Lab 4

Thomas Haase

September 26, 2025

Table of contents

Part 1

In this lab, we will use a data set containing a random sample of public Facebook posts by members of the U.S. Congress from 2017.1 Our broad objective in this first part of the lab is to explore what topics were discussed, and possible variation by party membership.

```r
library(quanteda)
library(topicmodels)
library(word2vec)

library(data.table)

library(easystats)

library(tibble)
library(kableExtra)
library(ggplot2)
library(tidytext)

setwd("~/Github/ML-Labs/5")
```

Task 1

Begin by importing `fb-congress-data3.csv`. Report basic information about the data set; how many rows and column it has, as well as the name of the variables.

```r
d <- read.csv("fb-congress-data3.csv")

tribble(
  ~Name,      ~Value,
  "Rows",     nrow(d) |> as.character(),
  "Columns", ncol(d) |> as.character()) |>
  kable()
```

Name	Value
Rows	6752
Columns	4

```r
varnames <- names(d)
```

Variables contained in the dataset: doc_id, screen_name, party, message

Task 2

As you may have noticed from your inspection in #1, this data set has yet to be pre-processed (it contains punctuation, etc.). Hence, that is what you shall do now. More specifically, perform the following steps:

```r
# i
corp <- d |>
  corpus(docid_field = "doc_id",
         text_field = "message",
         meta = list("screen_name",
                     "party"))

# ii
toks <- corp |>
  tokens(remove_punct = T,
         remove_numbers = T,
         remove_symbols = T,
         remove_url = T) |>
# iii
  tokens_remove(stopwords("english"))

# iv
toks[1:3]
```

```
Tokens consisting of 3 documents and 2 docvars.
1 :
 [1] "President"     "Trump"         "backs"         "Paris"
 [5] "Agreement"     "economic"      "environmental" "national"
 [9] "security"      "moral"         "disaster"      "United"
[ ... and 6 more ]

2 :
 [1] "Many"       "thanks"     "first"      "class"      "summer"
 [6] "interns"    "Washington" "hard"       "work"       "folks"
[11] "#GA03"

3 :
 [1] "economy"    "needs"      "shot"       "arm"        "spur"
 [6] "growth"     "Co-Chair"   "bipartisan" "Problem"    "Solvers"
[11] "Caucus"     "I’ve"
[ ... and 27 more ]
```

```
# v
dfm <- toks |> dfm() |>

# vi
  dfm_trim(min_termfreq = 5)
# only keep documents with more then 10 features
dfm <- dfm[which(rowSums(dfm)>=10), ]
```

