R Notebook

Preprocessing

```r
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------- tidyverse 1.2

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------- tidyverse_conflicts
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(tidytext)
library(tm)
```

```
## Loading required package: NLP

##
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```r
library(sotu)
library(igraph)
```

```
##
## Attaching package: 'igraph'

## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union

## The following objects are masked from 'package:purrr':
##
##     compose, simplify

## The following object is masked from 'package:tidyr':
##
##     crossing
```

```
## The following object is masked from 'package:tibble':
##
##     as_data_frame

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

```r
library(quanteda)
```

```
## Package version: 1.5.1

## Parallel computing: 2 of 4 threads used.

## See https://quanteda.io for tutorials and examples.

##
## Attaching package: 'quanteda'

## The following object is masked from 'package:igraph':
##
##     as.igraph

## The following objects are masked from 'package:tm':
##
##     as.DocumentTermMatrix, stopwords

## The following object is masked from 'package:utils':
##
##     View
```

```r
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```r
library(topicmodels)
library(readr)
library(SnowballC)
library(textdata)


platforms <- read_csv("C:/Users/Keysar Lab/Box Sync/UML/platforms.csv")
```

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

```r
unt <- platforms %>%
  unnest_tokens(output = word, input = platform)

as_tibble(stop_words)
```

```
## # A tibble: 1,149 x 2
##    word        lexicon
##    <chr>       <chr>
##  1 a           SMART
##  2 a's         SMART
##  3 able        SMART
##  4 about       SMART
##  5 above       SMART
##  6 according   SMART
##  7 accordingly SMART
##  8 across      SMART
##  9 actually    SMART
## 10 after       SMART
## # ... with 1,139 more rows
```

```r
stops <- unt %>%
  anti_join(stop_words,
          by = "word") # drop words in stop words data by "unjoining" them from the df

tidy_text<-stops[-grep("\\b\\d+\\b", stops$word),]

tidy_text$word <- gsub("\\s+","",tidy_text$word)

tidy_text$word <- gsub("\\s+","",tidy_text$word)


tidy_text<-tidy_text %>%
      mutate_at("word", funs(wordStem((.), language="en")))
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once per session.
```

Quick EDA

Observations from EDA

Republican narratives focus more on growth, taxes, market, and the economy. It also leans more nationalistic, since “nation” is a big key word. Democratic narraitves focus a lot more on workers, wages, and creating jobs, protecting and supporting, as well as people and families. Both narratives mention "America" a lot. It would be interesting to perform bigram analysis to see whether "America" relates to similar or different topics in both parties.

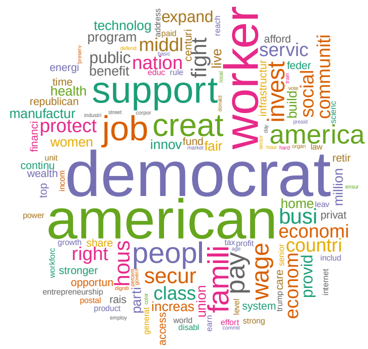
```
r_tidy <- tidy_text %>% filter(party == "republican")
d_tidy <- tidy_text %>% filter(party == "democrat")

r_freq <- r_tidy %>%
  count(word, sort = TRUE)

d_freq <- d_tidy %>%
  count(word, sort = TRUE)

set.seed(2305)
wordcloud(r_tidy$word,
          max.words = 120,
          colors = brewer.pal(8, "Dark2"),
          random.order = FALSE,
          rot.per = 0.30,
          random.color = TRUE,
          scale = c(2, 0.01))
```

```
## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,
## tm::stopwords())): transformation drops documents
```


```r
wordcloud(d_tidy$word,
          max.words = 120,
          colors = brewer.pal(8, "Dark2"),
          random.order = FALSE,
          rot.per = 0.30,
          random.color = TRUE,
          scale = c(2, 0.01))
```

```
## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
## transformation drops documents
```



Sentiment

Observations about sentiments

Overall, Democratic narratives tend to be more optimistic than Republican narratives, as seen by both an increased proportion in positive sentiment words and decreased proportion in negative sentiment words. However, without knowing exactly which topics these sentiments are referencing, it is difficult to know exactly what they are more optimistic about. Speculatively, it seems that Republicans might generally be operating on a more emotional, threat-based approach to politics (e.g. immigrants are stealing jobs! build a wall!) so I think that fits with my perceptions of the parties.

```
# overall sentiment, republicans

r_bing <- r_tidy %>%
  select(-party)%>%
  inner_join(get_sentiments("bing")) %>%
  count(sentiment, sort = TRUE)%>%
  mutate(party = "republican")
```

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

```
r_bing_perc <- r_bing%>%
  rowwise%>%
  mutate(percent = n/222)
```

```r
r_afinn <- r_tidy %>%
  select(-party)%>%
  inner_join(get_sentiments("afin"))
```

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

```r
# overall sentiment, democrats

d_bing <- d_tidy %>%
  select(-party)%>%
  inner_join(get_sentiments("bing")) %>%
  count(sentiment, sort = TRUE)%>%
  mutate(party = "democrat")
```

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

```r
d_bing_perc <- d_bing%>%
  rowwise%>%
  mutate(percent = n/248)


d_afinn <- d_tidy %>%
  select(-party)%>%
  inner_join(get_sentiments("afin"))
```

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

```r
# compare results

bing <- rbind(r_bing_perc,d_bing_perc)

bing
```

```
## Source: local data frame [4 x 4]
## Groups: <by row>
##
## # A tibble: 4 x 4
##   sentiment     n party      percent
##   <chr>     <int> <chr>        <dbl>
## 1 positive    137 republican   0.617
## 2 negative     85 republican   0.383
## 3 positive    177 democrat     0.714
## 4 negative     71 democrat     0.286
```

