hw3

Juan Vila

11/25/2019

Problem Set 3

Part 1 and 2

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
## Warning: package 'tidytext' was built under R version 3.5.2
```

```
# LOADING TEXTS
#1.
## FOR MAC:
texts <- file.path("~", "Documents","GitHub","problem-set-3", "Party Platforms Data")


# Now we can create our raw corpus, which we will preprocess in a moment
docs <- VCorpus(DirSource(texts))


docs <- tm_map(docs,removeNumbers)
docs <- tm_map(docs,removePunctuation)
docs <- tm_map(docs,content_transformer(tolower))
docs <- tm_map(docs,removeWords, stopwords("english"))


# MAybe we need a bit more cleaning of unique characters
for (j in seq(docs)) {
  docs[[j]] <- gsub("/", " ", docs[[j]])
  docs[[j]] <- gsub("’", " ", docs[[j]])
  docs[[j]] <- gsub("-", " ", docs[[j]])
  docs[[j]] <- gsub("\\|", " ", docs[[j]])
  docs[[j]] <- gsub("@", " ", docs[[j]])
  docs[[j]] <- gsub("\u2028", " ", docs[[j]])  # an ascii character that does not translate
}

for (j in seq(docs)) {
  docs[[j]] <- gsub("federal government", "federal_goverment", docs[[j]])
  docs[[j]] <- gsub("united states", "united_states", docs[[j]])
}
(docs <- tm_map(docs, PlainTextDocument))
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 3
```

```
docs <- tm_map(docs, stripWhitespace)
# manually removing words too for your specific purposes (e.g., our "representative" or "honourable" ex

docs <- tm_map(docs, removeWords, c("will","must","also"))
docs <- tm_map(docs, removeWords, c("democrats","republicans"))
```

We clean the following items: 1. numbers 2. Punctuation 3. We transform the data into lower letters 4. Stop words.

Also, we eliminate words that are repited but they don't have content in the analysis because are words that refer to the actions that they will do as will and must. Also, we eliminate the word also due to that the role that have in the phrases is to connect them. Additionally, we see that democrats and republicans also have a high frecuency especially in the case of democrats, this could imply that is their plataform is much more self-centered or triying to create a brand. Nevertheless, we decide to eliminate both, because as if even this word are high in the ranking this maybe bias the analysis of the content of the words. Finally, we decide also after analyzing some results to consider some words as one, for example Federal Goverment and United States, because this are one concept.

Part 3

After doing this we analyze the basic frequencies found in both corpus. We analyze the sixth most repeated word in the Corpus. Now we proceed to analyze them by each party.

Sixth more used words by Democrats

```
##    health   support   believe    people americans  american
##       130       123       117       111        94        86
```

In the case of the Democrats we found something very interesting they use three words that in the context of the text means the same, we are talking of people, american and americans which rank from fourth to six. Also, we can see that the most used words are health, support and belive. The word health is used in the context of healthcare policy, support in any policy focused in help or bussiness, students or person in any kind of things as paying loans, helping farmers, first home buyers, etc.

Sixth more used words by Republicans

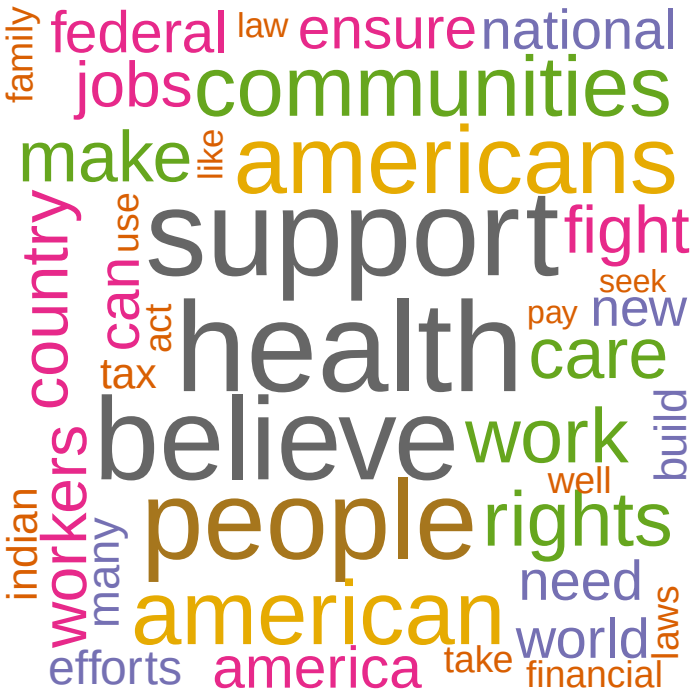
```
##   american government    federal    support     people   national
##        121        110        106        100         98         83
```

In the case of the republicans they use also people(5th), national(6th) and american(1th), this is interesting taking into consideration that not even in the most 15 words used the word amaricans(19th) is not appear. Specially taking into consideration the fact that the Republican Ideological agenda is always focused in a indvidualized view of how to construct a society and they nationalistic views. In other hand, we see that they made reference to "goverment"(2nd), and "federal"(3th), appears beetween the places 2nd until 3th, the context where they used this word reading them in the corpus is critizing big goverment. For example, "Big government undermines federalism through more than 1,100 grants-in-aid programs that consume more than one-sixth of the federal budget for matters that should be the exclusive responsibility of the states.". Finally, in the case of the case of support that have in the context depends in how the republican party supports different policies.

Now if we build the wordcloud for the Republican and Democratic corpus, we found the follwing result.Besides the six first we see the are words as ensure, take, jobs, work, rights, need, student, climate that star to show what are the elements that build the Democratic Plataform that made sense to the things that we

expect that a democrats would speek for and what are the agents and policy which their efforts if they are goverment will be focus. The same is true in the case of the republicans. Because we found words as economic, adminsitration, economy, private,tax, religuous, military,etc. This made sanse for what is expected how the policy that the Republican Party want to implement, policy related to free market economics, where the military have a predominant role in society and where the religious right, it's free to educate young people as they please.

Democrats agenda Frequency Wordcloud



Republicans agenda Frequency Wordcloud

Part 4 and 5

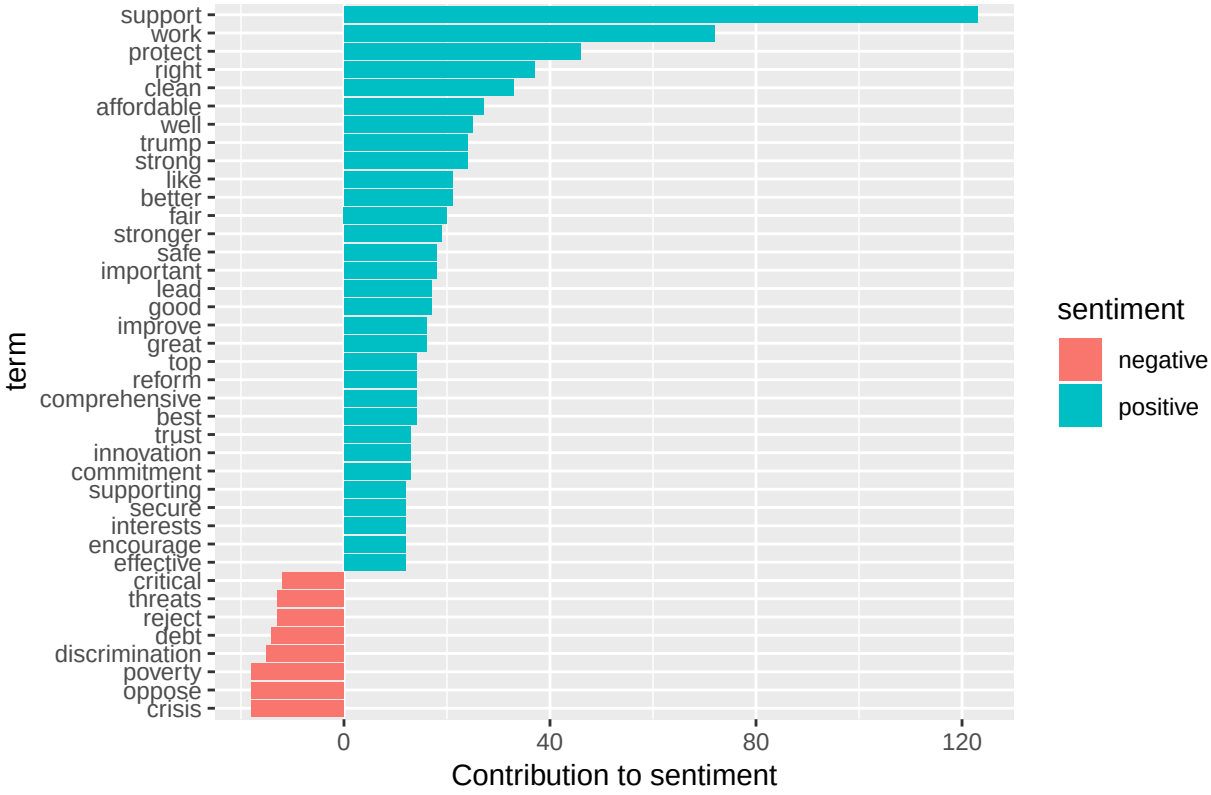
Bing

