Text as Data Problem Set 2

Ma Adelle Gia Arbo

17 October 2022

Contents

I. Research Question

The objective of this report is to answer the research question, "How did the intentions, motives, and views between Democrats and Republicans evolve in the United States using electoral manifestos in 2012, 2016, and 2020?" Specifically, this report wants to answer the following questions:

- 1. What are the prevailing topics from the United States electoral manifestos between the parties across the years?
- 2. What are the differences and similarities of the prevailing topics between the parties across the years?

These questions can be answered using topic modeling, an unsupervised machine learning method for classifying a set of documents, detecting words and phrases, and clustering word groups that best describe a set of documents. For the task at hand, two types of topic modeling methods will be implemented. First is the Latent Dirichlet Allocation (LDA), which treats each document as a mixture of topics and each topic as a mixture of words. Structural topic modeling (STM) will also be implemented which is similar to LDA but employs meta data about documents and allows correlations between topics. The results of each method will be compared to assess whether using one approach is better and more efficient and flexible than the other. Moreover, the discussion will include how specifying hyperparameters affect the findings.

With topic modeling, the prevailing and dominant topics in the manifestos between Democrats and Republicans can be observed over time. Furthermore, the differences and similarities between the parties' views will be identified.

II. Data Collection and Processing

The raw dataset on the United States manifestos in 2012, 2016, and 2018 is obtained from the WZB. The data collection and processing of the data will be discussed in this section. The final dataset consists a total of 12,121 documents between the Democratic Party and Republican Party across the last three electoral years.

```
# loading packages
if (!require("pacman")) install.packages("pacman")
pacman::p_load(manifestoR, quanteda, tidyr, purrr, ggplot2,
               tidytext, httr, rvest, readr, xml2, reshape2,
               stringr, stringi, dplyr, tibble, lexicon,
               NMF, topicmodels, LDAvis, stm)
```

In the code below, the manifestos of each party and year are downloaded and saved into a dataframe. For ease of use, the dataset from 2000 to 2020 is already saved in the data folder so as to avoid downloading the same data repeatedly.

```
# loop data collection

collect = FALSE

if (collect == TRUE) {
  mp_setapikey("manifesto_apikey.txt")

  mpds <- mp_maindataset()
  my_corpus <- mp_corpus(countryname == "United States" &
                           edate > as.Date("2000-11-07")) #change country year
  sample <- mpds %>%
    filter(countryname == "United States" &
             edate > as.Date("2000-11-07"))

  us_df <- tibble()

  for (i in names(my_corpus)) {
    doc <- my_corpus[[i]]
    doc_df <- as_tibble(as.data.frame(doc))
    doc_df$id <- i
    us_df <- rbind(us_df, doc_df)
  }
```

```
  us_df <- us_df %>%
    mutate(party = as.numeric(sub("_.*", "", id)),
           date = as.numeric(sub(".*_", "", id)))

  us_manifestos <- us_df %>%
  left_join(sample[,1:9], by = c("party"="party", "date"="date"))

  # save data
  write.csv(us_manifestos, "./data/us_manifestos.csv")
} else {
  us_manifestos <- read.csv("./data/us_manifestos.csv")
}
```

A total of 12,121 documents from the US manifestos for the last three electoral years will be processed and analyzed for topic modeling with 5750 documents for the Democratic Party and 6371 documents for the Republican Party. The frequency table is summarized below.

```
df <- us_manifestos %>%
  filter(edate >= as.Date("2012-11-06")) %>%
  filter(!is.na(text))

print(table(df$edate, df$partyname))
```

```
            Democratic Party Republican Party
  2012-11-06             1395              1793
  2016-11-08             1593              2289
  2020-11-03             2762              2289
```

```
print(table( df$partyname))
```

```
Democratic Party Republican Party
            5750              6371
```

III. Topic Modeling Results and Discussion

As mentioned above, the topic modeling task will involve two approaches: LDA and STM. Since there are two parties and three years in the dataset, it can be thought of that there are six sets of documents. For the pre-processing task, the document-feature matrix is created for each party and electoral year. Each dfmat is generated using the set of documents which are tokenized by removing English stopwords and punctuations, omitting obvious words like "united" and "state", stemming words, and setting the minimum term frequency to 10 (also a hyperparameter). The resulting six dfmats are appended into a single list, `dfmat_list`. This pre-processing task is implemented inside the loop of the code below.

1. Latent Dirichlet Allocation

The LDA model processes each dfmat in the `dfmat_list` in the loop. Each document-topics matrix (gamma) and topic-words matrix (beta) are converted into a dataframe and are tagged to the corresponding party and

year. Each converted dataframe is appended to the previous converted dataframe inside the loop to create a single dataframe called `doc_topics_df` and `topic_words_df`.

`doc_topics_df` is a dataframe in which every document for a particular party and year is a mixture of topics. Each document may contain words from several topics in particular proportions. For instance, if the number of topics (k) is set as 2, document 1 can be 80% topic1 and 20% topic2 while document 2 can be 40% topic1 and 60% topic2.

Meanwhile, `topic_words_df` is a dataframe in which every topic for a specific party and year is a mixture of words. For example, a two-topic model of US manifestos of the Republican Party in 2020 can have one topic about economic growth and another on healthcare. The most common words for the economic growth topic can be "GDP", "development", "inflation", while the healthcare topic can be "covid", "pandemic", and "health". It must be noted that words can be shared between topics in LDA.

```
# loop create dfmat and LDA by party and year

LDA = TRUE

if (LDA == TRUE) {
  dfmat_list <- list()
  doc_topics_df <- tibble()
  topic_words_df <- tibble()
  omit_words <- c("united", "state", "democrat", "republican",
                      "american", "america", "u.s")

  # create dfmat
  for (i in unique(df$edate)) {
    for (j in unique(df$partyname)) {
      df1 <- df %>%
      filter(partyname == j & edate == as.Date(i))

      dfmat <-  df1$text %>%
        tokens(remove_punct = T) %>%
        tokens_remove(pattern=stopwords("en")) %>%
        tokens_remove(omit_words) %>%
        tokens_wordstem() %>%
        dfm()  %>%
        dfm_trim(min_termfreq = 10)

      raw.sum=apply(dfmat,1,FUN=sum)
      dfmat=dfmat[raw.sum!=0,]

      print(sprintf("Created (%s, %s) dfmat using %s manifestos of the %s",
                    dim(dfmat)[1], dim(dfmat)[2], format(as.Date(i), format = "%Y"), j))

      dfmat_list <- append(dfmat_list, dfmat)


      # LDA model
      print(sprintf("Starting LDA model using %s manifestos of the %s",
                    format(as.Date(i), format = "%Y"), j))

      lda <- LDA(dfmat, control=list(seed=28), k=10) # change num of topics

      W <- lda@gamma # document-topic
```

```
      H <- lda@beta # topic-term

      doc_topics <- tidy(lda, matrix="gamma") %>%
        mutate(date = i, partyname = j)
      doc_topics_df <- rbind(doc_topics_df, doc_topics)


      topic_words <- tidy(lda, matrix="beta") %>%
        mutate(date = i, partyname = j)
      topic_words_df <- rbind(topic_words_df, topic_words)

      print(sprintf("Finished LDA model using %s manifestos of the %s",
                    format(as.Date(i), format = "%Y"), j))
    }
  }
  write.csv(doc_topics_df, "./data/doc_topics.csv")
  write.csv(topic_words_df, "./data/topic_words.csv")
} else {
  doc_topics_df <- read.csv("./data/doc_topics.csv")
  topic_words_df <- read.csv("./data/topic_words.csv")
}
```

```
[1] "Created (1386, 354) dfmat using 2012 manifestos of the Democratic Party"
[1] "Starting LDA model using 2012 manifestos of the Democratic Party"
[1] "Finished LDA model using 2012 manifestos of the Democratic Party"
[1] "Created (1760, 444) dfmat using 2012 manifestos of the Republican Party"
[1] "Starting LDA model using 2012 manifestos of the Republican Party"
[1] "Finished LDA model using 2012 manifestos of the Republican Party"
[1] "Created (1562, 384) dfmat using 2016 manifestos of the Democratic Party"
[1] "Starting LDA model using 2016 manifestos of the Democratic Party"
[1] "Finished LDA model using 2016 manifestos of the Democratic Party"
[1] "Created (2238, 516) dfmat using 2016 manifestos of the Republican Party"
[1] "Starting LDA model using 2016 manifestos of the Republican Party"
[1] "Finished LDA model using 2016 manifestos of the Republican Party"
[1] "Created (2721, 599) dfmat using 2020 manifestos of the Democratic Party"
[1] "Starting LDA model using 2020 manifestos of the Democratic Party"
[1] "Finished LDA model using 2020 manifestos of the Democratic Party"
[1] "Created (2238, 516) dfmat using 2020 manifestos of the Republican Party"
[1] "Starting LDA model using 2020 manifestos of the Republican Party"
[1] "Finished LDA model using 2020 manifestos of the Republican Party"
```

In topic modeling, the number of topics is considered as a hyperparameter as it affects the resulting topics. A hyperparameter is a parameter that is external to the model and is used to control the learning process. For example, a higher number of topics allows the identification of a more diverse and specific set of topics. In contrast, by decreasing this hyperparameter, the model can identify more general topics given the set of documents. In the LDA model below, the number of topics is arbitrarily set to 10. It must also be noted that the LDA model returns different results unless the `contol` is fixed by setting a similar seed every time the code is run.

After investigating the document-topics matrix, a 10-topic model identifies that each topic account for more or less than 10% per document. For example, the table below shows the share of each identified topic from the 2012 Democrats manifesto. Note that each manifesto of a party in a given electoral year is fed into the LDA model separately.

```
doc_topics_df %>%
  arrange(document, date, partyname) %>%
  mutate(topic_share = round(gamma,3)) %>%
  select(date, partyname, document, topic, topic_share) %>%
  head(10)
```

```
# A tibble: 10 x 5
   date       partyname        document topic topic_share
   <chr>      <chr>            <chr>    <int>       <dbl>
 1 2012-11-06 Democratic Party text1        1       0.099
 2 2012-11-06 Democratic Party text1        2       0.099
 3 2012-11-06 Democratic Party text1        3       0.1
 4 2012-11-06 Democratic Party text1        4       0.102
 5 2012-11-06 Democratic Party text1        5       0.1
 6 2012-11-06 Democratic Party text1        6       0.1
 7 2012-11-06 Democratic Party text1        7       0.1
 8 2012-11-06 Democratic Party text1        8       0.1
 9 2012-11-06 Democratic Party text1        9       0.099
10 2012-11-06 Democratic Party text1       10       0.1
```

Now, to investigate the underlying topics in each party's manifestos in the three recent electoral years, the results are plotted showing the top 10 terms with the highest per-topic-per-word probabilities in the topic. The `topic_words_df` is first tidied for plotting with `ggplot`.

```
topic_words_df1 <- topic_words_df %>%
  mutate(year = format(as.Date(date), format = "%Y")) %>%
  mutate(party = ifelse(partyname == "Republican Party", "Republicans", "Democrats")) %>%
  mutate(group = paste0(year, " ", party),
         color = ifelse(party == "Democrats", "#2c7fb8", "#de2d26")) %>%
  group_by(topic, date, year, partyname) %>%
  slice_max(beta, n = 10) %>%
  ungroup() %>%
  arrange(date, topic, partyname, -beta)
```

For the interpretation of results, it must be reiterated that LDA topic modeling is performed for each of the 6 sets of documents or manifesto per party and year. This means that topic 1 revealed from the 2012 Democrats' manifesto is unique to the "topic 1" identified from 2012 Republicans' manifesto and to "topic 1" from 2016 Democrats' manifesto and so on. Thus, each bar chart of top terms of topics per party and year will be interpreted separately, and the underlying topics will be compared later.

