Assignment 4

Steve Kerr (211924)

2023-11-20

Introduction

In this assignment, you are asked to use topic modelling to investigate manifestos from the manifesto project maintained by WZB. You can either use the UK manifestos we looked at together in class, or collect your own set of manifestos by choosing the country/countries, year/years and party/parties you are interested in. You should produce a report which includes your code, that addresses the following aspects of creating a topic model, making sure to answer the questions below.

This time, you will be assessed not only on whether the code gets the right result, but on how you understand and communicate your understanding of the modelling process and how this can answer your research question. The best research question is one that is interesting and answerable, but the most important thing is that the research question is answerable with the methods you choose.

You will also be assessed on the presentation of your results, and on the concision and readability of your code.

1. Data acquisition, description, and preparation

Bring together a dataset from the WZB. What years, countries and parties are included in the dataset? How many texts do you have for each of these?

Prepare your data for topic modelling by creating a document feature matrix. Describe the choices you make here, and comment on how these might affect your final result.

Setup

We start by loading our packages.

```
# TODO: discard irrelevant packages
# load packages
pacman::p_load(
  dplyr,
  ggplot2,
  gt,
  here,
  kableExtra,
  knitr,
  lubridate,
  manifestoR,
  quanteda,
  readr,
  tidyr
  #tidytext,
  #topicmodels,
  #LDAvis
  #stm
  )
```

Data acquisition

We'll be using WZB's `manifestoR` package to download and work with the manifesto data. In order to establish a connection with WZB's API, we first need to set our API key.

```
# define API key
mp_setapikey(here("Assignments_SK/Assignment-4/manifesto_apikey.txt"))
```

We limit our analysis to **the United Kingdom** and **the United States**. I examine the period **from XXXX to 2020** (all years for which data are available). Furthermore, I exclude parties with less than XXX over the XXX-year period. In total, the analysis includes **XXX parties**: the United Kingdom (X) and the United States (X). In total, our corpus consists of **XX party manifestos**. We start by checking for the availability of manifestos.

```
# list of countries
countries <- c("United States")

# list of parties
parties <- c("Democratic Party", "Republican Party") # US parties

# check for available docs
```

```
available_docs <- mp_availability(countryname %in% countries &
                                    partyname %in% parties)
```

```
Connecting to Manifesto Project DB API...
Connecting to Manifesto Project DB API... corpus version: 2023-1
Connecting to Manifesto Project DB API...
Connecting to Manifesto Project DB API... corpus version: 2023-1
Connecting to Manifesto Project DB API... corpus version: 2023-1
```

```
# select only annotated docs
annotated_docs <- available_docs |>
  filter(annotations == TRUE)
```

Next, we load our corpus with the manifestos we've identified as relevant.

```
# define file path for corpus
corpus_path <- here("Assignments_SK/Assignment-4/data/corpus.rds")

# download data if doesn't already exist
if (!file.exists(corpus_path)) {

  corpus <- mp_corpus(annotated_docs) # query data # codefilter = 401
  #saveRDS(corpus, corpus_path) # save data
  saveRDS(corpus, corpus_path)

} else corpus <- read_rds(corpus_path) # load data

# TODO: Make sure that no duplicates included
# TODO: Make sure that all texts are in English

# inspect corpus
corpus
```

```
<<ManifestoCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 10
```

```
head(content(corpus[[1]])) # view beginning of text of first manifesto
```

```
[1] "A New Covenant with the American People"
[2] "Two hundred summers ago, this Democratic Party was founded by the man whose burning pen
[3] "In 1992, the party Thomas Jefferson founded invokes his spirit of revolution anew."
[4] "Our land reverberates with a battle cry of frustration that emanates from America's very
[5] "America is on the wrong track."
[6] "The American people are hurting."
```

```r
NLP::meta(corpus[[1]]) # view meta data of first manifesto
```

```
manifesto_id               : 61320_199211
party                      : 61320
date                       : 199211
language                   : english
source                     : MARPOR
has_eu_code                : FALSE
is_primary_doc             : TRUE
may_contradict_core_dataset: FALSE
md5sum_text                : 1e1529bd0d2f579335c027865dfd0d22
url_original               : NA
md5sum_original            : NA
annotations                : TRUE
handbook                   : 1
is_copy_of                 : NA
title                      : Party statement of policies mirrors Clinton’s goals
id                         : 61320_199211
```