```r
mean(r_afinn$value)
```

```
## [1] 0.4326923
```

```r
mean(d_afinn$value)
```

```
## [1] 0.5873016
```

Topic models

Fitting the LDA model with k = 5

```r
# r <- readLines("C:/Users/beckylau/Box Sync/UML/r16.txt")
# d <- readLines("C:/Users/beckylau/Box Sync/UML/d16.txt")

r <- VCorpus(VectorSource(r_tidy$word))
d <- VCorpus(VectorSource(d_tidy$word))

library(topicmodels)
r_dtm <- DocumentTermMatrix(r)
d_dtm <- DocumentTermMatrix(d)

frequency_r_dtm <- sort(colSums(as.matrix(r_dtm)),
                  decreasing=TRUE) # add number of times each term is used, and sorting based on freque
head(frequency_r_dtm)
```

```
## american   govern    feder   nation  america     busi
##       40       31       28       25       24       23
```

```r
frequency_d_dtm <- sort(colSums(as.matrix(d_dtm)),
                  decreasing=TRUE) # add number of times each term is used, and sorting based on freque
head(frequency_d_dtm)
```

```
## democrat american   worker  support      job    creat
##       52       45       38       30       29       25
```

```r
#dealing with error message about "Each row of the input matrix needs to contain at least one non-zero

raw.sum=apply(d_dtm,1,FUN=sum) #sum by raw each raw of the table
d_dtm=d_dtm[raw.sum!=0,]

raw.sum=apply(r_dtm,1,FUN=sum) #sum by raw each raw of the table
r_dtm=r_dtm[raw.sum!=0,]


r_lda <- LDA(r_dtm, k = 5, control = list(seed = 1234))
d_lda <- LDA(d_dtm, k = 5, control = list(seed = 1234))

r_topics <- tidy(r_lda, matrix = "beta")
d_topics <- tidy(d_lda, matrix = "beta")
```

General trends in topics that emerge:

Observations from topic models, general trends

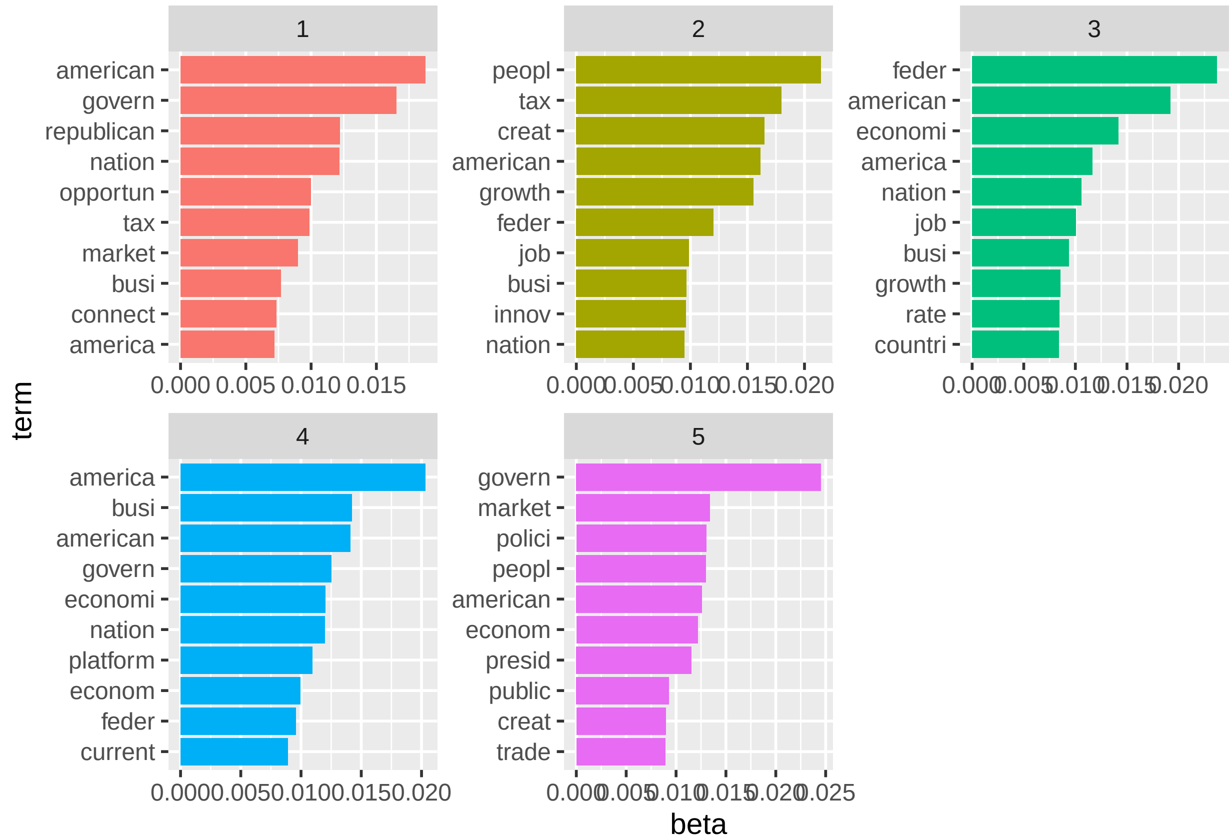
I think the two parties discuss similar topics but in a very different light. For example, Republicans talk about jobs in the context of growth, taxes, innovaction, business, and the econom, while Democrats talk about jobs in the context of workers, families, people, communities, fairness and social security.

Republican topics