```
## Current value sentiment Dem. Party:  -67
```

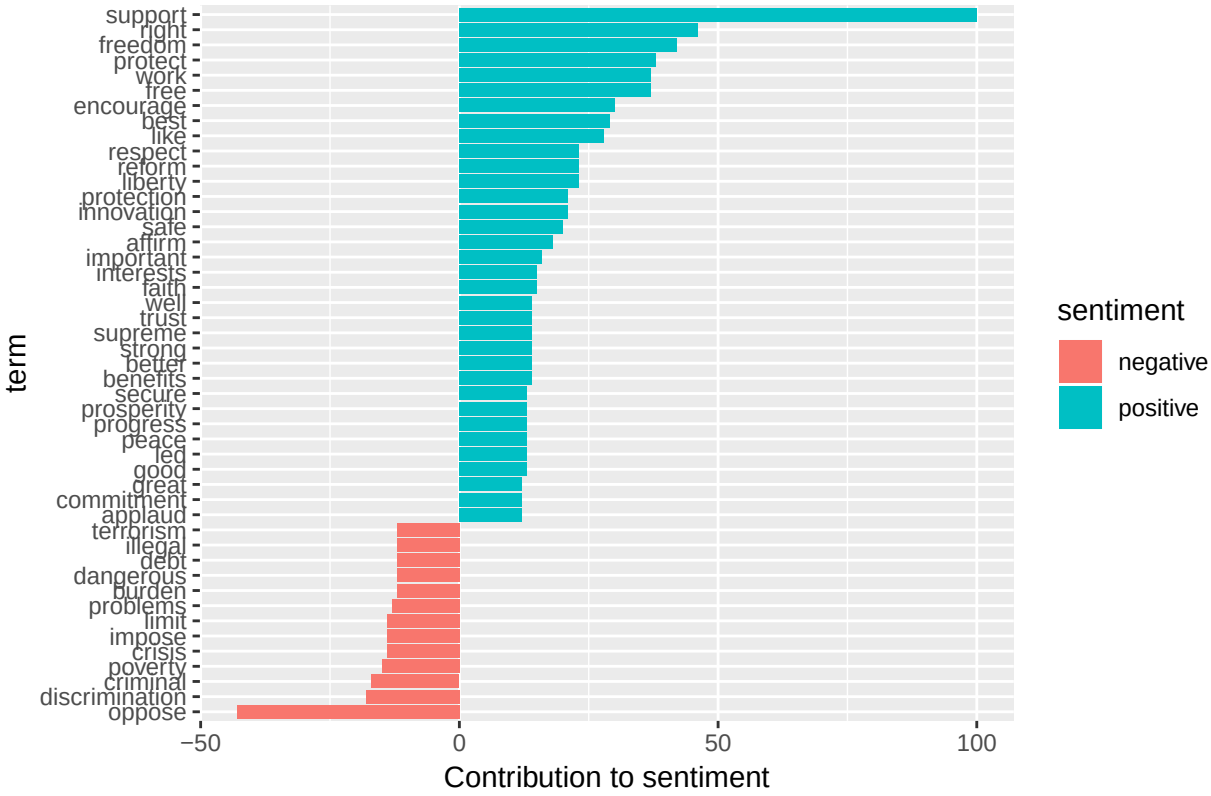
```
## Current value sentiment Rep. Party:  -182
```

We can see that in general both add value of the sentiment scores are negative meaning that in general are more negatives sentiments than positive. Nevertheless, we can see that the Rep. Party have negative score, which made sense considering that the Republican Party since 1980s have looked always for that old good days of 50s, this imply that in their speechs and plataform would likely to dimminish their positive view about the current situation. This is view the following graphs where we can see that they have more negatives words that have more than 12 counts in Corpus than the Democratas. This reenforce the overall result in the general score.

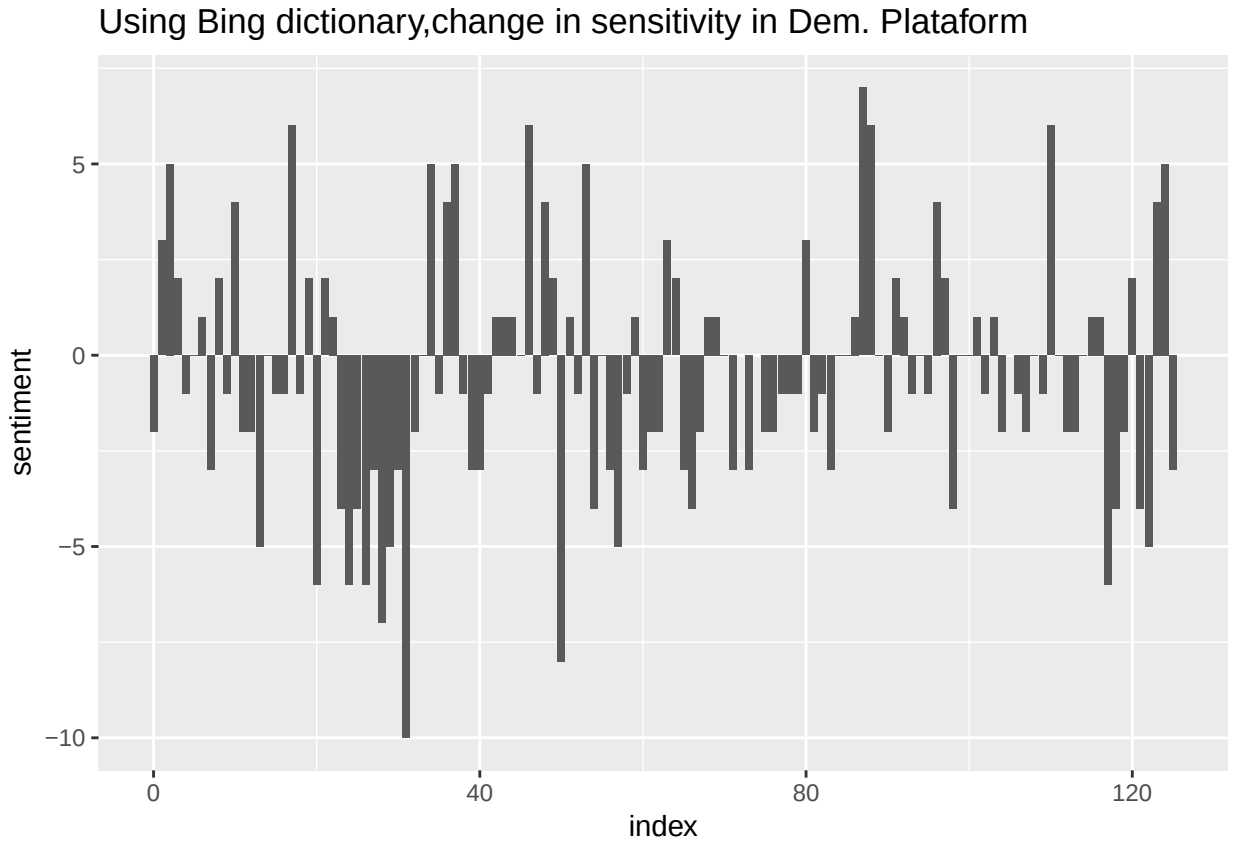
Dem. Plataform Contribution to Sentiment (More than 12 counts)

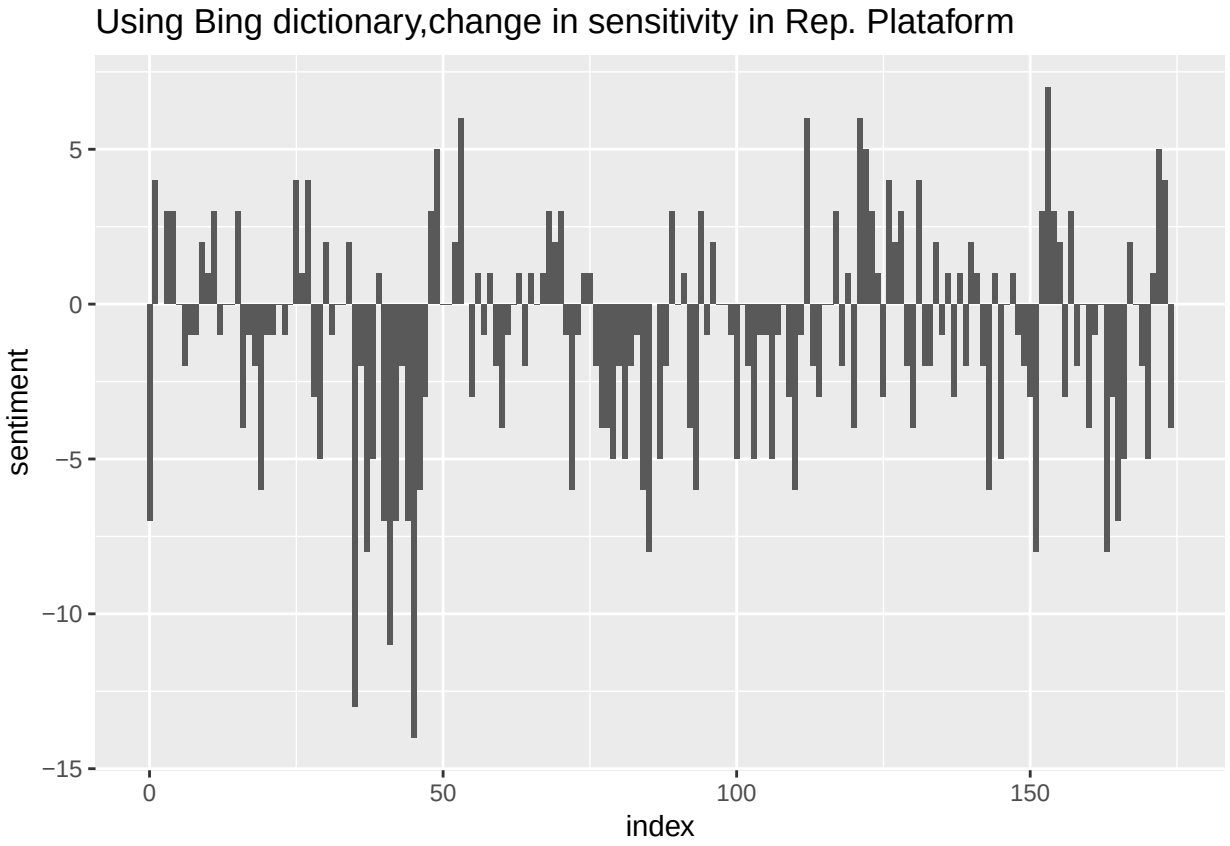