This section will now discuss the prevailing topics of the manifestos between the Democratic Party and the Republican Party for the electoral years 2012, 2016, and 2020.

2012: Obama vs. Romney In 2012, Democrats' manifesto appear to show Obama's focus on health care in topic 3, which was coined as "Obamacare" during Obama's term. Topic 7 seems to be about the tax scheme concerning middle class. Topic 9 pertains to the US economy. After suffering from the effects of recession which displaced many due to the housing bubble and mortgage crisis, the Democratic party during the 2012 elections espouses a comprehensive social protection program. It aims to expand health care (topic 3), provide jobs (topic 8), and generate businesses towards economic recovery (topic 9).

On the other hand, the Republican party doubles down on effective governance wherein they stress the importance of federalism, which is identified in topic 1 and 3. This is enshrined in the US Constitution supporting devolution of governance at the state-level and small government at the federal level. Topic 5 tackles education with words like "support", "school", "right", and "need".

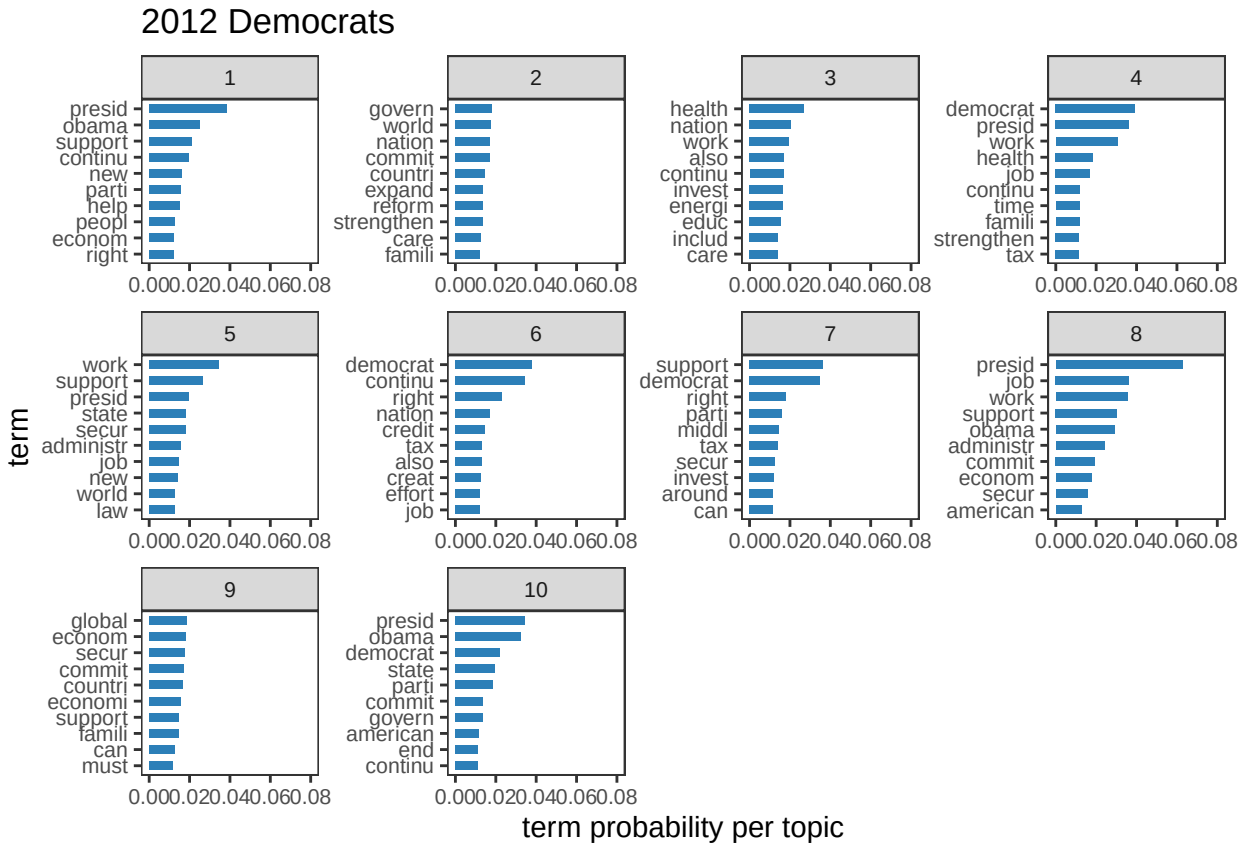
2016: Clinton vs. Trump As America recovered and continued to progress from the recession, the Democrats during the 2016 elections acknowledged that such growth cannot be sustained if it is not shared equitably. In the plot, the prevailing topics from Democrats' manifesto appear to focus on health policies as well as seen in topic 2 and 5. Major policy reforms sought for the protection of the vulnerable sectors of society such as women and workers as seen in topic 3 with words like "support", "work", and "protect". Another topic is revealed in 8 regarding education with words like "student" and "school" on promoting student support. The same can be said for topic 6 on educational support.

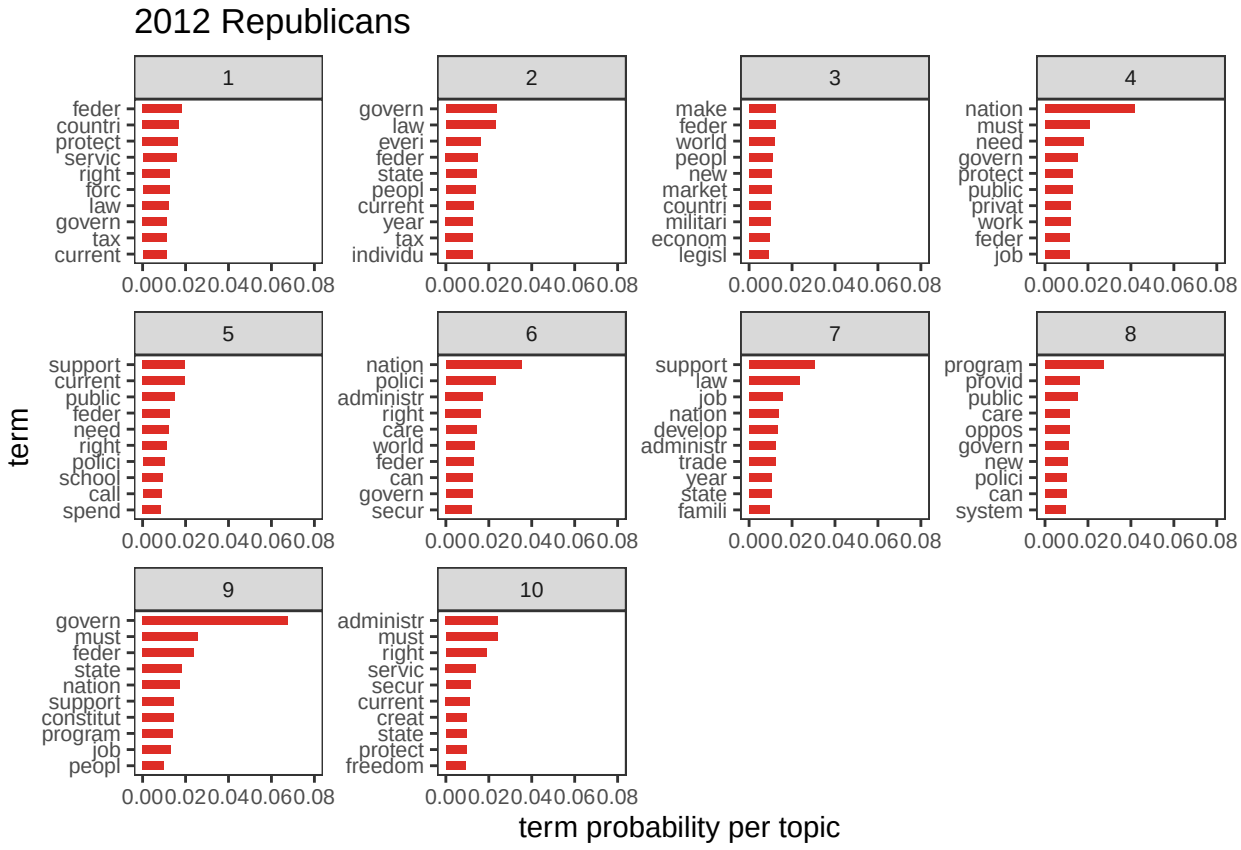
The Republican party on the other hand targeted their policies that hit much closer to home as the party revealed the vulnerabilities of the last administration wherein although recovered from the recession, American families continue to be deprived of services, support, and opportunities. This is embodied in topic 1 with words like "protect", "ensure", and "support". Similar to Democrats, topic 4 is about education rights with "educ", "children", and "famili". Republicans are calling for government to take boost and support law enforcement authorities so they protect American families which calls for more policing and protection. Furthermore, like in 2012, federalism is still an underlying topic in their manifesto as revealed in topic 7 and 10.

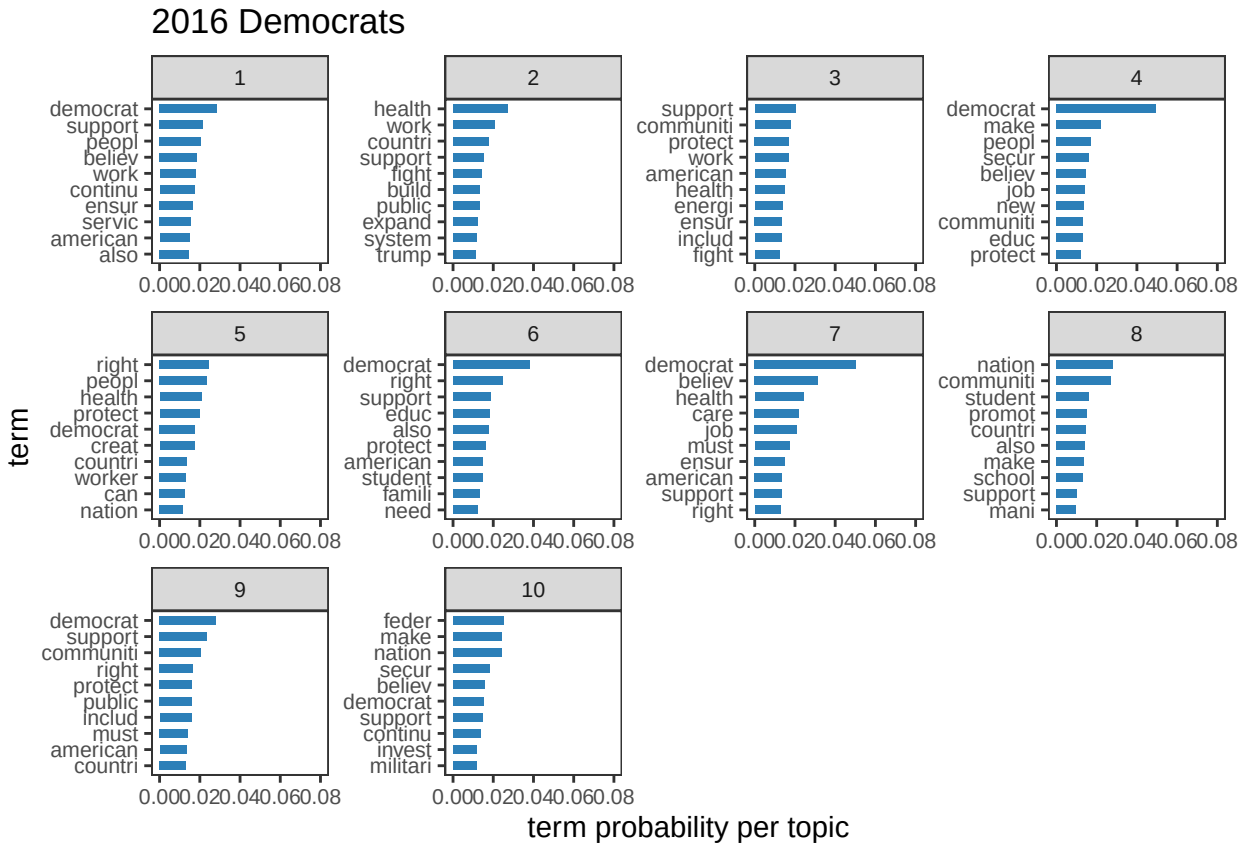
2020: Biden vs. Trump As expected, most topics revolve around health due to COVID-19 in the Democrats' manifesto as in topic 8 and 10. Topic 10 is specifically about health workers. Topic 5 is about access to education which can be inferred from words "support", "access", "fund", "increase", and "student". Topic 9 conveys a similar topic but can be more particular on back-to-school policies amid the health crisis. For the first time, the term "women" and "right" appeared in topic 7 which is identified to be on women rights.