```
r_topics
```

```
## # A tibble: 5,100 x 3
##    topic term      beta
##    <int> <chr>    <dbl>
##  1     1 20th  0.000364
##  2     2 20th  0.000199
##  3     3 20th  0.000229
##  4     4 20th  0.000630
##  5     5 20th  0.000596
##  6     1 21st  0.000782
##  7     2 21st  0.000891
##  8     3 21st  0.000372
##  9     4 21st  0.000898
## 10     5 21st  0.00109
## # ... with 5,090 more rows
```

```
r_top_terms <- r_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

r_top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```



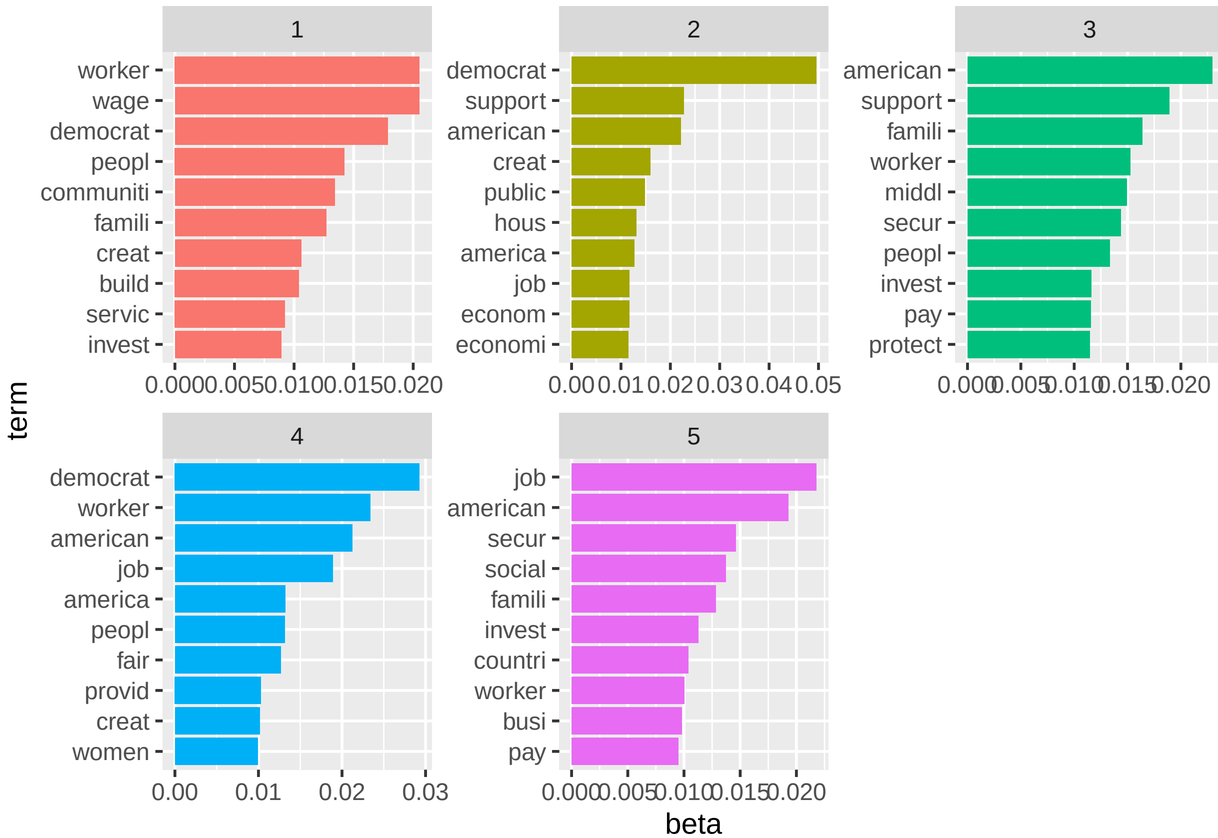
Democrat topics

```
d_topics
```

```
## # A tibble: 4,475 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     1 21st  0.00230
##  2     2 21st  0.000827
##  3     3 21st  0.00382
##  4     4 21st  0.00286
##  5     5 21st  0.000255
##  6     1 aapi  0.000465
##  7     2 aapi  0.000232
##  8     3 aapi  0.000222
##  9     4 aapi  0.0000229
## 10     5 aapi  0.00107
## # ... with 4,465 more rows
```

```
d_top_terms <- d_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

d_top_terms %>%
```

```r
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```



Fitting topic models at different levels of k

```r
#first, split data into training and testing sets

length_r <- length(r)
cat_r <- as.factor(c(rep('Train',ceiling(length_r*0.8)),
                rep('Test',length_r-ceiling(length_r*0.8))))
split_r <- split(r,cat_r)
split_r
```

