# 2022 State of State Speech Comparison between Democrats and Republicans

Gamze Bilsen

## Background

I got the pdf's of 43 states governor's state of the state speeches in 2022. I separated the states into their respective political categories: there were only 18 republican states among the 43, so I randomly eliminated some states from democratic states to end with 18 state of the state speeches for each category (democratic and republican).

I changed the pdf's into csv's, then combined the csv's together given each governor/states' political leaning - this was done using python.

My hypothesis is that democratic and republican states' speeches will be different. I'm mostly interested in seeing if one or the other talk about the climate compared to other words, and if so, among who does it the most - my hypothesis being democratic states will talk about that more compared to republican states. This is under the assumption that different governors of each party can be grouped well together - which is highly deductive.

## Bag of words and removal of upper case, white space, punctuation, and word stemming.

```
#Preliminary Code
stopifnot(require(wordcloud))
stopifnot(require(tm))
stopifnot(require(dplyr))
stopifnot(require(data.table))
stopifnot(require(SnowballC))
stopifnot(require(lsa))
stopifnot(require(broom))
stopifnot(require(scales))
library(tidyverse)
library(tidytext)
library(stringr)
library(wordcloud)
library(tm)
#Rweka wouldn't download
```

```
setwd("~/Documents/columbia/spring/adv analytics/lab2v2/data22/")
dem <- read.csv('dem/dem.csv', header=TRUE)
rep <- read.csv('rep/rep.csv', header=TRUE)
```

```r
dem$X0 <- gsub("&amp", " ", dem$X0)
dem$X0 <- gsub("(RT|via)((?:\\b\\W*@\\w+)+)", " ", dem$X0)
dem$X0 <- gsub("@\\w+", " ", dem$X0)
dem$X0 <- gsub("[[:punct:]]", " ", dem$X0)
dem$X0 <- gsub("[[:digit:]]", " ", dem$X0)
dem$X0 <- gsub("http\\w+", " ", dem$X0)
dem$X0 <- gsub("[ \t]{2,}", " ", dem$X0)
dem$X0 <- gsub("^\\s+|\\s+$", " ", dem$X0)
dem$X0 <- gsub("state", " ", dem$X0,ignore.case = TRUE)
dem$X0 <- gsub("governor|will|also|address|www|delivers", " ", dem$X0,ignore.case = TRUE)
dem_text <- sapply(dem$X0, function(row) iconv(row, "latin1", "ASCII", sub=""))
dem_2 <- paste(unlist(dem_text), collapse =" ")
dem_2 <- Corpus(VectorSource(dem_2))
dem_2 <- tm_map(dem_2, PlainTextDocument)
dem_2 <- tm_map(dem_2, removePunctuation)
dem_2 <- tm_map(dem_2, content_transformer(tolower))
dem_2 <- tm_map(dem_2, removeWords, stopwords("english"))
```

```r
rep$X0 <- gsub("&amp", " ", rep$X0)
rep$X0 <- gsub("(RT|via)((?:\\b\\W*@\\w+)+)", " ", rep$X0)
rep$X0 <- gsub("@\\w+", " ", rep$X0)
rep$X0 <- gsub("[[:punct:]]", " ", rep$X0)
rep$X0 <- gsub("[[:digit:]]", " ", rep$X0)
rep$X0 <- gsub("http\\w+", " ", rep$X0)
rep$X0 <- gsub("[ \t]{2,}", " ", rep$X0)
rep$X0 <- gsub("^\\s+|\\s+$", " ", rep$X0)
rep$X0 <- gsub("state", " ", rep$X0,,ignore.case = TRUE)
rep$X0 <- gsub("governor|will|also|address|www|delivers", " ", rep$X0,ignore.case = TRUE)
rep_text <- sapply(rep$X0, function(row) iconv(row, "latin1", "ASCII", sub=""))
rep_2 <- paste(unlist(rep_text), collapse =" ")
rep_2 <- Corpus(VectorSource(rep_2))
rep_2 <- tm_map(rep_2, PlainTextDocument)
rep_2 <- tm_map(rep_2, removePunctuation)
rep_2 <- tm_map(rep_2, content_transformer(tolower))
rep_2 <- tm_map(rep_2, removeWords, stopwords("english"))
```

## Relative word frequencies for each bag of words & Comparison

```r
dem.dtm <- TermDocumentMatrix(dem_2)
dem.m <- as.matrix(dem.dtm)
dem.v <- sort(rowSums(dem.m),decreasing=TRUE)
dem.d <- data.frame(word = names(dem.v),freq=dem.v)
head(dem.d, 15)
```

```
##               word freq
## new            new  327
## people      people  244
## can            can  235
## one            one  216
## work          work  211
```

```
## year         year  178
## years       years  178
## every       every  173
## families families  156
## time         time  156
## million   million  147
## just         just  145
## make         make  144
## know         know  143
## need         need  137
```

```
rep.dtm <- TermDocumentMatrix(rep_2)
rep.m <- as.matrix(rep.dtm)
rep.v <- sort(rowSums(rep.m),decreasing=TRUE)
rep.d <- data.frame(word = names(rep.v),freq=rep.v)
head(rep.d, 15)
```

```
##                 word freq
## can              can  287
## year            year  244
## work            work  209
## people        people  200
## new              new  194
## years          years  192
## one              one  181
## dakota        dakota  170
## make            make  169
## last            last  158
## million      million  156
## education education  148
## just            just  147
## today          today  144
## time            time  138
```

Comparing the top 15 words of democrats vs republicans, they seem very similar as they both have can, will, work, people and year/years: the state of the state speeches of both seem to more or less use a lot of the same words that are geared towards a broad range of audiences. With regards to the differences they have, democrats use "families" more than republicans, which is contrary to the general reputation of republicans being more family oriented. Republican's also have a lot of dakota, meaning the governor's from the dakota's most likely speak of their state a lot in their speeches.

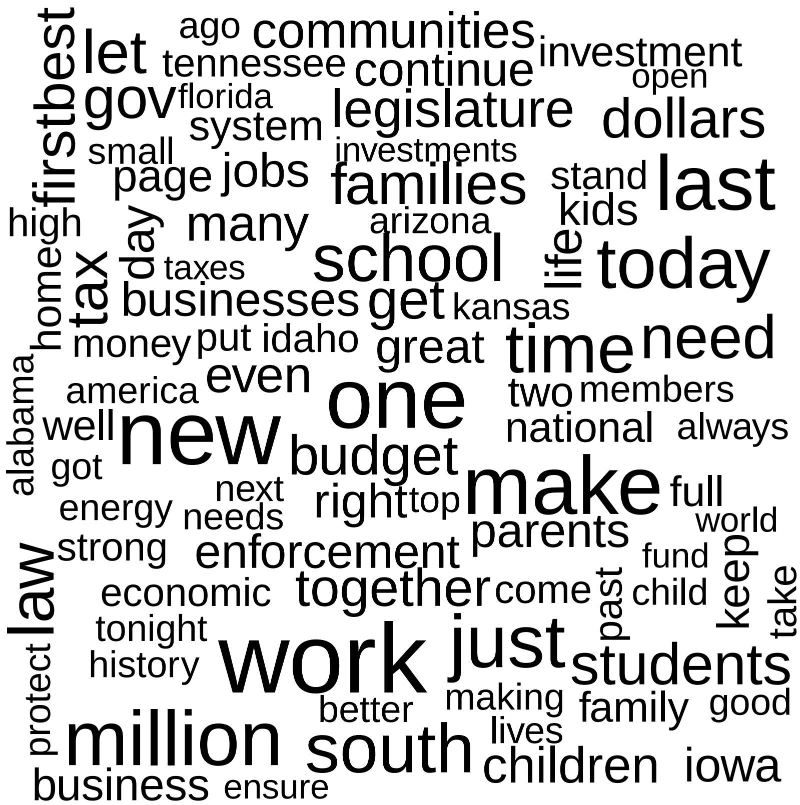
```
all.corpus <- c(dem_2, rep_2)
all.corpus <- Corpus(VectorSource(all.corpus))
all.tdm <- TermDocumentMatrix(all.corpus)
all.m <- as.matrix(all.tdm)
all.df = as.data.frame(all.m)
all.df = all.df[,c(1,4)]
colnames(all.df) <- c("dem", "rep")
df <- cbind(names = rownames(all.df), all.df)
rownames(df) <- 1:nrow(df)
```