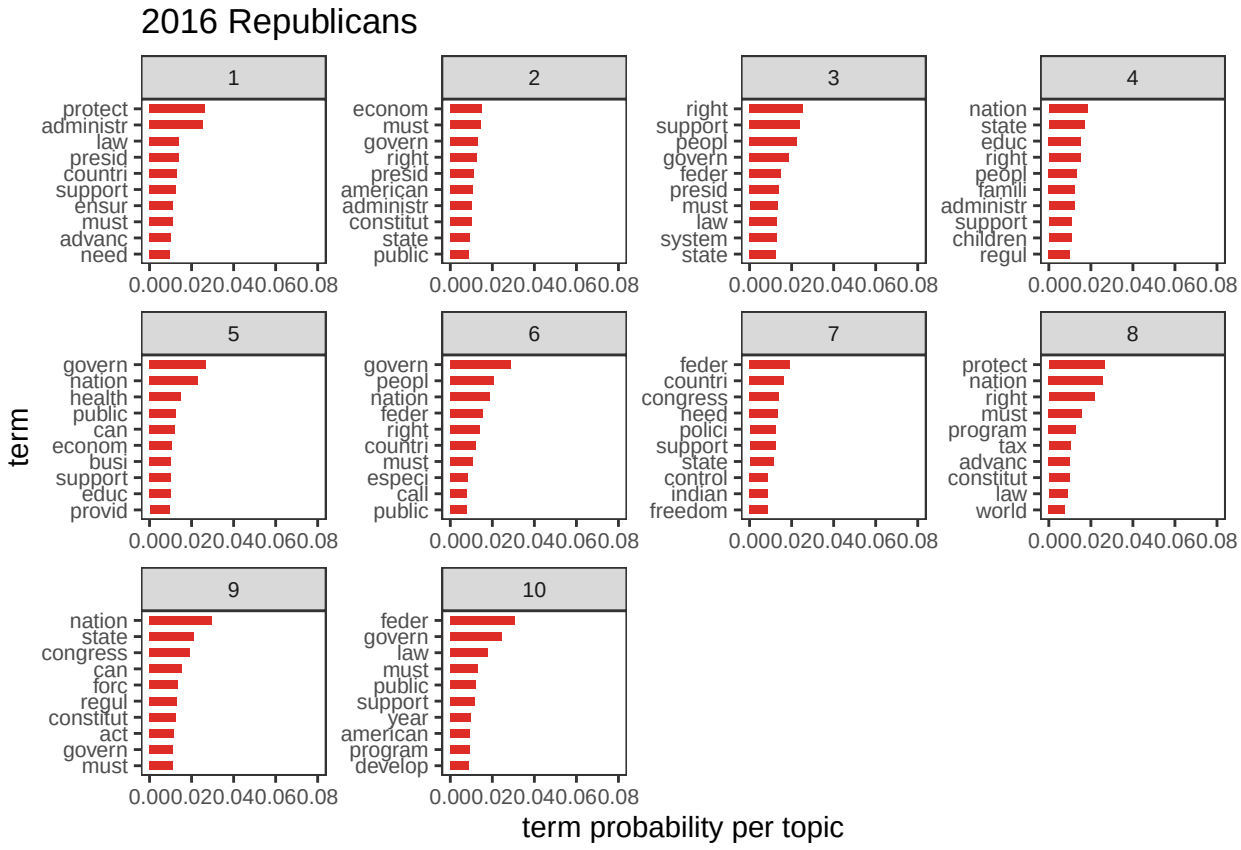
Similarly, Republicans focused on education (topic 4) as well as public health (topic 5). Furthermore, federalism is again a topic in in topic 7 and 10.

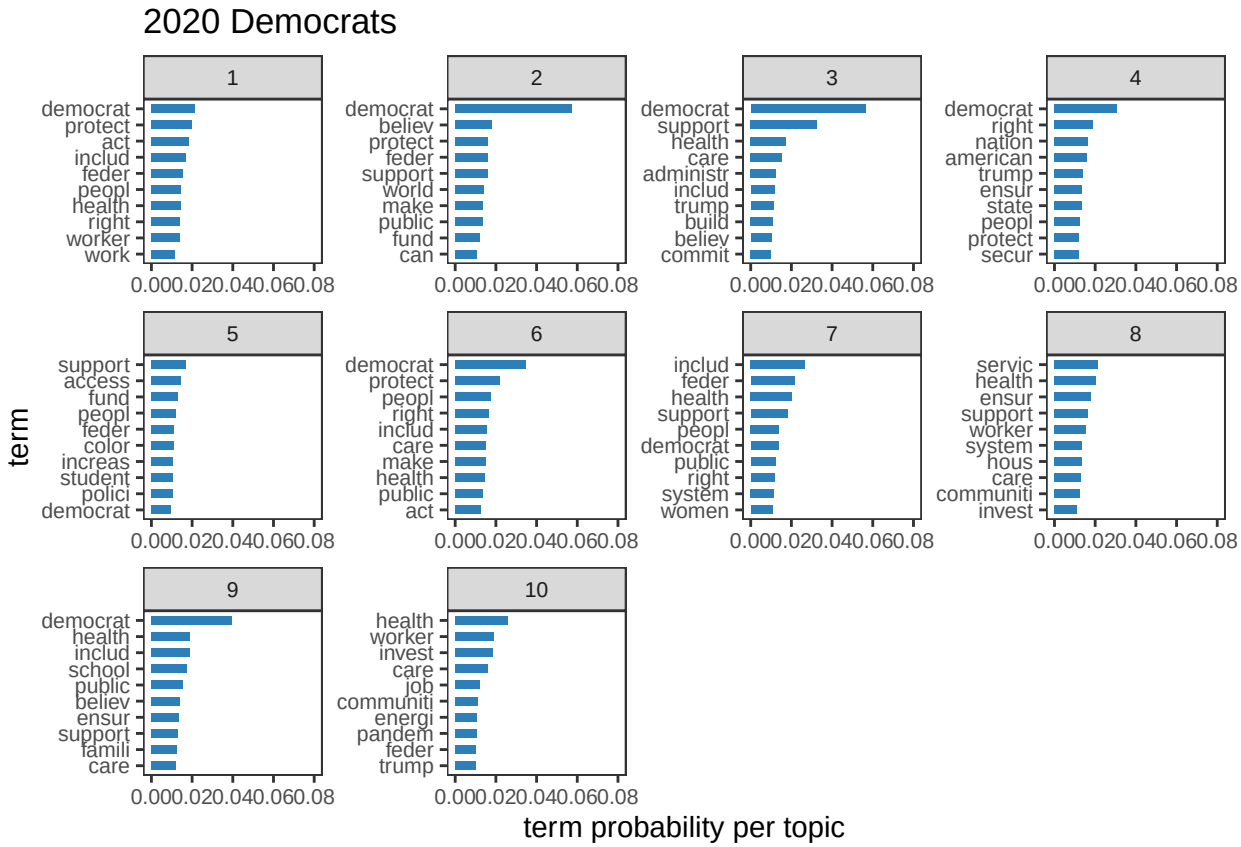
```
for (i in unique(topic_words_df1$group)){
    p<-topic_words_df1 %>%
      filter(group == i) %>%
      mutate(term = reorder_within(term, beta, topic)) %>%
      ggplot(aes(beta, term, fill = color)) +
      geom_bar(stat="identity", width = .5) +
      facet_wrap(~ topic, scales = "free", ncol=4) +
      xlim(0, 0.08) +
      scale_y_reordered() +
      labs(title = i) +
      scale_fill_identity() +
      theme_test() +
      theme(axis.text.y = element_text(size = 8),
            axis.text.x = element_text(size = 8),
            strip.text.x = element_text(size = 8)) +
    labs(y = "term", x = "term probability per topic")
    print(p)
    ggsave(paste0("./plots/lda_",i, ".png"))
}
```