```
## $Test
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 497
##
## $Train
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 1989
```

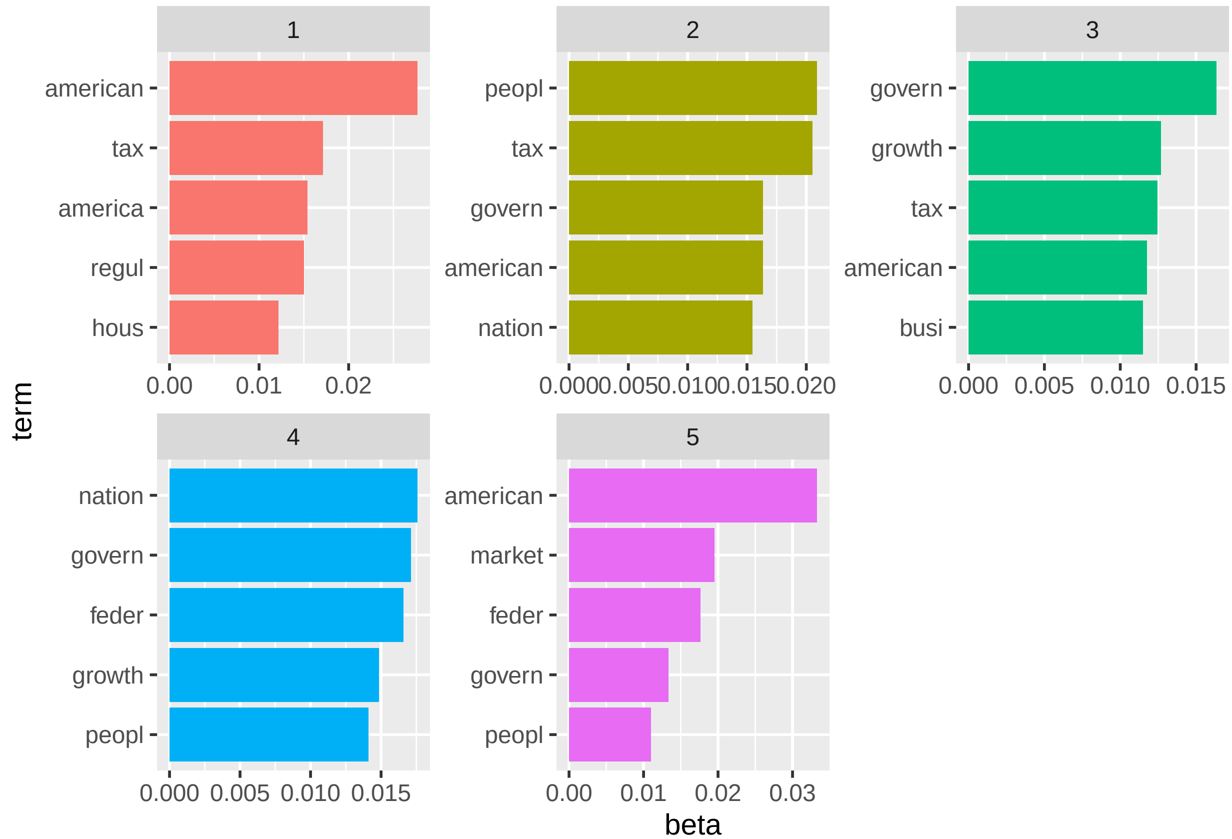
```r
length_d <- length(d)
cat_d <- as.factor(c(rep('Train',ceiling(length_d*0.8)),
                    rep('Test',length_d-ceiling(length_d*0.8))))
split_d <- split(d,cat_d)
split_d
```

```
## $Test
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 497
##
## $Train
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 1990
```

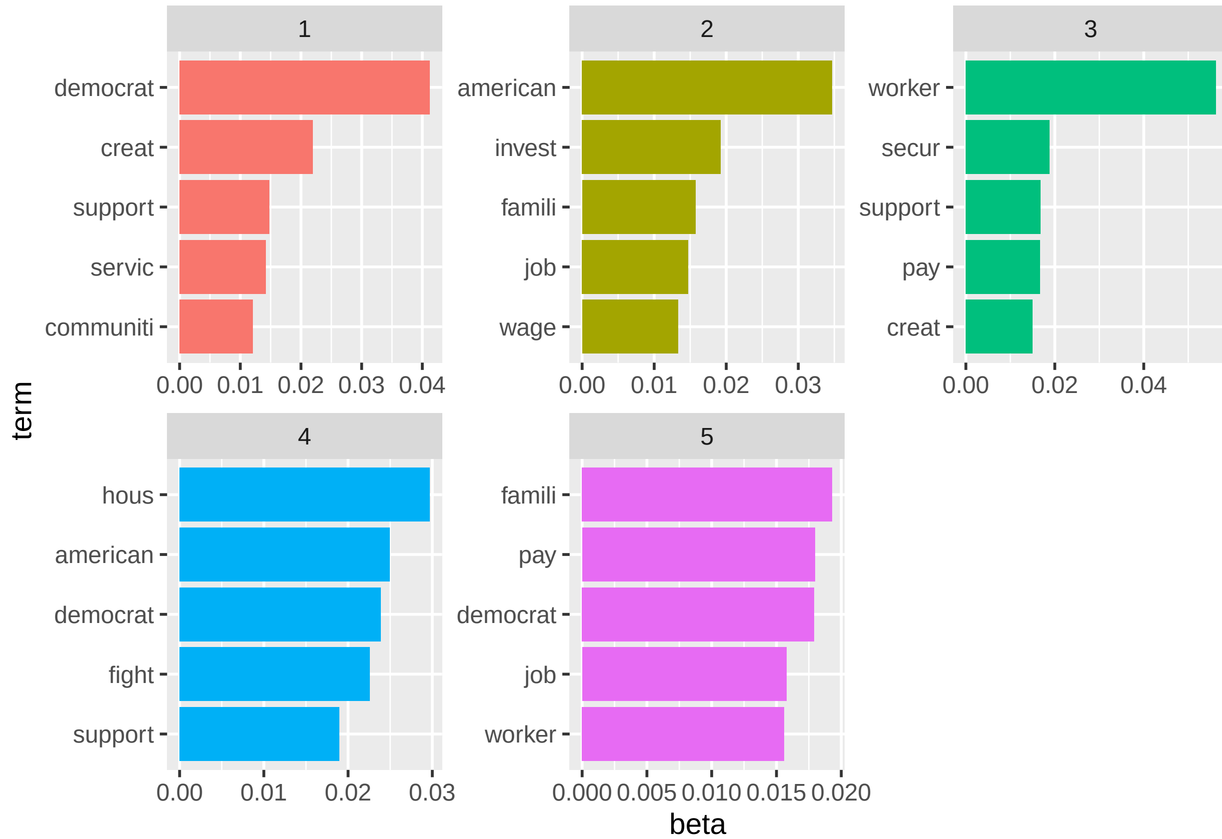
```r
r_dtm_train <- DocumentTermMatrix(split_r$Train)
d_dtm_train <- DocumentTermMatrix(split_d$Train)
r_dtm_test <- DocumentTermMatrix(split_r$Test)
d_dtm_test <- DocumentTermMatrix(split_d$Test)