Task 3

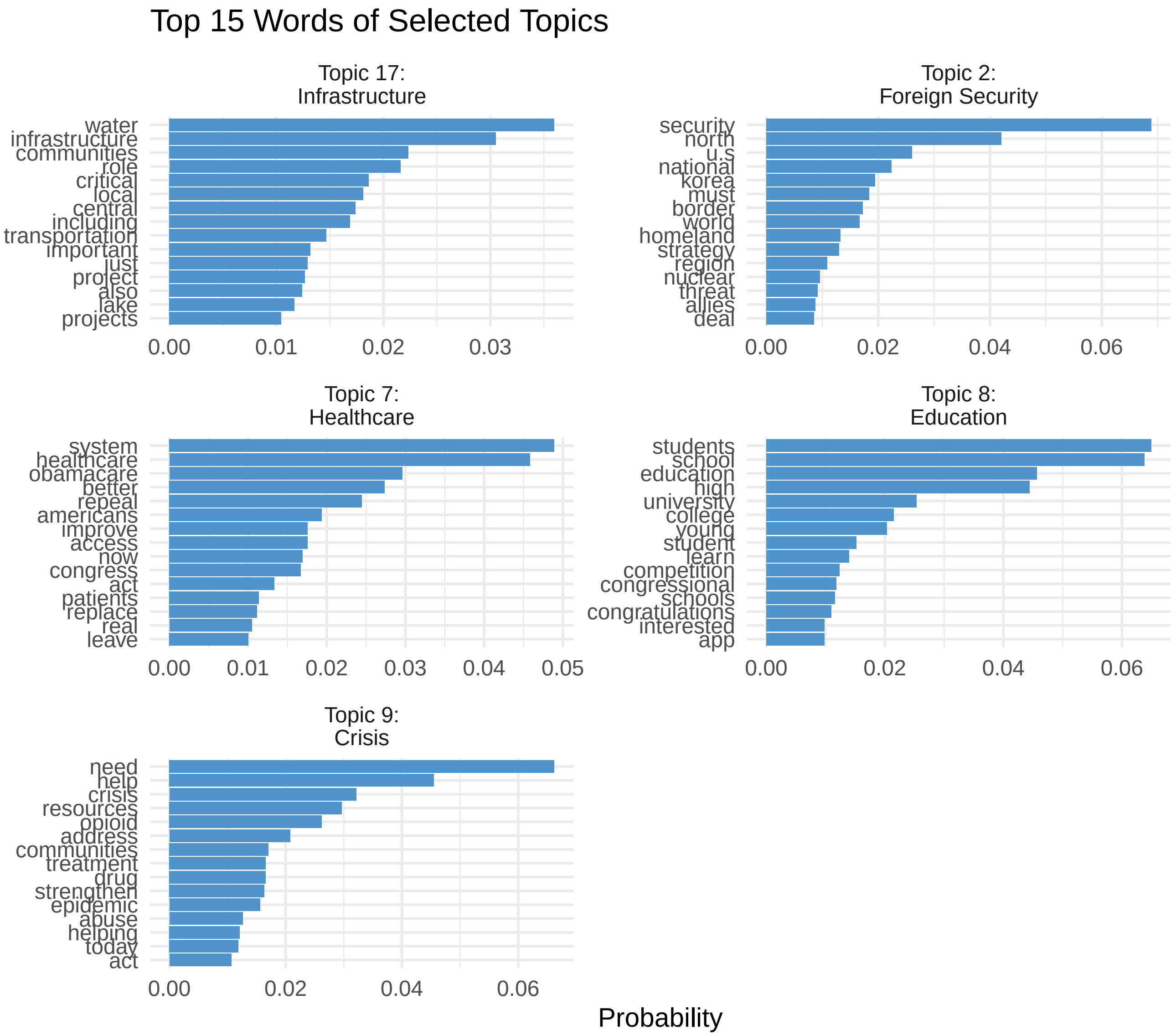
Now we are ready to do some topic modeling! To do so, we will use the topicmodels package, and the function LDA(). Set x to your document-term-matrix and specify method="Gibbs" (note: Gibbs is the name of a particular estimation procedure; see the Appendix of the lecture for more details). Set the number of iterations to 1000, and specify a seed number to ensure replicability (hint: to specify iterations and seed number, use the control argument). Finally, set the number of topics, K=50 With these settings specified, start the estimation. This could take a minute or two.

```
set.seed(5)

K <- 50

if(!file.exists("lda.rds")){
  lda <- LDA(x = dfm, k = K,
             method="Gibbs",
             control=list(iter = 1000,
                          seed = 5,
                          verbose = 1))
  saveRDS(lda, file = "lda.rds")
} else {
  lda <- readRDS("lda.rds")
}
```

Task 4

Once the estimation is finished, use the `get_terms()` function to extract the 15 words with the highest probability in each topic. In a real research setting, we would carefully examine each of the topics. Here, I only ask you to briefly skim them, and then focus on 5 that (i) you think are interesting, (ii) has a clear theme, and (iii) are clearly distinct from the other topics. Provide a label to each of those based on the top 15 words. Complementing your label, please also provide a bar chart displaying on the y-axis the top 15 words, and on the x-axis their topic probabilities. Hint: you can retrieve each topic's distribution over words using topicmodels's function posterior.3 Lastly, please also report a general assessment—based on your skim—about the general quality of the topics; do most of them appear clearly themed and distinct, or are there a lot of "junk" topics?

```
# get_terms(lda,15)

words <- data.table(topic = 1:K,
                    posterior(object = lda)$terms) |>
  melt.data.table(id.vars = 'topic')
words <- words[order(value,decreasing = T)]

t15 <- words[,head(.SD,15),by='topic']
t15 <- t15[topic %in% c(2,7,8,9,17)]

# plot
topic_labels <- c(
  "2"  = "Topic 2:\nForeign Security",
  "7"  = "Topic 7:\nHealthcare",
  "8"  = "Topic 8:\nEducation",
  "9"  = "Topic 9:\nCrisis",
  "17" = "Topic 17:\nInfrastructure"
)

# Apply topic labels
t15[, topic_label := topic_labels[as.character(t15$topic)]][!is.na(topic_label)] |>
  ggplot(aes(x = value, y = reorder_within(variable, value, topic_label))) +
  geom_bar(stat = 'identity', position = 'dodge', fill = "steelblue3") +
  facet_wrap(~topic_label, scales = 'free', ncol = 2) +
  scale_y_reordered() +
  labs(title = "Top 15 Words of Selected Topics",
       x = "Probability", y = "") +
  theme_minimal()
```



There are a lot of junk topics, but a few of them are very distinct.