## Word Cloud of each

I wanted to see the words that have above 50 frequency in word clouds for each to perform a better comparison between the two.

```
wordcloud(df$names, df$dem, min.freq=50, random.color=T,
ordered.colors=T) #democrats
```


```
wordcloud(df$names, df$rep, min.freq=50, random.color=T,
ordered.colors=T) #republicans
```




Right off the bat, it seems that Democrats seem to the same words more often. Democrats also use 'new' a lot compared to Republicans' top usage of 'people'. Looking at the wordcloud, we see families in Republican usage as a top word as well which I wasn't able to uncover from just looking at the top 15 words.

Looking primarily at the Democratic wordcloud, I was able to decipher words around jobs/economy (investments, workforce..), budgeting, infrastructure, education/schools, health/pandemic. I also see words such as future, forward, community and support, words that have more hopefully together type of social/political meaning to them. "Every" is also one of the largest words, most likely emphasizing things such as everyone in the state benefiting from their work. 'Climate' is mentioned at least 50 times which is a hopefuly sign, however it's relatively very small compared to other topics mentioned. There are likely other words that are related to what the state can do with regards to climate change mitigation/adaptation such as adjusting the infrastructure: most likely the usage of the word "infrastructure" could also be related to the climate as the infrastructure bill has a lot of emphasis on just transition - additional analysis is needed.

Looking primarily at the Republican wordcloud, I was similarly able to decipher words around education, jobs/economy, budgeting. I see words such as protect, ensure and stand, which is in line with the Republican reputation of being protective. Nation also has a larger font in Rebublican word cloud which is in line with their populist ideology of being nationalistic and protective. Republican speeches also tend to have more allusion to the state names compared to Democratic states: tt seems Republican's tend to want to address the state by name more compared to Democratic states who tend to have words related to outside the state name. Lastly, there are no words around climate but there is a large emphasis on the word infrastructure.

So overall, it seems they both talk a lot about the economy, budgeting, education, infrastructure and the health system. In their wording, it seems Republicans tend to want to emphasize their state in their speeches compared to Democrats.

I want to see the word climate's # frequencies in each speech: it seems below that republicans don't mention it once. This could be due to an error with the lemmatization, however, even if it is, there's a significant difference between the two since we could reliably assume a similar lemmatization error would be present for the democratic speeches.

Republicans mention nation almost twice as much as Democrats.

```
df %>% filter(names=='climate')
```

```
##     names dem rep
## 1 climate  50   0
```

```
df %>% filter(names=='nation')
```

```
##    names dem rep
## 1 nation  53 116
```

## Further differentiation using distinctive words of each

```
summing = function(x) x/sum(x, na.rm=T)
df.2 = apply(all.df, 2, summing)
df.2 <- cbind(names = rownames(df.2), df.2)
rownames(df.2) <- 1:nrow(df.2)
total <- merge(df,df.2,by="names")
i <- c('dem.y', 'rep.y')
total[ , i] <- apply(total[ , i], 2,
                    function(x) as.numeric(as.character(x)))
total$dem.over.rep = (total$dem.y) - (total$rep.y)
sort.OT <- total[order(total$dem.over.rep) , ]
(sort.OT[1:15, ])
```

```
##             names dem.x rep.x        dem.y       rep.y dem.over.rep
## 2708       dakota     1   170 2.630333e-05 0.003945506 -0.003919202
## 10155       south     4   134 1.052133e-04 0.003109987 -0.003004773
## 7362        north     7   126 1.841233e-04 0.002924316 -0.002740193
## 6973     missouri     0    85 0.000000e+00 0.001972753 -0.001972753
## 6220          law    53   125 1.394076e-03 0.002901107 -0.001507031
## 5343        idaho     0    62 0.000000e+00 0.001438949 -0.001438949
## 6027       kansas     0    58 0.000000e+00 0.001346114 -0.001346114
## 10801   tennessee     1    59 2.630333e-05 0.001369323 -0.001343019
## 795       arizona     0    57 0.000000e+00 0.001322905 -0.001322905
## 7574     oklahoma     0    57 0.000000e+00 0.001322905 -0.001322905
## 7189       nation    53   116 1.394076e-03 0.002692227 -0.001298151
## 6970  mississippi     1    55 2.630333e-05 0.001276487 -0.001250184
## 301       alabama     0    52 0.000000e+00 0.001206861 -0.001206861
## 10469    students    54   110 1.420380e-03 0.002552974 -0.001132594
## 3649  enforcement    29    81 7.627966e-04 0.001879917 -0.001117121
```

```
rev.sort.OT <- total[rev(order(total$dem.over.rep) ), ]
rev.sort.OT[1:15, ]
```

```
##              names dem.x rep.x       dem.y        rep.y dem.over.rep
## 7281           new   327   194 0.008601189 4.502518e-03  0.004098671
## 8059        people   244   200 0.006418013 4.641771e-03  0.001776241
## 5253       housing    68     5 0.001788626 1.160443e-04  0.001672582
## 4033      families   156   106 0.004103319 2.460139e-03  0.001643181
## 12523      workers    91    35 0.002393603 8.123100e-04  0.001581293
## 7594           one   216   181 0.005681519 4.200803e-03  0.001480716
## 2207   commonwealth    56     0 0.001472986 0.000000e+00  0.001472986
## 5953          jobs   129    84 0.003393130 1.949544e-03  0.001443586
## 3817         every   173   137 0.004550476 3.179613e-03  0.001370863
## 8447      prepared    63    13 0.001657110 3.017151e-04  0.001355395
## 5382      illinois    54     4 0.001420380 9.283543e-05  0.001327544
## 2058       climate    50     0 0.001315167 0.000000e+00  0.001315167
## 8468         press    52     3 0.001367773 6.962657e-05  0.001298147
## 8132          phil    50     1 0.001315167 2.320886e-05  0.001291958
## 9649         scott    50     5 0.001315167 1.160443e-04  0.001199122
```