raw.sum=apply(r_dtm_train,1,FUN=sum) #sum by raw each raw of the table
r_dtm_train=r_dtm_train[raw.sum!=0,]

raw.sum=apply(d_dtm_train,1,FUN=sum) #sum by raw each raw of the table
d_dtm_train=d_dtm_train[raw.sum!=0,]

raw.sum=apply(r_dtm_test,1,FUN=sum) #sum by raw each raw of the table
r_dtm_test=r_dtm_test[raw.sum!=0,]

raw.sum=apply(d_dtm_test,1,FUN=sum) #sum by raw each raw of the table
d_dtm_test=d_dtm_test[raw.sum!=0,]

# fit model at different levels of k with training sets
r_lda5 <- LDA(r_dtm_train, k = 5, control = list(seed = 1234))
d_lda5 <- LDA(d_dtm_train, k = 5, control = list(seed = 1234))
r_lda10 <- LDA(r_dtm_train, k = 10, control = list(seed = 1234))
d_lda10 <- LDA(d_dtm_train, k = 10, control = list(seed = 1234))
r_lda25 <- LDA(r_dtm_train, k = 25, control = list(seed = 1234))
d_lda25 <- LDA(d_dtm_train, k = 25, control = list(seed = 1234))
```

```r
#presenting terms that are most common within each topic for k = 5
r_topics5 <- tidy(r_lda5, matrix = "beta")
d_topics5 <- tidy(d_lda5, matrix = "beta")

r_top_terms5 <- r_topics5 %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

r_top_terms5 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
```

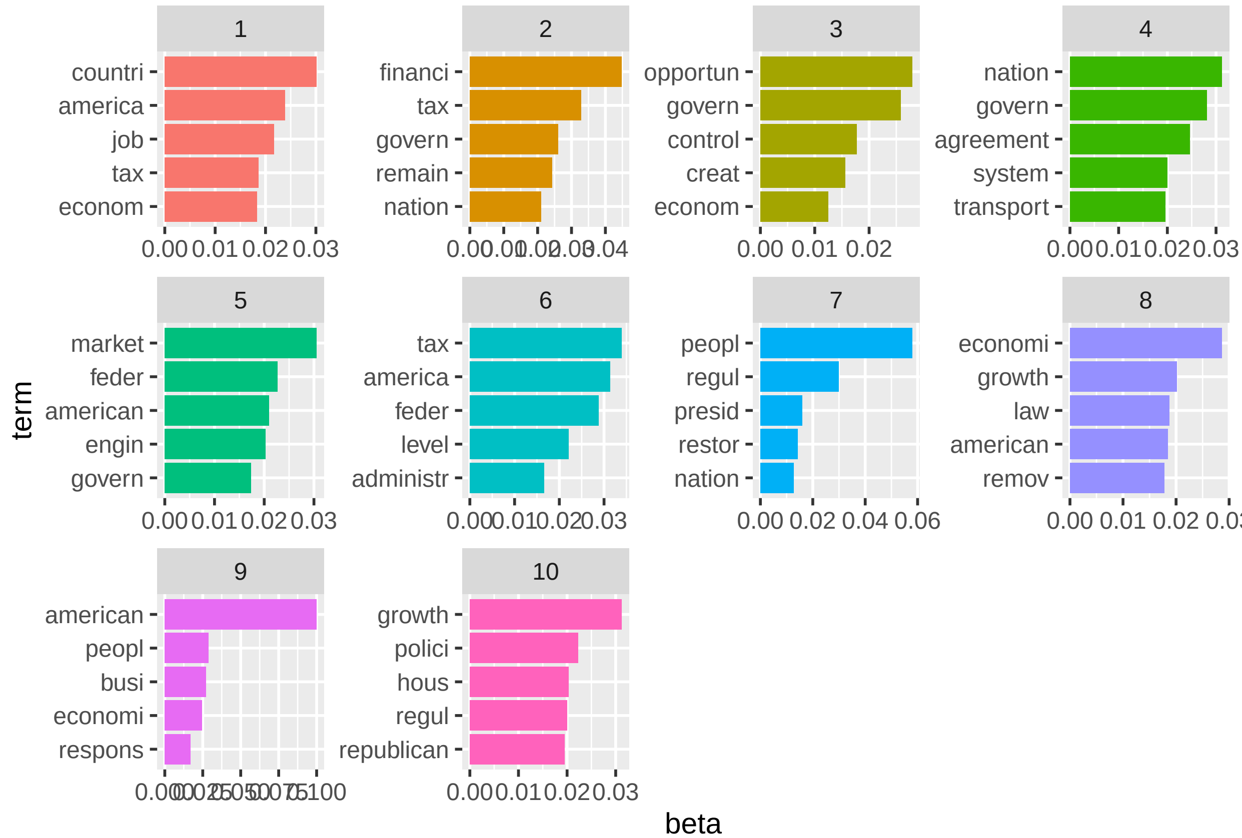
```r
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```




```r
d_top_terms5 <- d_topics5 %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

d_top_terms5 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```

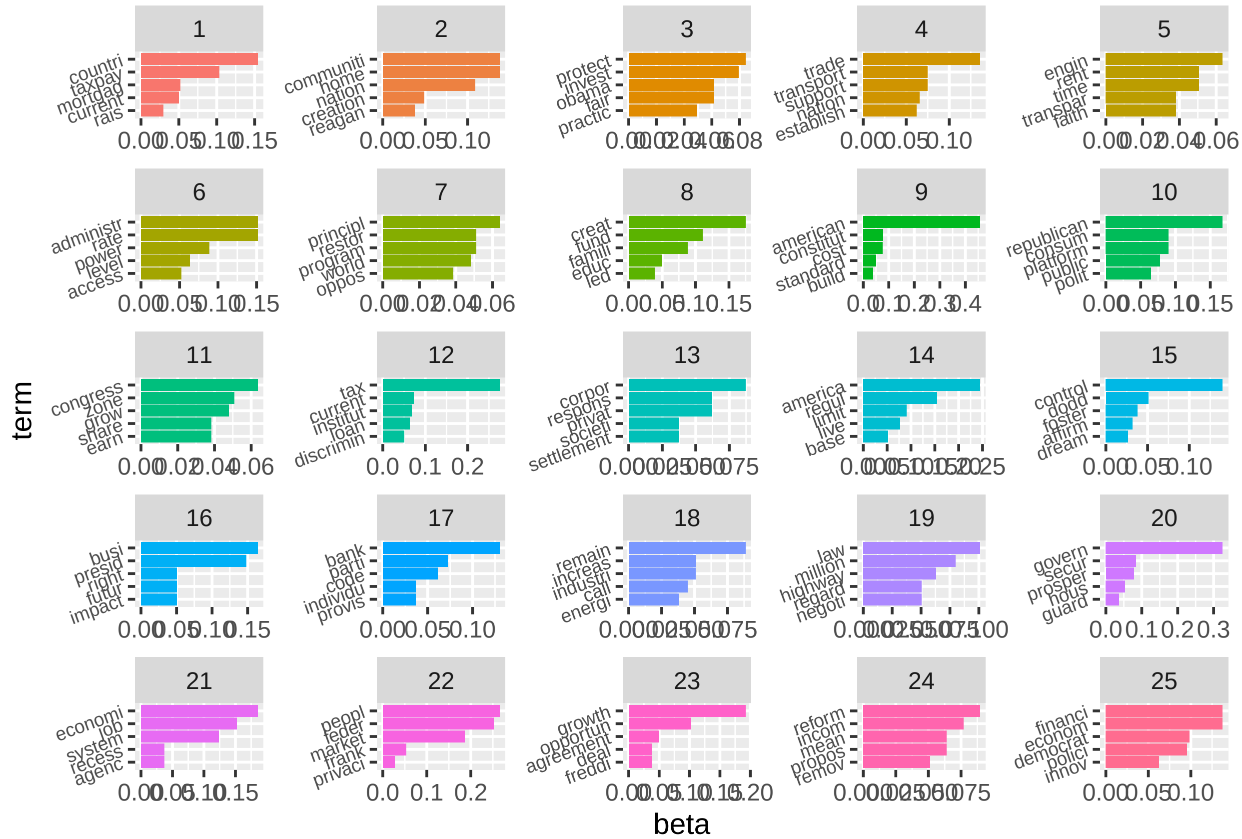