Next, we load the main dataset which contains meta data for each manifesto as well as labels for which passages correspond with which topics.

```r
# define file path for main dataset
mp_path <- here("Assignments_SK/Assignment-4/data/mp_data.rds")

# download data if doesn't already exist
if (!file.exists(mp_path)) {

  mp_df <- mp_maindataset() # query data
  saveRDS(mp_df, mp_path) # save data

} else mp_df <- read_rds(mp_path) # load data

# clean main dataset
```

```r
mp_df <- mp_df |>
  filter(countryname %in% countries & partyname %in% parties)

# TODO: Make sure that no duplicates are included

head(mp_df)
```

```
# A tibble: 6 x 175
  country countryname   oecdmember eumember edate        date party partyname
    <dbl> <chr>              <dbl>    <dbl> <date>      <dbl> <dbl> <chr>
1      61 United States          0        0 1920-11-02 192011 61320 Democratic ~
2      61 United States          0        0 1920-11-02 192011 61620 Republican ~
3      61 United States          0        0 1924-11-04 192411 61320 Democratic ~
4      61 United States          0        0 1924-11-04 192411 61620 Republican ~
5      61 United States          0        0 1928-11-06 192811 61320 Democratic ~
6      61 United States          0        0 1928-11-06 192811 61620 Republican ~
# i 167 more variables: partyabbrev <chr>, parfam <dbl>, candidatename <chr>,
#   coderid <dbl>, manual <dbl>, coderyear <dbl>, testresult <dbl>,
#   testeditsim <dbl>, pervote <dbl>, voteest <dbl>, presvote <dbl>,
#   absseat <dbl>, totseats <dbl>, progtype <dbl>, datasetorigin <dbl>,
#   corpusversion <chr>, total <dbl>, peruncod <dbl>, per101 <dbl>,
#   per102 <dbl>, per103 <dbl>, per104 <dbl>, per105 <dbl>, per106 <dbl>,
#   per107 <dbl>, per108 <dbl>, per109 <dbl>, per110 <dbl>, per201 <dbl>, ...
```

