Assignment3

Shuyi Yu

11/27/2019

Question1

```
file <- read.csv(file = "/Users/Shuyi/Desktop/AS3code/platforms.csv", header = TRUE, sep = ",")
demCorpus <- VCorpus(DirSource("/Users/Shuyi/Desktop/AS3code/d16"))
repCorpus <- VCorpus(DirSource("/Users/Shuyi/Desktop/AS3code/r16"))
```

Question2

```
demCorpus <- tm_map(demCorpus, tolower)
demCorpus <- tm_map(demCorpus, removeWords, stopwords("SMART"))
demCorpus <- tm_map(demCorpus, removeWords, c("democrats", "americans", "american", "america"))
demCorpus <- tm_map(demCorpus, removeNumbers)
demCorpus <- tm_map(demCorpus, removePunctuation)

repCorpus <- tm_map(repCorpus, tolower)
repCorpus <- tm_map(repCorpus, removeWords, stopwords("SMART"))
repCorpus <- tm_map(repCorpus, removeWords, c("republican", "americans", "american", "america"))
repCorpus <- tm_map(repCorpus, removeNumbers)
repCorpus <- tm_map(repCorpus, removePunctuation)
```

Question3

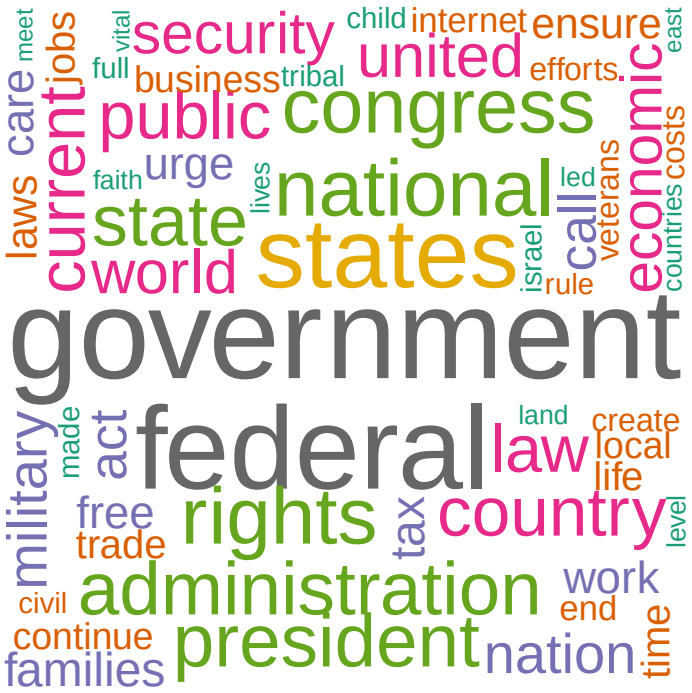
First we draw the wordcloud for democratic platform, then we draw the wordcloud for republican platform. The special words for democratic platform are health, care, communities, public, fight, jobs, worker, energe, education; the special words for republican platform are government, administration, president, congress, military, security, law, economic, trade, tax, families; the common words for both parties are federal, nation, rights, world, etc. From these, we can tell democrats care more about public service and republicans care more about national security and economy.

```
demPlain <- tm_map(demCorpus, PlainTextDocument)
demDTM <- DocumentTermMatrix(demPlain)
demFreq <- sort(colSums(as.matrix(demDTM)), decreasing=TRUE)

set.seed(1234)
layout(matrix(c(1, 2), nrow=2), heights=c(1, 6))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Democratic Party Platform Wordcloud")
wordcloud(names(demFreq), demFreq,
          min.freq = 1, # terms used at least once
          max.words = 300, # 300 most frequently used terms
          random.order = FALSE, # centers cloud by frequency, > = center
          rot.per = 0.30, # sets proportion of words oriented horizontally
          colors = brewer.pal(8, "Dark2")
)
```

Democratic Party Platform Wordcloud


```
repPlain <- tm_map(repCorpus, PlainTextDocument)
repDTM <- DocumentTermMatrix(repPlain)
repFreq <- sort(colSums(as.matrix(repDTM)), decreasing=TRUE)

set.seed(1234)
layout(matrix(c(1, 2), nrow=2), heights=c(1, 6))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Republican Party Platform Wordcloud")
wordcloud(names(repFreq), repFreq,
          min.freq = 1, # terms used at least once
          max.words = 300, # 300 most frequently used terms
          random.order = FALSE, # centers cloud by frequency, > = center
          rot.per = 0.30, # sets proportion of words oriented horizontally
          colors = brewer.pal(8, "Dark2")
)
```