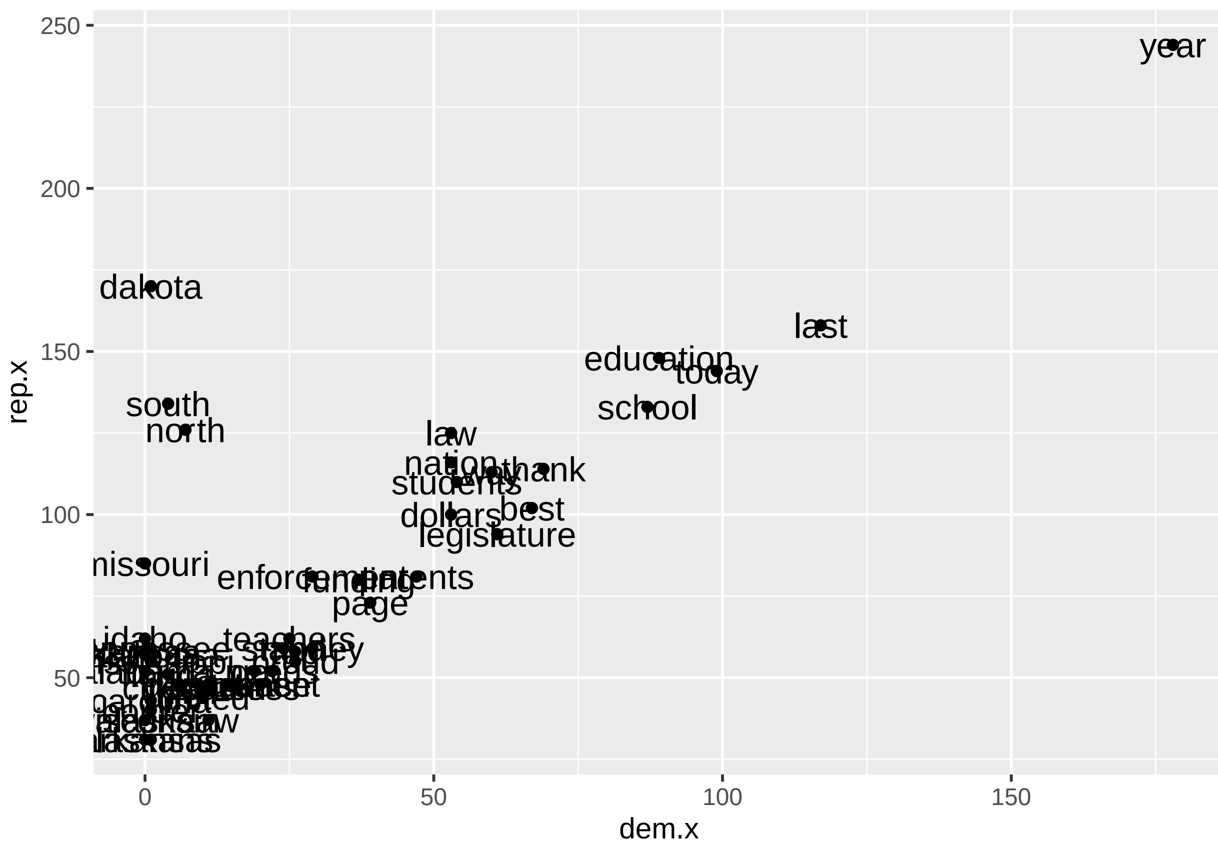
The first dataframe is words that Republicans use more often than Democrats - this seems to be mostly their respective state names which is in line with the Republican word cloud that had a lot of state words. There's also 'law' that is more often used, which goes in line with the law&order emphasis of the Republican populist ideology.

While most commonly used Republican words are more or less never used in Democratic speeches, except 'law', top Democratic words are also popular words in Republican speeches except for 'commonwealth' which is likely from the commonwealth states and 'housing'. Housing could be a heavier emphasis from Democratic governors due to their having larger cities wind tend to require more housing investment.

```r
to_graph = sort.OT[1:50, ]
q = qplot(dem.x, rep.x, data = to_graph)
q<- q + geom_text(aes(label=names), size = 4.5)
```

I also wanted to look at the correlation between the two - it seems that education, law, enforcement, nation, dollars, students are couple of the words that I was able to decipher which Republicans say more often the Democrats.

```r
to_graph = sort.OT[1:50, ]
q = qplot(dem.x, rep.x, data = to_graph)
q + geom_text(aes(label=names), size = 4.5)
```




## Statistical tests of association between the bags of words

So far, speeches seems to be similar to each other with regards to the main topics they talk about, and despite Republicans emphasizing their respective states more. However, I want to look at both cosine similarity and chi squared test to check this result.

```r
library(lsa)
cosine <- as.data.frame(cosine(all.m))[c(1,4),c(1,4)]
colnames(cosine) <- c("dem", "rep")
cosine
```

```
##         dem       rep
## 1 1.0000000 0.8871894
## 4 0.8871894 1.0000000
```

```r
ctable <- table(all.m)
chisq.test(ctable)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  ctable
## X-squared = 7083821, df = 150, p-value < 2.2e-16
```

Looking at the cosine, it seems that they are very similar given the 0.865 cosine. The chi-squared test is also statistically significant meaning both political parties' governors' aggregated speeches are not independent.

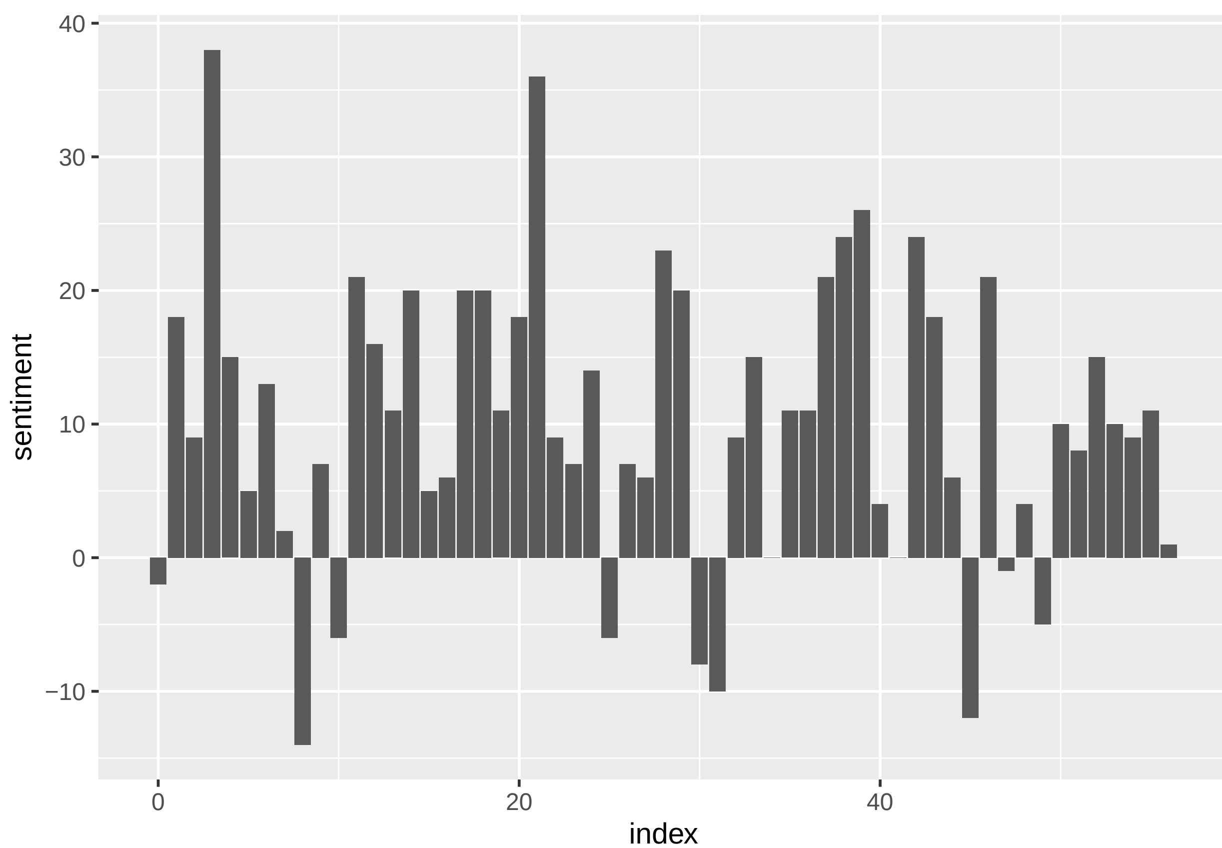
## Sentiment analysis of the bags of words