Rep. Plataform Contribution to Sentiment (More than 12 counts)



This overall negativity is noted when we bunch the merged word with the dictionary and see how the score evolves throught the document, we see that are parts where the maximium negativity comes in the indexes 30 for Democrats and 50 for Republicans, but in this parts Republicans are more way negative than Democrats.





AFINN

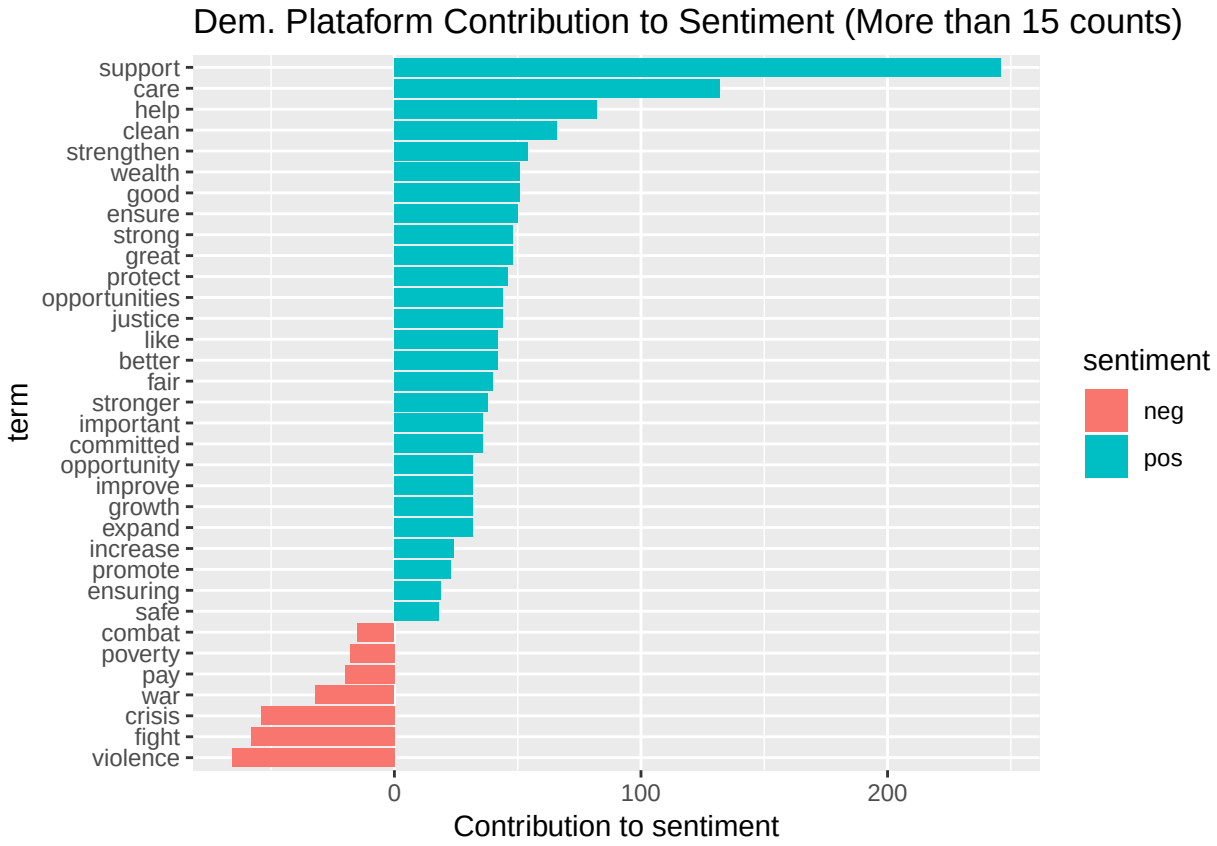
One of the major limitations of the Bing dictionary is that not take into consideration that even if a word is negative or positive they can have a different weight depending of the word. For taking that into consideration we now use the AFINN dictionary. At firs, we found that the value sentiment for this dictionary are positive in aggregate, but we still see that the Democratic Party look is more positive in general.

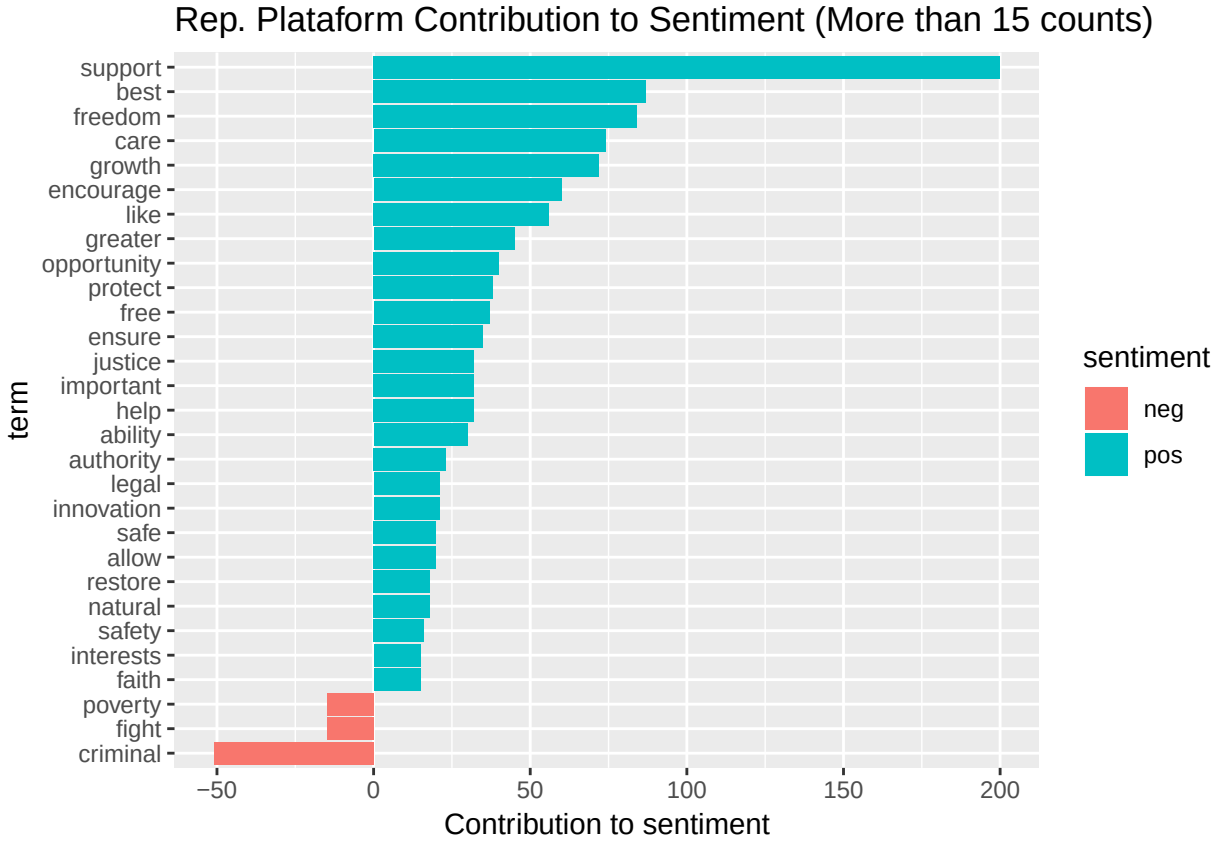
```
## Current value sentiment Dem. Party:  1270
```