Republican Party Platform Wordcloud

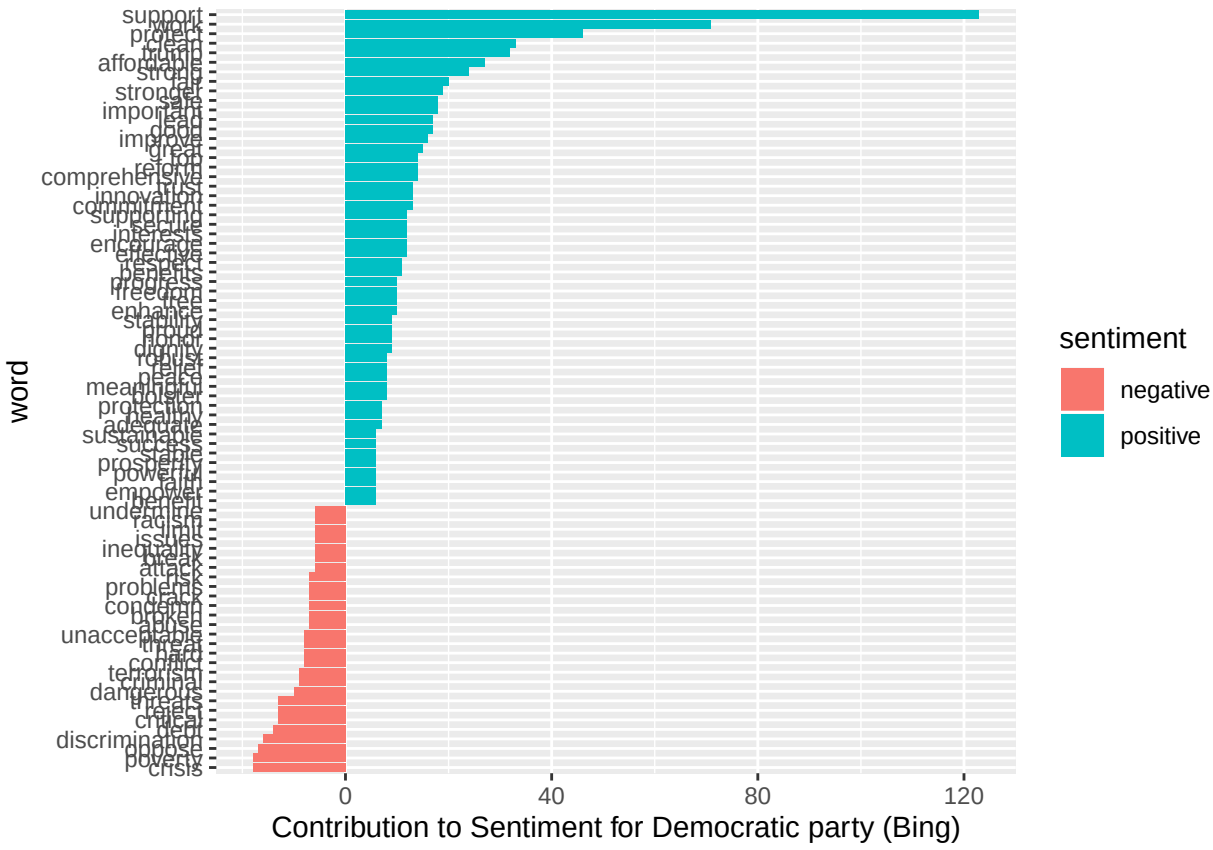


Question4

```
bing <- get_sentiments("bing")
demWord <- data.frame(word = as.character(names(demFreq)), freq = demFreq, stringsAsFactors=FALSE)
demBing <- demWord %>% inner_join(bing, by = "word")
demBing <- demBing[order(-demBing$freq),]
demBingScore <- (sum(demBing[demBing$sentiment == "positive",]$freq)
                 -sum(demBing[demBing$sentiment == "negative",]$freq))
demBingScore
```

```
## [1] 433
```

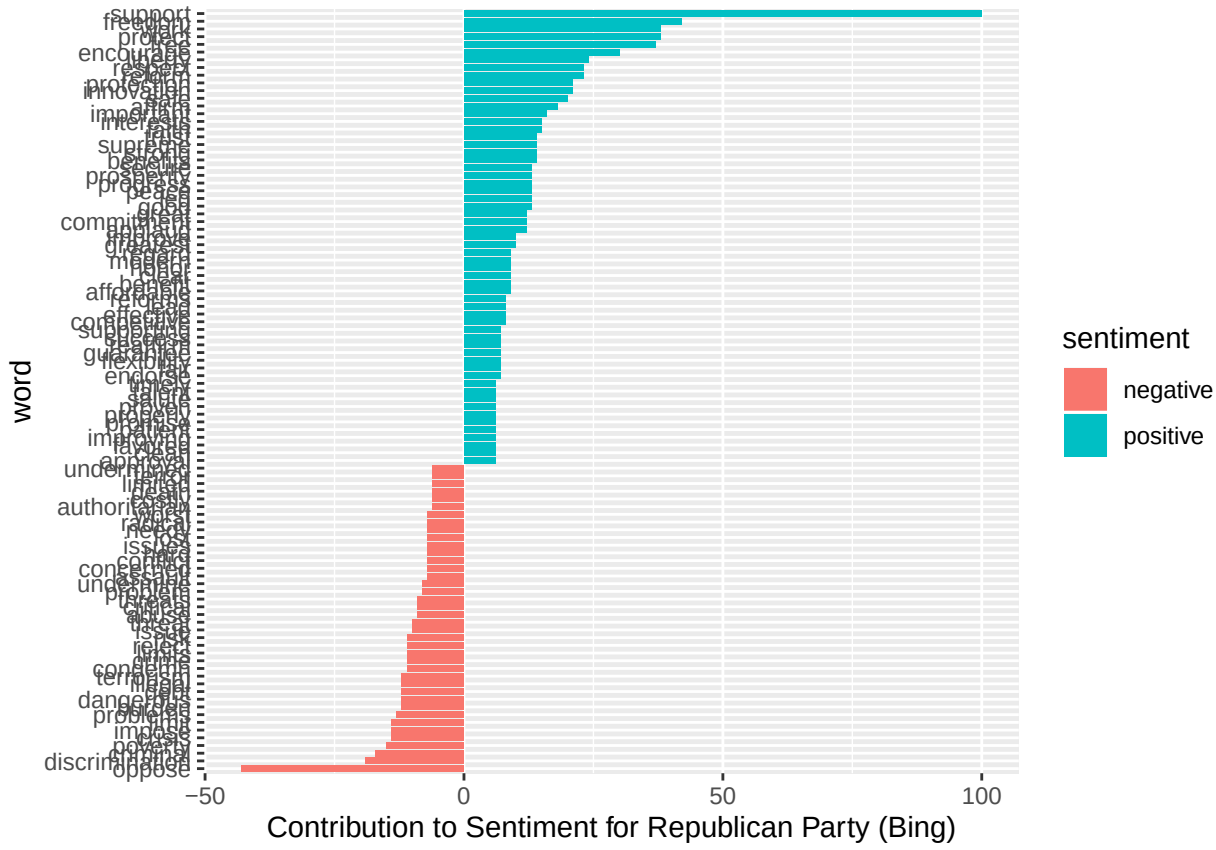
```
demBing %>%
  filter(freq > 5) %>%
  mutate(freq = ifelse(sentiment == "negative", -freq, freq)) %>%
  mutate(word = reorder(word, freq)) %>%
  ggplot(aes(word, freq, fill = sentiment)) +
  geom_col() +
  coord_flip() +
  labs(y = "Contribution to Sentiment for Democratic party (Bing)")
```




```
repWord <- data.frame(word = as.character(names(repFreq)), freq = repFreq, stringsAsFactors=FALSE)
repBing <- repWord %>% inner_join(bing, by = "word")
repBing <- repBing[order(-repBing$freq),]
repBingScore <- (sum(repBing[repBing$sentiment == "positive",]$freq)
                 -sum(repBing[repBing$sentiment == "negative",]$freq))
repBingScore
```

```
## [1] 166
```

