UML - Problem Set 5

Jesus Pacheco

11/20/2019

PREPROCESSING & (light) EDA

1. The data

```
platforms <- read_csv("/Users/JesusPacheco/GitHub/Problem_Set_5/Party Platforms Data/platforms.csv")
```

```
## Parsed with column specification:
## cols(
##   party = col_character(),
##   platform = col_character()
## )
```

```
txt_files2 <- file.path("~", "Downloads", "txt_files")
#dir(txt_files2)
#corps <- VCorpus(DirSource(txt_files2, encoding = "UTF-8"))
#corps <- VCorpus(DirSource(txt_files2))
corps <- VCorpus(VectorSource(platforms$platform))
#writeLines(as.character(corps[[1]]))
```

2. Preprocessing

```
corps <- tm_map(corps, content_transformer(tolower)) #Convert to lowercase
corps <- tm_map(corps, removePunctuation) #Remove the numbers
corps <- tm_map(corps, removeNumbers) #Remove all punctuation
corps <- tm_map(corps, stripWhitespace) #Remove the whitespac
corps <- tm_map(corps, removeWords, stopwords("english")) #Remove the stopwordse
corps <- tm_map(corps, removeWords, c("will", "also")) #This words are used a lot
for (j in seq(corps)) {
  corps[[j]] <- gsub("health care", "health_care", corps[[j]])
  corps[[j]] <- gsub("<U+2014>", " ", corps[[j]])
}

corps <- tm_map(corps, PlainTextDocument)
#writeLines(as.character(corps[[2]]))
#The corpus seem quite clean with this simple preprocessing,
dtm_d <- DocumentTermMatrix(corps[1])
dtm_r <- DocumentTermMatrix(corps[2])
```

3. Wordcloud

```
freq_d <- sort(colSums(as.matrix(dtm_d)),
                  decreasing=TRUE)
freq_r <- sort(colSums(as.matrix(dtm_r)),
                  decreasing=TRUE)
head(freq_d)
```

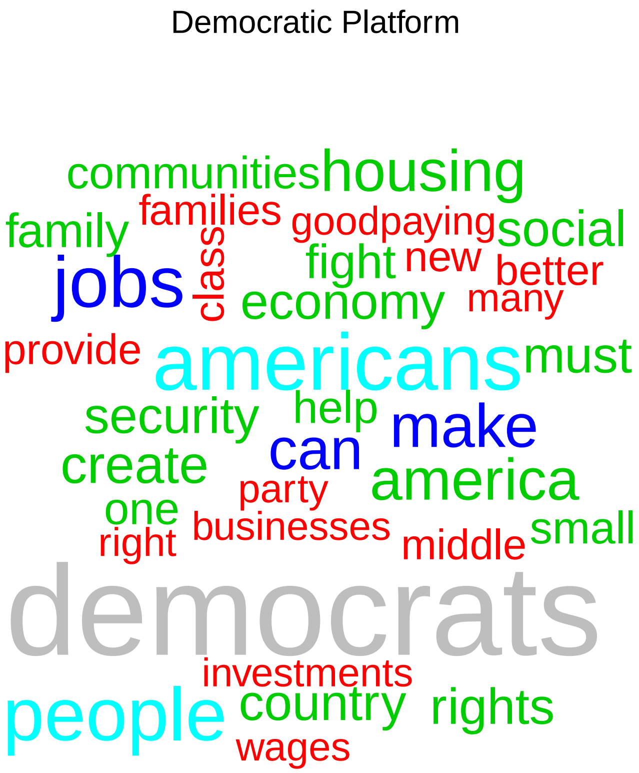
```
## democrats    workers    believe americans    people    support
##        46         36         29        26        24         24
```

```
head(freq_r)
```