I first wanted to see if which party tends to be more negative.

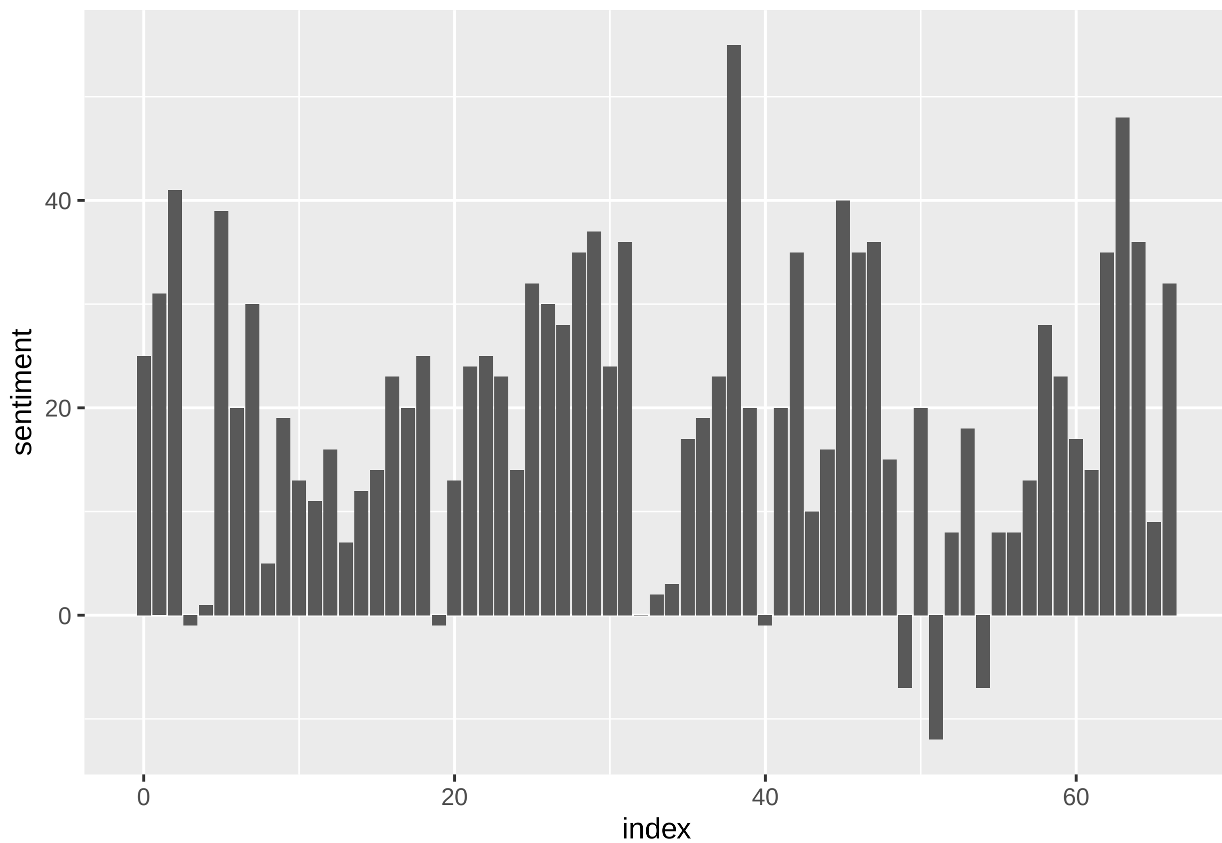
```
rep.words = as.data.frame(rep.m)
rep.words <- cbind(names = rownames(rep.words), rep.words)
rownames(rep.words) <- 1:nrow(rep.words)

dem.words = as.data.frame(dem.m)
dem.words <- cbind(names = rownames(dem.words), dem.words)
rownames(dem.words) <- 1:nrow(dem.words)
```

```
setwd("~/Documents/columbia/spring/adv analytics/lab2v2/data22/")
dem <- read.csv('dem/dem.csv', header=TRUE)
rep <- read.csv('rep/rep.csv', header=TRUE)

dem$X0 <- gsub("&amp", " ", dem$X0)
dem$X0 <- gsub("(RT|via)((?:\\b\\W*@\\w+)+)", " ", dem$X0)
dem$X0 <- gsub("@\\w+", " ", dem$X0)
dem$X0 <- gsub("[[:punct:]]", " ", dem$X0)
dem$X0 <- gsub("[[:digit:]]", " ", dem$X0)
dem$X0 <- gsub("http\\w+", " ", dem$X0)
dem$X0 <- gsub("[ \t]{2,}", " ", dem$X0)
dem$X0 <- gsub("^\\s+|\\s+$", " ", dem$X0)
dem$X0 <- gsub("state", " ", dem$X0)
dem$X0 <- gsub("governor", " ", dem$X0)

rep$X0 <- gsub("&amp", " ", rep$X0)
rep$X0 <- gsub("(RT|via)((?:\\b\\W*@\\w+)+)", " ", rep$X0)
rep$X0 <- gsub("@\\w+", " ", rep$X0)
rep$X0 <- gsub("[[:punct:]]", " ", rep$X0)
rep$X0 <- gsub("[[:digit:]]", " ", rep$X0)
rep$X0 <- gsub("http\\w+", " ", rep$X0)
rep$X0 <- gsub("[ \t]{2,}", " ", rep$X0)
rep$X0 <- gsub("^\\s+|\\s+$", " ", rep$X0)
rep$X0 <- gsub("state", " ", rep$X0)
rep$X0 <- gsub("governor", " ", rep$X0)

tokenize_d <- tibble(line=1:4530,text=dem$X0)
data(stop_words)
to_sent_d <- tokenize_d %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words)
tokenize_r <- tibble(line=1:5358,text=rep$X0)
to_sent_r <- tokenize_r %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words)
```

```r
#Code from here: https://www.tidytextmining.com/sentiment.html
dem_sentiment <- to_sent_d %>%
  inner_join(get_sentiments("bing")) %>%
  count(index = line %/% 80, sentiment) %>%
  pivot_wider(names_from = sentiment, values_from = n, values_fill = 0) %>%
  mutate(sentiment = positive - negative)

rep_sentiment <- to_sent_r %>%
  inner_join(get_sentiments("bing")) %>%
  count(index = line %/% 80, sentiment) %>%
  pivot_wider(names_from = sentiment, values_from = n, values_fill = 0) %>%
  mutate(sentiment = positive - negative)

ggplot(dem_sentiment, aes(index, sentiment)) +
  geom_col(show.legend = FALSE)
```