Task 5

Out of the 5 topics that you labeled, select two which you think are particularly interesting. For these two, identify the three documents which have the highest proportion assigned of this topic (hint 1: use topicmodels's `posterior()` to extract documents' distribution over topics | hint 2: to identify the document ids which correspond to each row of what you extract from posterior(), you can use `ldaobject@documents`. See help file for more details.), and do a qualitative inspection (= 2 × 3 documents to read). Does your readings corroborate your labels? Are they about what you expected?

```
documents <- data.table(doc_id = lda@documents,
                        posterior(object = lda)$topics)
# Assign topics as column names
```

```r
colnames(documents)[2:ncol(documents)] <- paste0('Topic',colnames(documents)[2:ncol(documents)])



top_docs <- documents[
  order(Topic8,decreasing = T)][
    1:3,doc_id] |>
  as.integer()

tibble(ID = as.character(top_docs),
       Education = as.character(corp)[top_docs]) |>
  kable() |>
  kable_styling(bootstrap_options = c("striped", "hover",
                                      "condensed", "responsive")) |>
  column_spec(1, bold = TRUE, width = "3em") |>
  column_spec(2, width = "40em")
```

ID	Education
3937	Thompson Valley High School is a local example of Career and Technical Education in action. The school combines traditional classroom education with applied learning. Yesterday, I got to attend a cooking, entrepreneurship, and agriculture class alongside students. As the top-Democratic member of the Early Childhood and K-12 subcommittee on the Education and Workforce Committee, I've worked hard to support Career and Technical Education programs. This spring, the Education and Workforce committee debated and the House of Representatives approved a reauthorization of the Perkins Career and Technical Education program that will support innovation and bring our CTE programs into the 21st century. I want to thank the educators who are on the front lines teaching Career and Technical Education, helping to prepare students for the real-world and good jobs. http://www.reporterherald.com/news/education/ci_31312475/polis-visits-career-classes-thompson
3933	I visited Lake Minneola High School, Weeki Wachee High School, Crystal River High School, Lecanto High School, Wildwood Middle High School, Tavares High School and Lake Weir High School to congratulate the winning students. This is the biggest turnout we've had so far - 86 students participated in this year's competition. The spectacular artwork and record turnout is a reflection of the emphasis that our schools are placing on the arts.
274	Announcing the 2017 Congressional Art Competition! Your students are invited to participate in my annual Congressional Art Competition–now open to all high school students who reside in the 11th District. Entries can include paintings, drawings, prints, and more. We have had a blast with this competition in the past and are really looking forward to the entries this year! Winners will be chosen by a panel of art professionals, and there will be a reception for all the students who enter–as well as their teachers and families–on April 29, 2017. The over-all winner of our district's competition will receive two round-trip tickets to the National Reception in Washington, DC, a $3,000 scholarship to a prestigious Southeastern art college, and their art will be displayed for one year in the U.S. Capitol! You can find more information on my website below. Note that the DEADLINE to enter this year is Friday April 21, 2017.

```r
top_docs <- documents[
  order(Topic17,decreasing = T)][
    1:3,doc_id] |>
  as.integer()

tibble(ID = as.character(top_docs),
       Infrastructure = as.character(corp)[top_docs]) |>
  kable() |>
  kable_styling(bootstrap_options = c("striped", "hover",
                                      "condensed", "responsive")) |>
  column_spec(1, bold = TRUE, width = "3em") |>
  column_spec(2, width = "40em")
```