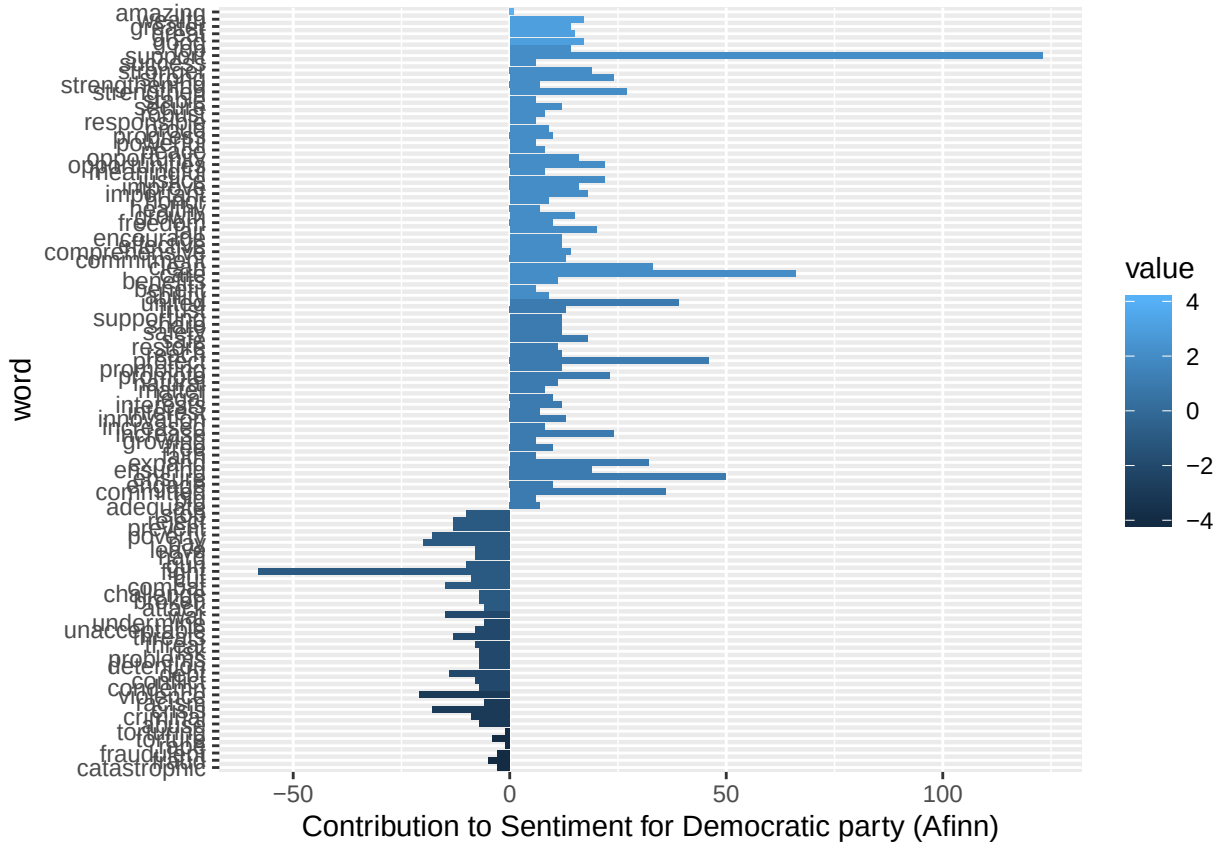
```
repBing %>%
  filter(freq > 5) %>%
  mutate(freq = ifelse(sentiment == "negative", -freq, freq)) %>%
  mutate(word = reorder(word, freq)) %>%
  ggplot(aes(word, freq, fill = sentiment)) +
  geom_col() +
  coord_flip() +
  labs(y = "Contribution to Sentiment for Republican Party (Bing)")
```




```
afinn <- get_sentiments("afinn")
demAfinn <- demWord %>% inner_join(afinn, by = "word")
demAfinn <- demAfinn[order(-demAfinn$value),]
demAfinn$total <- with(demAfinn, freq * value)
demAfinnScore <- (sum(demAfinn[demAfinn$value>0,]$total)
                  +sum(demAfinn[demAfinn$value<0,]$total))
demAfinnScore
```

```
## [1] 1082
```

