Abstract:

It is our belief that as citizens it is our job to stay informed of issues that are affecting our country. The foundation of American democracy is built upon checks and balances in such a way that ensures no person or institution can rise to the level of tyranny. Our democracy comes under attack when our political institutions and leaders willfully and treacherously attempt to disarm us of our democratic rights through disinformation, smear campaigns, and outright lies. Democracy flourishes in the hands of informed citizens and our current administration is trying to destroy our ability to be informed. To do this we need to collect and analyze large amounts of data and be able to draw our own conclusions

Therefore, in this project, we have downloaded and analyze all of Trump's tweets (as sourced from http://trumptwitterarchive.com/archive which contains over 35,000 entries for us to analyze). We started by pre-processing (tokenizing and creating a corpus) all of the tweets so that we could further analyze the language he uses. We used the corpus to analyze the sentiment of individual tweets against buzz words & sensitive topics such as ”guns”, ”women”, ”immigrants”, etc to see how his sentiment changed as newsworthy events took place in the past. Finally, we produced some interactive graphs based on the results sourced from the above to visualize our findings in a intuitive & accessible format

Proposed Methods:

First we found an archive of all of trump’s tweets going back to the start of his twitter account
We downloaded them to a csv
We cleaned up the csv by changing the dates to the correct date format that R can read
We forced all text to character type to make it easier for R to handle
We forced all numeric non-date columns to integer type

Second we created a column that has a running total that resets every new day so we can properly graph tweets by day
We also calculated the polarity of each tweet and put it into its own corresponding column using the sentimentR R package
The sentiment function returns results for different sentiment lexicons but we choose to go with the standard Jockers 2017 dictionary
Storing the polarity score and graphing it allows us to view his sentiment for each individual tweet

Third we decided to pull out specific polarity scores for current buzzwords
We settled on the following buzz words:

- Liberal
- Democrat
- Republican
- Gun
- Immigrant
- Women
- Latino
- North Korea

To find those buzz words we used a regex find on our original tweet csv and returned the row number
Having the row numbers we subsetted the original csv to a specific sentiment data frame
We plotted these as a line graph corresponding to the date and a positive and negative score

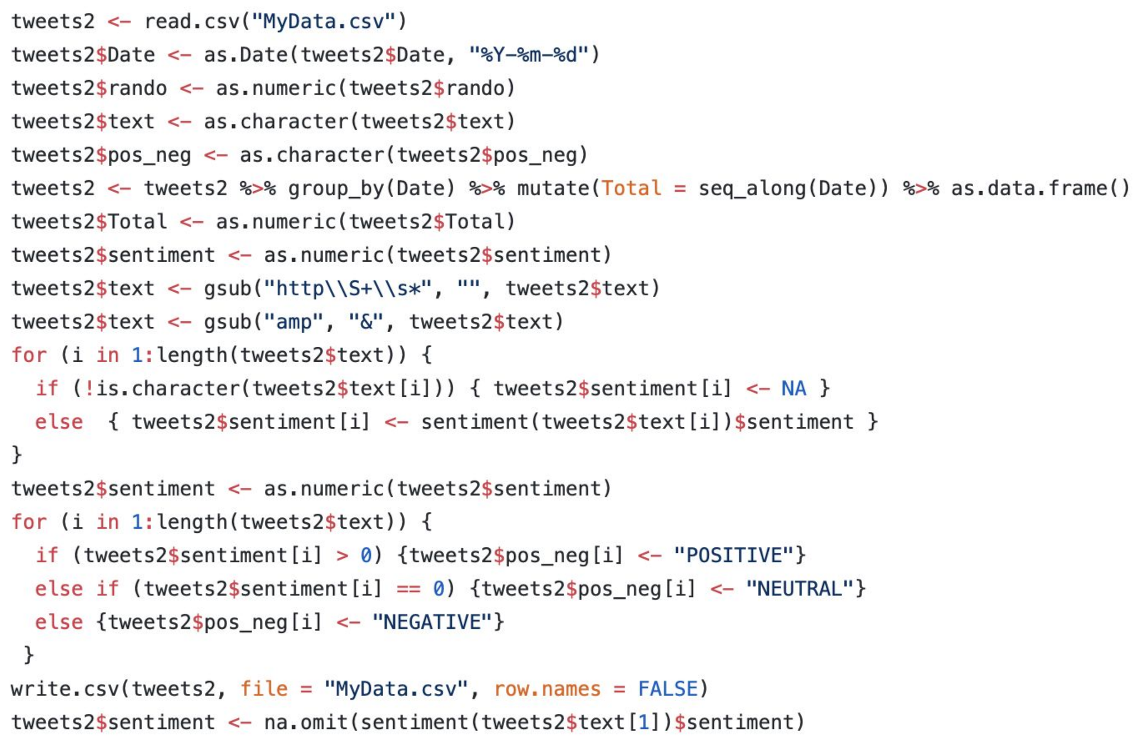
Fourth we copied his tweet text column to a .txt file as required by the TM (text-mining) R package
The TM package creates a corpus for us to use that:

- Removes urls
- Fixes text encoding errors
- Removes special characters
- Removes punctuation
- Removes stop words
- Removes numbers
- Removes white space

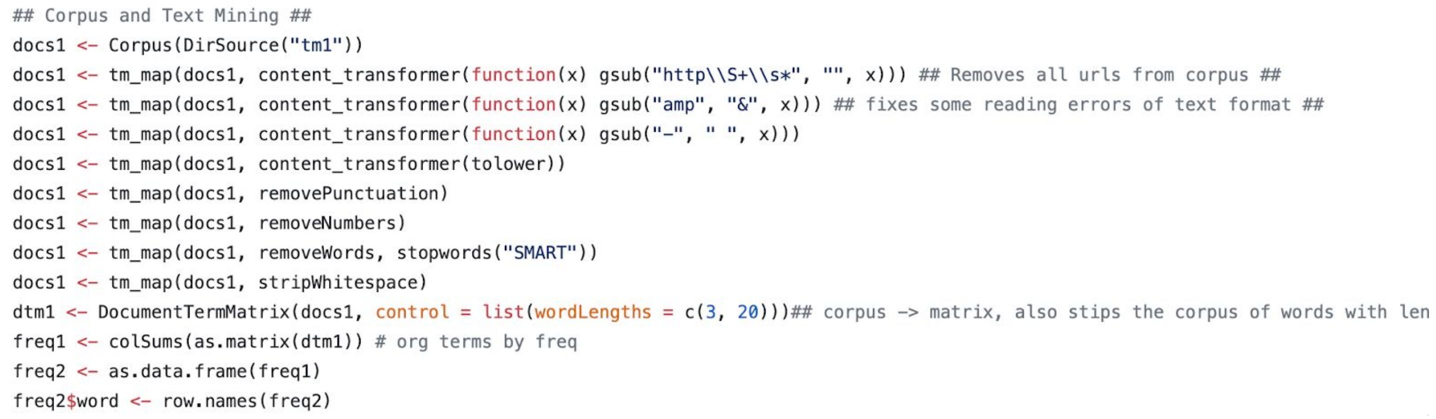
It then creates a document term matrix that allows us to access word counts, frequency of words, and co-occurrence if we choose
We were going to use this for the word count per tweet table and a word cloud but we didn’t have time to tie it together