ID	Infrastructure
6440	From harmful algal blooms in Owasco Lake, to bursting pipes in Syracuse, to sewer problems in Oswego, to aging filtration systems in communities around the district – our region faces water infrastructure challenges like never before. Today I stood in at the northern shore of Owasco Lake with a bipartisan group of local leaders to highlight my commitment to strengthening water infrastructure for CNY. I was proud to lead bipartisan letters in Congress to double the Clean & Drinking Water Funds, invest in Harmful Algal Bloom research, and to prioritize programs for small and rural communities to develop wastewater and drinking water infrastructure.
2068	I was pleased to join Transportation and Infrastructure Chairman Shuster, Subcommittee Chairman Graves, and my Florida delegation colleagues to discuss the upcoming WRDA bill. This important piece of legislation has authorized critical Everglades restoration projects, including the Central Everglades Planning Project and the Picayune Strand. It has also authorized much needed upgrades and expansions to PortMiami and Port Everglades. Because of these projects, we are able to restore and protect the Everglades for future generations, as well as provide our ports with adequate resources to fit the needs of our growing economy. This roundtable is just the beginning of the process, but I appreciate that the Transportation and Infrastructure Committee recognizes how important this bill is to Florida. I thank Chairman Shuster for his leadership throughout the years on ensuring WRDA properly served our community. I remain committed to working with Speaker Ryan, Chairman Shuster, and Subcommittee Chairman Graves to ensure the 2018 WRDA bill reflects Florida's infrastructure needs.
2037	This is great news in our efforts to keep Lake Erie healthy and safe! Lake Erie is one of Ohio's most precious and important natural resources. Many lakeshore communities rely on it for clean and affordable drinking water and the lake plays a key role in Ohio's economy. For the last few years, I have been working with the Ohio EPA to prevent the U.S. Army Corps of Engineers from dumping dredged sediment from the Cuyahoga River that contains contaminants called polychlorinated biphenyls, or PCBs, into Lake Erie. Elevated PCB levels can lead to contamination among the lake's fish populations. I'm very happy to see this result!

For both topics the three selected prototypical-tweets adress gunviolence and infrasstructure. The model seems very convincing, since the content of the inspected texts meet my expectation.

Task 6

Now, estimate a topic model—as in #3—but with K=3 instead. Extract the top 15 words from each topic, (try to) label each, and then make an assessment of the overall quality of them. To further explore the quality of this topic model, reconsider the documents you read in #5: extract the distribution over topics for these documents (from your new K=3 model). How well does this topic model capture the theme of these documents? Based on your analysis, which of the two K's do you prefer? Motivate.

```r
# estimate LDA
set.seed(5)
K <- 3

if(!file.exists("lda2.rds")){
  lda <- LDA(x = dfm, k = K,
             method="Gibbs",
             control=list(iter = 1000,
                          seed = 5,
                          verbose = 1))
  saveRDS(lda, file = "lda2.rds")
} else {
  lda <- readRDS("lda2.rds")
}


# get top words
get_terms(lda,15) |>
  kable(col.names = c("Healthcare & Tax",
                      "Veterans & Community",
                      "National Government"),
        caption = "Top 15 Terms of Topic Model (K = 3)")
```

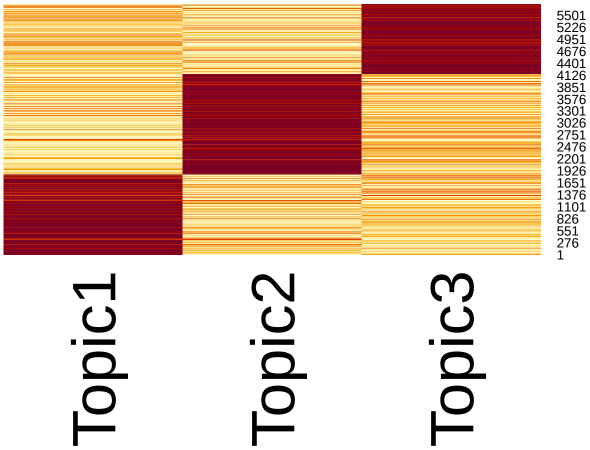
Healthcare & Tax	Veterans & Community	National Government
health	today	president
care	veterans	trump
can	great	u.s
bill	day	federal
act	community	national
tax	week	must
americans	office	congress
families	thank	years
help	service	law
new	district	house
make	washington	american
people	congressional	continue
work	local	security
need	one	states

Healthcare & Tax	Veterans & Community	National Government
house	students	committee

Table 2: Top 15 Terms of Topic Model (K = 3)

It seems like multiple topics got recognized as one from the model. There are probably more then 3 larger topics adressed in the texts.

```
documents <- data.table(doc_id = lda@documents,
                        posterior(object = lda)$topics)
# Assign topics as column names
colnames(documents)[2:ncol(documents)] <- paste0('Topic',colnames(documents)[2:ncol(documents)])

set.seed(5)

documents[,-"doc_id"] |>
  as.matrix() -> mat

par(mar=c(5,10,4,2))
mat[order(apply(mat, 1, which.max)), ] |>
  heatmap(Rowv = NA, Colv = NA, margins = c(10, 1))
```



The heatmap, that ordered the documents by maximum probability, shows, that the distribution over Topics is quite distinct. Only few texts have high probability in all topics. The K= 50 model captured unnecessary topics, while the K = 3 one still shows many stripes in its heatmap. These could probably make up new topics! The optimum lies somewhere inbetween 3 and 50. Just from the Interpretation the 50 topic model is more powerful because it captures more distinctions.