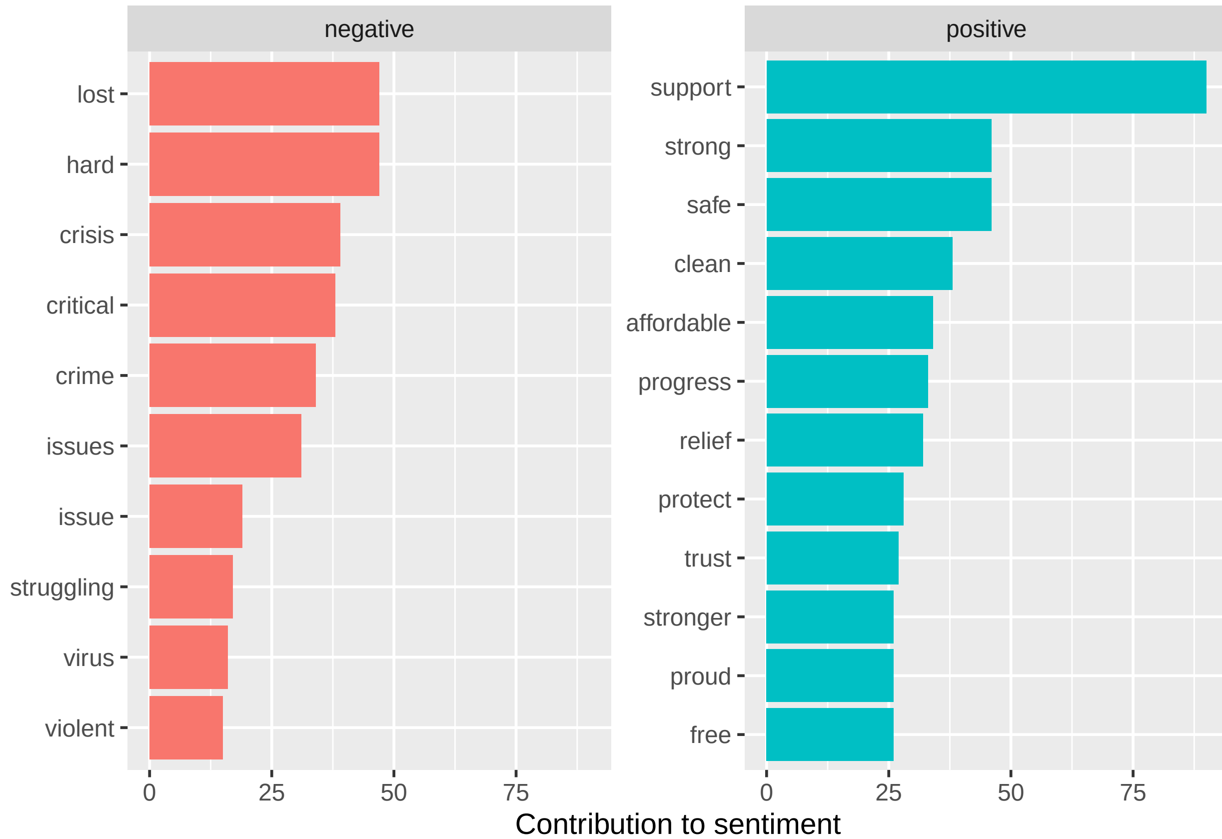

```r
ggplot(rep_sentiment, aes(index, sentiment)) +
  geom_col(show.legend = FALSE)
```



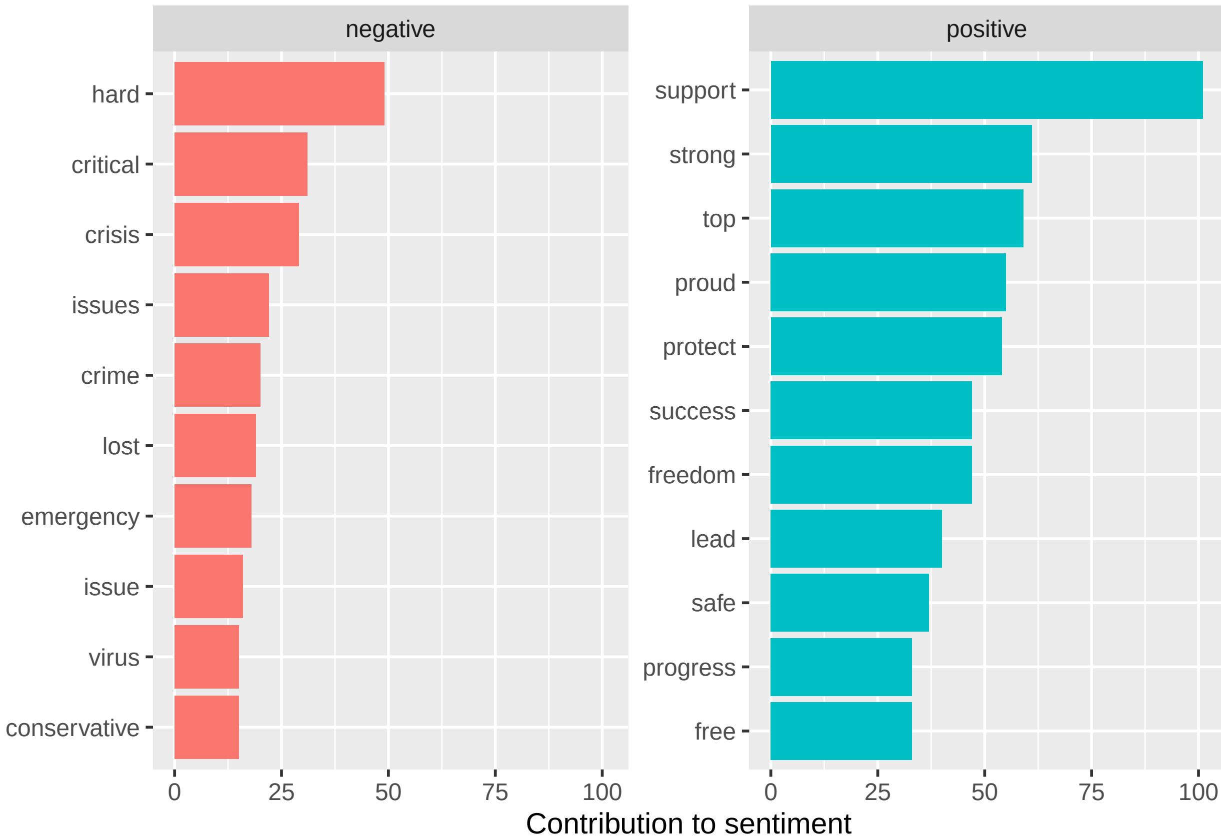

It seems that democrats tend to be more negative then republicans in their state of the state speeches - however, this doesn't consider the not words, thus additional analysis is needed.