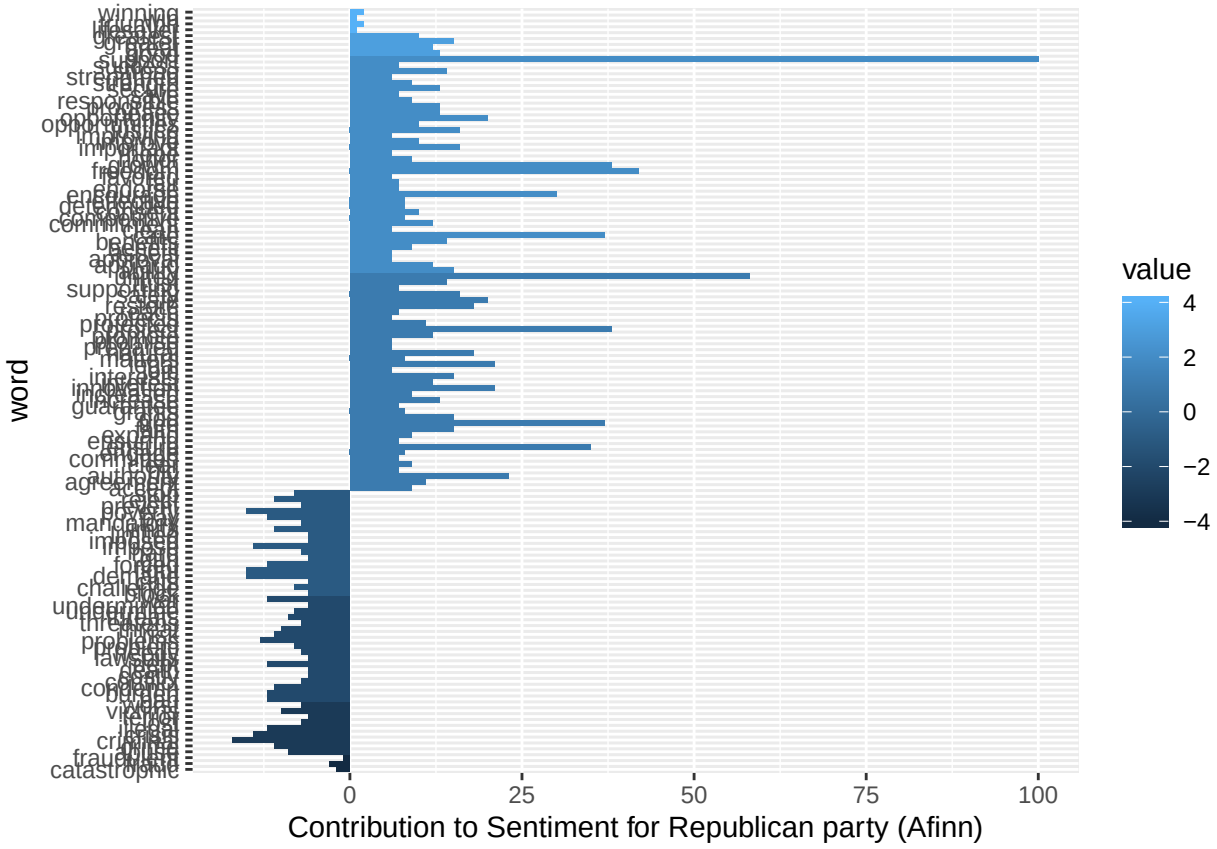
```
demAfinn %>%
  filter(freq >5 | abs(value)==4) %>%
  mutate(freq = ifelse(value<0, -freq, freq)) %>%
  mutate(word = reorder(word, value)) %>%
  ggplot(aes(word, freq, fill = value)) +
  geom_col() +
  coord_flip() +
  labs(y = "Contribution to Sentiment for Democratic party (Afinn)")
```




```
repAfinn <- repWord %>% inner_join(afinn, by = "word")
repAfinn <- repAfinn[order(-repAfinn$value),]
repAfinn$total <- with(repAfinn, freq * value)
repAfinnScore <- (sum(repAfinn[repAfinn$value>0,]$total)
                  +sum(repAfinn[repAfinn$value<0,]$total))
repAfinnScore
```

```
## [1] 700
```

```
repAfinn %>%
  filter(freq >5 | abs(value)==4) %>%
  mutate(freq = ifelse(value<0, -freq, freq)) %>%
  mutate(word = reorder(word, value)) %>%
  ggplot(aes(word, freq, fill = value)) +
  geom_col() +
  coord_flip() +
  labs(y = "Contribution to Sentiment for Republican party (Afinn)")
```



Question5

Generally, we can see that the democratic platform has a higher sentiment score than the republican platform. Thus democratic party tends to be more optimistic about the future. Yes, this does comport with my perception of the parties, as I perceive democratic party as more progressive than republican party.