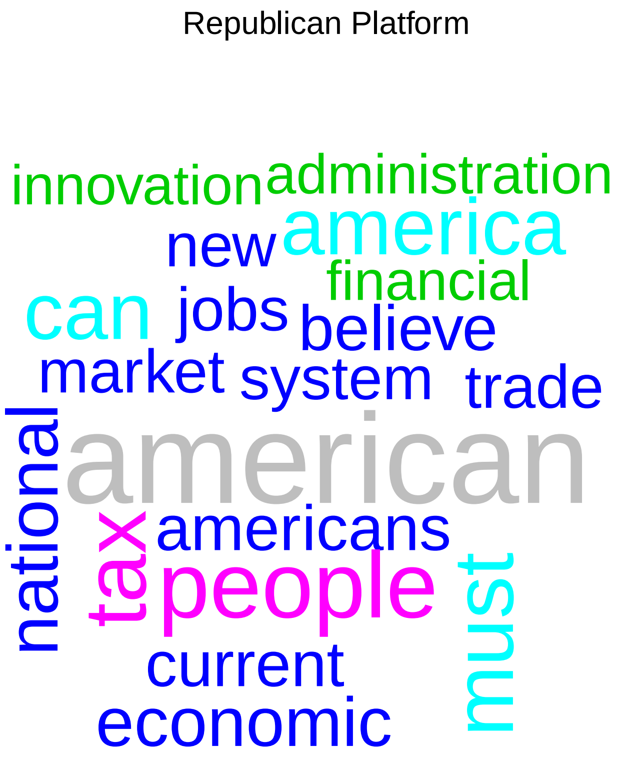
```
##   american    federal government    economy     people        tax
##         28         26         25         19         19         18
```

```
set.seed(123)

layout(matrix(c(1, 2), nrow=2), heights=c(1, 4))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Democratic Platform")
wordcloud(names(freq_d), freq_d, min.freq = 10, main="Democrats", colors = palette())
```



```
layout(matrix(c(1, 2), nrow=2), heights=c(1, 4))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Republican Platform")
wordcloud(names(freq_r), freq_r, min.freq = 10, main="Republicans", colors = palette())
```



At this early stage we see that a lot of common terms are used in both platforms: America, the economy and jobs. But we can also get a sense of the different subjects in their platforms. Democrats use frequently terms such as rights, goodpaying, wages and housing; while republicans use tax, markets and trade. Nothing too surprising from my (limited) understanding of American politics.