Task 7

Continuing with the topic model you concluded the most appropriate, perform the following sets of analyses:

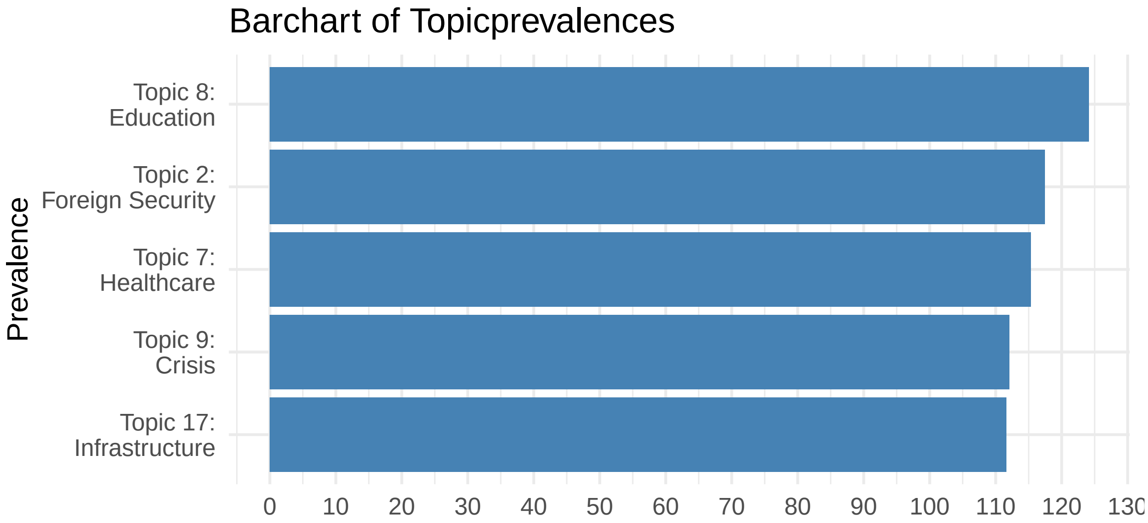
- i. Compute the prevalence of each topic, across all documents. Report which is the most prevalent topic, overall, and then report—in the form of a single plot; e.g., a bar chart—the prevalence of the topics you labeled.

-ii. Compare the prevalence on your labeled topics between democrats and republicans. You can for example fit a fractional regression model using glm(family="quasibinomial") or using t-tests of difference in means. Interpret.

```r
# load 50 topic lda
lda <- readRDS("lda.rds")
K <- 50


documents <- data.table(doc_id = lda@documents,
                        posterior(object = lda)$topics)
colnames(documents)[2:ncol(documents)] <- paste0('Topic',colnames(documents)[2:ncol(documents)])

# query for most prevalent topic
top_topic <- documents[, -"doc_id"] |>
  colSums() |>
  which.max() |>
  names()

# report
cat("Most prevalent topic:", top_topic, "\n")
```

```
Most prevalent topic: Topic11
```

```r
# plot labled topics
data.table(Topic = documents[, -"doc_id"] |>
             colSums() |>
             names(),
           Prevalence = documents[, -"doc_id"] |>
             colSums())[
               sub("Topic", "", Topic) %in% names(topic_labels)
               ][
                 , Topic := topic_labels[sub("Topic", "", Topic)]
                 ] |>
ggplot(aes(x = reorder(Topic, Prevalence),
           y = Prevalence)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(
    x = "Prevalence", y = "",
    title = "Barchart of Topicprevalences"
  ) +
  scale_y_continuous(breaks = seq(0, 150, 10)) +
  theme_minimal()
```