Results

To start with we needed to perform some pre-processing on the dataset we retrieved from TrumpTwitterArchive in order to prepare it for analysis & extract features for visualization. Some snippets of the code involved in this are as follows:



```
tweets2 <- read.csv("MyData.csv")
tweets2$Date <- as.Date(tweets2$Date, "%Y-%m-%d")
tweets2$rando <- as.numeric(tweets2$rando)
tweets2$text <- as.character(tweets2$text)
tweets2$pos_neg <- as.character(tweets2$pos_neg)
tweets2 <- tweets2 %>% group_by(Date) %>% mutate(Total = seq_along(Date)) %>% as.data.frame()
tweets2$Total <- as.numeric(tweets2$Total)
tweets2$sentiment <- as.numeric(tweets2$sentiment)
tweets2$text <- gsub("http\\S+\\s*", "", tweets2$text)
tweets2$text <- gsub("amp", "&", tweets2$text)
for (i in 1:length(tweets2$text)) {
  if (!is.character(tweets2$text[i])) { tweets2$sentiment[i] <- NA }
  else  { tweets2$sentiment[i] <- sentiment(tweets2$text[i])$sentiment }
}
tweets2$sentiment <- as.numeric(tweets2$sentiment)
for (i in 1:length(tweets2$text)) {
  if (tweets2$sentiment[i] > 0) {tweets2$pos_neg[i] <- "POSITIVE"}
  else if (tweets2$sentiment[i] == 0) {tweets2$pos_neg[i] <- "NEUTRAL"}
  else {tweets2$pos_neg[i] <- "NEGATIVE"}
 }
write.csv(tweets2, file = "MyData.csv", row.names = FALSE)
tweets2$sentiment <- na.omit(sentiment(tweets2$text[1])$sentiment)
```



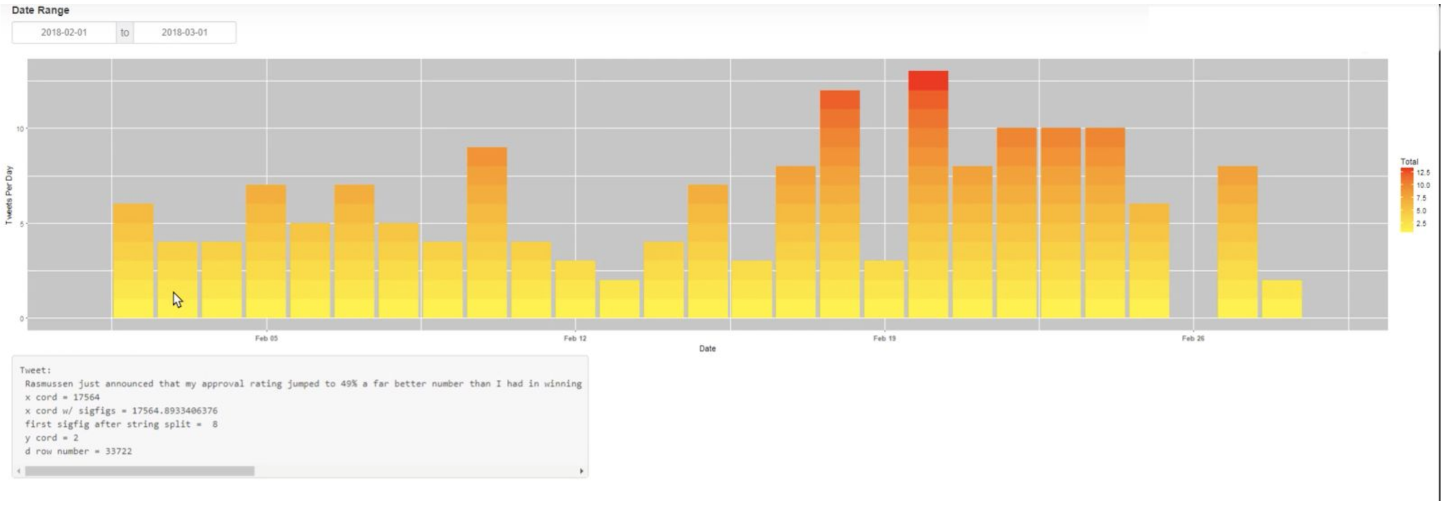
```
## Corpus and Text Mining ##
docs1 <- Corpus(DirSource("tm1"))
docs1 <- tm_map(docs1, content_transformer(function(x) gsub("http\\S+\\s*", "", x))) ## Removes all urls from corpus ##
docs1 <- tm_map(docs1, content_transformer(function(x) gsub("amp", "&", x))) ## fixes some reading errors of text format ##
docs1 <- tm_map(docs1, content_transformer(function(x) gsub("-", " ", x)))
docs1 <- tm_map(docs1, content_transformer(tolower))
docs1 <- tm_map(docs1, removePunctuation)
docs1 <- tm_map(docs1, removeNumbers)
docs1 <- tm_map(docs1, removeWords, stopwords("SMART"))
docs1 <- tm_map(docs1, stripWhitespace)
dtm1 <- DocumentTermMatrix(docs1, control = list(wordLengths = c(3, 20)))## corpus -> matrix, also stips the corpus of words with len
freq1 <- colSums(as.matrix(dtm1)) # org terms by freq
freq2 <- as.data.frame(freq1)
freq2$word <- row.names(freq2)
```