SENTIMENT ANALYSIS

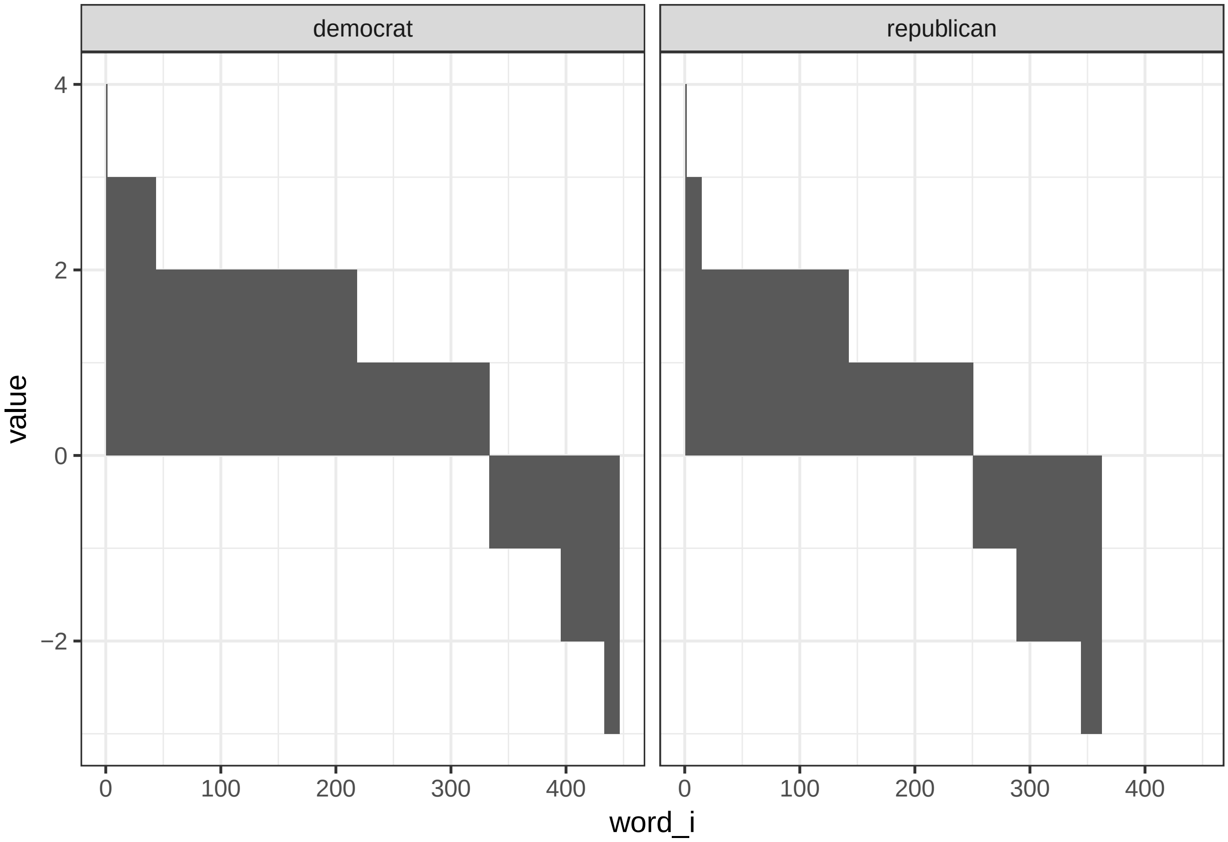
4. Bing and AFFIN

The sentiment scores using the AFFIN dictinary look like this:

```
#get_sentiments("afinn")

#I'm using the "untidy" version of the text since we only care about words that match the dictionary
#The unnest_tokens function is making the terms lower case
platforms %>% group_by(party) %>%
  unnest_tokens(word, platform, token = "words") %>% inner_join(get_sentiments("afinn")) %>%
  arrange(-value) %>%
  mutate(word_i = 1:n()) %>%
  ggplot(aes(word_i, value)) + geom_col() +
  facet_wrap(~party) + theme_bw()
```

```
## Joining, by = "word"
```



The next table present the positive/negative sentiments using the Bing dictionary

```
#get_sentiments("bing")
bing <- platforms %>% group_by(party) %>%
  unnest_tokens(word, platform, token = "words") %>% inner_join(get_sentiments("bing")) %>% ungroup()
```

```
## Joining, by = "word"
```

```
prop.table(table(bing$sentiment, bing$party))
```

```
##
##              democrat republican
##   negative 0.1486989  0.1920694
##   positive 0.3729864  0.2862454
```

5. Interpretation

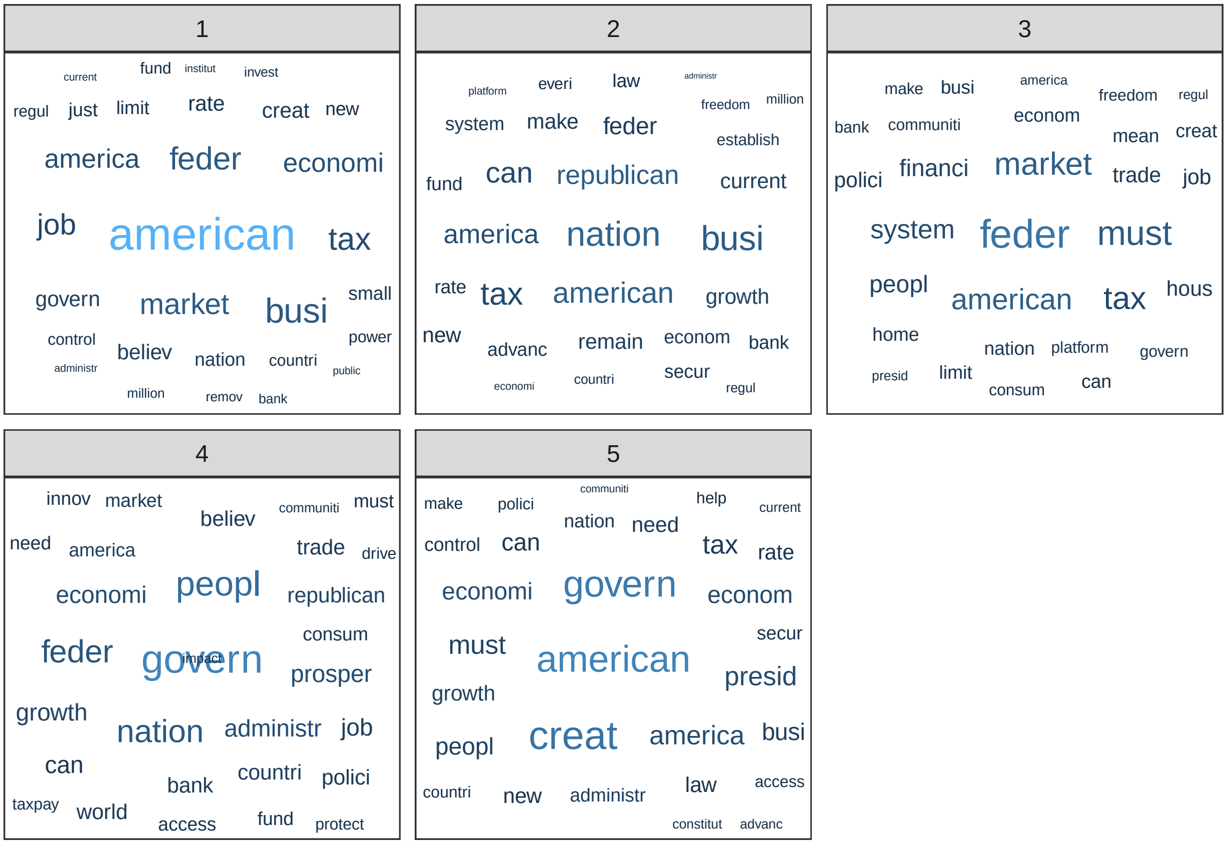
The barplot of the sentiment analysis using “AFINN” displays (ordered) words in the x-axis and their “sentiment values” in the y-axis. We can see from the barplot, that the democratic platform has more words in general, and it is easy to see that the positive terms are significantly more in their platform. The negative terms are more alike, but with higher scores (in absolute values) for the republican platform (terms such as ‘evil’, ‘criminal’, ‘crisis’). The proportions table displaying the sentiment using the bing dictionary shows that the 14.9 percent of the recognized terms have a negative sentiment in the democratic platform, and 19.2 in the republican one. From these results, we can conclude that the republican platform has a more negative (pessimistic) sentiment. This sentiment could be interpreted as a pessimistic vision of the future from the republicans, but also due to a pessimistic view of the status quo and 8 years of a democratic presidency.