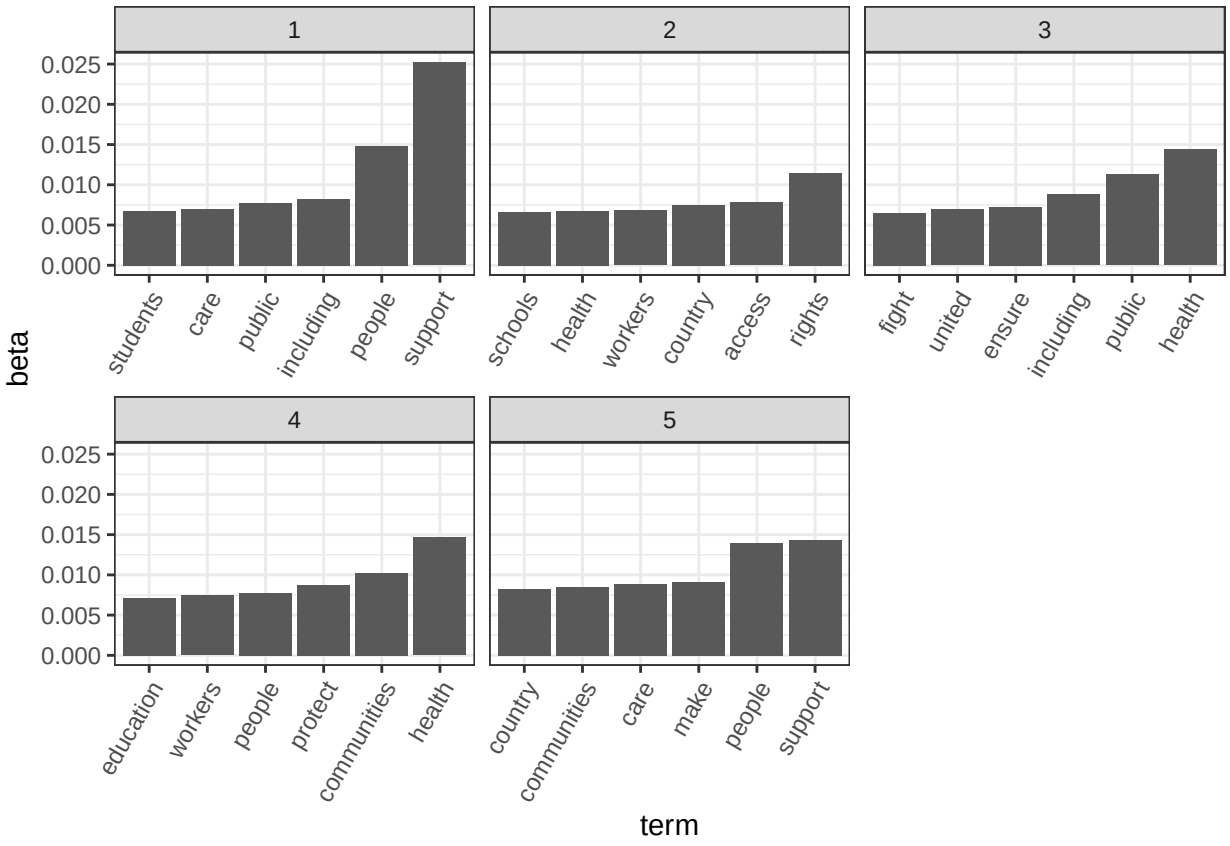
Question6

```
demlda5 <- LDA(demDTM, k = 5, control = list(seed = 1234))
demlda5terms <- tidy(demlda5)
demlda5topterms <- demlda5terms %>%
  group_by(topic) %>%
  top_n(6, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
demlda5topterms
```

```
## # A tibble: 30 x 3
##    topic term         beta
##    <int> <chr>       <dbl>
##  1     1 support   0.0252
##  2     1 people    0.0148
##  3     1 including 0.00824
##  4     1 public    0.00772
##  5     1 care      0.00702
##  6     1 students  0.00673
```

```
##  7     2 rights    0.0114
##  8     2 access    0.00782
##  9     2 country   0.00748
## 10     2 workers   0.00683
## # ... with 20 more rows
```

```
theme_set(theme_bw())
demlda5topterms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity") +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free_x") +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```

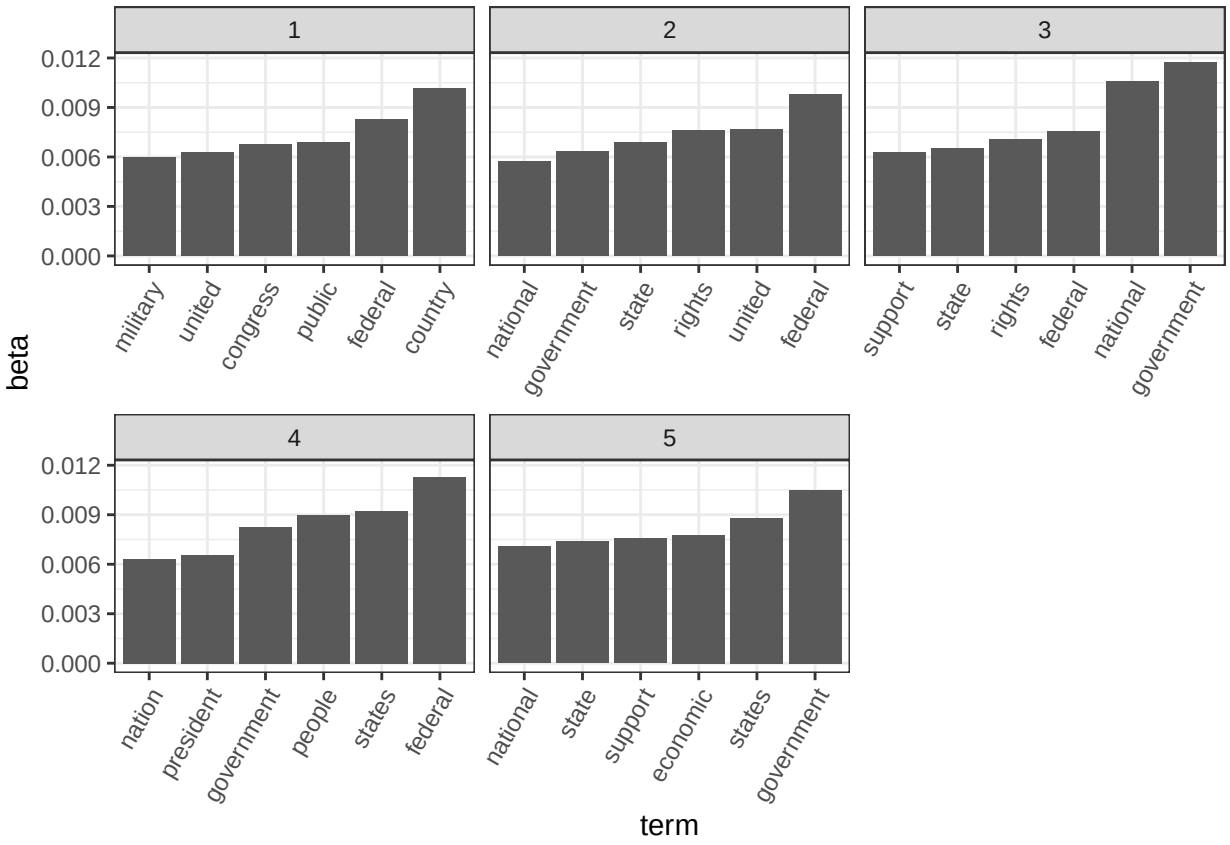



```
replda5 <- LDA(repDTM, k = 5, control = list(seed = 1234))
replda5terms <- tidy(replda5)
replda5topterms <- replda5terms %>%
  group_by(topic) %>%
  top_n(6, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
replda5topterms
```

```
## # A tibble: 30 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     1 country  0.0102
```

```
##  2     1 federal  0.00828
##  3     1 public   0.00689
##  4     1 congress 0.00676
##  5     1 united   0.00626
##  6     1 military 0.00599
##  7     2 federal  0.00977
##  8     2 united   0.00770
##  9     2 rights   0.00762
## 10     2 state    0.00689
## # ... with 20 more rows
```

```
theme_set(theme_bw())
replda5topterms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity") +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free_x") +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



Question7

Generally the two parties talk about different topics. The democratic platform focuses on topics like education, healthcare, community support, worker rights, etc. The republican platform focuses on topics like military, economy, government functionality, etc.