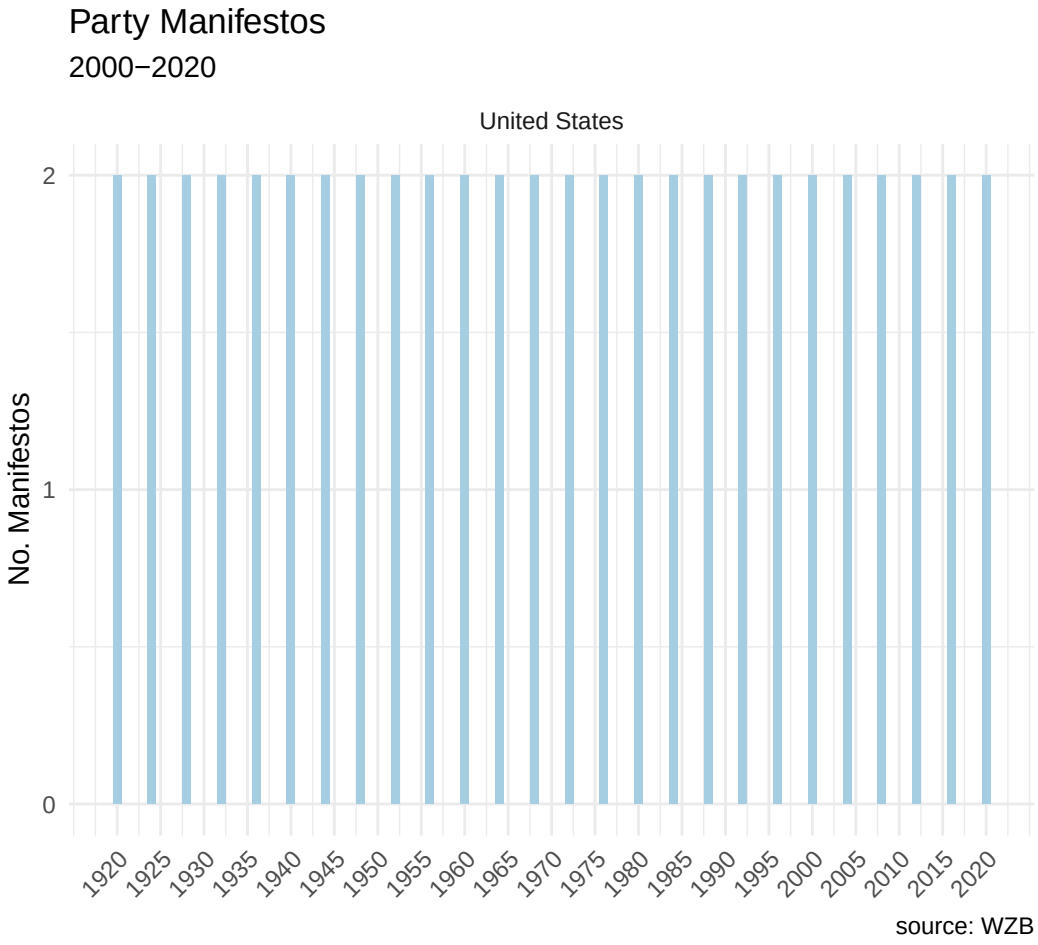
Data description

Here's a breakdown of countries and parties contained in our dataset.

```r
# TODO: make table look nice
# table of countries and parties
mp_df |>
  count(countryname, partyname) |>
  arrange(countryname, desc(n)) |>
  select(-n) |>
  gt(
    rowname_col = "partyname",
    groupname_col = "countryname",
    row_group_as_column = TRUE
  )
```

United States	Democratic Party Republican Party

```r
# TODO: fix plot
# plot number of manifestos by country over time
# TODO: make sure that those with annotations are taken into consideration (corpus vs main
mp_df |>
  mutate(year = year(edate)) |>
  count(year, countryname) |>
  ggplot(aes(x = year, y = n, fill = countryname)) +
  geom_col(width = 1) +
  scale_fill_brewer(palette = "Paired") +
  labs(
    title = "Party Manifestos",
    subtitle = "2000-2020",
    x = NULL,
    y = "No. Manifestos",
    fill = NULL,
    caption = "source: WZB"
  ) +
  facet_wrap(~ countryname, ncol = 1) +
  scale_x_continuous(breaks = seq(1920, 2020, 5)) +
  scale_y_continuous(breaks = seq(0, 6, 1)) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none"
    )
```



We can observe differences in election cycles between the United States and the United Kingdom. Wheras the US has a predictable election cycle with elections scheduled to take place once every four years, the United Kingdom's parliamentarian systems lends itself to a much more erratic election schedule (?). Note that the number of manifestos issued by major parties (as defined per our inclusion criteria) remains the same across time for the United States (2), while this figure fluctuates in the United Kingdom (?).

