Lab 4

Thomas Haase

September 25, 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.

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

library(data.table)

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.

```
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

```
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:

```
# 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)[,c(8,15,17,23,33)] |>
  as_tibble()
```

```
# A tibble: 15 x 5
   `Topic 8`       `Topic 15`     `Topic 17`     `Topic 23` `Topic 33`
   <chr>           <chr>          <chr>          <chr>      <chr>
 1 students        president      water          first      veterans
 2 school          trump          infrastructure violence   va
 3 education       administration communities    americans  care
 4 high            trump's        role           must       health
 5 university      donald         critical       thoughts   medical
 6 college         trump's        local          victims    benefits
 7 young           j              central        congress   provide
 8 student         obama          including      lives      receive
 9 learn           white          transportation gun        ensure
10 competition     actions        important      attack     access
11 congressional   signed         just           prayers    affairs
12 schools         vice           project        action     veteran
13 congratulations undermine      also           families   choice
14 interested      rule           lake           others     program
15 app             putting        projects       responders services
```

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

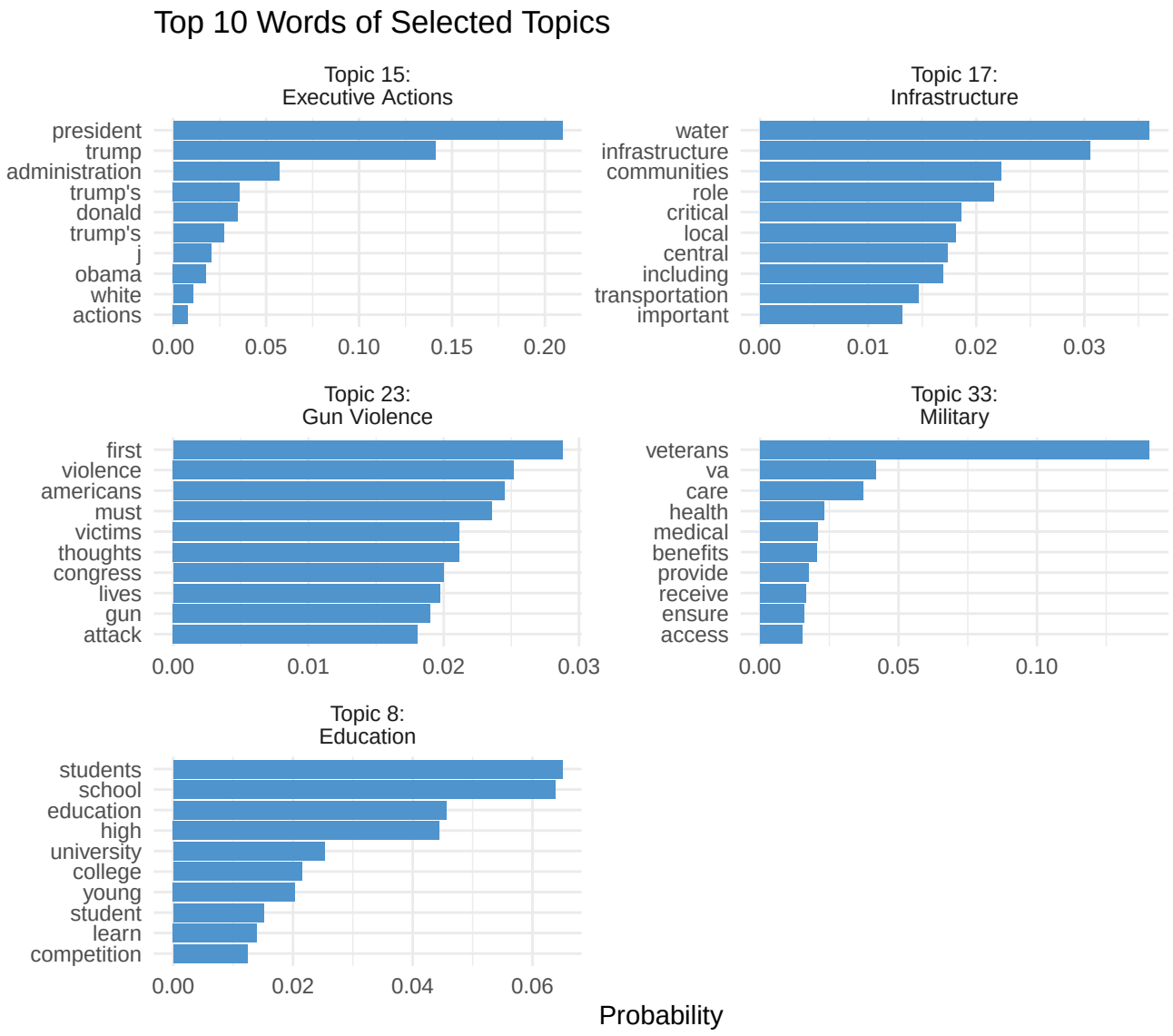
```
       topic   variable        value
       <int>     <fctr>        <num>
    1:    15  president 2.092810e-01
```

```
     2:    29   american 1.815100e-01
     3:    35        new 1.728059e-01
     4:    29     people 1.485333e-01
     5:    20        tax 1.449705e-01
    ---
274196:    11 vegetation 1.482492e-05
274197:    11    klamath 1.482492e-05
274198:    11       ptsd 1.482492e-05
274199:    11   tuskegee 1.482492e-05
274200:    11   naismith 1.482492e-05
```

```r
t10 <- topic_distr_over_words[order(-value), .SD[1:10], by = topic]

# plot
topic_labels <- c(
  "8" = "Topic 8:\nEducation",
  "15" = "Topic 15:\nExecutive Actions",
  "17" = "Topic 17:\nInfrastructure",
  "23" = "Topic 23:\nGun Violence",
  "33" = "Topic 33:\nMilitary"
)


# Apply topic labels
t10[, topic_label := topic_labels[as.character(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 10 Words of Selected Topics",
       x = "Probability", y = "") +
  theme_minimal()
```