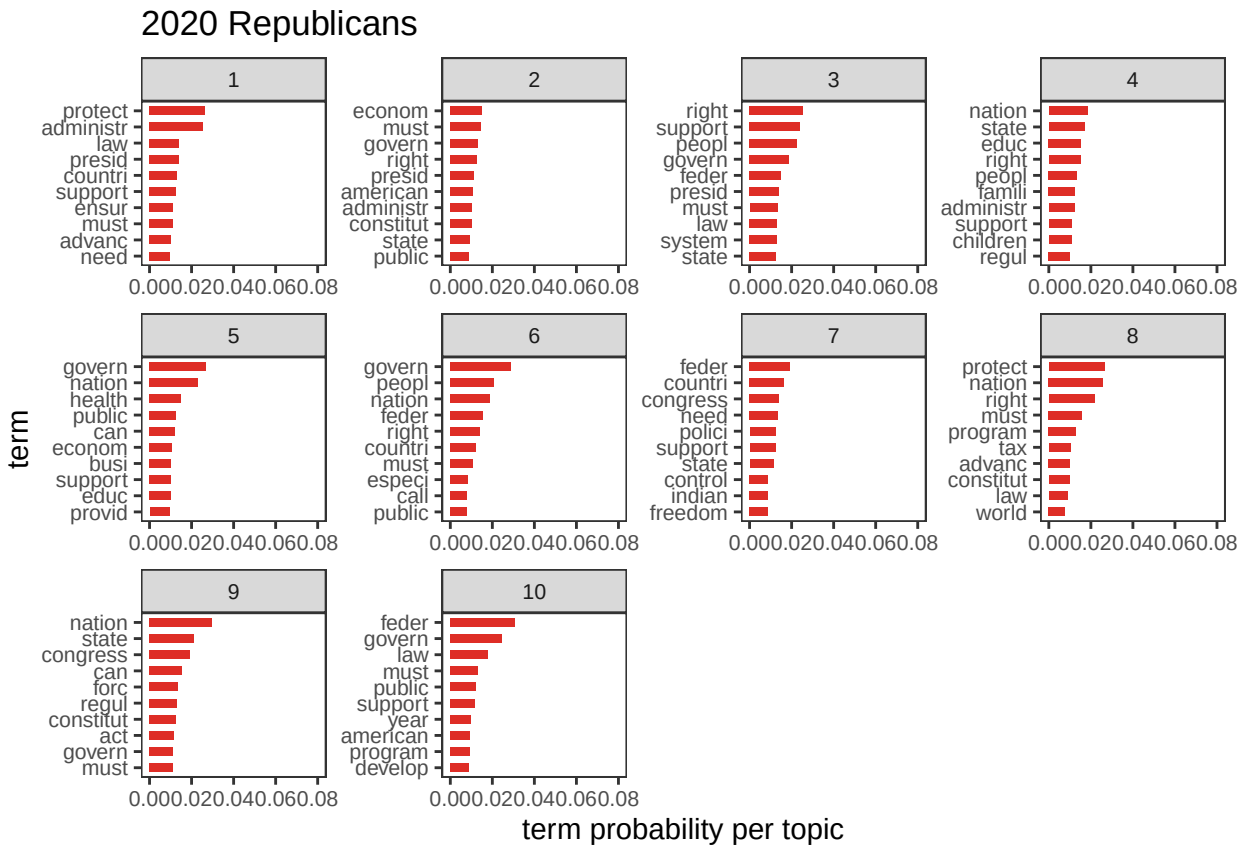

2012 Democrats
1
presid
obama
support
continu
new
parti
help
peopl
econom
right
2
govern
world
nation
commit
countri
expand
reform
strengthen
care
famili
3
health
nation
work
also
continu
invest
energi
educ
includ
care
4
democrat
presid
work
health
job
continu
time
famili
strengthen
tax
5
work
support
presid
state
secur
administr
job
new
world
law
6
democrat
continu
right
nation
credit
tax
also
creat
effort
job
7
support
democrat
right
parti
middl
tax
secur
invest
around
can
8
presid
job
work
support
obama
administr
commit
econom
secur
american
9
global
econom
secur
commit
countri
economi
support
famili
can
must
10
presid
obama
democrat
state
parti
commit
govern
american
end
continu
0.00 0.02 0.04 0.06 0.08
term
term probability per topic


2012 Republicans
1
feder
countri
protect
servic
right
forc
law
govern
tax
current
2
govern
law
everi
feder
state
peopl
current
year
tax
individu
3
make
feder
world
peopl
new
market
countri
militari
econom
legisl
4
nation
must
need
govern
protect
public
privat
work
feder
job
5
support
current
public
feder
need
right
polici
school
call
spend
6
nation
polici
administr
right
care
world
feder
can
govern
secur
7
support
law
job
nation
develop
administr
trade
year
state
famili
8
program
provid
public
care
oppos
govern
new
polici
can
system
9
govern
must
feder
state
nation
support
constitut
program
job
peopl
10
administr
must
right
servic
secur
current
creat
state
protect
freedom
0.00 0.02 0.04 0.06 0.08
term
term probability per topic


2016 Democrats
1
democrat
support
peopl
believ
work
continu
ensur
servic
american
also
2
health
work
countri
support
fight
build
public
expand
system
trump
3
support
communiti
protect
work
american
health
energi
ensur
includ
fight
4
democrat
make
peopl
secur
believ
job
new
communiti
educ
protect
5
right
peopl
health
protect
democrat
creat
countri
worker
can
nation
6
democrat
right
support
educ
also
protect
american
student
famili
need
7
democrat
believ
health
care
job
must
ensur
american
support
right
8
nation
communiti
student
promot
countri
also
make
school
support
mani
9
democrat
support
communiti
right
protect
public
includ
must
american
countri
10
feder
make
nation
secur
believ
democrat
support
continu
invest
militari
0.00 0.02 0.04 0.06 0.08
term
term probability per topic


2016 Republicans
1
protect
administr
law
presid
countri
support
ensur
must
advanc
need
2
econom
must
govern
right
presid
american
administr
constitut
state
public
3
right
support
peopl
govern
feder
presid
must
law
system
state
4
nation
state
educ
right
peopl
famili
administr
support
children
regul
5
govern
nation
health
public
can
econom
busi
support
educ
provid
6
govern
peopl
nation
feder
right
countri
must
especi
call
public
7
feder
countri
congress
need
polici
support
state
control
indian
freedom
8
protect
nation
right
must
program
tax
advanc
constitut
law
world
9
nation
state
congress
can
forc
regul
constitut
act
govern
must
10
feder
govern
law
must
public
support
year
american
program
develop
0.00
0.02
0.04
0.06
0.08
term
term probability per topic


2020 Democrats
1
democrat
protect
act
includ
feder
peopl
health
right
worker
work
2
democrat
believ
protect
feder
support
world
make
public
fund
can
3
democrat
support
health
care
administr
includ
trump
build
believ
commit
4
democrat
right
nation
american
trump
ensur
state
peopl
protect
secur
5
support
access
fund
peopl
feder
color
increas
student
polici
democrat
6
democrat
protect
peopl
right
includ
care
make
health
public
act
7
includ
feder
health
support
peopl
democrat
public
right
system
women
8
servic
health
ensur
support
worker
system
hous
care
communiti
invest
9
democrat
health
includ
school
public
believ
ensur
support
famili
care
10
health
worker
invest
care
job
communiti
energi
pandem
feder
trump
0.00 0.02 0.04 0.06 0.08
term
term probability per topic



2. Structural Topic Modeling

In this section, structural topic model will be implemented using the `stm` package. STM offers more flexibility in the analysis such as explicitly modeling which variables influence the prevalence of topics. It also allows for topic correlation which will be shown later.

To start STM, each document-feature matrix which were created in the code above is converted to a `stm` format then STM is applied to identify 15 topics in the loop. For this task, the number of topics is arbitrarily chosen to be 15 since decreasing the number of topics results to longer time for the model to converge. The results of the model for each set of documents is saved as a list in `model_list`, so that it can easily be called later on.

Similar to the task that was done in LDA, each set of documents or manifesto of a party in a given electoral year is fed into STM, giving six separate results. In STM, the hyperparameter number of topics is set to 15. To visualize these results, a built-in function of the `stm` package allows for topic exploration to analyze the topic-words matrix. With `plot.STM`, the topic distribution is shown together with the most common words for each topic in one plot.

This is basically similar to the topic exploration that was done with LDA but the plots are easier to interpret as it is arranged by the expected topic proportions and with the top common words already provided for each topic.

In the next discussion, the results will be interpreted based on the top 3 topics of each manifesto of a party in an electoral year with the three most common words in that topic. (See Appendix 1 for the plots showing the top 5 most common word for all 15 topic per manifesto.)

2012: Obama vs. Romney Starting with the Democrats' manifesto, the three main topics relate to proposed institutional reforms (topic 6), job creation (topic 9), and healthcare (topic 8). This result is quite similar to what was found in LDA. Meanwhile, Republicans focus on topics relating to federalism (topic 14) with words like "constitution" and "power" like in LDA. One topic that first appeared here is on tech regulation (topic 2).

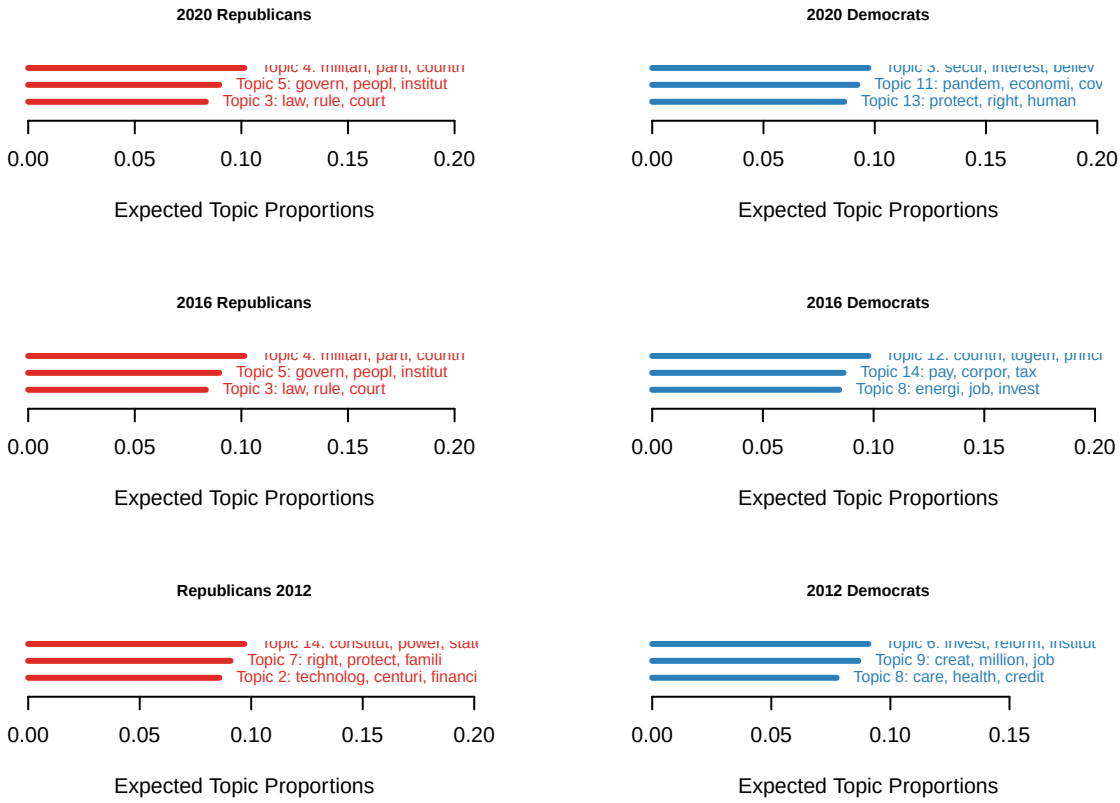
2016: Clinton vs. Trump In 2016, for the Democratic party, topic 14 relates to tax policies (topic 14) and job creation (topic 8). For the Republican party, topics again are about federalism and legal framework (topic 4, 5, 3).

2020: Biden vs. Trump Finally, in 2020, the Democratic Party tackled health policies to combat covid-19 pandemic (topic 11) as well as topics on human rights (topic 13). On the other hand, the `stm` results identified similar topics on federalism and law as in the 2016 manifesto for the Republican party.