```
#presenting terms that are most common within each topic for k = 10
r_topics10 <- tidy(r_lda10, matrix = "beta")
d_topics10 <- tidy(d_lda10, matrix = "beta")

r_top_terms10 <- r_topics10 %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

r_top_terms10 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```

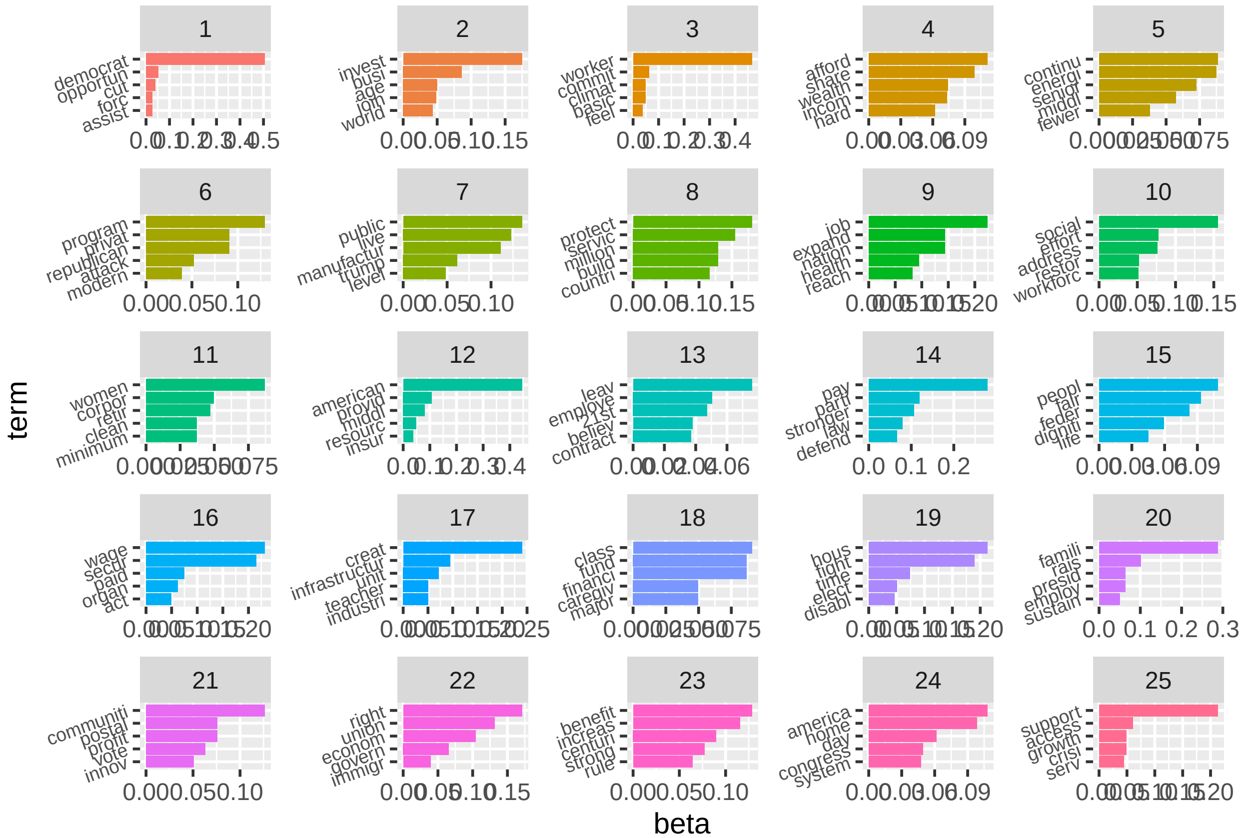



```r
d_top_terms10 <- d_topics10 %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

d_top_terms10 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```