Question8

First we fit the k=10 and k=25 models for democratic platform. Then we fit the k=10 and k=25 models for republican platform.

```
demlda10 <- LDA(demDTM, k = 10, control = list(seed = 1234))
demlda10terms <- tidy(demlda10)
demlda10topterms <- demlda10terms %>%
  group_by(topic) %>%
  top_n(6, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
demlda10topterms
```

```
## # A tibble: 60 x 3
##    topic term         beta
##    <int> <chr>       <dbl>
##  1     1 support   0.0178
##  2     1 people    0.0133
##  3     1 care      0.00759
##  4     1 including 0.00737
##  5     1 public    0.00681
##  6     1 jobs      0.00615
##  7     2 rights    0.0155
##  8     2 country   0.00803
##  9     2 workers   0.00723
## 10     2 schools   0.00686
## # ... with 50 more rows
```

```
demlda25 <- LDA(demDTM, k = 25, control = list(seed = 1234))
demlda25terms <- tidy(demlda25)
demlda25topterms <- demlda25terms %>%
  group_by(topic) %>%
  top_n(6, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
demlda25topterms
```

```
## # A tibble: 150 x 3
##    topic term         beta
##    <int> <chr>       <dbl>
##  1     1 support   0.0147
##  2     1 people    0.0137
##  3     1 including 0.00710
##  4     1 public    0.00694
##  5     1 jobs      0.00649
##  6     1 federal   0.00636
##  7     2 rights    0.0114
##  8     2 country   0.00866
##  9     2 health    0.00823
## 10     2 national  0.00618
## # ... with 140 more rows
```

```
replda10 <- LDA(repDTM, k = 10, control = list(seed = 1234))
replda10terms <- tidy(replda10)
replda10topterms <- replda10terms %>%
```

```
  group_by(topic) %>%
  top_n(6, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
replda10topterms
```

```
## # A tibble: 60 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     1 congress 0.00889
##  2     1 federal  0.00880
##  3     1 public   0.00663
##  4     1 states   0.00661
##  5     1 country  0.00624
##  6     1 united   0.00581
##  7     2 federal  0.0103
##  8     2 rights   0.00759
##  9     2 united   0.00714
## 10     2 support  0.00663
## # ... with 50 more rows
```

```
replda25 <- LDA(repDTM, k = 25, control = list(seed = 1234))
replda25terms <- tidy(replda25)
replda25topterms <- replda25terms %>%
  group_by(topic) %>%
  top_n(6, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
replda25topterms
```

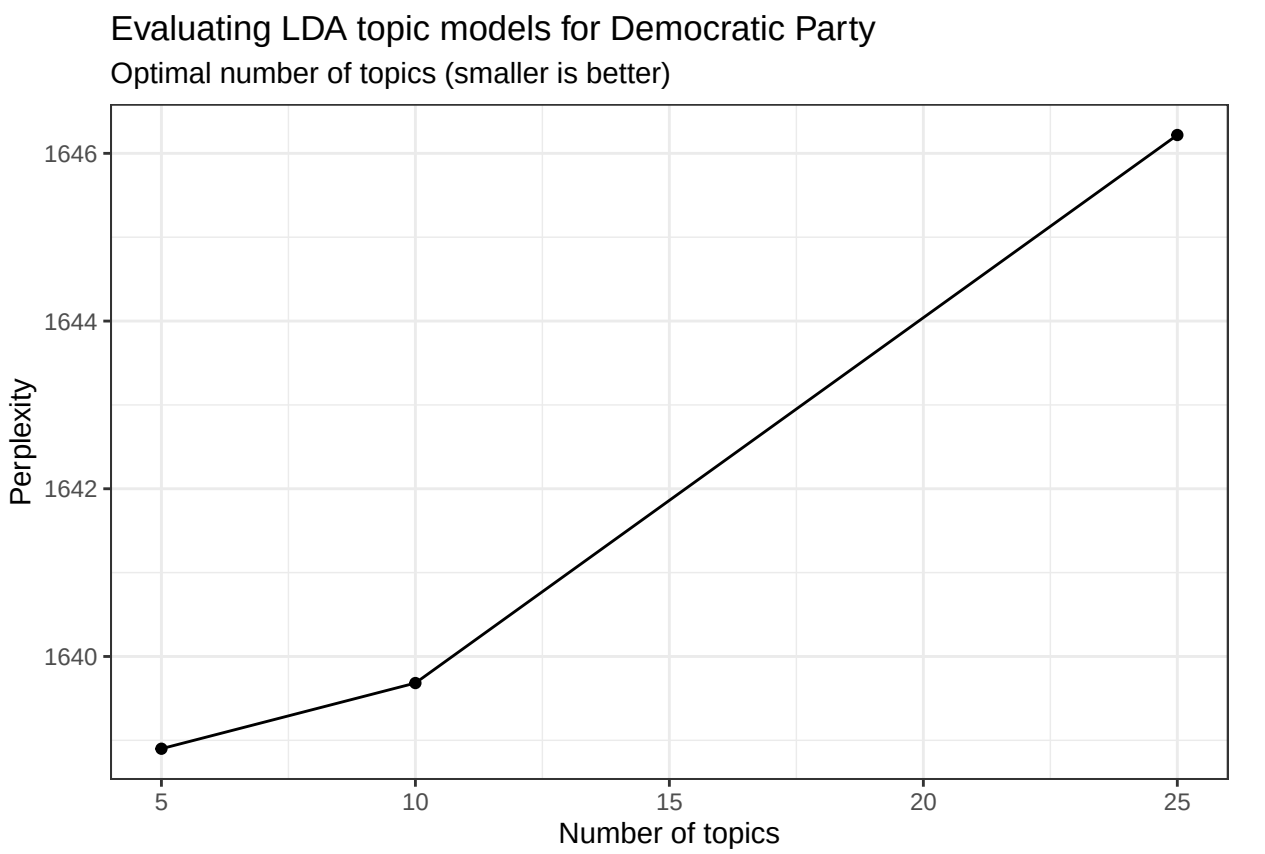
```
## # A tibble: 150 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     1 federal    0.00930
##  2     1 states     0.00675
##  3     1 congress   0.00648
##  4     1 public     0.00609
##  5     1 united     0.00523
##  6     1 country    0.00518
##  7     2 federal    0.0109
##  8     2 rights     0.00757
##  9     2 support    0.00693
## 10     2 government 0.00657
## # ... with 140 more rows
```