TOPIC MODELS

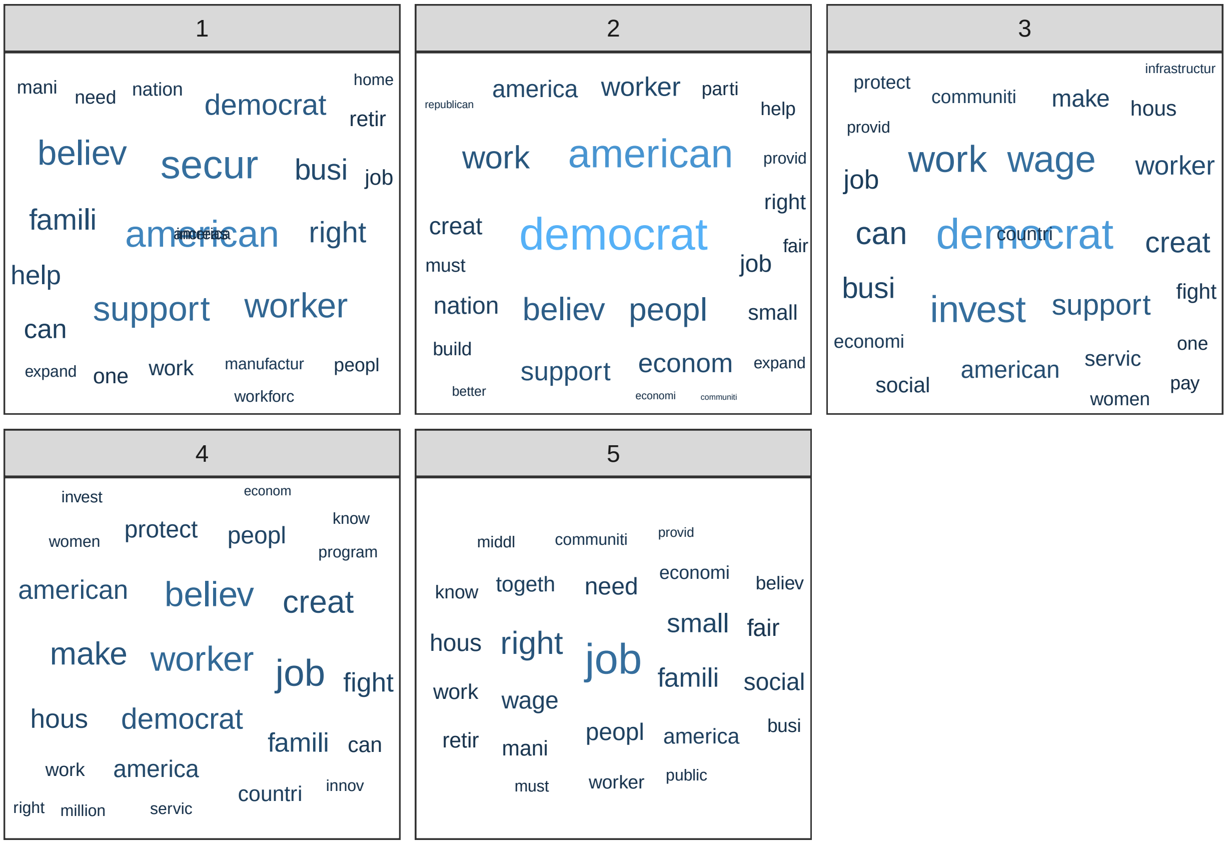
6. LDA

```
corps_stem <- tm_map(corps, stemDocument)
dtm_d_st <- DocumentTermMatrix(corps_stem[1])
dtm_r_st <- DocumentTermMatrix(corps_stem[2])
#writeLines(as.character(corps_stem[[2]]))
rep_lda <- LDA(dtm_r_st, k = 5, control = list(seed = 1234))
dem_lda <- LDA(dtm_d_st, k = 5, control = list(seed = 1234))


#WORDCLOUDS
LDA_r <- tidy(rep_lda) %>%
  group_by(topic) %>% top_n(30, beta)  %>% arrange(topic, -beta)
LDA_r %>% ggplot(aes(label=term,size=beta, col=beta)) + geom_text_wordcloud_area() +
   facet_wrap(~topic) + theme_bw()
```



```
LDA_d <- tidy(dem_lda) %>%
  group_by(topic) %>% top_n(25, beta)  %>% arrange(topic, -beta)
LDA_d %>% ggplot(aes(label=term,size=beta, col=beta)) + geom_text_wordcloud_area() +
   facet_wrap(~topic) + theme_bw()
```