```
## Current value sentiment Rep. Party:  891
```

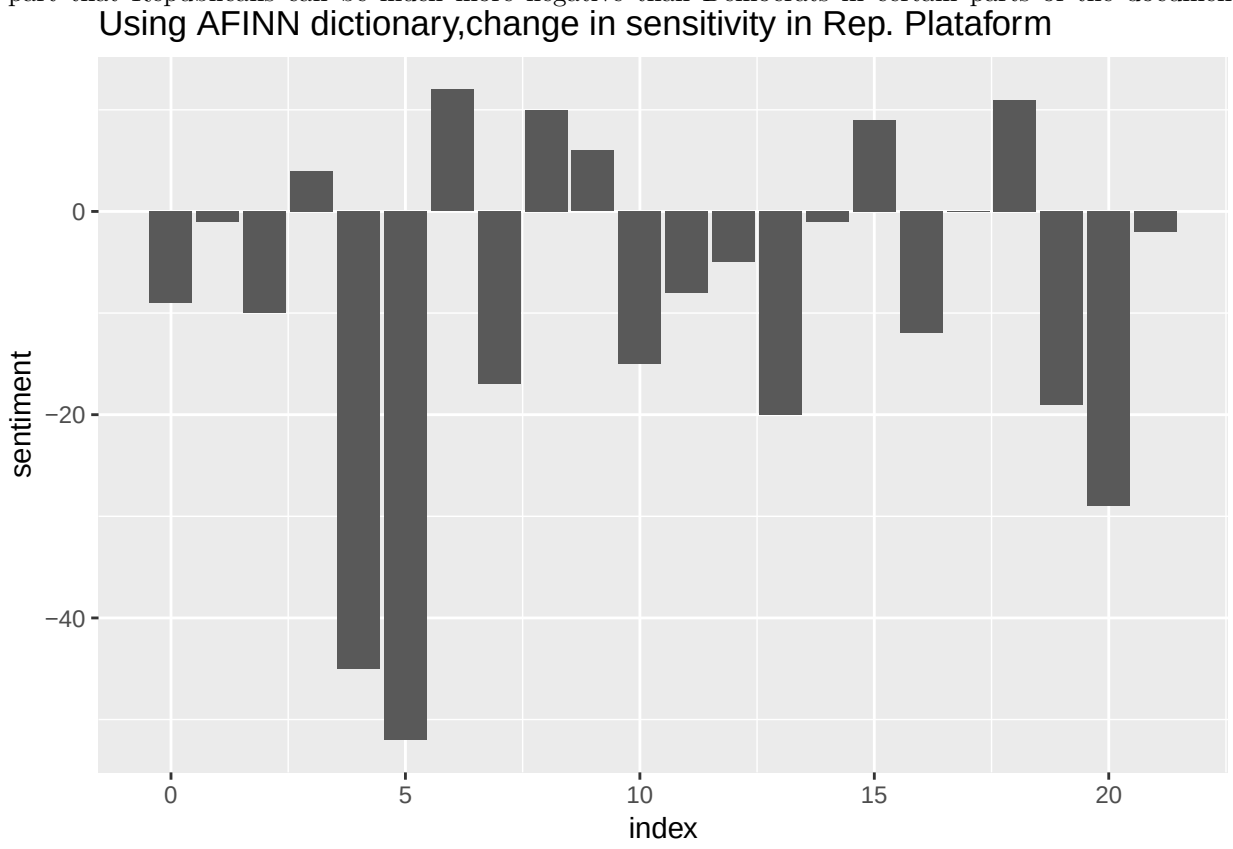
If we see the added sentiment by type of word with more of 15 counts we see a switch for the results of this same graphs in the previpus part. We found that in general Democrats uses negatives words that have a much higher score of negativity than Republicans, for example Violence that is the more aggregated negative word that is used by the Democrats, is not used by Republicans. In other hand, Republicans tend to use positive words what have a higher count and wheight as support, freedom and best.

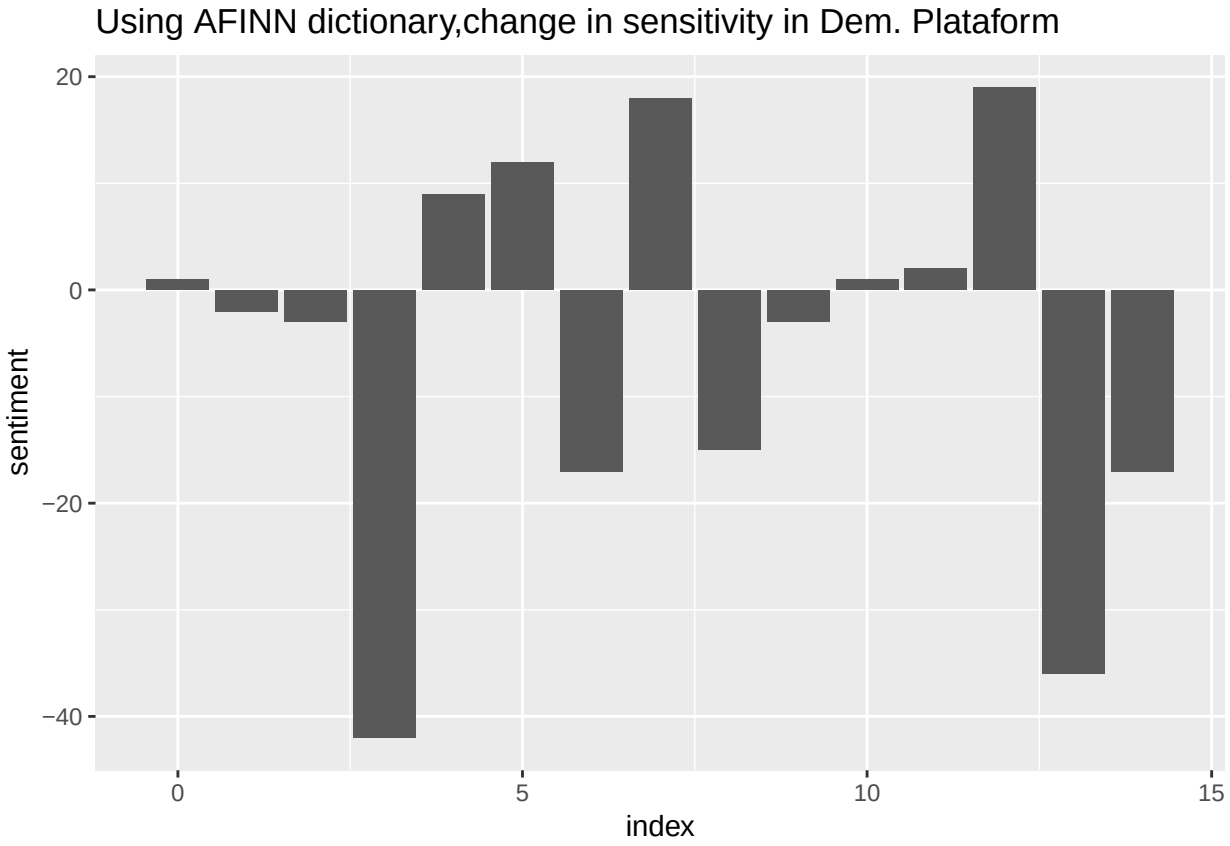
In the case of Democrats we see a great ammount of usage of words as strengthen, protect, opportunities and ensure, that is consistent with what we think about democrats, in the sense that as they are more pro big goverment in a certain way they are looking for more protection of states of nature and look throught goverment intervention.


Dem. Plataform Contribution to Sentiment (More than 15 counts)
term
support
care
help
clean
strengthen
wealth
good
ensure
strong
great
protect
opportunities
justice
like
better
fair
stronger
important
committed
opportunity
improve
growth
expand
increase
promote
ensuring
safe
combat
poverty
pay
war
crisis
fight
violence
0
100
200
Contribution to sentiment
sentiment
neg
pos


Rep. Plataform Contribution to Sentiment (More than 15 counts)
term
support
best
freedom
care
growth
encourage
like
greater
opportunity
protect
free
ensure
justice
important
help
ability
authority
legal
innovation
safe
allow
restore
natural
safety