Question9

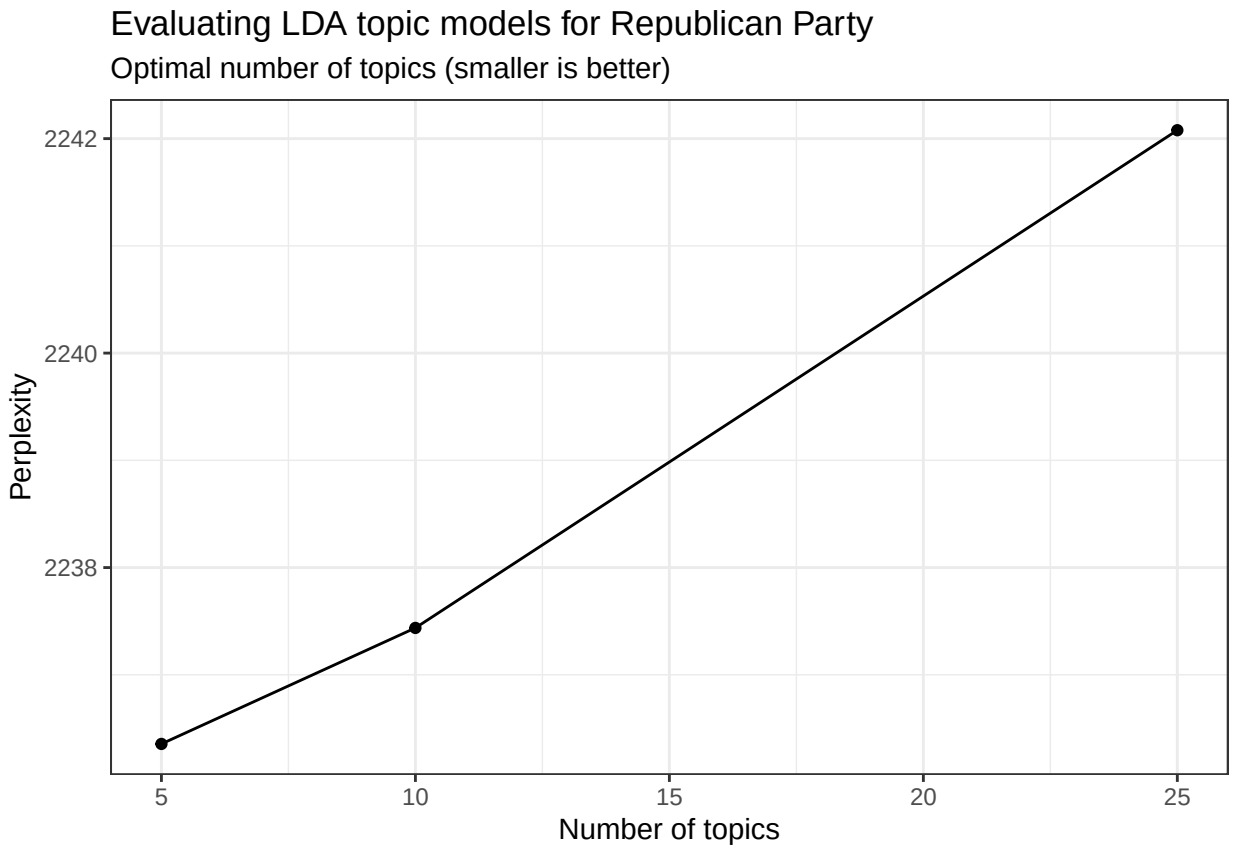
The perplexity scores are calculated for k=5, k=10, k=25 models and both platforms. I should say the result is quite strange. But for both platforms, k=5 models perform best.

```
demperp5 <- perplexity(demlda5)
demperp10 <- perplexity(demlda10)
demperp25 <- perplexity(demlda25)
tibble(k = c(5,10,25),
       perplex = c(demperp5, demperp10, demperp25)) %>%
```

```
  ggplot(aes(k, perplex)) +
  geom_point() +
  geom_line() +
  labs(title = "Evaluating LDA topic models for Democratic Party",
       subtitle = "Optimal number of topics (smaller is better)",
       x = "Number of topics",
       y = "Perplexity")
```