I next wanted to see the most common positive and negative words for each.

```
d_word_counts <- to_sent_d %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()

d_word_counts %>%
  group_by(sentiment) %>%
  slice_max(n, n = 10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(x = "Contribution to sentiment",
       y = NULL)
```




```
r_word_counts <- to_sent_r %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()

r_word_counts %>%
  group_by(sentiment) %>%
  slice_max(n, n = 10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(x = "Contribution to sentiment",
       y = NULL)
```



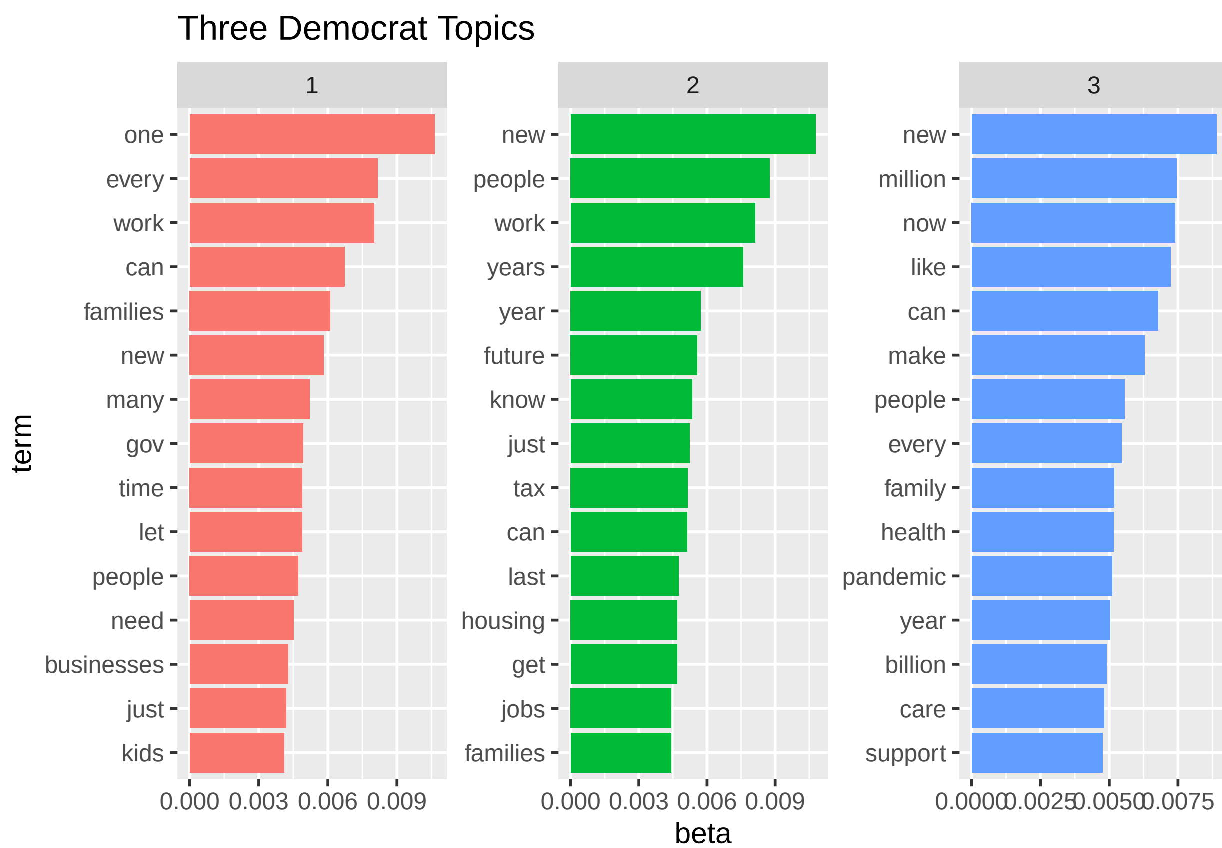

For Democrats, negative words are mostly vague without the additional context. However, we can see that there is some emphasis on crime and covid. With regards to positive words, it seems that they are using support, being strong against these negative words/hardships. Affordable is also the top 5th word.

Republicans similarly use vague words however they havemuch less negative words compared to Democrats. Oddly enough the sentiment analysis identifies conversative as a negative word. In either case, they seem to highlight the same issues as Democrats with the emergency/virus words and crime. With regards to positive words, they similarly use protection, progress, proudness and freedom.

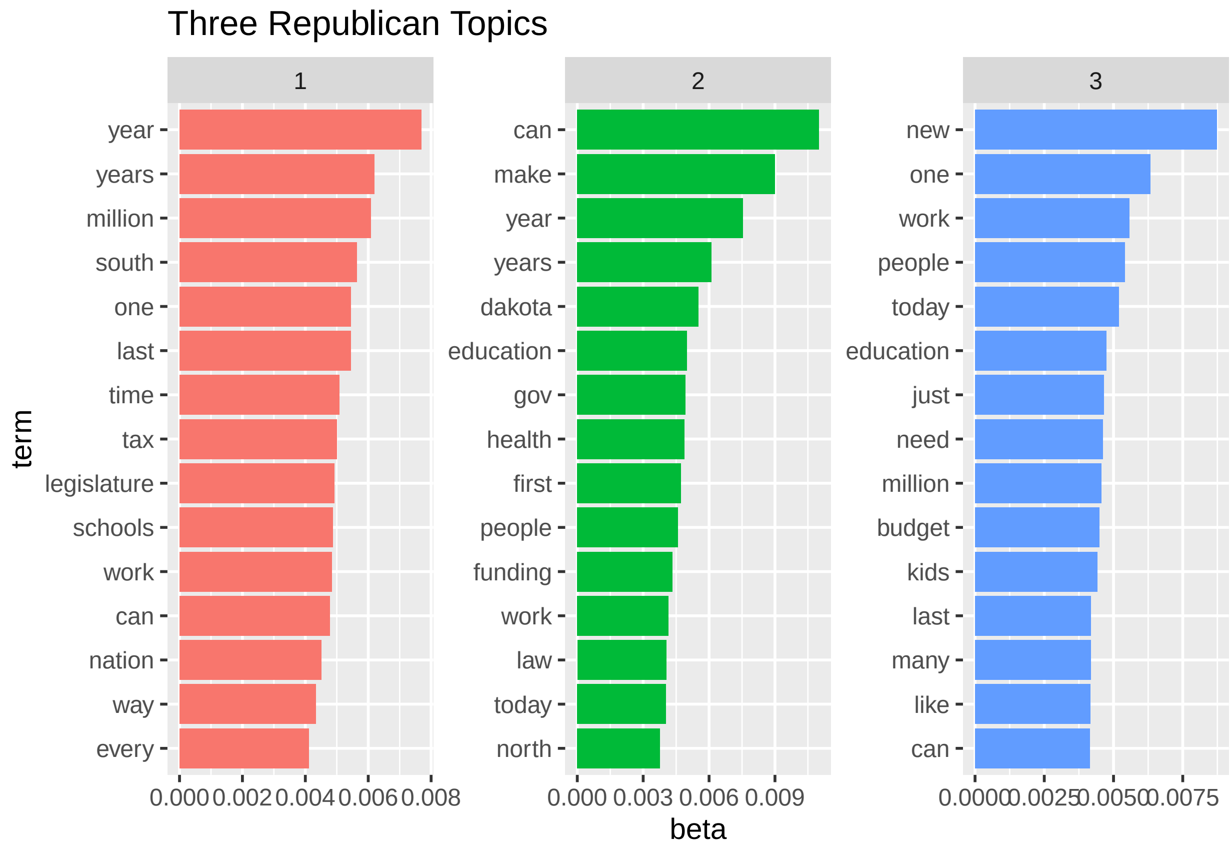
## Topic Modelling

```
library(topicmodels)
dem.dtm <- DocumentTermMatrix(dem_2)
d_lda <- LDA(dem.dtm, k = 3, control = list(seed = 1234))
d_topics <- tidy(d_lda, matrix = "beta")

d_top_terms <- d_topics %>%
  group_by(topic) %>%
  slice_max(beta, n = 15) %>%
  ungroup() %>%
  arrange(topic, -beta)

d_top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
```

```
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered() + ggtitle('Three Democrat Topics')
```




```
rep.dtm <- DocumentTermMatrix(rep_2)
r_lda <- LDA(rep.dtm, k = 3, control = list(seed = 1234))
r_topics <- tidy(r_lda, matrix = "beta")


r_top_terms <- r_topics %>%
  group_by(topic) %>%
  slice_max(beta, n = 15) %>%
  ungroup() %>%
  arrange(topic, -beta)

r_top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered() + ggtitle('Three Republican Topics')
```