7. Interpreting LDA

In the interpretation of the topics, the difference in the topics are not as clear within the platforms, as they are comparing among platforms. First, let's look within the platforms:

Republicans: Topic 1 seems to be related to business because of terms such as job, economy, markets and business. Topic 4 might be related to the world economy with terms such as growth, prosperity and trade. The rest might be more government and administrative-related. As I mentioned, the topics are not very interpretable to me at this point within the same platform. Surprisingly, we don't much topics about military and security, but a lot of security and law related ones.

Democrats: Topic 3 contains ideas related to 'support' and social protection: expand, ensure, strengthen and communities. The rest are very heavily leaning towards the job market.

The differences in the topics among both platforms are somewhat. While democrats include topics of rights, protection and supporting communities; republicans include topics of security, law and government. They both care about jobs and the economy.

8. Six more models

```r
#The 6 LDAs
for (i in c(5,10,25)) {
  assign(paste0("rep_lda_",i), LDA(dtm_r_st, k = i, control = list(seed = 1234)))
  assign(paste0("dem_lda_",i), LDA(dtm_d_st, k = i, control = list(seed = 1234)))
}
```

The results of the k=5 topic models were previously shown and discussed in question 6, the results of the k=10 topic is shown and discussed in question 10. So, I'll present here a look at the results from the k=25 model for each party platform. The 3 most common terms for each topic should do it, but the point should be made that interpretability is just a lost cause here. For republicans:

```
terms(rep_lda_25, 3)
```

```
##      Topic 1    Topic 2   Topic 3    Topic 4  Topic 5    Topic 6
## [1,] "american" "nation"  "feder"    "govern" "american" "american"
## [2,] "feder"    "america" "market"   "nation" "govern"   "govern"
## [3,] "market"   "busi"    "american" "feder"  "america"  "nation"
##      Topic 7  Topic 8  Topic 9    Topic 10 Topic 11   Topic 12
## [1,] "govern" "govern" "american" "govern" "american" "feder"
## [2,] "tax"    "tax"    "govern"   "feder"  "govern"   "creat"
## [3,] "feder"  "feder"  "america"  "peopl"  "busi"     "administr"
##      Topic 13   Topic 14   Topic 15   Topic 16   Topic 17   Topic 18
## [1,] "american" "american" "american" "american" "american" "peopl"
## [2,] "nation"   "peopl"    "nation"   "govern"   "busi"     "tax"
## [3,] "creat"    "feder"    "peopl"    "creat"    "peopl"    "must"
##      Topic 19 Topic 20   Topic 21   Topic 22   Topic 23   Topic 24
## [1,] "feder"  "american" "american" "american" "american" "govern"
## [2,] "nation" "economi"  "nation"   "govern"   "busi"     "market"
## [3,] "tax"    "market"   "creat"    "nation"   "creat"    "busi"
##      Topic 25
## [1,] "market"
## [2,] "nation"
## [3,] "tax"
```

We can see with this table, that the topics start to become somewhat repetitive (govern, tax and federation). And we see the same pattern in the democrat table:

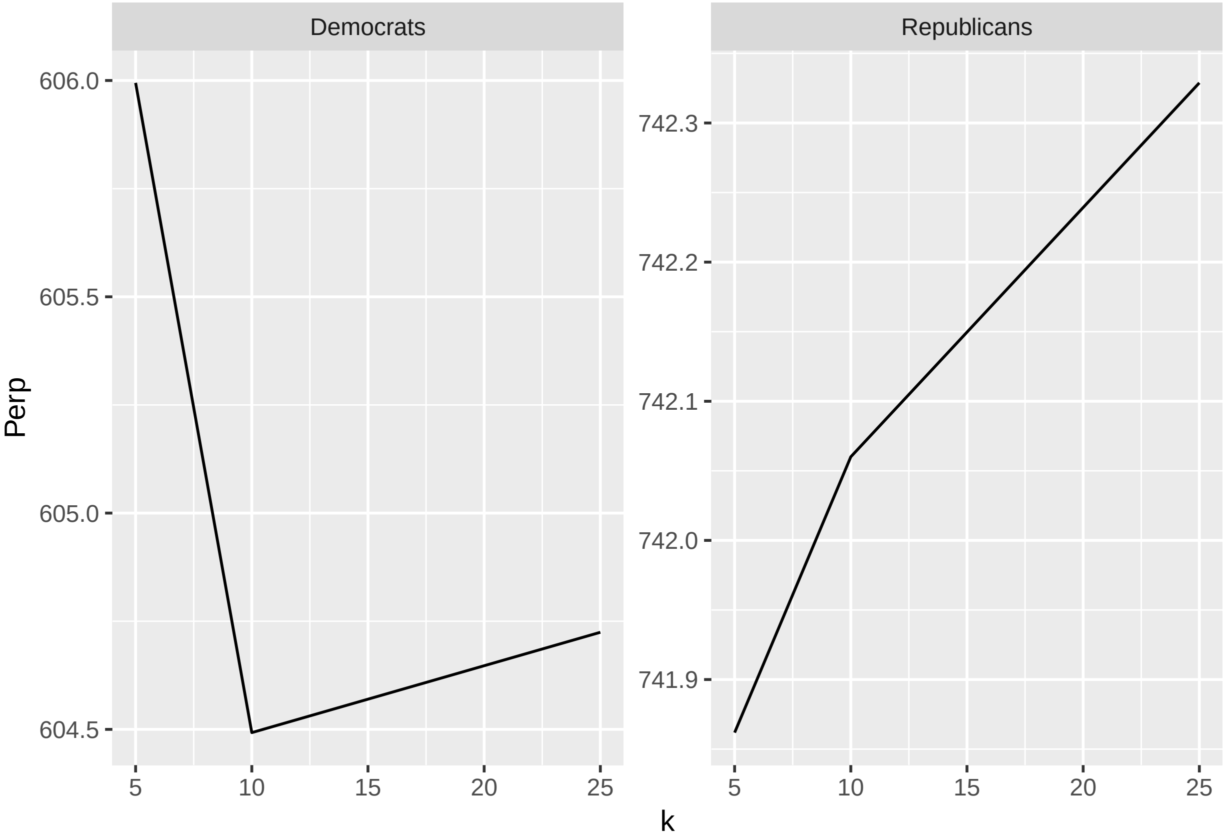
```
terms(dem_lda_25, 3)
```

```
##      Topic 1    Topic 2    Topic 3    Topic 4  Topic 5  Topic 6
## [1,] "american" "american" "democrat" "worker" "job"    "democrat"
## [2,] "worker"   "democrat" "work"     "job"    "right"  "work"
## [3,] "support"  "work"     "support"  "believ" "famili" "right"
##      Topic 7   Topic 8  Topic 9    Topic 10  Topic 11  Topic 12
## [1,] "work"    "peopl"  "believ"   "worker"  "make"    "democrat"
## [2,] "support" "wage"   "american" "support" "job"     "creat"
## [3,] "worker"  "famili" "work"     "work"    "support" "make"
##      Topic 13   Topic 14   Topic 15   Topic 16   Topic 17   Topic 18
## [1,] "american" "democrat" "american" "democrat" "democrat" "believ"
## [2,] "work"     "work"     "right"    "american" "american" "support"
## [3,] "democrat" "believ"   "famili"   "job"      "worker"   "worker"
##      Topic 19   Topic 20   Topic 21   Topic 22   Topic 23   Topic 24
## [1,] "american" "democrat" "make"     "american" "worker"   "democrat"
## [2,] "democrat" "worker"   "american" "support"  "believ"   "worker"
## [3,] "worker"   "job"      "believ"   "job"      "american" "american"
##      Topic 25
## [1,] "democrat"
## [2,] "believ"
## [3,] "american"
```