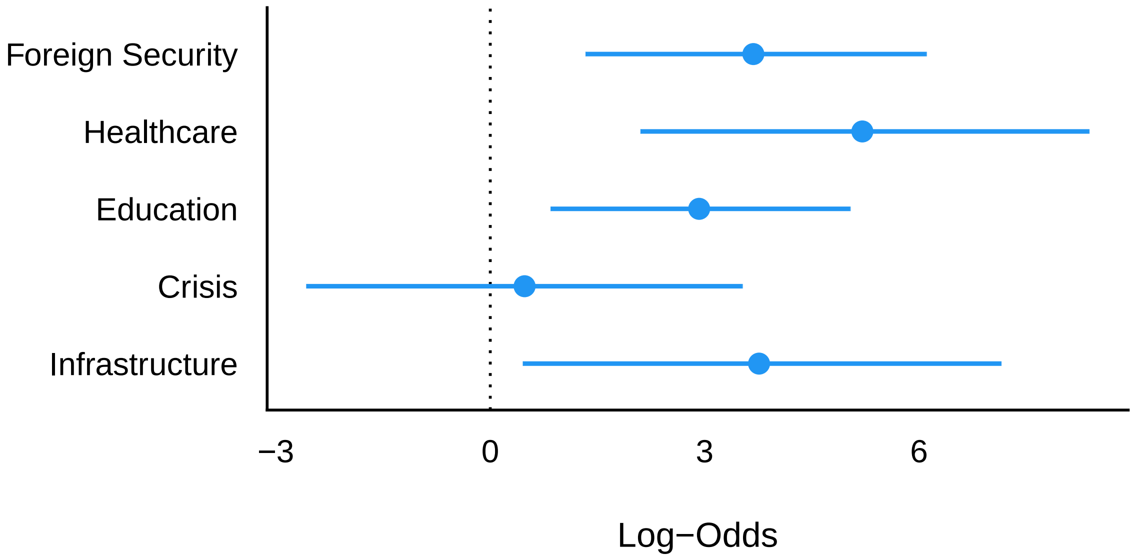


```
# prepare data
documents <- data.table(
  doc_id = lda@documents,
  posterior(lda)$topics
)
colnames(documents)[2:ncol(documents)] <- paste0('Topic',colnames(documents)[2:ncol(documents)])
# add partyinformation
documents <- documents[, party := docvars(dfm)$party |>
                         factor(levels = c("Democrat",
                                           "Republican"))]
# filter data
documents <- documents[, .(doc_id,party,
                           Topic2,Topic7,
                           Topic8,Topic9,
                           Topic17)]
setnames(documents,
         c("Topic2","Topic7","Topic8",
           "Topic9","Topic17"),
         c("Foreign Security",
           "Healthcare",
           "Education",
           "Crisis",
           "Infrastructure"))

# calculate model
glm <- glm(party ~ `Foreign Security` + Healthcare +
             Education + Crisis + Infrastructure,
           family = "binomial",
           data = documents)

# because quasibinomial was suggested but
# Im using binomial lets check overdispersion
```

```
check_overdispersion(glm)
```

```
# Overdispersion test

 dispersion ratio = 1.000
          p-value =  0.88
```

```
glm |>
  parameters(exponentiate = F) |>
  plot()
```



```
glm |> summary()
```

```
Call:
glm(formula = party ~ `Foreign Security` + Healthcare + Education + 
    Crisis + Infrastructure, family = "binomial", data = documents)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -0.3688     0.0742  -4.970 6.68e-07 ***
`Foreign Security`   3.6819     1.2136   3.034  0.00242 ** 
Healthcare           5.2070     1.5988   3.257  0.00113 ** 
Education            2.9235     1.0677   2.738  0.00618 ** 
Crisis               0.4806     1.5508   0.310  0.75662    
Infrastructure       3.7617     1.7024   2.210  0.02713 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7965.7  on 5747  degrees of freedom
Residual deviance: 7937.4  on 5742  degrees of freedom
AIC: 7949.4

Number of Fisher Scoring iterations: 4
```

The topics Foreign Security, Healthcare, Education and Infrastructure can distinguish between both parties statistically significant. Adressing the topic “Crisis” can not be used to distinguish between parties.

Task 8

BONUS (not obligatory; suggestion; do after you have completed the rest of the lab). As a bonus exercise—to expose you to the traditional computer sciency way of selecting the number of topics, K-you shall consider a data-driven approach, relying on the measure of hold-out likelihood (or, perplexity as its also called). To do so, do the following:

- i. Split your document term matrix into two (a training and test set); 80/20 division.
- ii. Write a loop which in each iteration (a) estimates a topic model using a particular K, and then (b) computes (and stores) its perplexity using the topicmodels function perplexity(), which takes as input the model object and the test document-term-matrix (note: the document-term-matrix needs to be transformed into a particular format: use … for this).
- iii. Consider the following range of K: $3, 10, 25, 50, 75, 100, 200$, and run the loop. This may take a few minutes. Once the loop has finished, plot your results (x-axis: K, y-axis: perplexity).

Interpret. Based on this, what is a reasonable K?