Looking at the Democratic topics, we can see that the first is likely about families and working. The next one is more about businesses, taxation and jobs. The last one is more about healthcare and the pandemic.

Looking at the Republican topics, we can see that the first is likely about taxation and legislation. The next one is more about funding needed for things such as education and law enforcement. The last one seems to be a combination of the two with emphasis on budgeting and education.

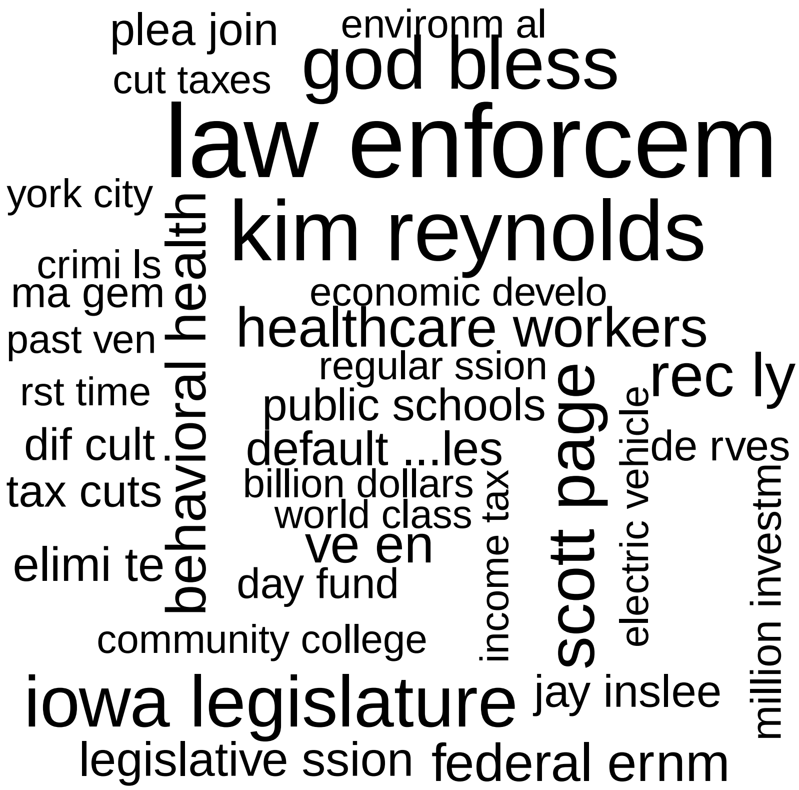
## Bigram Analysis

Code is from here: https://bookdown.org/Maxine/tidy-text-mining/tokenizing-by-n-gram.html

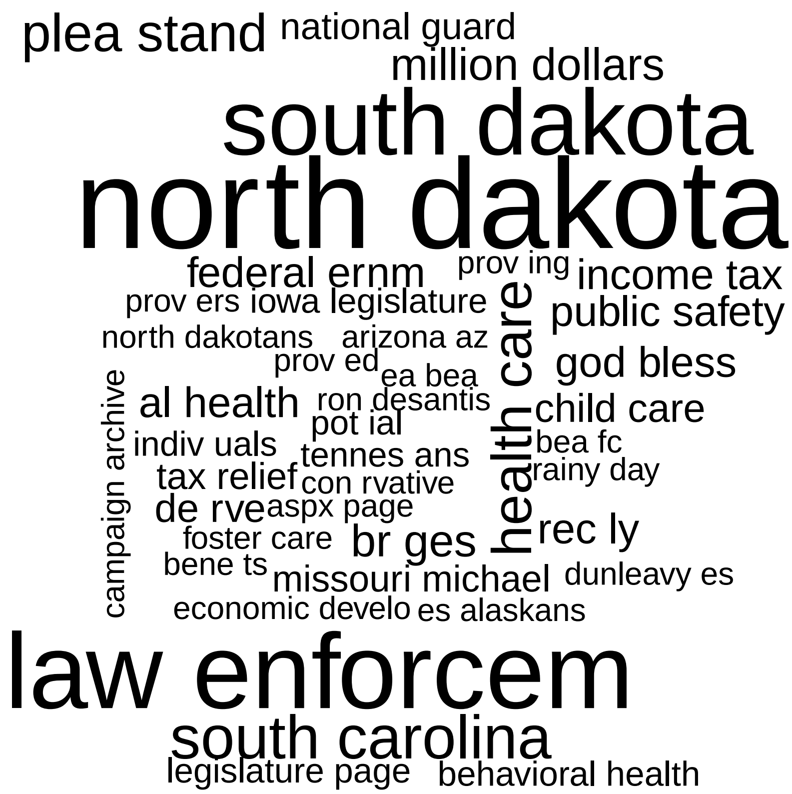
```
tokenize_r$text <- gsub("&amp", " ", tokenize_r$text)
tokenize_r$text <- gsub("(RT|via)((?:\\b\\W*@\\w+)+)", " ", tokenize_r$text)
tokenize_r$text <- gsub("@\\w+", " ", tokenize_r$text)
tokenize_r$text <- gsub("[[:punct:]]", " ", tokenize_r$text)
tokenize_r$text <- gsub("[[:digit:]]", " ", tokenize_r$text)
tokenize_r$text <- gsub("http\\w+", " ",tokenize_r$text)
tokenize_r$text <- gsub("[ \t]{2,}", " ", tokenize_r$text)
tokenize_r$text <- gsub("^\\s+|\\s+$", " ", tokenize_r$text)
tokenize_r$text <- gsub("NA}state|watch|address", " ", tokenize_r$text,ignore.case = TRUE)
tokenize_r$text <- gsub("governor", " ", tokenize_r$text,ignore.case = TRUE)
tokenize_r$text <- gsub("delivers|subscribe|version|lady|id|kim reynolds", " ", tokenize_r$text,ignore.
tokenize_r$text <- gsub("press|release|news|speech|releases|lieute|phil|gov|www|pm|site|sites|delivery|
```

```
r_bigrams <- tokenize_r %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2)
r_bigrams <- r_bigrams %>%
  count(bigram, sort = TRUE) %>%
  separate(bigram, into = c("word1", "word2"), sep = " ")%>%
  filter(!word1 %in% stop_words$word,
         !word2 %in% stop_words$word) %>%
  unite(bigram, c(word1, word2), sep = " ")

tokenize_d$text <- gsub("&amp", " ", tokenize_d$text)
tokenize_d$text <- gsub("(RT|via)((?:\\b\\W*@\\w+)+)", " ", tokenize_d$text)
tokenize_d$text <- gsub("@\\w+", " ", tokenize_d$text)
tokenize_d$text <- gsub("[[:punct:]]", " ", tokenize_d$text)
tokenize_d$text <- gsub("[[:digit:]]", " ", tokenize_d$text)
tokenize_d$text <- gsub("http\\w+", " ",tokenize_d$text)
tokenize_d$text <- gsub("[ \t]{2,}", " ", tokenize_d$text)
tokenize_d$text <- gsub("^\\s+|\\s+$", " ", tokenize_d$text)
tokenize_d$text <- gsub("state", " ", tokenize_d$text,ignore.case = TRUE)
tokenize_d$text <- gsub("watch", " ", tokenize_d$text,ignore.case = TRUE)
tokenize_d$text <- gsub("address", " ", tokenize_d$text,ignore.case = TRUE)
tokenize_d$text <- gsub("NA", " ", tokenize_d$text,ignore.case = TRUE)
tokenize_d$text <- gsub("governor", " ", tokenize_d$text,ignore.case = TRUE)
tokenize_d$text <- gsub("delivers", " ", tokenize_d$text,ignore.case = TRUE)
tokenize_d$text <- gsub("press|release|news|speech|releases|lieute|phil|gov|www|pm|site|sites|delivery|

d_bigrams <- tokenize_d %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2)
d_bigrams <- d_bigrams %>%
  count(bigram, sort = TRUE) %>%
  separate(bigram, into = c("word1", "word2"), sep = " ")%>%
  filter(!word1 %in% stop_words$word,
         !word2 %in% stop_words$word) %>%
  unite(bigram, c(word1, word2), sep = " ")
```

```
d<-data.frame(d_bigrams)
d<-d[2:8613,]
wordcloud(d$bigram, d$n, min.freq=7, random.color=T,
ordered.colors=T)
```




```
r<-data.frame(r_bigrams)
r<-r[2:9336,]
wordcloud(r$bigram, r$n, min.freq=13, random.color=T,
ordered.colors=T)
```