Women rights

Another relevant topic of the results of STM is on women rights. Both women violence appeared as a topic in 2012 Democrats' and Republicans' manifestos. It is also again a prevailing topic by Democrats in 2016 and 2020 and is associated with health and safety and gun violence, respectively. (See Appendix 2)

```
par(mfrow=c(3,2))
par(bty="n",col="#de2d26",lwd=3)
plot.STM(model_list[[6]],type="summary", label="frex", n = 3, main = "2020 Republicans",
     topics=c(4,5,3), text.cex=.8, width=30, cex.main =.8)
par(bty="n",col="#2c7fb8",lwd=3)
plot.STM(model_list[[5]],type="summary", label="frex", n = 3, main = "2020 Democrats",
     topics=c(3, 11, 13), text.cex=.8, width=30, cex.main =.8)
par(bty="n",col="#de2d26",lwd=3)
plot.STM(model_list[[4]],type="summary", label="frex", n = 3, main = "2016 Republicans",
     topics=c(4,5,3), text.cex=.8, width=30, cex.main =.8)
par(bty="n",col="#2c7fb8",lwd=3)
plot.STM(model_list[[3]],type="summary", label="frex", n = 3, main = "2016 Democrats",
     topics=c(12,14,8), text.cex=.8, width=30, cex.main =.8)
par(bty="n",col="#de2d26",lwd=3)
plot.STM(model_list[[2]],type="summary", label="frex", n = 3, main = "Republicans 2012",
    topics=c(14,7,2), text.cex=.8, width=30, cex.main =.8)
par(bty="n",col="#2c7fb8",lwd=3)
plot.STM(model_list[[1]],type="summary", label="frex", n = 3, main = "2012 Democrats",
     topics=c(6,9,8), text.cex=.8, width=30, cex.main =.8)
```



IV. Conclusion

The findings show that the Democratic Party's manifestos have focused on public health policies even prior to the COVID-19 pandemic. Meanwhile, the Republican Party have constantly mentioned topics relating to federalism in their manifestos. Throughout the years, both parties have campaigned for educational rights, access to school, and student support. The relevant topic on human rights, gun violence, and women safety are also identified in both parties' manifestos.

Moreover, the results between LDA and STM have shown to be consistent. For example, both approach identified topics on healthcare, job creation, and reforms for the Democratic Party in 2012. In addition, the topic on federalism was also revealed in both topic modeling methods for the Republican Party. Both approach were able to identify clear topics but using STM showed more flexibility and efficiency in the analysis.

With topic modeling, this report was able to answer the research question on how the intentions, motives, and views between Democrats and Republicans have evolved in the United States using electoral manifestos in 2012, 2016, and 2020. A more detailed discussion on the topic modeling pipeline, hyperparameter tuning, and research recommendation is discussed in the next section.

V. Research Recommendation

Modify data processing pipeline

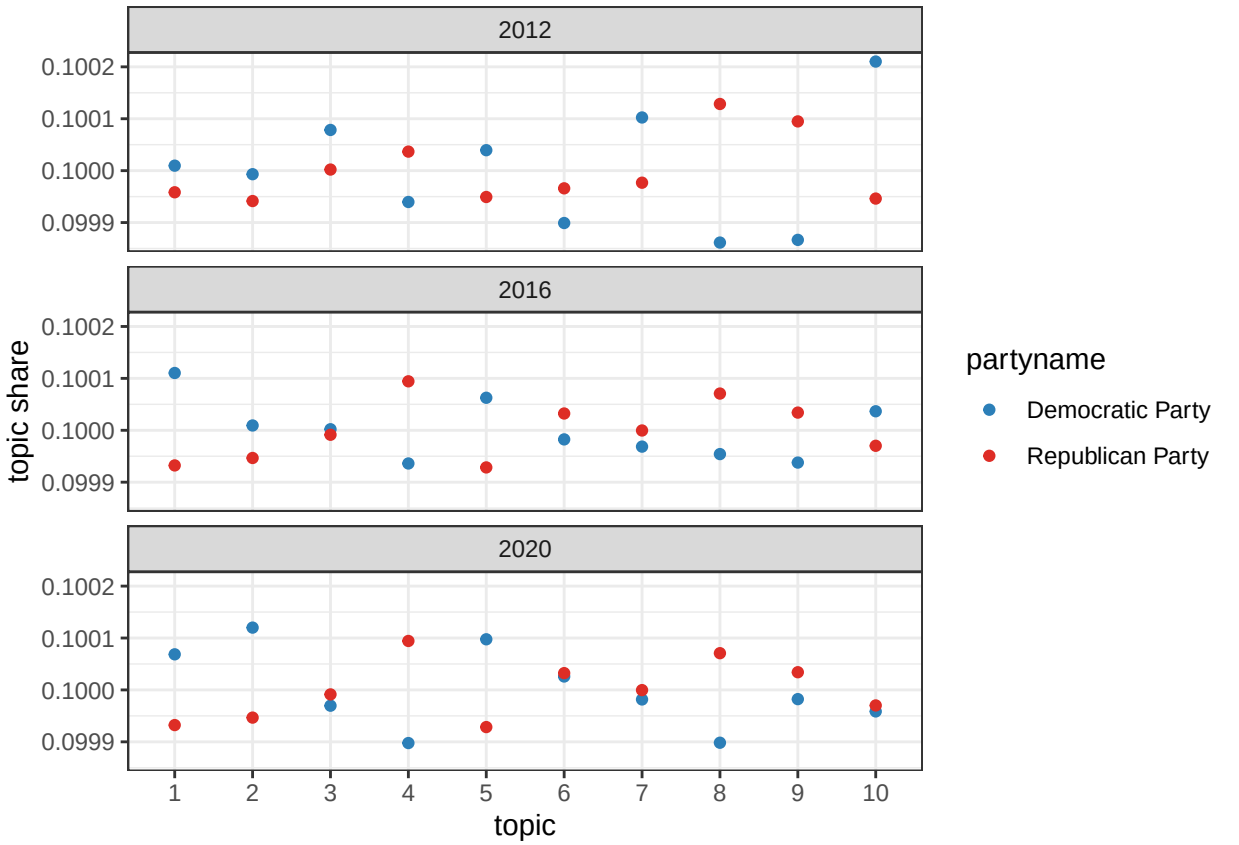
One main recommendation to answer the research questions is to reconsider how the data was processed and used for topic modeling. As the research question deals with a time series type of data for three electoral

years, he total 12,121 documents can be fed into the topic model as a whole with k =10, instead of feeding each of the 6 sets of documents or manifesto per party and year separately. The results of the single approach can be merged with the original dataset to tag the document-topics matrix with the corresponding party and year.

With this approach, the share of topics can be plotted between Democrats and Republicans over time as shown in the plot below.

```
df2 <- df %>%
  mutate(document = paste0("text",1:nrow(df))) %>%
  filter(!is.na(text))

dfmat2 <-  df2$text %>%
        tokens(remove_punct = T) %>%
        tokens_remove(pattern=stopwords("en")) %>%
        tokens_remove(omit_words) %>%
        tokens_wordstem() %>%
        dfm()  %>%
        dfm_trim(min_termfreq = 10)
raw.sum=apply(dfmat2,1,FUN=sum)
dfmat2=dfmat2[raw.sum!=0,]

lda2 <- LDA(dfmat2, control=list(seed=28), k=10)
```

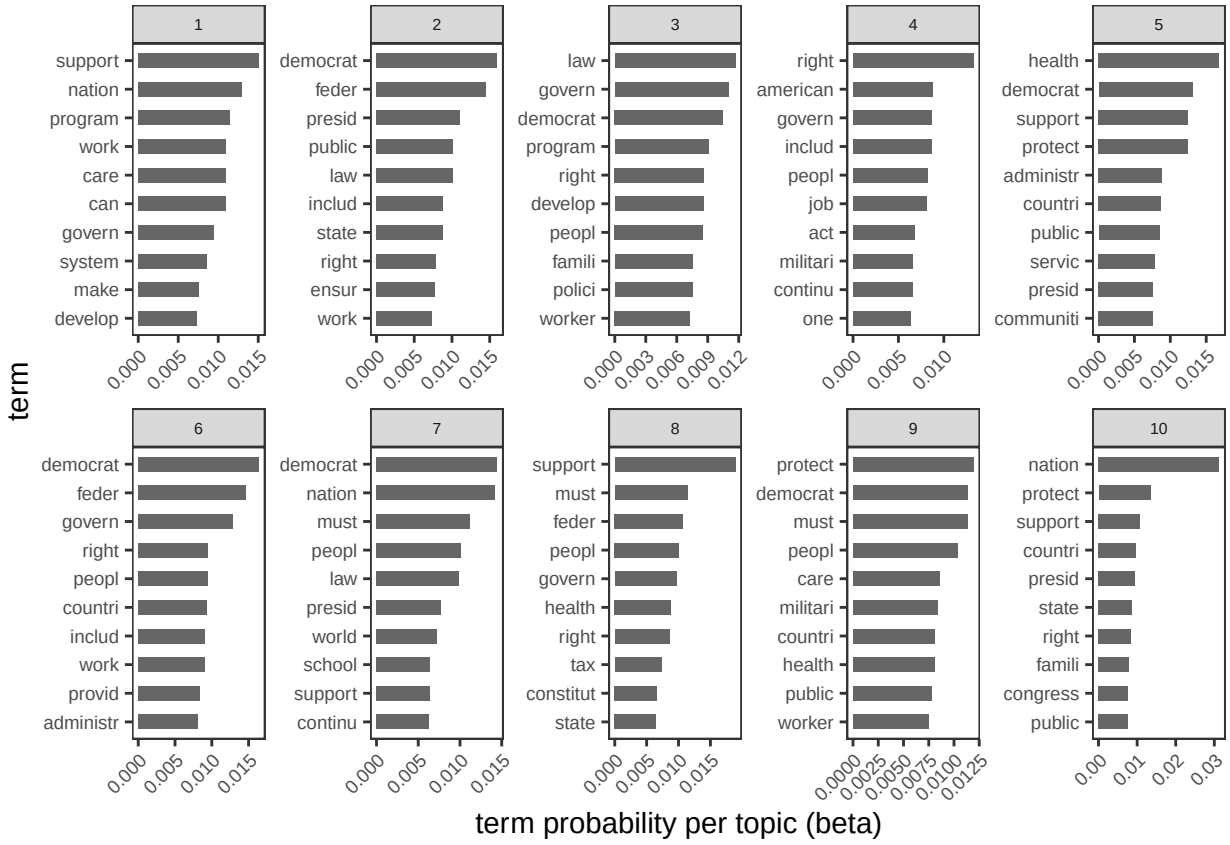
```
doc_topics2 <- tidy(lda2, matrix="gamma") %>%
  left_join(df2, by="document") %>%
  select(c(document, topic, gamma, partyname, date, edate)) %>%
  mutate(year = format(as.Date(edate), format = "%Y"))

yp_topics <- doc_topics2 %>%
  group_by(year, partyname, topic) %>%
  summarise(gamma = sum(gamma)) %>%
  group_by(year, partyname,) %>%
  mutate(year_share = gamma/sum(gamma)) %>%
  ungroup() %>%
  mutate(topic = factor(topic))

yp_topics %>%
  ggplot(aes(x=topic, y = year_share, group = partyname, color = partyname)) +
  geom_point() +
  scale_color_manual(values = c("#2c7fb8", "#de2d26")) +
  theme_bw() +
  facet_wrap(~year, ncol=1) +
  labs(x = "topic", y = "topic share")
```




```
ggsave(paste0("./plots/lda_year_party.png"))
```

Moreover, since the 10 topics are consistent across all manifestos, the prevalent topics can be easily identified using the single topic-words matrix from all manifestos. In this way, the same topics can be plotted across time and can be compared between parties. The plot above can then be easily associated with the topics identified in the plot below.

```
topic_words2 <- tidy(lda2, matrix="beta") %>%
  group_by(topic) %>%
  slice_max(beta, n = 10) %>%
  ungroup() %>%
  arrange(topic, -beta)

topic_words2 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = "grey40")) +
  geom_bar(stat="identity", width = .5) +
  facet_wrap(~ topic, scales = "free", ncol=5) +
  scale_y_reordered() +
  scale_fill_identity() +
  theme_test() +
  labs(y = "term", x = "term probability per topic (beta)") +
  theme(axis.text.y = element_text(size = 7),
        axis.text.x = element_text(size = 7),
        strip.text.x = element_text(size = 6)) +
  scale_x_continuous(guide = guide_axis(angle = 45))
```




```
ggsave(paste0("./plots/lda_year_party_topics.png"))
```

Although the approach described makes the analysis and data visualization much simpler, the pipeline adopted in this report also has its own advantage as it already assumes that the time and party variables have an (correlational) effect on the topics in each manifesto. This makes the resulting set of topics revealed independent per set of documents.