9. Perplexity

```
perplexity <- data.frame("Party"=c(rep("Republicans",3),
                          rep("Democrats",3)), "k" = c(5,10,25), "Perp" = 0)
perplexity$Perp[1] <- perplexity(rep_lda_5)
perplexity$Perp[2] <- perplexity(rep_lda_10)
perplexity$Perp[3] <- perplexity(rep_lda_25)
perplexity$Perp[4] <- perplexity(dem_lda_5)
perplexity$Perp[5] <- perplexity(dem_lda_10)
perplexity$Perp[6] <- perplexity(dem_lda_25)

perplexity %>% ggplot(aes(x=k, y=Perp)) + geom_line() + facet_wrap(~Party, scales="free")
```

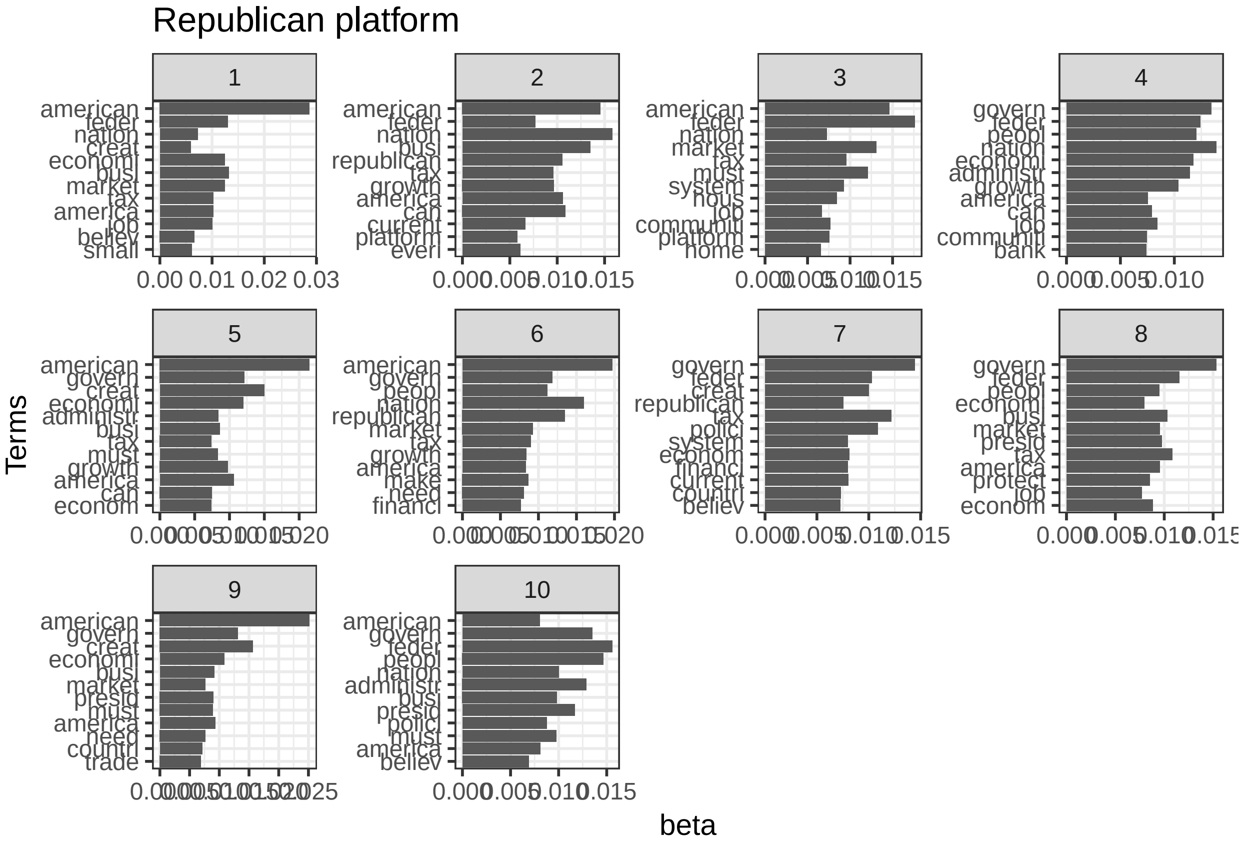


The computation of the perplexity measure is throwing some odd results, especially for the republican platforms where we see an increase in perplexity (the model is explaining less with more topics). It might be a flaw in my calculation, I did try the screeplot for more values of k and the perplexity is generally increasing as well.