```r
#presenting terms that are most common within each topic for k = 25

r_topics25 <- tidy(r_lda25, matrix = "beta")
d_topics25 <- tidy(d_lda25, matrix = "beta")

r_top_terms25 <- r_topics25 %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

r_top_terms25 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()+ theme(axis.text.y = element_text(angle = 20, size = 7))
```




```r
d_top_terms25 <- d_topics25 %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

d_top_terms25 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()+ theme(axis.text.y = element_text(angle = 20, size = 7))
```



Perplexity

Observations about perplexity

We want to minimize perplexity. For both republican and democrat data, k = 25 is optimal. However, the differences between different k are small.

```
perplexity(r_lda5, newdata = r_dtm_test)
```

```
## [1] 540.1267
```

```
perplexity(r_lda10, newdata = r_dtm_test)
```

```
## [1] 545.2889
```

```
perplexity(r_lda25, newdata = r_dtm_test)
```

```
## [1] 530.4225
```

```
perplexity(d_lda5, newdata = d_dtm_test)
```

```
## [1] 451.2888
```

```r
perplexity(d_lda10, newdata = d_dtm_test)
```

```
## [1] 453.9317
```

```r
perplexity(d_lda25, newdata = d_dtm_test)
```

```
## [1] 443.3811
```

barplot of k = 10 for each party

Observations of k = 10

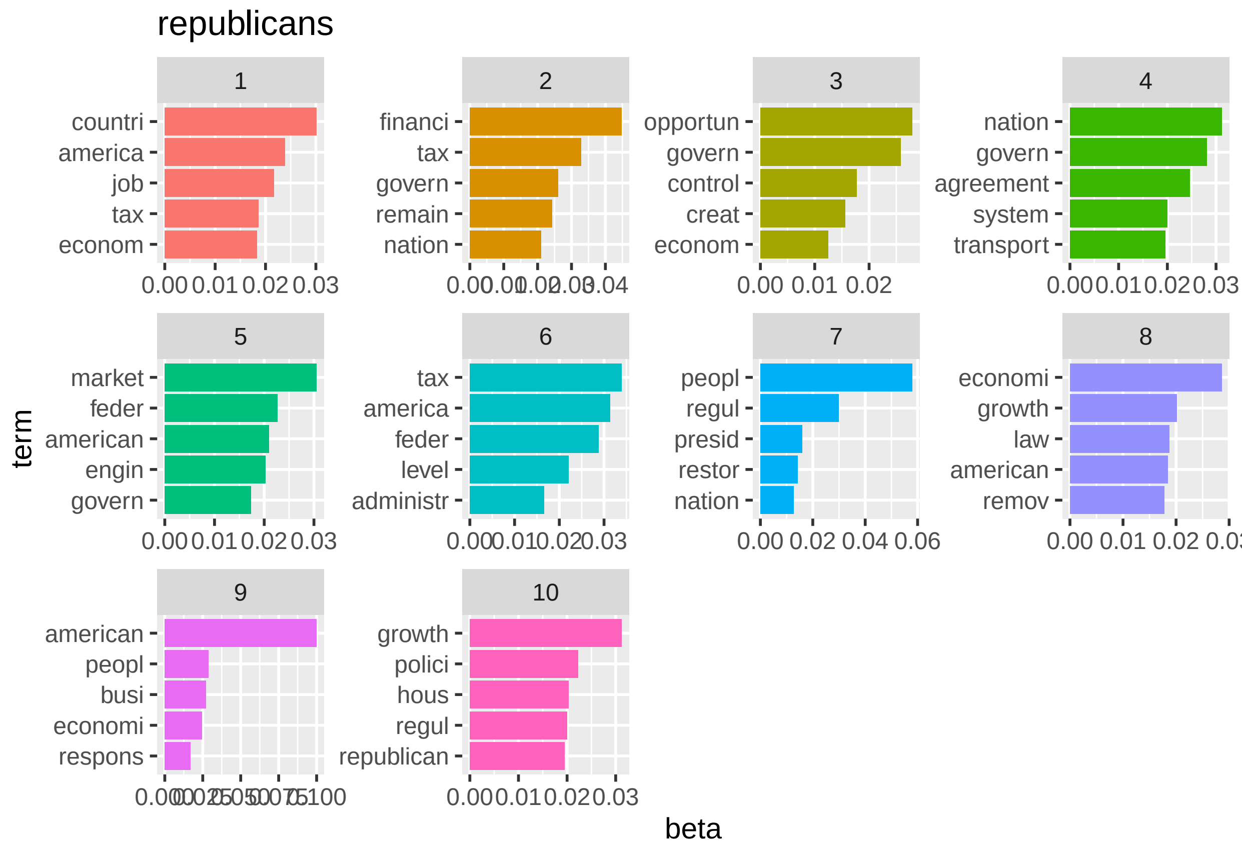
Similar as before, Democrats focus a lot more on workers, welfare/benefits, and families. There is a lot of focus on creating jobs. Interestingly, two kinds of workers were the highlight of two separate topics - teachers in topic 5, and manufacturers in topic 7. It'll be interesting to see why there was a focus on these two kinds of jobs in these Democratic narratives during that time.

Republicans still focus a lot more on the market, taxes, and the economy.

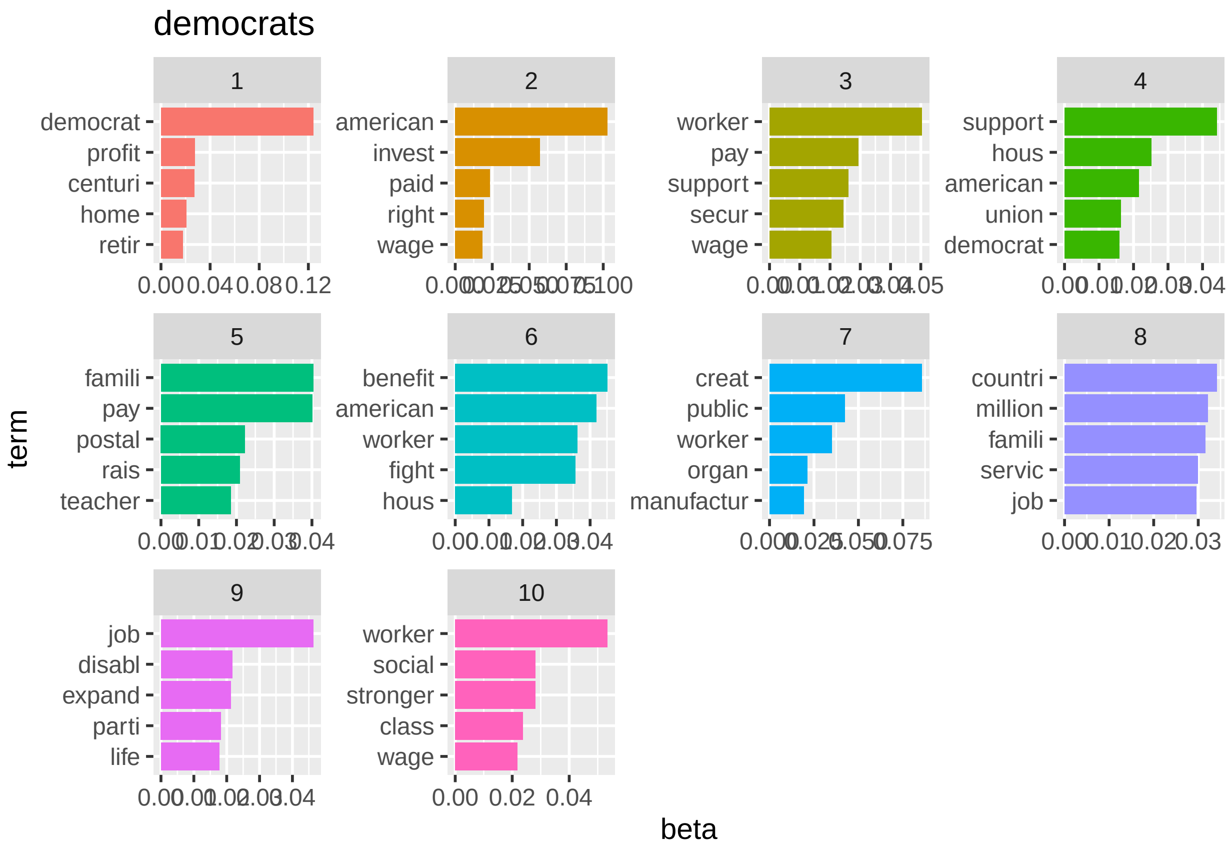