Furthermore, extending the number timeframe of the dataset can reveal and understand the evolution of topics further. Initially, five electoral years were considered for analysis but unfortunately, the data for the Democratic Party in 2008 only contained one document. Due to this, the analysis was limited on the three recent electoral years but this can be resolved by processing the data so that each row will consist of each line in the manifesto.

Hyperparameter tuning

Another recommendation is to use hyperparameter tuning to best identify the number number of topics, k, for topic modeling. For this report, k is only arbitrarilty chosen, but for future reserach the number of topics should be decided based on statistical fit and interpretability of topics. K determines the values of gamma and beta wherein low values of gamma indicate documents should be composed of few topics and low values of beta indicate topics should be composed of few terms.

Further topic exploration

Finally, as discussed above, exploring STM and the functions of the `stm` package such as computing the correlations betweeen topics, topic estimates of document-topic proportions, and examining the exclusivity vs. semantic coherence (See Appendix 2).

VI. Appendix

1. STM functions

The `stm` library also offers more builtin functions than `lda`. For example, the top words of the 2012 Democratic Party's manifesto is printed below by the highest conditional probability for each topic. FREX shows which words are comparatively common for a topic and exclusive for that topic compared to other topics. This will be used to make the interpretation of each topic more unique to each other.

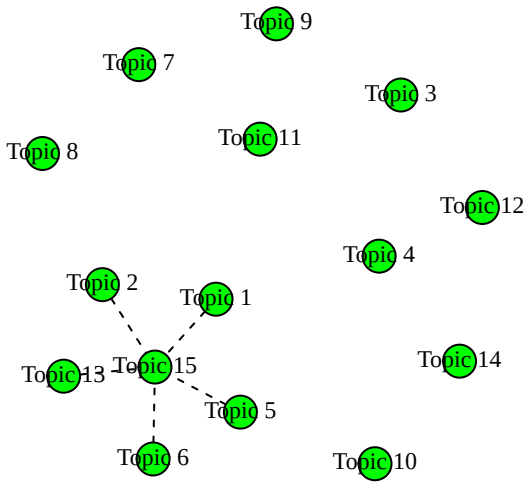
```
labelTopics(model_list[[1]])
```

```
Topic 1 Top Words:
     Highest Prob: peopl, right, democrat, also, nation, american, parti
     FREX: right, peac, peopl, understand, recogn, live, univers
     Lift: univers, understand, recogn, peac, israel, scienc, live
     Score: univers, right, understand, peopl, peac, recogn, israel
Topic 2 Top Words:
     Highest Prob: presid, obama, rule, strengthen, govern, financi, economi
     FREX: rule, strengthen, took, wall, play, financi, place
     Lift: wall, transpar, rule, place, account, took, strengthen
     Score: wall, rule, everyon, obama, presid, took, strengthen
Topic 3 Top Words:
     Highest Prob: world, middl, class, make, polici, back, us
     FREX: middl, polici, class, world, progress, foreign, us
     Lift: foreign, polici, progress, middl, class, allianc, world
     Score: foreign, fail, class, middl, world, polici, us
Topic 4 Top Words:
     Highest Prob: law, women, act, support, effort, protect, oppos
     FREX: law, oppos, women, violenc, enforc, act, equal
     Lift: enforc, equal, violenc, oppos, respect, law, freedom
     Score: enforc, law, women, equal, violenc, act, oppos
Topic 5 Top Words:
     Highest Prob: help, must, ensur, strong, worker, free, work
     FREX: strong, free, must, help, poverti, ensur, labor
     Lift: labor, free, strong, maintain, citi, must, poverti
     Score: labor, must, strong, help, free, safeti, poverti
Topic 6 Top Words:
     Highest Prob: support, invest, continu, reform, provid, democrat, fight
     FREX: invest, reform, institut, infrastructur, provid, support, standard
     Lift: financ, standard, process, institut, infrastructur, train, invest
     Score: financ, support, reform, invest, infrastructur, provid, institut
Topic 7 Top Words:
     Highest Prob: secur, new, america, chang, region, build, reduc
     FREX: chang, america, secur, climat, region, partnership, build
     Lift: climat, chang, practic, emerg, africa, natur, partnership
     Score: climat, chang, secur, global, trade, america, region
Topic 8 Top Words:
```

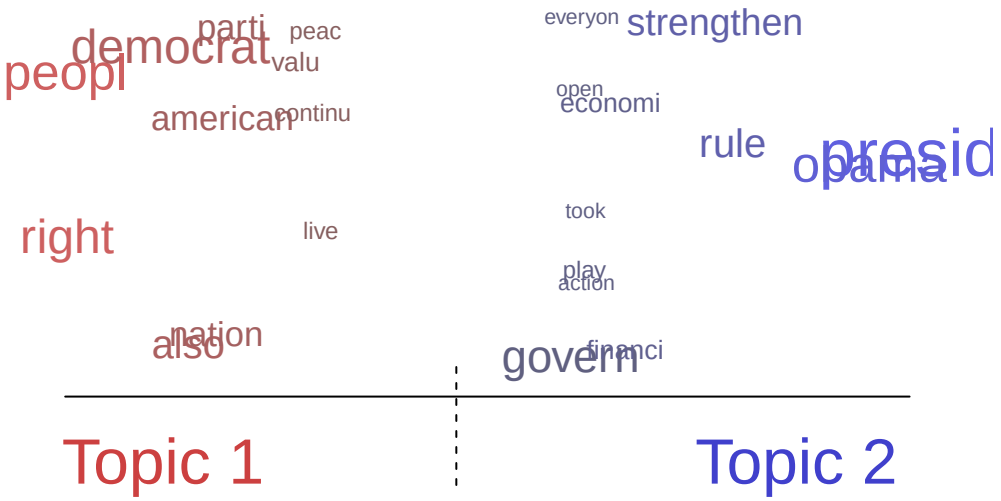
```
     Highest Prob: health, care, famili, busi, access, work, veteran
     FREX: care, health, credit, insur, veteran, access, busi
     Lift: abl, compani, mortgag, care, insur, credit, health
     Score: abl, health, care, credit, insur, busi, compani
Topic 9 Top Words:
     Highest Prob: job, creat, year, million, program, student, home
     FREX: creat, million, job, student, year, good, program
     Lift: loan, million, creat, student, thousand, good, job
     Score: loan, job, creat, million, student, year, save
Topic 10 Top Words:
     Highest Prob: commit, presid, obama, end, war, nuclear, romney
     FREX: war, end, commit, mitt, weapon, romney, iraq
     Lift: hiv, mitt, war, weapon, end, iraq, romney
     Score: hiv, commit, war, end, weapon, mitt, romney
Topic 11 Top Words:
     Highest Prob: presid, tax, cut, work, obama, administr, pay
     FREX: cut, pay, tax, first, relief, bill, strategi
     Lift: strategi, pay, bill, cut, first, relief, sign
     Score: strategi, tax, cut, pay, first, obama, presid
Topic 12 Top Words:
     Highest Prob: state, time, move, forward, togeth, continu, step
     FREX: forward, move, time, togeth, puerto, status, face
     Lift: forward, status, puerto, rico, move, togeth, face
     Score: forward, move, status, puerto, rico, togeth, time
Topic 13 Top Words:
     Highest Prob: countri, communiti, need, econom, take, made, intern
     FREX: countri, need, communiti, great, restor, serv, take
     Lift: serv, centuri, 21st, countri, restor, contribut, great
     Score: serv, countri, communiti, need, safe, centuri, great
Topic 14 Top Words:
     Highest Prob: public, advanc, innov, organ, technolog, threat, nation
     FREX: organ, innov, advanc, network, public, technolog, traffick
     Lift: traffick, network, terrorist, drug, crimin, organ, innov
     Score: traffick, network, public, crimin, terrorist, organ, al-qaeda
Topic 15 Top Words:
     Highest Prob: can, believ, democrat, street, work, like, administr
     FREX: believ, can, street, democrat, like, work, administr
     Lift: street, believ, can, like, democrat, exist, parti
     Score: street, can, believ, democrat, like, parti, work
```

The correlation across the topics can also be visualized. Using the same manifesto, topics 15, 1, 2, 6, 5 and 13 appear to be correlated. Two topics (topic 1 and 2) can be compared wherein words are plotted with size proportional to their use within the topic and the x-axis shows how close the terms are to one topic over the other (y-axis is random).

```
topic_correlation<-topicCorr(model_list[[1]])
plot(topic_correlation)
```

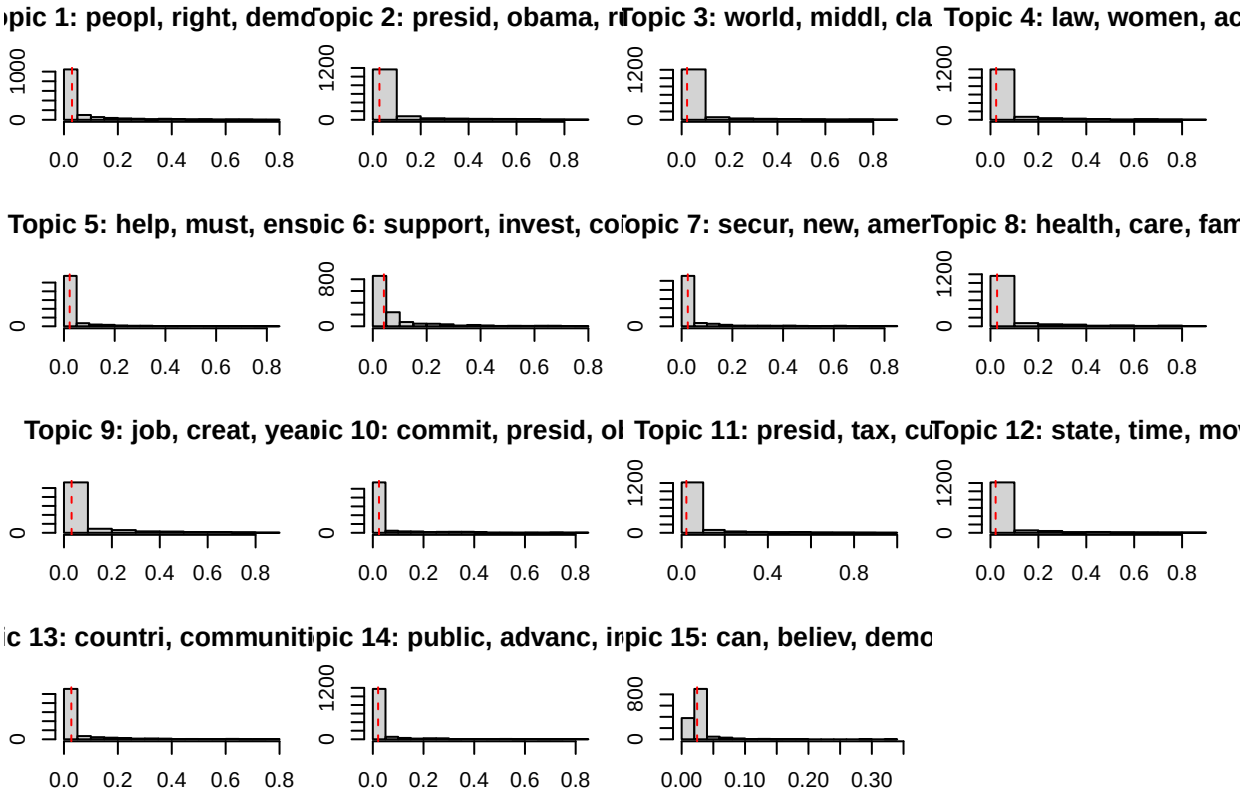



```
plot(model_list[[1]],
     type="perspectives",
     topics=c(1, 2),
     plabels = c("Topic 1","Topic 2"))
```



In addition, the topic estimates of document-topic proportions is plotted using 2012 Democrats' manifesto. This plot basically tells us which topics are coming from which documents. As expected, each topic has no relation or very little relation with several documents.