We intended to have some interactive graphs to display the following information in a time-series based format:

1. Processed version of the tweet (tokenized and pre-processed)
2. Buzz words contained in the tweet
3. Individual word count corresponding to the tweet
4. Overall sentiment of the tweet
5. Change in sentiment toward key topics mentioned in his tweets over time

a. examples of the above are: “democrats”, “immigrants”, “guns”, etc.

We decided to render components 1-3 in a interactive bar graph :

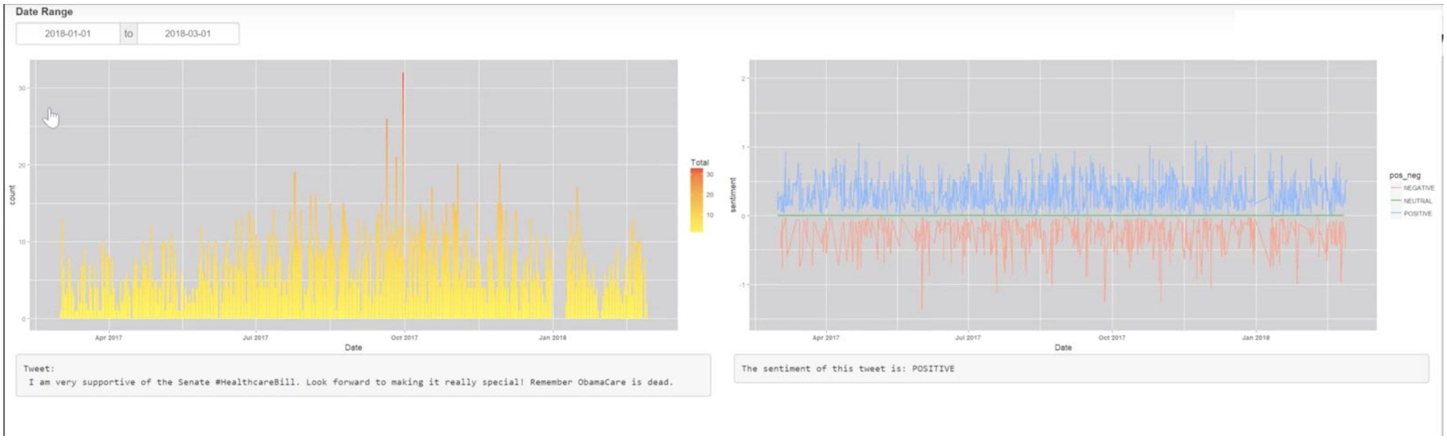

Date Range

2018-02-01 to 2018-03-01

```
Tweet:
 Rasmussen just announced that my approval rating jumped to 49% a far better number than I had in winning
 x cord = 17564
 x cord w/ sigfigs = 17564.8933406376
 first sigfig after string split =  8
 y cord = 2
 d row number = 33722
```

Figure 1: Processed tweet visualized in interactive bar graph

Component 4 is rendered as a line graph which is tied to the selected tweet as follows :