Looking at the word clouds of bigrams, we can see that the pdf to csv likely didn't do a great job or lemmatization didn't work well as well. There are a lot of words that should be one such as 'indiv uals' that were separated, likely leading to mistaken results from unique words and bigrams.

However, looking at what we do have, we can see that clean energy is a top topic of conversation for Democrats which is inline with their usage of climate a lot prior and also my hypothesis. There doesn't seem to be anything related to climate change in the bigram analysis for republicans as well: it seems that their governors on average have been more focused on health, taxation and public safety.

Democrats seem to place a heavy emphasis on climate change (yay), health, much more, taxation and economy.

Lastly, because it's a bit hard to decipher using the word clouds, I wanted to lastly compare the distinctive bigrams of each.

```
bigramdf <- merge(d,r,by='bigram')
colnames(bigramdf) <- c('bigram','dem','rep')
bigramdf$dem.over.rep = (bigramdf$dem) - (bigramdf$rep)
sort.OT <- bigramdf[order(bigramdf$dem.over.rep) , ]
(sort.OT[1:25, ])
```

```
##                 bigram dem rep dem.over.rep
## 971       south dakota   1  67          -66
## 591       law enforcem  27  79          -52
## 736         plea stand   2  32          -30
## 667    million dollars   8  26          -18
## 505         income tax   7  24          -17
```

```
## 208       con rvative   1  15          -14
## 743           pot ial   4  17          -13
## 380      federal ernm  11  23          -12
## 396       foster care   4  15          -11
## 317          elem ary   1  11          -10
## 73           bene ts   4  13           -9
## 244      cyber curity   1  10           -9
## 286   economic develo   7  15           -8
## 820            rec ly  14  22           -8
## 930  school districts   3  11           -8
## 1030    supreme court   4  12           -8
## 1036       task force   4  12           -8
## 1052       tax relief  11  19           -8
## 75        bi partisan   1   8           -7
## 111    budget surplus   1   8           -7
## 926     school choice   2   9           -7
## 48           att ion   4  10           -6
## 134       care system   2   8           -6
## 259          de rved   1   7           -6
## 350  executive budget   1   7           -6
```

```
rev.sort.OT <- bigramdf[rev(order(bigramdf$dem.over.rep) ), ]
rev.sort.OT[1:25, ]
```

```
##                    bigram dem rep dem.over.rep
## 176          clean energy  20   2           18
## 651          massachu tts  15   1           14
## 1147         unpreced ed  18   8           10
## 779        property taxes  14   4           10
## 786         public health  12   3            9
## 760              prev ion  12   3            9
## 15              al health  32  23            9
## 852          repre ntative  14   6            8
## 458     healthcare workers  12   4            8
## 258               de rve  28  20            8
## 637           lower costs   8   1            7
## 842          regular ssion   7   1            6
## 718              past ven   7   1            6
## 680        moving forward   8   2            6
## 623           local ernm   7   1            6
## 603      legislative ssion  10   4            6
## 417          future ready   7   1            6
## 303       education system  11   5            6
## 260              de rves   8   2            6
## 195      community college   7   1            6
## 1176           viol crime   8   3            5
## 1128 transportation system   6   1            5
## 778          property tax  11   6            5
## 450         health rvices  11   6            5
## 334          environm al   7   2            5
```

The first dataframe is the bigrams republicans use more often than democrats. It seems that law enforcement, million dollars (likely new investments the government has done is highlighted), schooling/education related issues, cyber security, economic development/taxation and security tend to be highlighted.

The second dataframe shows that Democrats tend to highlight clearn energy more often than Republicans - it seems it was also mentioned twice for republicans. They also tend to place an emphasis on health, taxation/costs, education and transportation. They also mention violent crimes and the environment.

It seems overall that Democrats tend to place more emphasis on climate change mitigation/adaptation efforts, specifically in the form of clean energy as it's mentioned 18x more often than Republican's speeches. With regards to other topics, it seems all state governors understandably tend to place an emphasis on healthcare, education (with democrats talking more about community colleges compared to Republicans talking more about elementary school/younger schooling) and taxation. It seems that Democrats tend to talk on issues that are more in line with urban areas such as transportation and housing than Republicans.

```
filter(bigramdf, grepl("energy", bigram, ignore.case = TRUE))
```

```
##                   bigram dem rep dem.over.rep
## 1           clean energy  20   2           18
## 2          energy especia   1   1            0
## 3        energy isgrowing   1   1            0
## 4          energy sources   2   1            1
## 5        renewable energy   2   3           -1
## 6 supportenergy sources   1   1            0
## 7            wind energy   5   1            4
```

Climate wouldn't show up, likely because there isn't bigrams that contain it in the Republican text based on the bag of words analysis. So I wanted to look at energy related text, and it seems overall Democrats talk more about it although Republicans also tend to mention it as well.