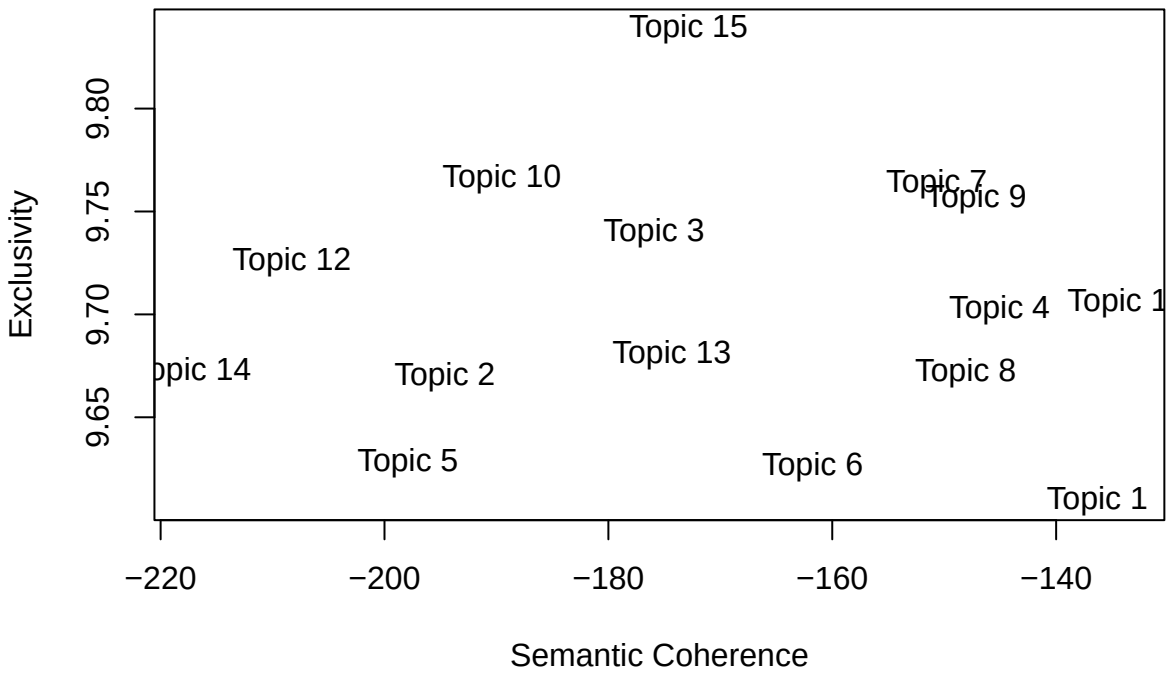
```
# plot(model_list[[1]], type = "hist", topics = sample(1:15, size = 15))
plot(model_list[[1]], type="hist")
```



Using `topicQuality`, the plot of semantic coherence vs. exclusivity of topics in the 2012 Democrats' manifesto is given below. Semantic coherence measures how coherent topics are, or how often terms describing a topic co-occur. Exclusivity measures how exclusive topics are or how much they differ from each other and describe different things.

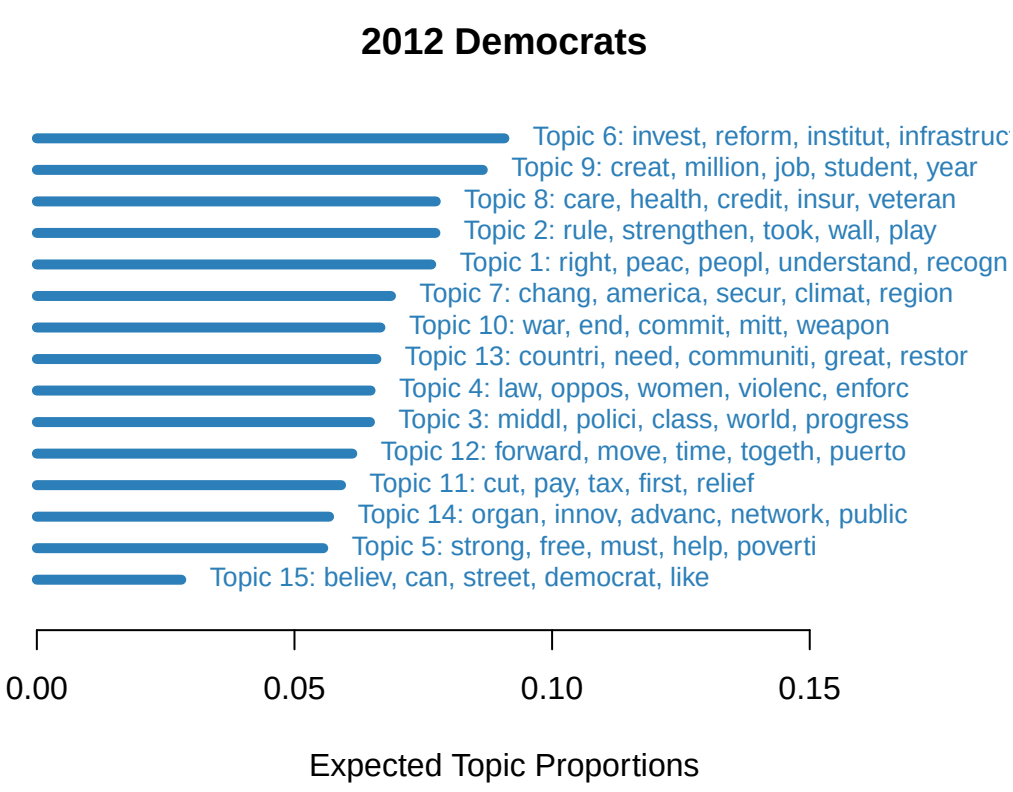
```
topicQuality(model=model_list[[1]], documents=dfm_stm_list[[1]]$documents)
```

```
 [1] -136.3100 -194.6037 -175.9260 -145.0520 -197.9109 -161.7569 -150.6700
 [8] -148.0860 -147.1607 -189.5163 -133.6732 -208.2623 -174.3291 -217.2114
[15] -172.8444
 [1] 9.609207 9.669494 9.739603 9.702083 9.627720 9.625868 9.763253 9.671788
 [9] 9.756122 9.765698 9.705767 9.725709 9.680164 9.672155 9.838986
```

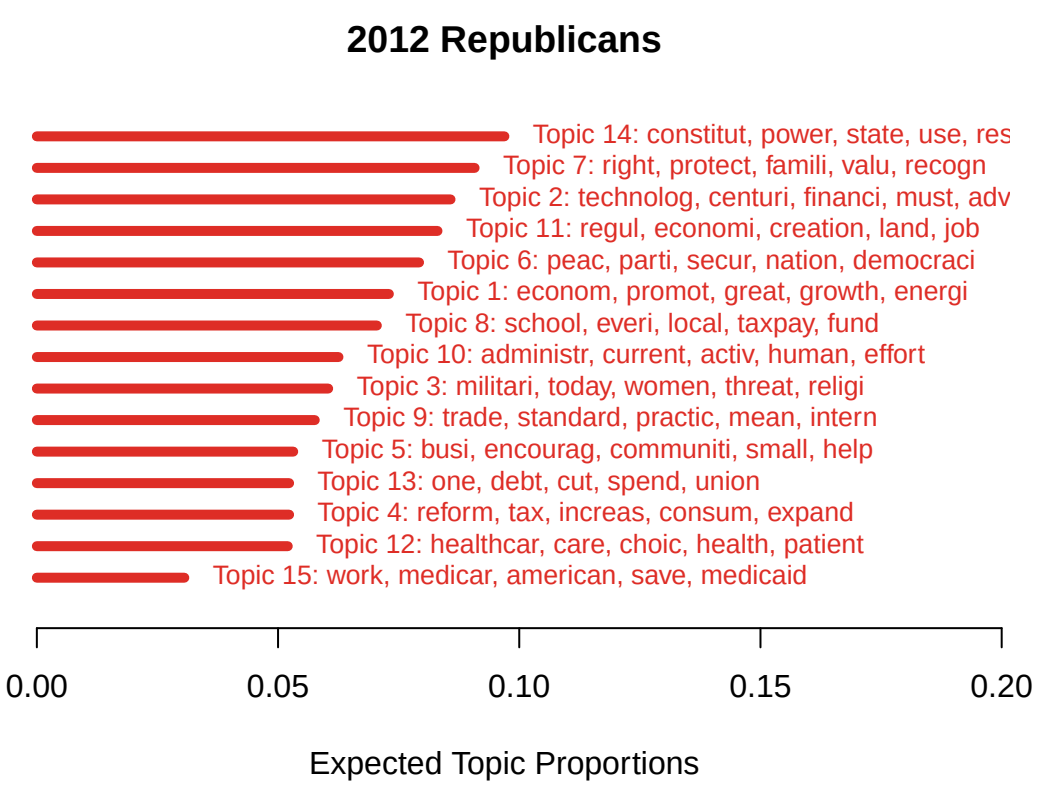


2. STM Topics per party and year

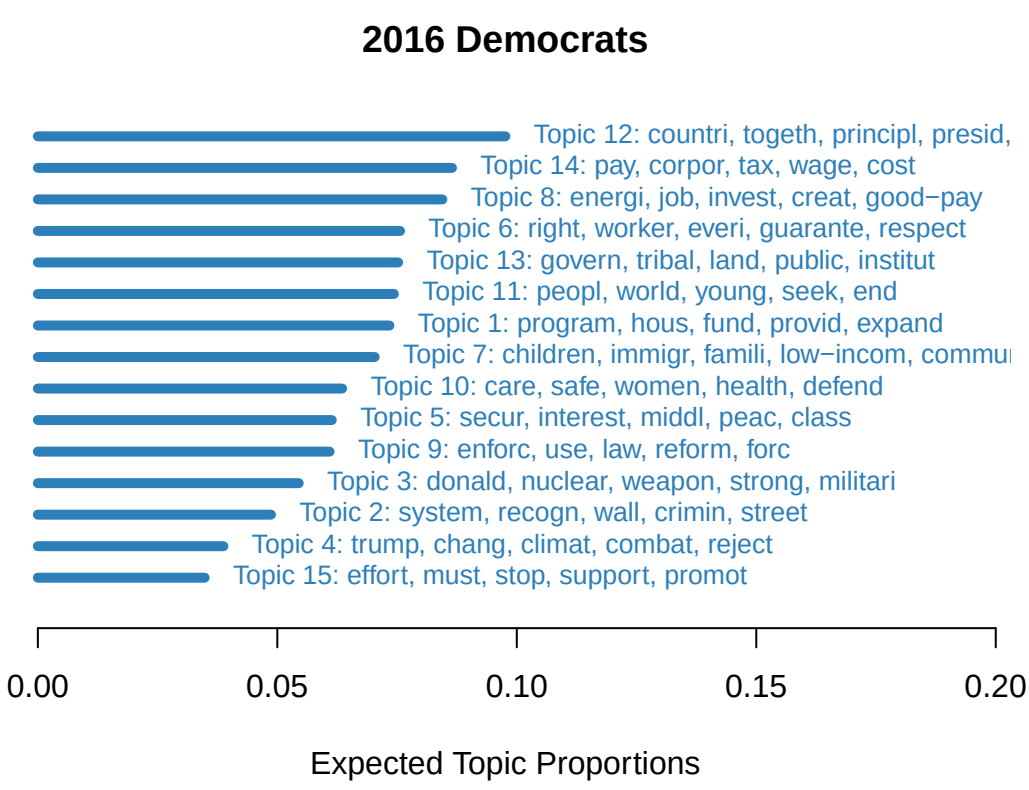
```
par(bty="n",col="#2c7fb8",lwd=5)
plot.STM(model_list[[1]],type="summary", label="frex", n = 5, main = "2012 Democrats",
         width=50, text.cex=.8)
```