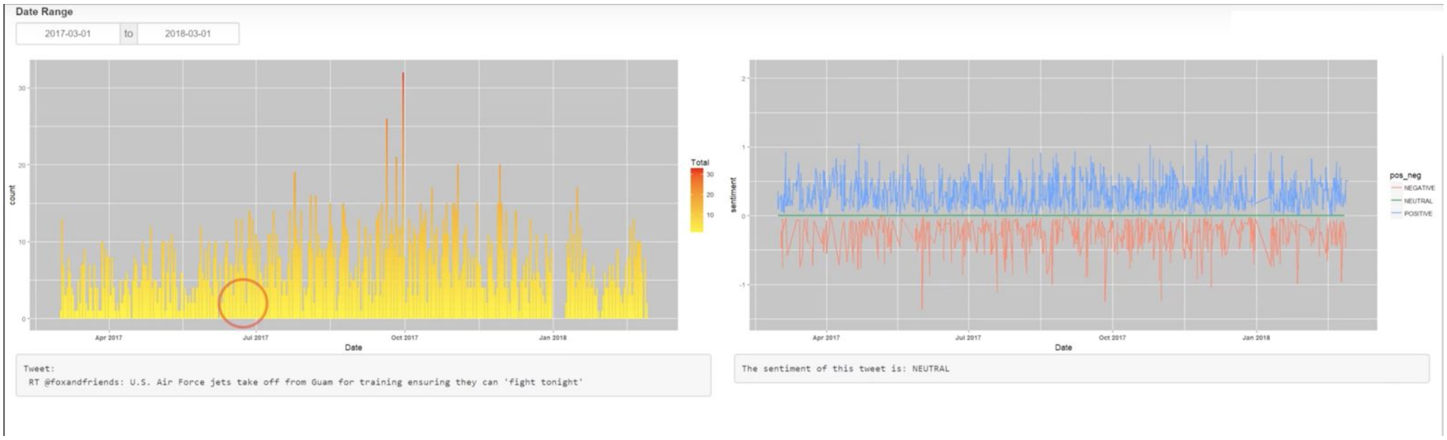

Date Range

2018-01-01 to 2018-03-01

```
Tweet:
 I am very supportive of the Senate #HealthcareBill. Look forward to making it really special! Remember ObamaCare is dead.
```

```
The sentiment of this tweet is: POSITIVE
```

Figure 2: Tweet with POSITIVE sentiment and corresponding visualization

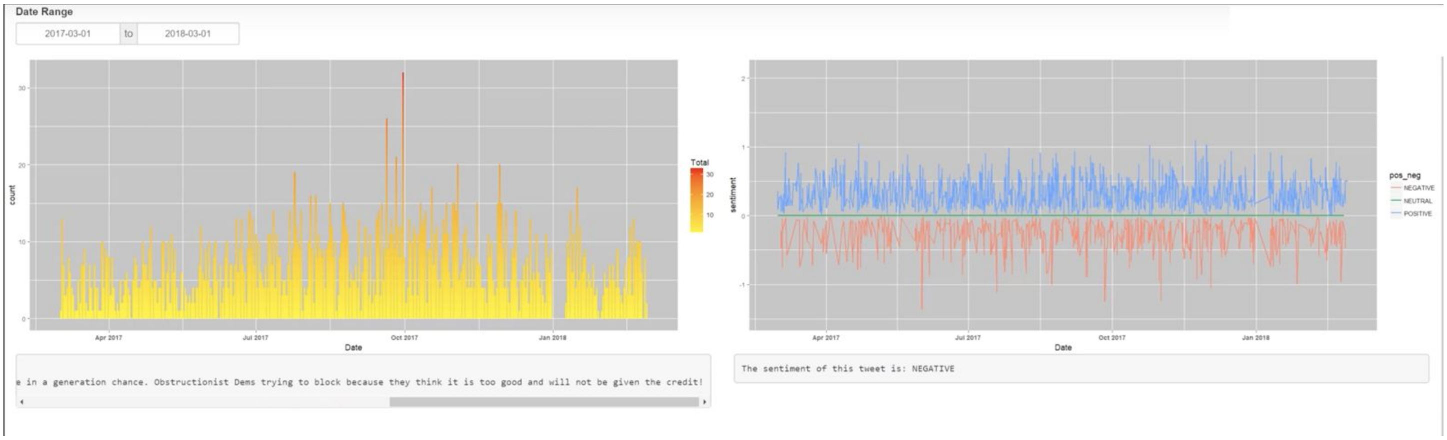

Date Range

2017-03-01 to 2018-03-01

```
Tweet:
 RT @foxandfriends: U.S. Air Force jets take off from Guam for training ensuring they can 'fight tonight'
```

```
The sentiment of this tweet is: NEUTRAL
```