interests
faith
poverty
fight
criminal
−50
0
50
100
150
200
Contribution to sentiment
sentiment
neg
pos

Finally, when we analyze how the the sentiment change through the document. We found that are parts of the plataforms where the overall sentiment is negative. Nevertheless, we keep finding as in the previous part that Republicans can be much more negative than Democrats in certain parts of the documents.

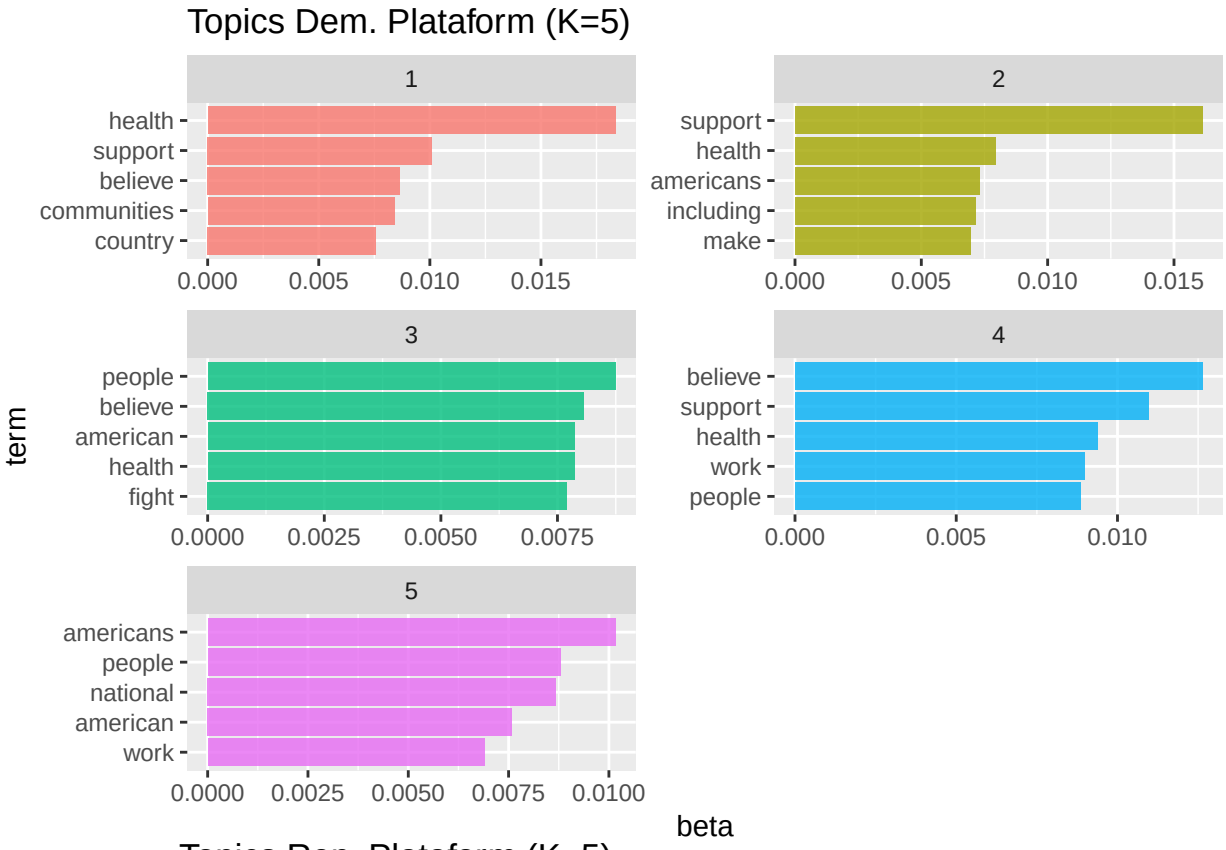


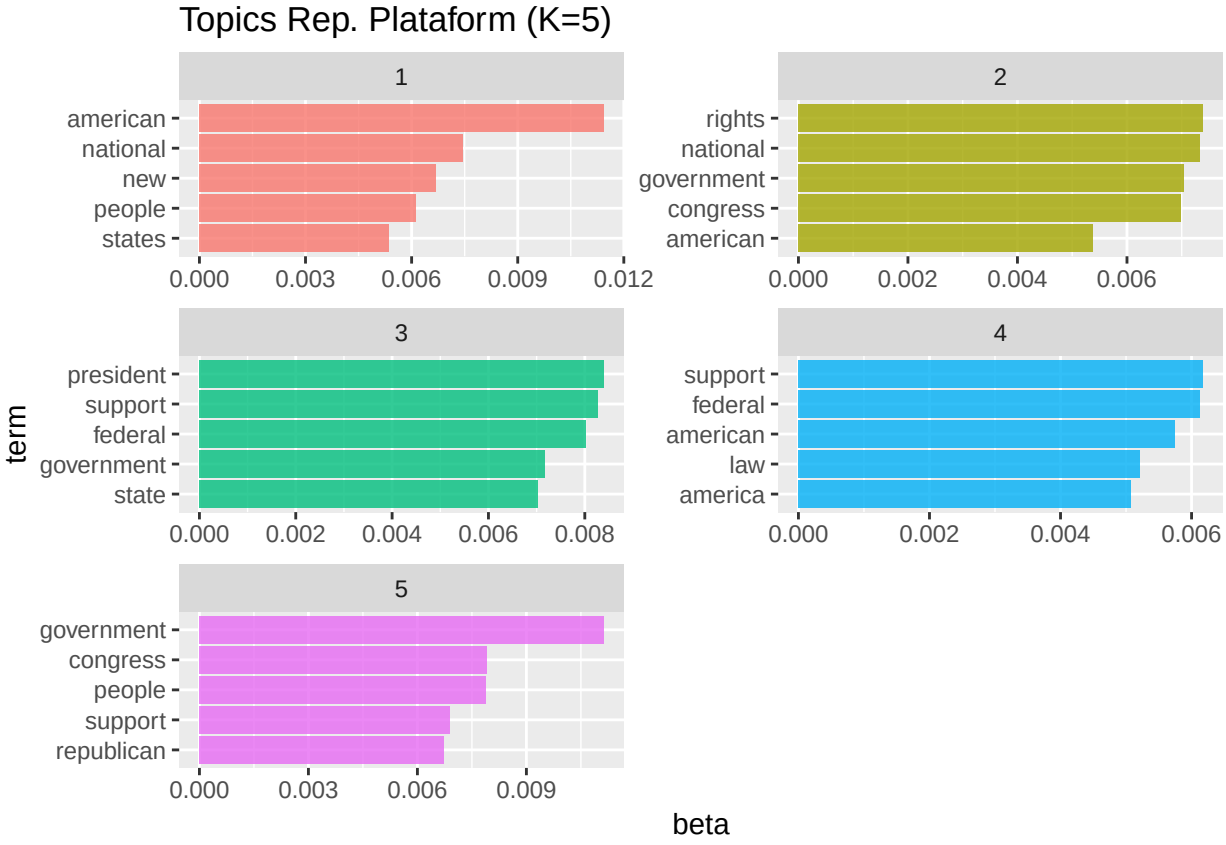


From this part we can see that in part the evidence presented is kind of contradictory when we look the overall use of words positive and negative. This contradiction is found in the result that republicans tend to use less negative words with a higher wheight than Democrats. Nevertheless, this contradiction is only in this point and the other elements presented show that Democrats are more optimistic about the future.

Part 6 and 7

We can see that in general both plataforms have different views. In the case of democrats we found that they have focus in concepts as Health, Support, Work and People. This made us think that the plataform of the Democrat party is more close to protect people and communities. In the case of the Republican Party, we can see that is more about the relationship between government and people. Which now this made sense in the case of them. Because, a great part of the Republican agenda is “protect” people from the Government.

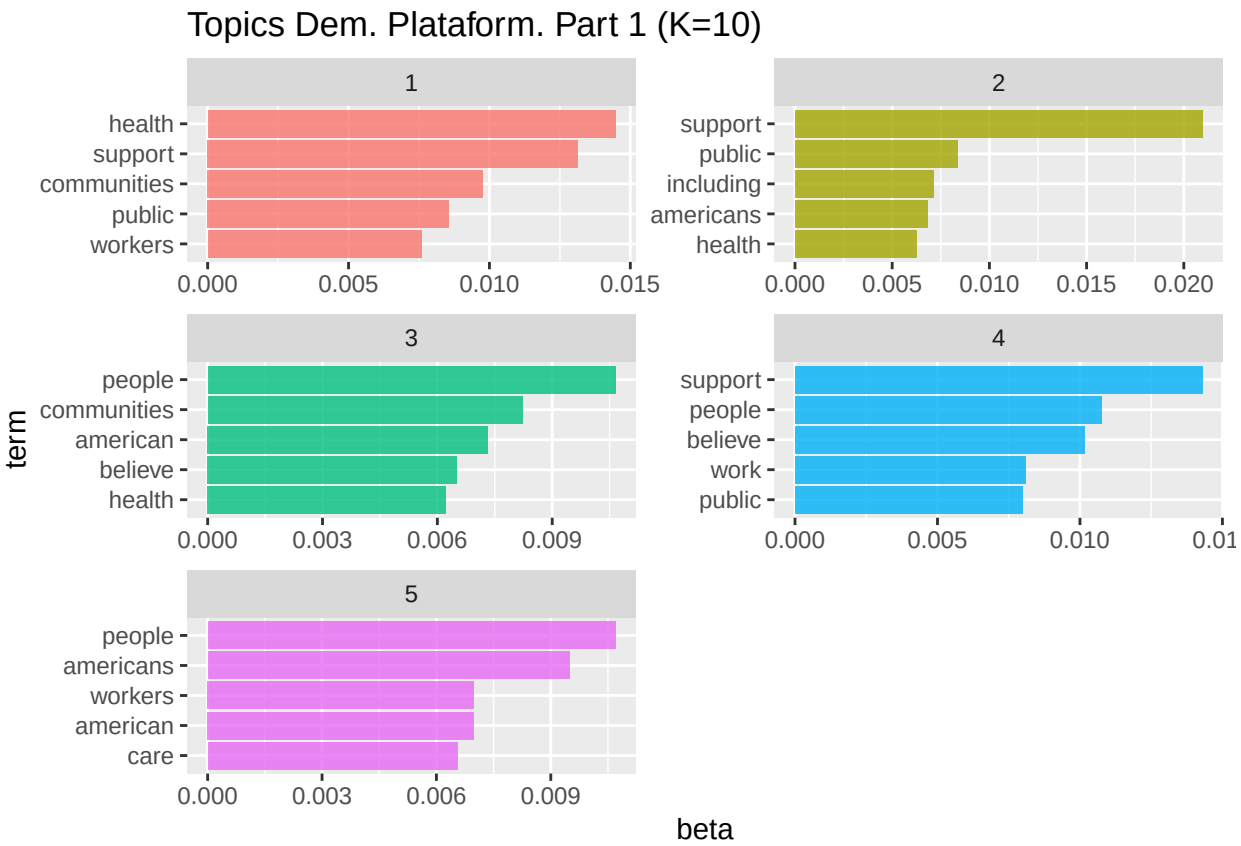

Topics Dem. Plataform (K=5)
1
health
support
believe
communities
country
0.000
0.005
0.010
0.015
2