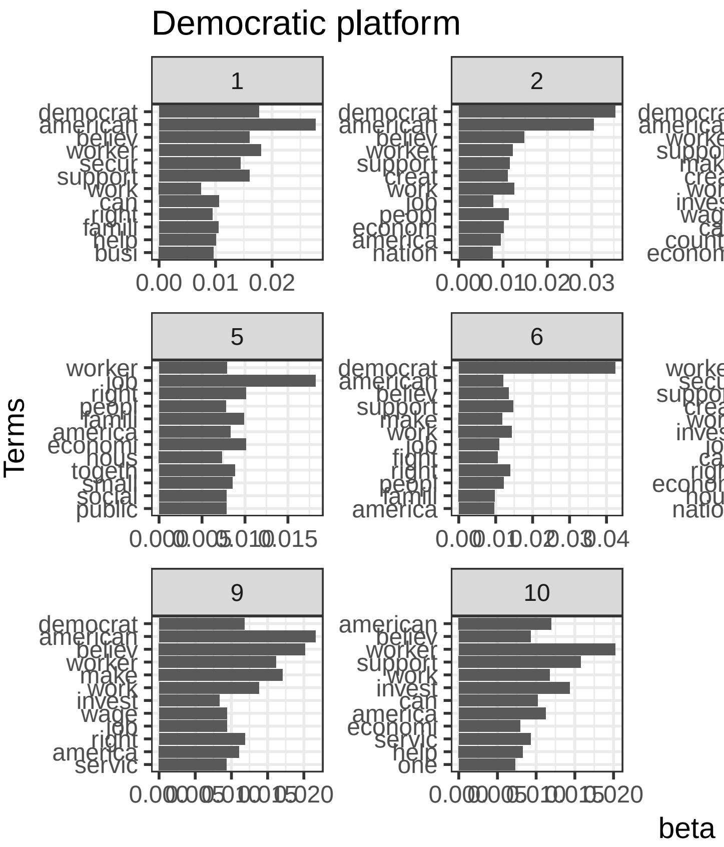
Two important things I would note: a) there might be a different optimal k for each model, republicans seem to need less topics to model their corpus, and b) the loss of interpretability with the increase of k does not seem to be compensated (at least as much) with gains from "perplexity", hence less topic seem to be enough to model the platforms in this case.

10. The k=10 model

```
#Republican topic models
tidy(rep_lda_10) %>%
      group_by(topic) %>% top_n(12, beta)  %>% arrange(topic, -beta) %>%
      ggplot(aes(x= reorder(term, beta), beta)) +
           geom_col() + facet_wrap(~topic, scales="free") + coord_flip() +
           labs(title="Republican platform", x="Terms") +
           theme(text = element_text(size=10)) + theme_bw()
```



```
#Democratic topic models
tidy(dem_lda_10) %>%
      group_by(topic) %>% top_n(12, beta)  %>% arrange(topic, -beta) %>%
      ggplot(aes(x= reorder(term, beta), beta)) +
           geom_col() + facet_wrap(~topic, scales="free") + coord_flip() +
           labs(title="Democratic platform", x="Terms") +
           theme(text = element_text(size=10)) + theme_bw()
```



Building on the previous question, display a barplot of the k = 10 model for each party, and offer some general inferences as to the main trends that emerge. Are there similar themes between the parties? Do you think k = 10 likely picks up differences more efficiently? Why or why not As we sort of already saw in the k=5 model, the interpretability of the topics becomes more complicated with the increase of k. We see a lot of repeated common terms across topics and the “uniqueness” of the topics becomes harder to see in both platforms. Even some topics start to become very similiar across platforms with terms such as economy, America, nation, support and government. Overall, I would say that k=10 is not picking up differences more efficiently than k=5 and the interpretation is quickly becoming more difficult with the increase of topics.

11. Conclusion

In this hypothetical case, I would support the democratic party. Naturally, ideology plays an important role here, and I would support a government that acknowledges disparities and unequal opportunities and propose policies in favor of support of those communities. The general sentiment of the democratic party might not be altogether positive, but rather less pessimistic than the republican. Its brand is more of protecting communities and ensuring a more comprehensive concept of rights to the American people. Governments whose main priority is the rule of law, as we have seen throughout history, could lead to conservativism and concentration of privileges. As I mentioned before, I do not believe the more negative sentiment of the republican platform is due to pessimism of the future but a pessimistic view of the present, and I believe, this could be dangerous and can drive people to make decisions based on fear.