such as jobs, security, and health care. My second word cloud used a TF-IDF matrix to rank words based on their importance to the document rather than their frequency count. Unlike the first word cloud, this model focused less on general political issues. Instead, it showcased the importance of more specific policies that relate to current events during this period—including terrorism, Afghanistan, Iraq, and immigration.

My second method involved studying the relationships between important words. The first part of this method used word associations to show how often certain terms appear together. This section yielded information regarding the possible attitudes that U.S. presidents may have towards issues such as the economy, health care, and immigration. For instance, the results seem to indicate a desire to promote job creation, improve and modernize the education system, expand health care coverage, and secure the U.S. borders. The second part of this method involved building word correlation plots that show the correlations between frequent terms. This section provided further insights regarding the president's possible attitudes on certain issues. For instance, the correlation between "business" and "change" may refer to proposed economic reforms designed to help American businesses. The strongest correlation was between "build" and "America," suggesting an overall desire to make America stronger.

The final method involved grouping terms into clusters using the k-means clustering algorithm. I limited the terms to include only those that appear in every document. In my initial approach, I used the formula $k \approx \sqrt(n/2)$ to select the number of clusters—which resulted in a k-value of 4. But after plotting the clustering model, I found that there was significant overlap between two of the clusters. Thus, I used the elbow method to select a more ideal number of clusters. By using a k-value of 3, I found that the new clustering model contained much less overlap between clusters. And by analyzing the results, I found that the terms were neatly grouped into 3 clusters based on their frequency ranges.

Throughout this analysis, I noticed several limitations with my methods. One of these issues involved my second word cloud, which contains specific terms such as Iraq, Afghanistan, and terrorism. This word cloud grouped all terms together regardless of the time period they were used. Some of these words refer to specific events which occurred within the past 15 years or so—

meaning that they were not relevant during this entire 28-year period. Although this is a minor issue, it can still serve as inspiration for future analysis. To address this limitation, I suggest conducting text mining studies for each president separately instead of grouping their speeches together. This will help to make the word choices specific to that particular president and time period. But more importantly, it will allow us to analyze the important issues for each president and how the opinion towards these issues changed over time.

However, the biggest challenge of this analysis involved determining word context. For nearly every component of the study, I could only make inferences about the importance and meaning of each word due to the lack of context and word order information. This problem is rooted in the bag-of-tokens approach itself, since it focuses solely on analyzing the words while disregarding the word order and meaning of each term (Bramer, 2016). Still, some of my methods provided a greater degree of insight than others with regards to finding possible word meanings. For example, the word associations offered strong evidence regarding the usage of words such as jobs and immigration—which is largely due to the inclusion of related terms. However, the word cloud provided no information about the meanings of such words since it focused only on word frequency. Another problem is the inability to recognize synonyms, since it causes some terms to have lower frequencies than they actually should. For instance, the word “school” would likely have a higher frequency if related terms such as “education” were recognized as its synonyms. Since these problems are common within natural language processing, I expect that the improvement of NLP technology will provide the tools needed to solve these problems.

References

Basic Text Mining with R. (n.d.). Retrieved February 5, 2018, from https://rstudio-pubs-static.s3.amazonaws.com/132792_864e3813b0ec47cb95c7e1e2e2ad83e7.html

Bati, F. (2015, Fall). Clustering. Lecture presented at UMUC. Retrieved July 9, 2017.

Bramer, M. (2016). *Principles of Data Mining*. Retrieved February 7, 2018.

DeSilver, D. (2017, February 15). U.S. students' academic achievement still lags that of their peers in many other countries. Retrieved February 15, 2018, from http://www.pewresearch.org/fact-tank/2017/02/15/u-s-students-internationally-math-science/

Feinerer, I., & Hornik, K. (2017, March 2). Package ‘tm’. Retrieved February 13, 2018, from https://cran.r-project.org/web/packages/tm/tm.pdf

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques* (3rd ed.). Retrieved February 9, 2018.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. Retrieved June 28, 2017.

Maceli, M. (2016, July 19). Introduction to Text Mining with R for Information Professionals. Retrieved July 18, 2017, from http://journal.code4lib.org/articles/11626

Tatman, R. (2017, July 20). State of the Union Corpus (1989 - 2017). Retrieved February 9, 2018, from https://www.kaggle.com/rtatman/state-of-the-union-corpus-1989-2017/data

Text Mining Analysis Using R: Analyzing Course Descriptions (2018). Retrieved February 5, 2018.