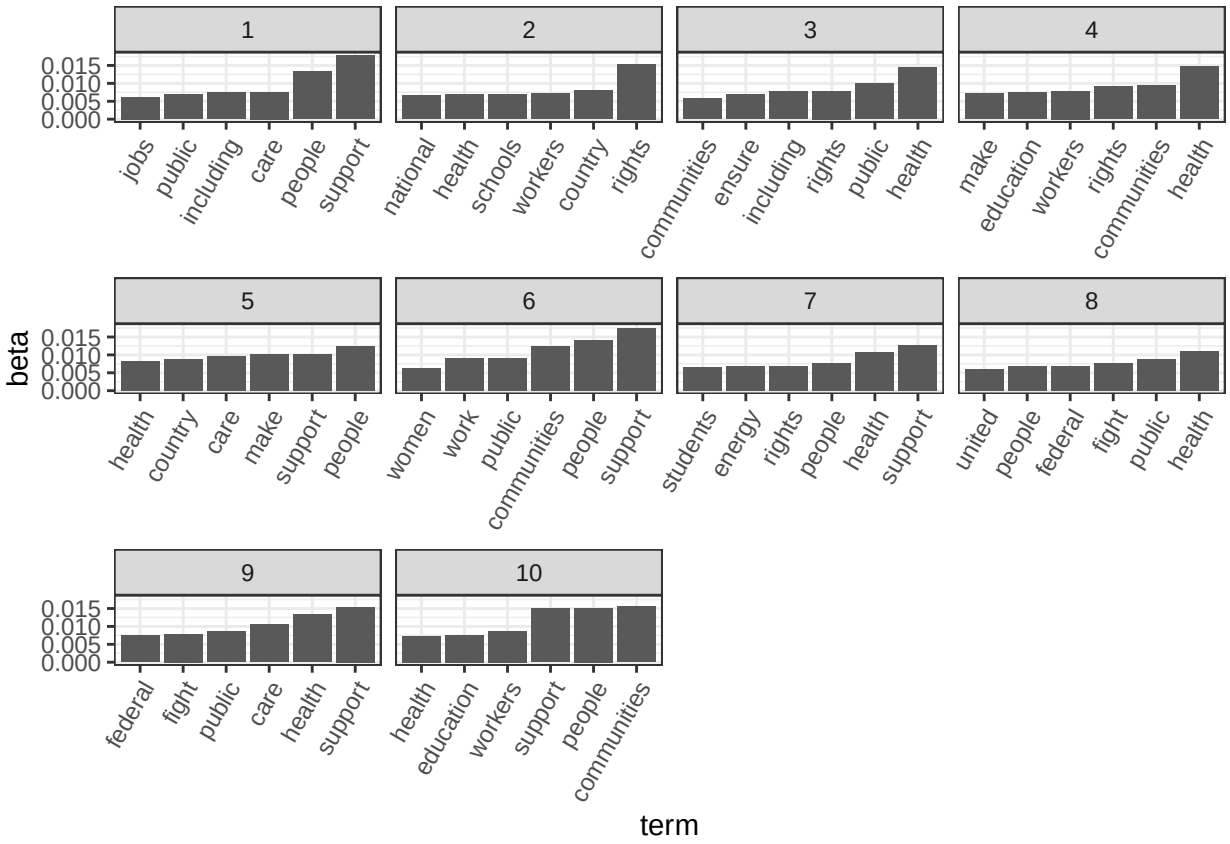

```
repperp5 <- perplexity(replda5)
repperp10 <- perplexity(replda10)
repperp25 <- perplexity(replda25)
tibble(k = c(5,10,25),
       perplex = c(repperp5, repperp10, repperp25)) %>%
  ggplot(aes(k, perplex)) +
  geom_point() +
  geom_line() +
  labs(title = "Evaluating LDA topic models for Republican Party",
       subtitle = "Optimal number of topics (smaller is better)",
       x = "Number of topics",
       y = "Perplexity")
```



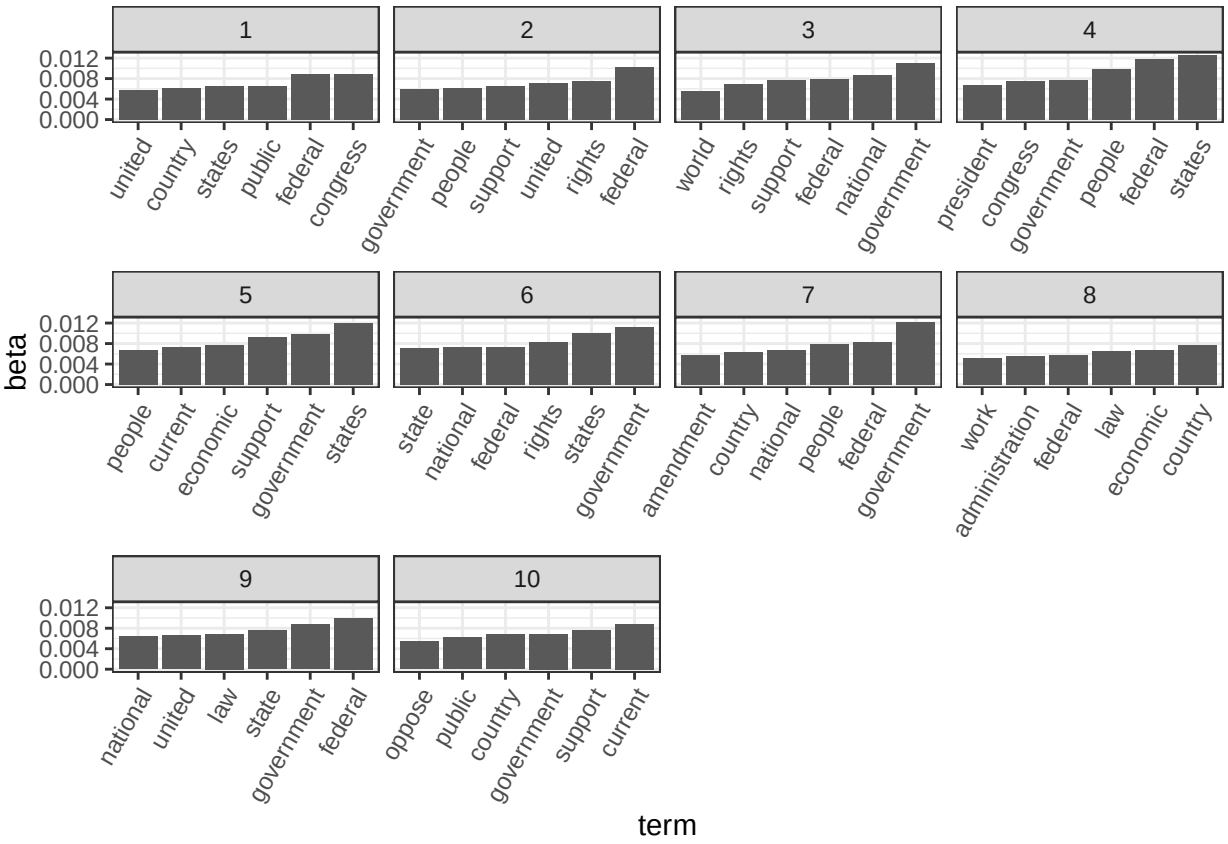
Question10

It seems no similar themes exist between the two parties. The k=10 model generates many overlapping themes, thus I suspect the k=5 model would do better.

```
theme_set(theme_bw())
demlda10topterms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity") +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free_x") +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```




```
theme_set(theme_bw())
replda10topterms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity") +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free_x") +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



Question11

I would support the democratic party. I believe uneven redistribution is the single most important problem facing the country. Only the democratic party has the will, methods and optimism to tackle it.