support
health
americans
including
make
0.000
0.005
0.010
0.015
3
people
believe
american
health
fight
0.0000
0.0025
0.0050
0.0075
4
believe
support
health
work
people
0.000
0.005
0.010
5
americans
people
national
american
work
0.0000
0.0025
0.0050
0.0075
0.0100
term
beta

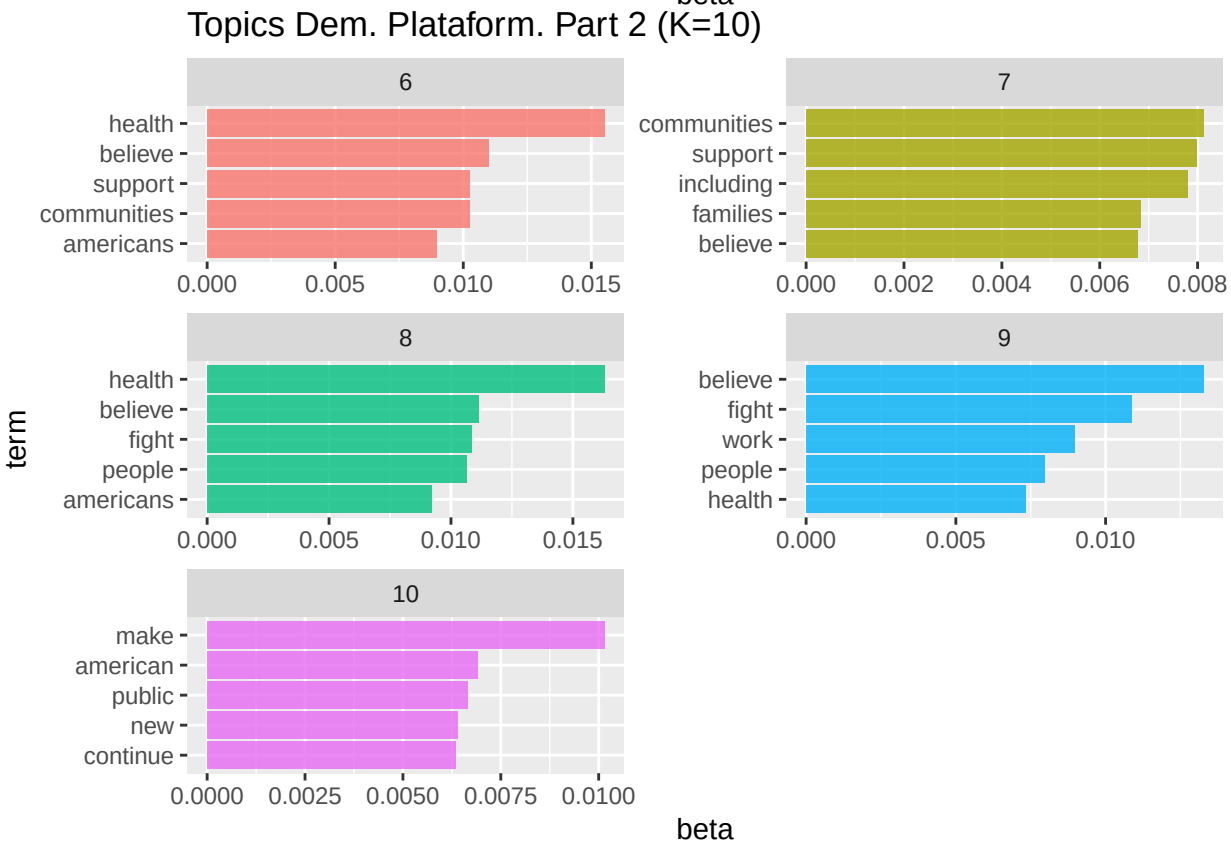

Topics Rep. Plataform (K=5)
1
american
national
new
people
states
0.000
0.003
0.006
0.009
0.012
2
rights
national
government
congress
american
0.000
0.002
0.004
0.006
3
president
support
federal
government
state
0.000
0.002
0.004
0.006
0.008
4
support
federal
american
law
america
0.000
0.002
0.004
0.006
5
government
congress
people
support
republican
0.000
0.003
0.006
0.009
term
beta

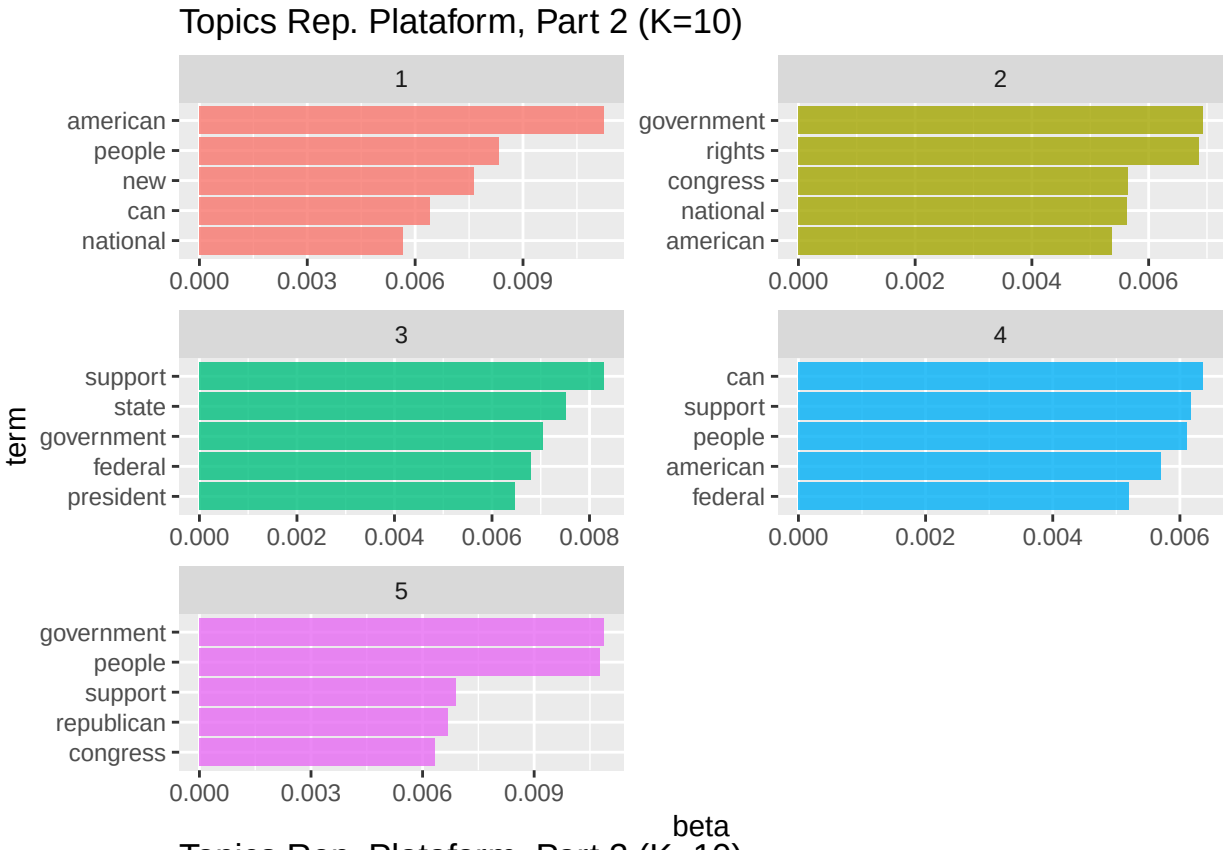
Part 8

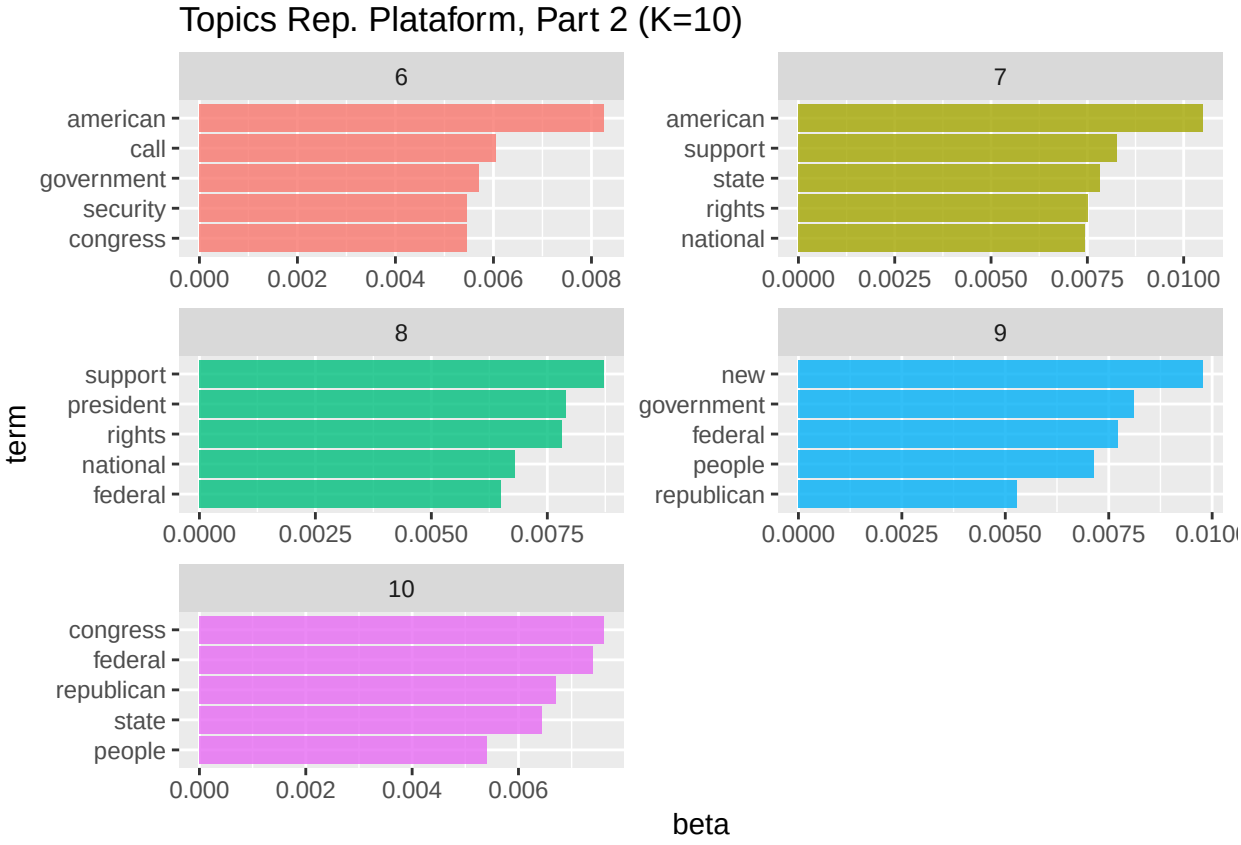
In general, if we look K=10 for both partys we notice that the terms that appears are mostly lineal combinations of the ones that that thet appear when we use K=5, this would imply that is redundant to use K=10, because we can have a more parsimonius explication of what we had found. For K=25, is exactly the same situation.

K=10

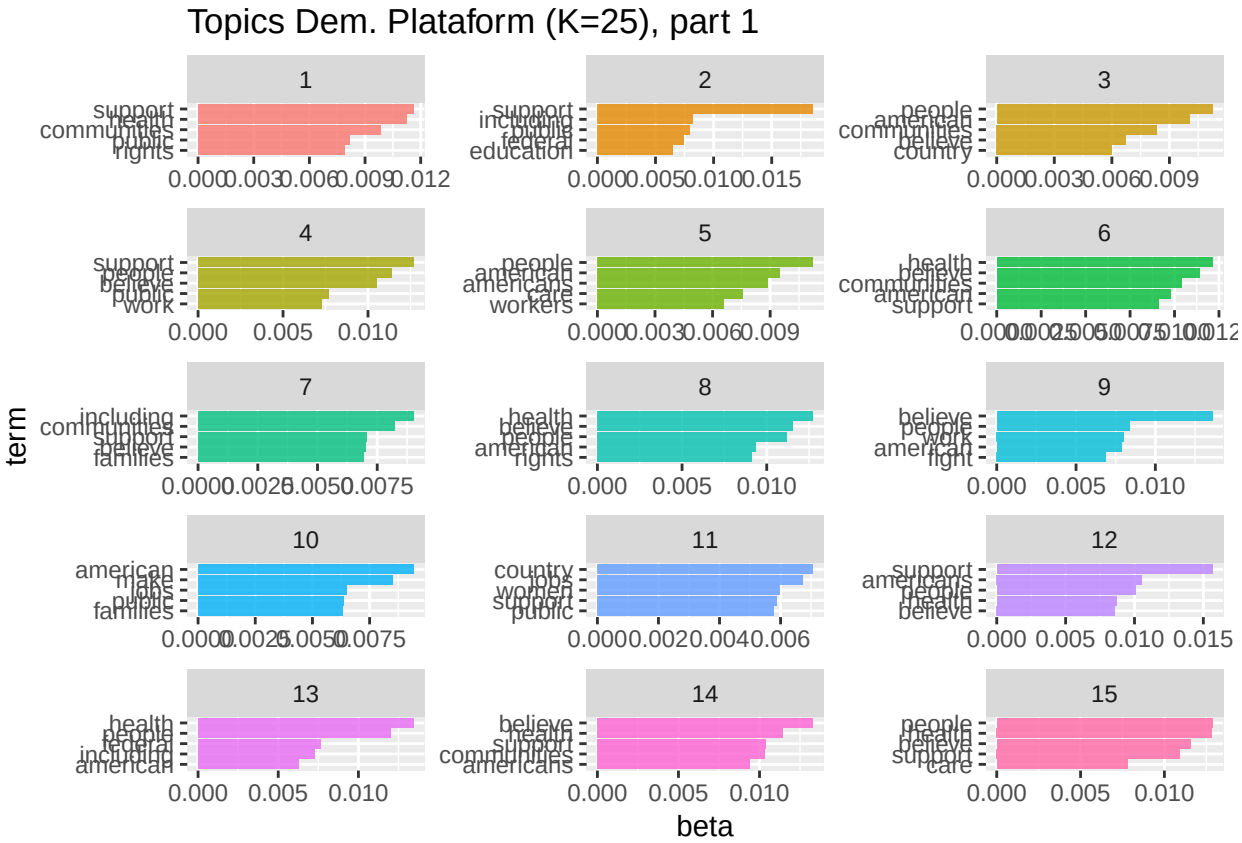

Topics Dem. Plataform. Part 1 (K=10)
1
health
support
communities
public
workers
0.000
0.005
0.010
0.015
2
support
public
including
americans
health
0.000
0.005
0.010
0.015
0.020
3
people
communities
american
believe
health
0.000
0.003
0.006
0.009
4
support
people
believe
work
public
0.000
0.005
0.010
0.01
5
people
americans
workers
american
care
0.000
0.003
0.006
0.009
term
beta

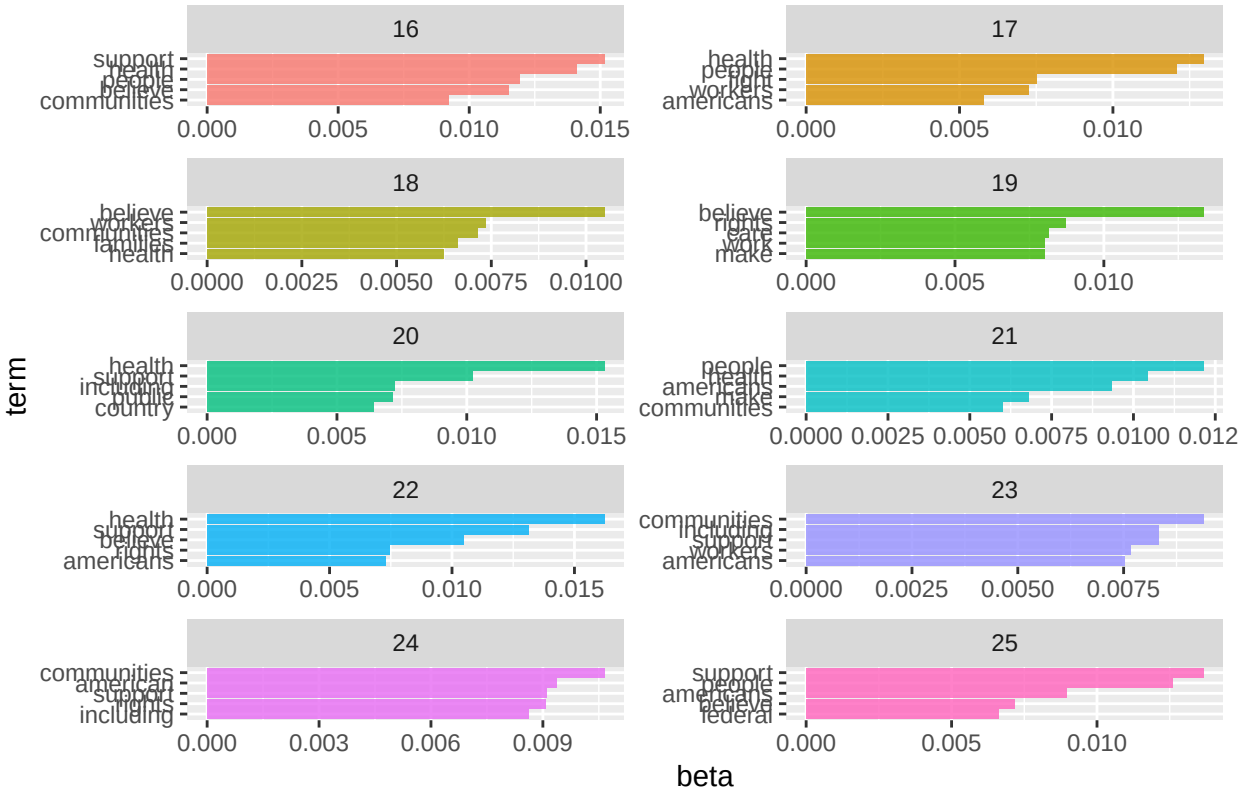