Figure 3: Tweet with NEUTRAL sentiment and corresponding visualization



Date Range

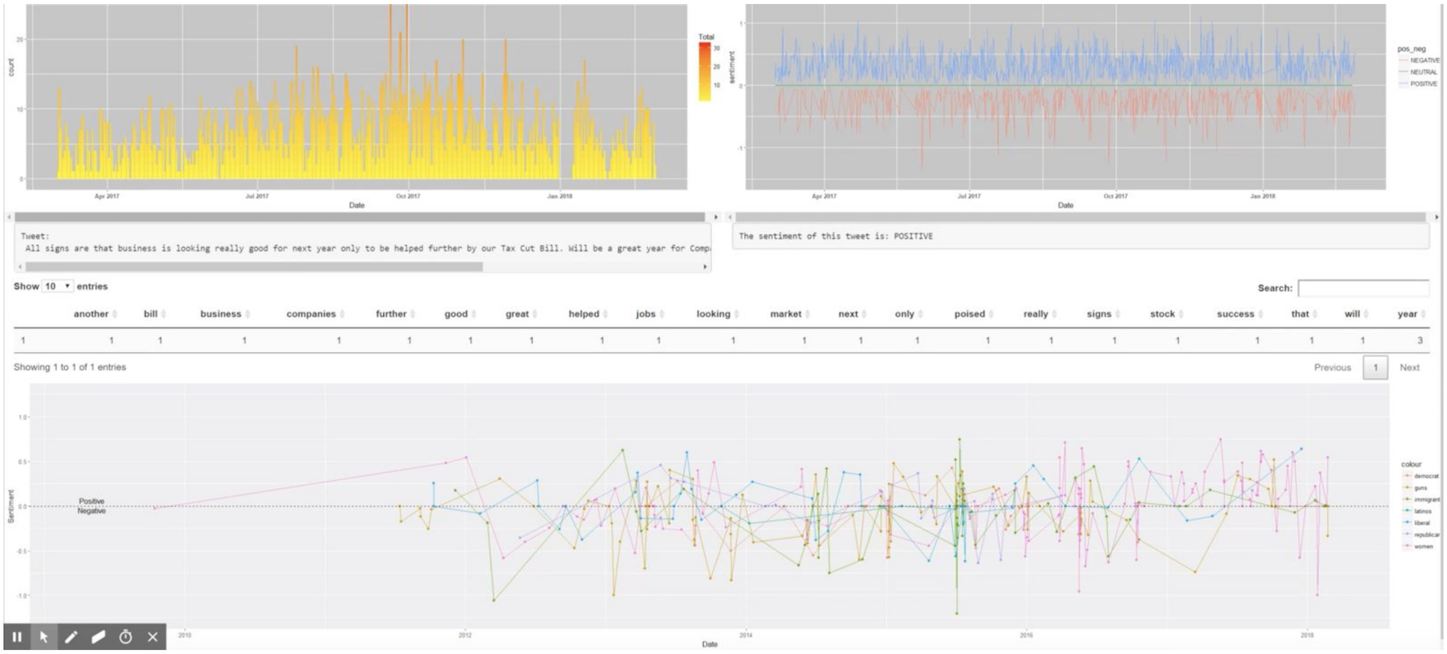
2017-03-01 to 2018-03-01

```
e in a generation chance. Obstructionist Dems trying to block because they think it is too good and will not be given the credit!
```

```
The sentiment of this tweet is: NEGATIVE
```

Figure 4: Tweet with NEGATIVE sentiment and corresponding visualization

Finally we display component 5 as a line graph with various indicators for the topics :



```
Tweet:
 All signs are that business is looking really good for next year only to be helped further by our Tax Cut Bill. Will be a great year for Comp
```

```
The sentiment of this tweet is: POSITIVE
```

Show 10 entries

Search:

	another	bill	business	companies	further	good	great	helped	jobs	looking	market	next	only	poised	really	signs	stock	success	that	will	year
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3

Showing 1 to 1 of 1 entries

Previous 1 Next

Figure 5: Line graph with lines marking changing sentiments on certain buzz-words