There are a lot of junk topics.

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 ($= 2 \times 3$ 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(Topic23,decreasing = T)][
    1:3,doc_id] |>
  as.integer()

tibble(ID = as.character(top_docs),
       Gunviolence = 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	Gunviolence
155	Today marks five years since a gunman walked into Sandy Hook Elementary School and took the lives of twenty children and six educators. It will forever be one of the saddest days in our nation's history. The events of December 14th, 2012 are still as shocking and horrific today as they were five years ago. The thought that somebody, anybody, could commit such a heinous act was unconscionable then, as it is now. I still pray for the families of those school children and educators who were killed that day. Their lives will forever be ingrained in our national consciousness. While we mourn the twenty-six lives lost five years ago today, we also mourn the loss of the 120,000 other Americans who have been killed by guns since that terrible tragedy. This year has been especially deadly; there have been more mass shootings in 2017 than there have been days, and three of the five deadliest mass shootings in our recent history have taken place in the past year. Gun violence is a national crisis, and we cannot let another year go by without a response. It is long past time that Congress pass legislation to address the gun violence epidemic in our country. Yet, just last week Republicans in the House passed the dangerous Concealed Carry Reciprocity Act to make it easier to carry loaded weapons in public spaces. As we remember the twenty-six innocent lives lost in Newtown, I urge my colleagues in Congress to commit to passing commonsense gun safety legislation to prevent future tragedies related to gun violence.

2907	One year ago, on a tragic night in the middle of Pride month, 49 innocent Americans were slaughtered at Pulse nightclub in Orlando. I vividly remember the images on the news and the horrible sinking of my heart as we realized the gunman targeted an LGBTQ club. Nearly all of those killed were LGBTQ, more than half of the victims were part of Central Florida's vibrant Puerto Rican diaspora. The LGBTQ community was heartbroken. The Latino community was shaken. And our entire nation was at yet another crossroads when it came to the issue of gun violence. One year later, our nation's wounds have yet to heal, and we have yet to take meaningful action to honor the victims. The Pulse nightclub shooting was the deadliest mass shooting by a single shooter in American history, and despite Democratic filibusters and sit-ins, Congressional Republicans have spent the last year blocking common sense gun reform. I'm talking about background checks. I'm talking about banning large ammunition magazines designed for shooting en masse. These are simply common-sense reforms. They are not controversial, they will protect the most vulnerable among us, and they do not violate the Second Amendment. How many prayers must be said? How many vigils must we have? How many moments of silence before we get serious about stopping gun violence in this country? Inaction is acceptance. I hope my Republican colleagues can get serious soon, before more tragedy strikes.
6208	This morning Americans awoke to news of the deadliest mass shooting in our nation's history. My thoughts and prayers are with the families of the victims and the survivors of this horrific mass shooting in Las Vegas, Nevada. The frequency of mass shootings are far too common in our country. No one deserves to bear witness to the type of carnage that occurred in Las Vegas last night.

```
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!

The