Topics Dem. Plataform. Part 2 (K=10)
6
health
believe
support
communities
americans
0.000
0.005
0.010
0.015
7
communities
support
including
families
believe
0.000
0.002
0.004
0.006
0.008
8
health
believe
fight
people
americans
0.000
0.005
0.010
0.015
9
believe
fight
work
people
health
0.000
0.005
0.010
10
make
american
public
new
continue
0.0000
0.0025
0.0050
0.0075
0.0100
term
beta

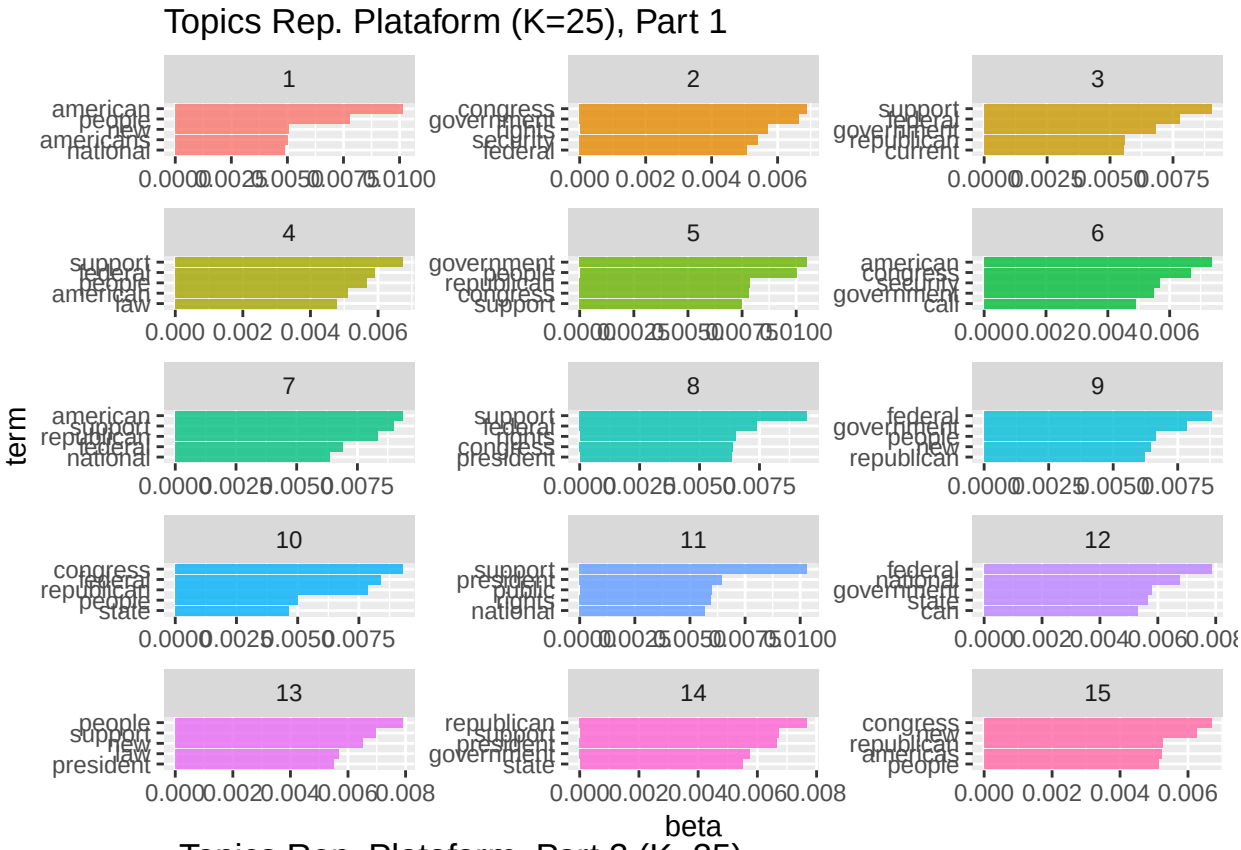

Topics Rep. Plataform, Part 2 (K=10)
1
american
people
new
can
national
0.000
0.003
0.006
0.009
2
government
rights
congress
national
american
0.000
0.002
0.004
0.006
3
support
state
government
federal
president
0.000
0.002
0.004
0.006
0.008
4
can
support
people
american
federal
0.000
0.002
0.004
0.006
5
government
people
support
republican
congress
0.000
0.003
0.006
0.009
term
beta

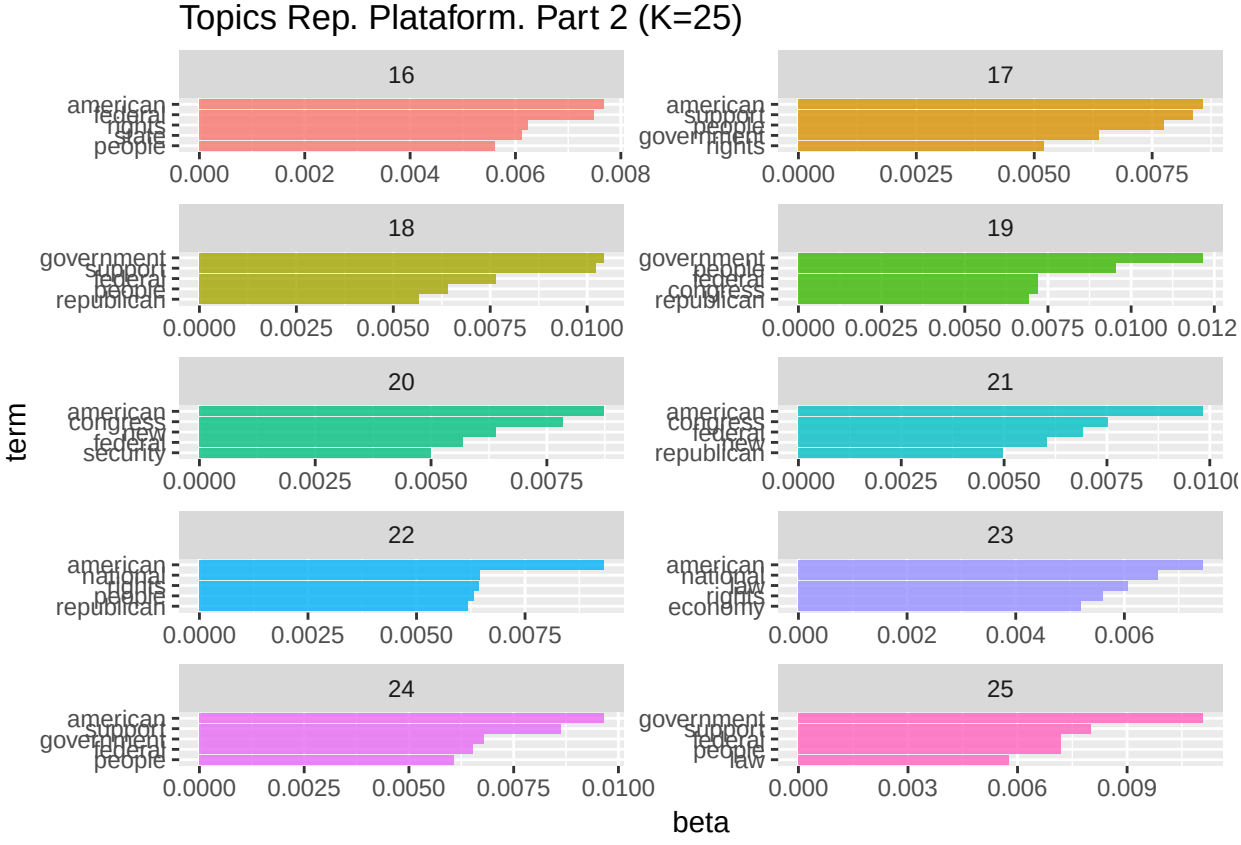

Topics Rep. Plataform, Part 2 (K=10)
6
american
call
government
security
congress
0.000
0.002
0.004
0.006
0.008
7
american
support
state
rights
national
0.0000
0.0025
0.0050
0.0075
0.0100
8
support
president
rights
national
federal
0.0000
0.0025
0.0050
0.0075
9
new
government
federal
people
republican
0.0000
0.0025
0.0050
0.0075
10
congress
federal
republican
state
people
0.000
0.002
0.004
0.006
term
beta

K=25


Topics Dem. Plataform (K=25), part 1
term
beta

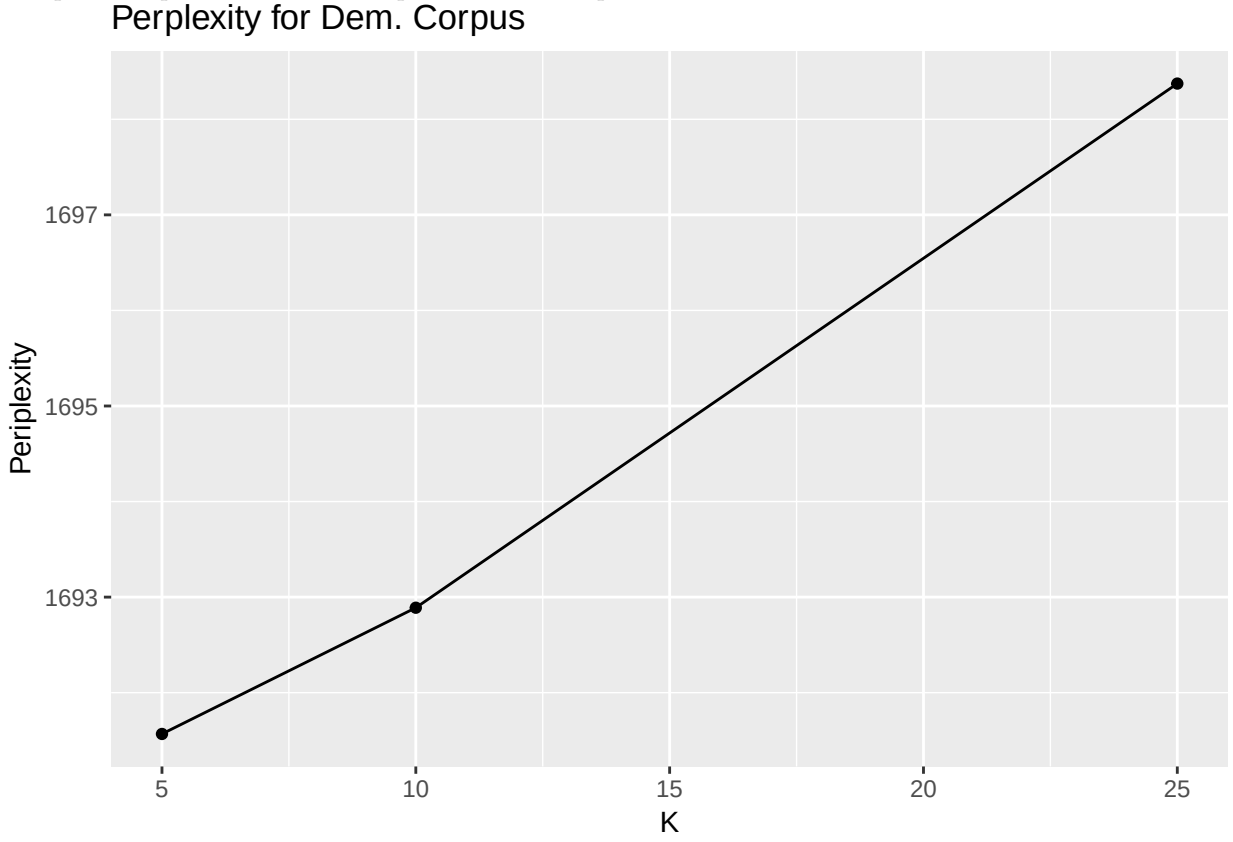

Topics Dem. Plataform (K=25), part 2
term
beta

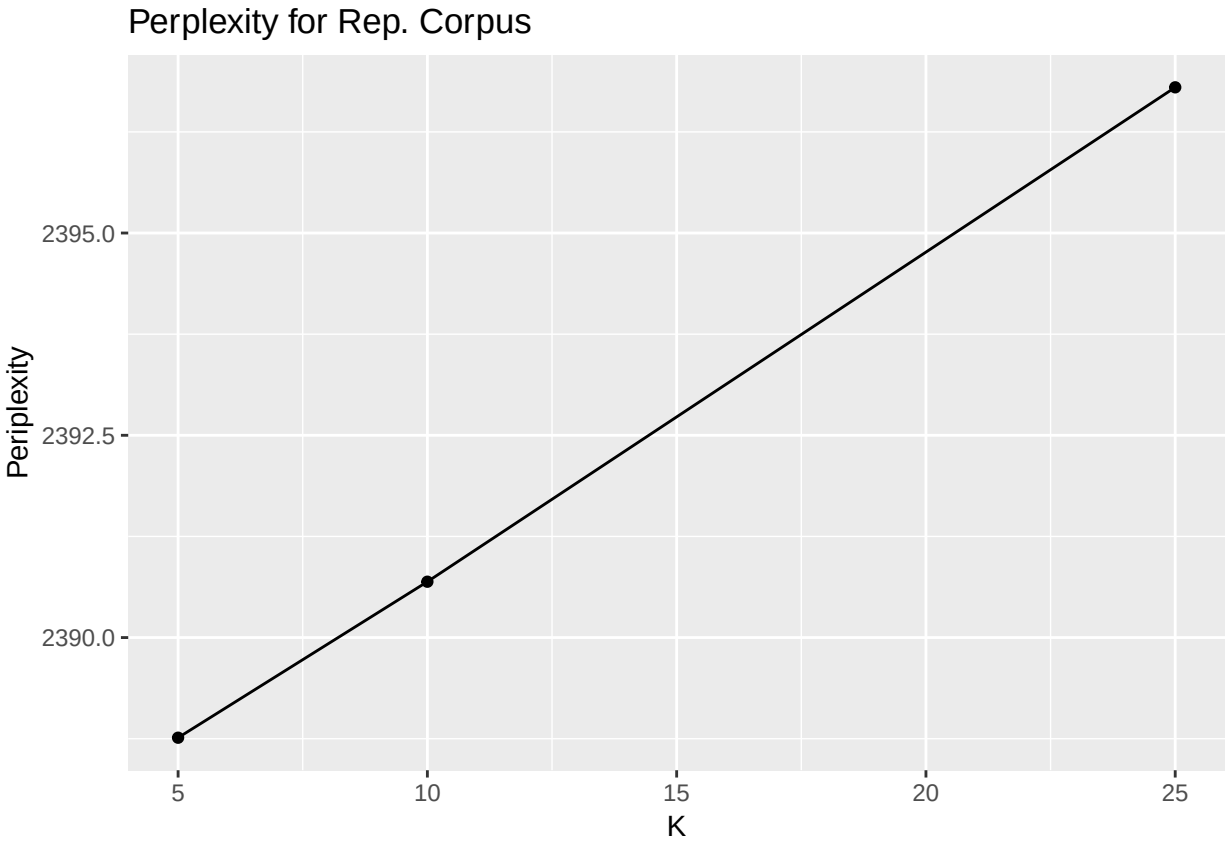

Topics Rep. Plataform (K=25), Part 1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
term
beta


Topics Rep. Plataform. Part 2 (K=25)
16
17
18
19
20
21
22
23
24
25
term
beta

Part 9

We can see that for any of both plataforms, the optimal K is 5. This is related to the issue that as we see in the previous parts that the concepts in each focal point that we had found are redundant as we increase k.





Part 10

As we see in the previous part, we can see that k=10 it's not the optimal number of K because the periplexity index minimize at k=5. When we made the graph of bars of the different focal points that we found is that at K=10, they tend to intersect to the concepts what we find in K=5, then form a efficeincy stand point use 10 k's don't made sense.

Part 11

In this case i would support the democrats, because the overall tone of that the government must protect the peoples and communities make sanse with my own valoric framwork. This is notice in the types of words that are selected in the different analysis that we made through this documents. For example, care, health, communitie between other are in coneection with my own personal belief. In other hand, the general tone of the Republican Corpus of protecting people from goverment is not correct in the current context. This is notorious when we see the sentiment analysis, al least from my perspective one explination of the more negativity of the democrats in the usage of high negative words is currect context of how the public affair are today, the diagnostic is grim and as is grim requires the usage of this type of words.

In other comments, i notice lately the word belive in the Democrat Corpus was used in a kind similar as Must in Republican, so i should had removed from the analysis.