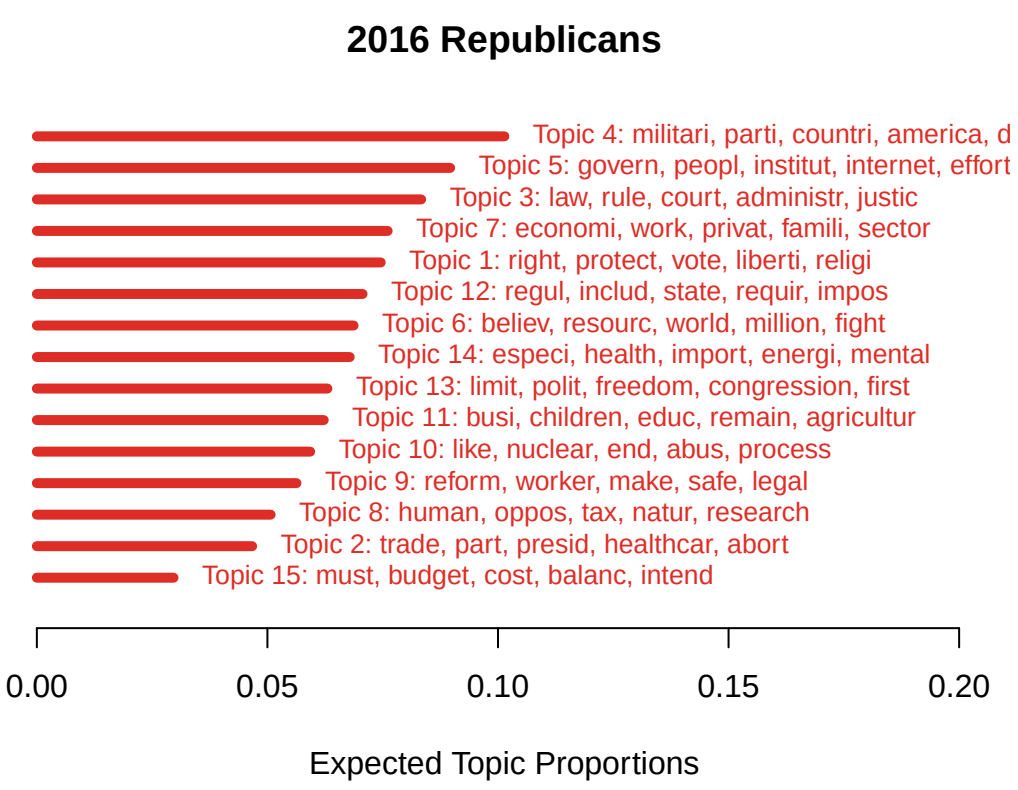

```
par(bty="n",col="#de2d26",lwd=5)
plot.STM(model_list[[2]],type="summary", label="frex", n = 5, main = "2012 Republicans",
        width=50, text.cex=.8)
```



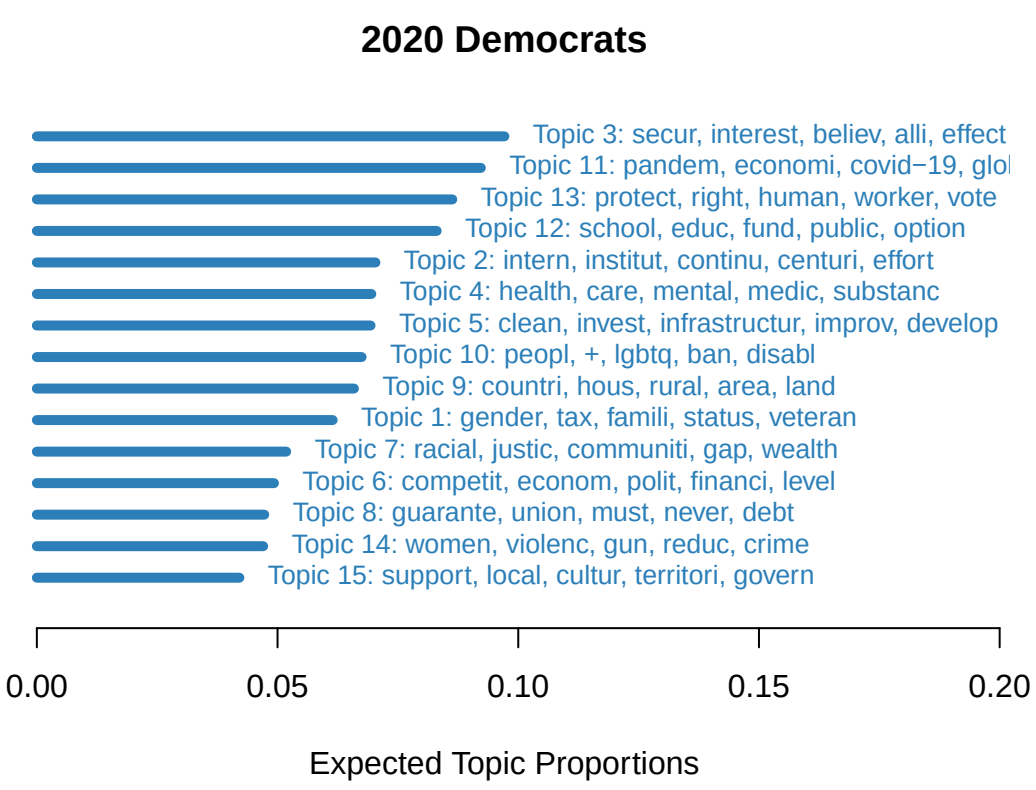

```
par(bty="n",col="#2c7fb8",lwd=5)
plot.STM(model_list[[3]],type="summary", label="frex", n = 5, main = "2016 Democrats",
         width=50, text.cex=.8)
```



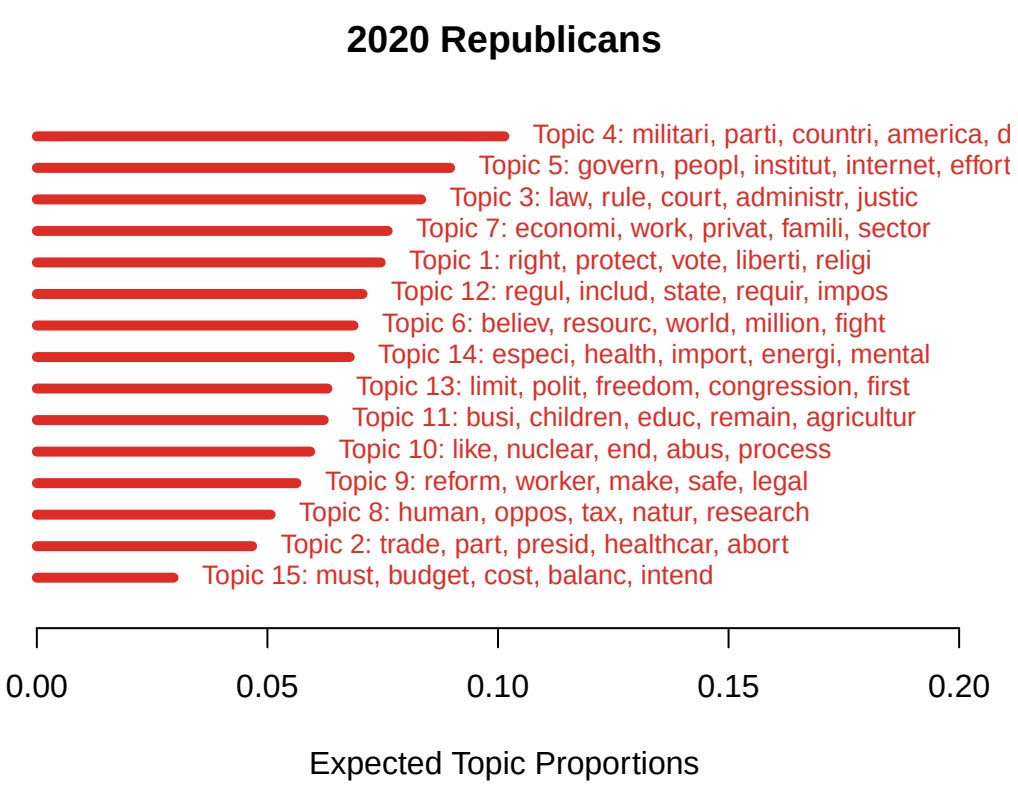

```
par(bty="n",col="#de2d26",lwd=5)
plot.STM(model_list[[4]],type="summary", label="frex", n = 5, main = "2016 Republicans",
         width=50, text.cex=.8)
```




```
par(bty="n",col="#2c7fb8",lwd=5)
plot.STM(model_list[[5]],type="summary", label="frex", n = 5, main = "2020 Democrats",
         width=50, text.cex=.8)
```




```
par(bty="n",col="#de2d26",lwd=5)
plot.STM(model_list[[6]],type="summary", label="frex", n = 5, main = "2020 Republicans",
         width=50, text.cex=.8)
```


2020 Republicans
Topic 4: militari, parti, countri, america, d
Topic 5: govern, peopl, institut, internet, effort
Topic 3: law, rule, court, administr, justic
Topic 7: economi, work, privat, famili, sector
Topic 1: right, protect, vote, liberti, religi
Topic 12: regul, includ, state, requir, impos
Topic 6: believ, resourc, world, million, fight
Topic 14: especi, health, import, energi, mental
Topic 13: limit, polit, freedom, congression, first
Topic 11: busi, children, educ, remain, agricultur
Topic 10: like, nuclear, end, abus, process
Topic 9: reform, worker, make, safe, legal
Topic 8: human, oppos, tax, natur, research
Topic 2: trade, part, presid, healthcar, abort
Topic 15: must, budget, cost, balanc, intend
0.00
0.05
0.10
0.15
0.20
Expected Topic Proportions