Although Republicans and Democrats are both interested in growth, Democrats tend to discuss growth in light of creating jobs and how that benefits people and workers, while Republicans may focus less onthe people and more on the economy and how that benefits the nation and businesses.

It seems that k = 10 is a more parsimonious way of understanding the topics than k = 25. Although k = 25 has the lowest perplexity score, it may be uninformative to interpret topics that are too specific. However, that heavily depends on our research question.

```r
r_top_terms10 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()+
  ggtitle("republicans")
```




```
d_top_terms10 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()+
  ggtitle("democrats")
```




```r
r_documents <- tidy(r_lda10, matrix = "gamma")
r_documents
```

```
## # A tibble: 19,810 x 3
##    document topic  gamma
##    <chr>    <int>  <dbl>
##  1 1            1 0.102
##  2 2            1 0.0998
##  3 3            1 0.0989
##  4 4            1 0.100
##  5 5            1 0.0999
##  6 6            1 0.101
##  7 7            1 0.0990
##  8 8            1 0.0992
##  9 9            1 0.0997
## 10 10           1 0.0995
## # ... with 19,800 more rows
```

```r
d_documents <- tidy(d_lda10, matrix = "gamma")
d_documents
```

```
## # A tibble: 19,880 x 3
##    document topic  gamma
##    <chr>    <int>  <dbl>
##  1 1            1 0.106
```

```
##   2 2             1 0.0994
##   3 3             1 0.0991
##   4 4             1 0.0995
##   5 5             1 0.0996
##   6 6             1 0.104
##   7 7             1 0.0991
##   8 8             1 0.0992
##   9 9             1 0.100
## 10 10             1 0.0990
## # ... with 19,870 more rows
```

Conclusion

Hypothetically, I would still support the Democratic party, mostly because of its focus on "people" and "welfare" on policy. I generally don't think nationalistic narratives that the Republican party advocates for makes much sense, especially as the world becomes more globalized.