Data preparation

Next, we prepare our data for topic modeling by transforming our corpus into a document feature matrix. As part of pre-processing, we …. Note that we also…. This may impact our final result by ….

```r
# interested in free market economy
mp_describe_code(code = "401")
```

```
Connecting to Manifesto Project DB API...
code: 401
title: Free Market Economy
description_md: Favourable mentions of the free market and free market capitalism as an
economic model. May include favourable references to:

-   Laissez-faire economy;

-   Superiority of individual enterprise over state and control systems;

-   Private property rights;

-   Personal enterprise and initiative;

-   Need for unhampered individual enterprises.
```

```r
tibble(sentence = content(corpus[[1]]),
       topic_code = codes(corpus[[1]])
       ) |>
  mutate(
    manifesto_id = "placeholder",
    sentence_id = row_number()
    ) |>
  relocate(manifesto_id, contains("sentence"), topic_code)
```

```
# A tibble: 444 x 4
  manifesto_id sentence                                    sentence_id topic_code
  <chr>        <chr>                                             <int> <chr>
1 placeholder  A New Covenant with the American People               1 <NA>
2 placeholder  Two hundred summers ago, this Democratic~            2 202
3 placeholder  In 1992, the party Thomas Jefferson foun~            3 201
4 placeholder  Our land reverberates with a battle cry ~            4 605
5 placeholder  America is on the wrong track.                       5 000
6 placeholder  The American people are hurting.                     6 606
7 placeholder  The American Dream of expanding opportun~            7 410
8 placeholder  Middle class families are working hard, ~            8 603
9 placeholder  Poverty has exploded.                                9 606
```

```
10 placeholder  Our people are torn by divisions.                  10 606
# i 434 more rows
```

```
# initialize empty df to store text data
text_df <- data.frame()

# for each text in corpus....
for (i in 1:length(corpus)) {
  temp_df <- as.data.frame(corpus[[i]], with.meta = TRUE)
  text_df <- bind_rows(text_df, temp_df)
}

text_tidy <- text_df |>
  filter(cmp_code == 401) |>
  mutate(text_id = glue::glue("{manifesto_id}_{pos}")) |>
  select(text_id, manifesto_id, pos, text, cmp_code, party, title, date) |>
  distinct()

head(text_tidy)
```

```
          text_id manifesto_id pos
1  61320_199211_50 61320_199211  50
2 61320_199211_223 61320_199211 223
3 61320_199211_305 61320_199211 305
4 61320_199211_314 61320_199211 314
5 61320_200411_538 61320_200411 538
6 61320_200411_539 61320_200411 539

1
2
3
4                                                We support tenant management and ownership,
5                                                                                          W
6 Government's responsibility is to create an environment that will promote private sector i
  cmp_code party
1      401 61320
2      401 61320
3      401 61320
4      401 61320
5      401 61320
6      401 61320
                                                                                    title
```

```
1                         Party statement of policies mirrors Clinton's goals
2                         Party statement of policies mirrors Clinton's goals
3                         Party statement of policies mirrors Clinton's goals
4                         Party statement of policies mirrors Clinton's goals
5 Strong at Home, Respected in the World. The Democratic Platform for America
6 Strong at Home, Respected in the World. The Democratic Platform for America
    date
1 199211
2 199211
3 199211
4 199211
5 200411
6 200411
```

Next, we perform pre-processing on our text data.

```
library(quanteda)

# pre-processing
dfmat <- text_tidy$text |>
  tokens(
    remove_punct = TRUE,
    remove_numbers = TRUE,
    remove_symbols = TRUE
    ) |> # TODO: remove more?
  tokens_remove(pattern = stopwords("en")) |>
  tokens_wordstem() |>
  dfm()  # TODO: dfm_trim more?
  #dfm_trim(min_termfreq = 5) # remove infrequent terms

rownames(dfmat) <- text_tidy$text_id # add IDs to row names

head(dfmat)
```

```
Document-feature matrix of: 6 documents, 1,397 features (99.30% sparse) and 0 docvars.
                 features
docs              believ free enterpris power market forc privat sector engin
  61320_199211_50      1    1         1     1      1    1      0      0     0
  61320_199211_223     0    0         0     0      0    0      1      1     1
  61320_199211_305     0    0         0     0      0    0      0      0     0
  61320_199211_314     0    0         0     0      0    0      0      0     0
  61320_200411_538     1    0         0     0      0    0      1      1     1
```

```
  61320_200411_539      0    0         0     0      0    0      1      1     0
                 features
docs              economi
  61320_199211_50        0
  61320_199211_223       1
  61320_199211_305       0
  61320_199211_314       0
  61320_200411_538       0
  61320_200411_539       1
[ reached max_nfeat ... 1,387 more features ]
```

2. Research question

> Describe a research question you want to explore with topic modelling. Comment on how answerable this is with the methods and data at your disposal.

For my research question, I am interested in exploring the extent to which climate-related issues were included in various parties' manifestos. Specifically, I am interested in whether a party's willingness to express a stance on climate change is more related to their political leaning or the country.

3. Topic model development

> Create a topic model using your data. Explain to a non-specialist what the topic model does. Comment on the choices you make here in terms of hyperparameter selection and model choice. How might these affect your results and the ability to answer your research question?

```r
library(topicmodels)
#library(lda)
#library(LDAvis)

# define number of topics
n_topics <- 5

tictoc::tic() # timer start
lda <- LDA(dfmat, n_topics) # TODO: experiment with number of topics
tictoc::toc() # timer stop
```

```
0.164 sec elapsed
```

```r
library(tidytext)
print(dim(lda@gamma))
```

```
[1] 642   5
```

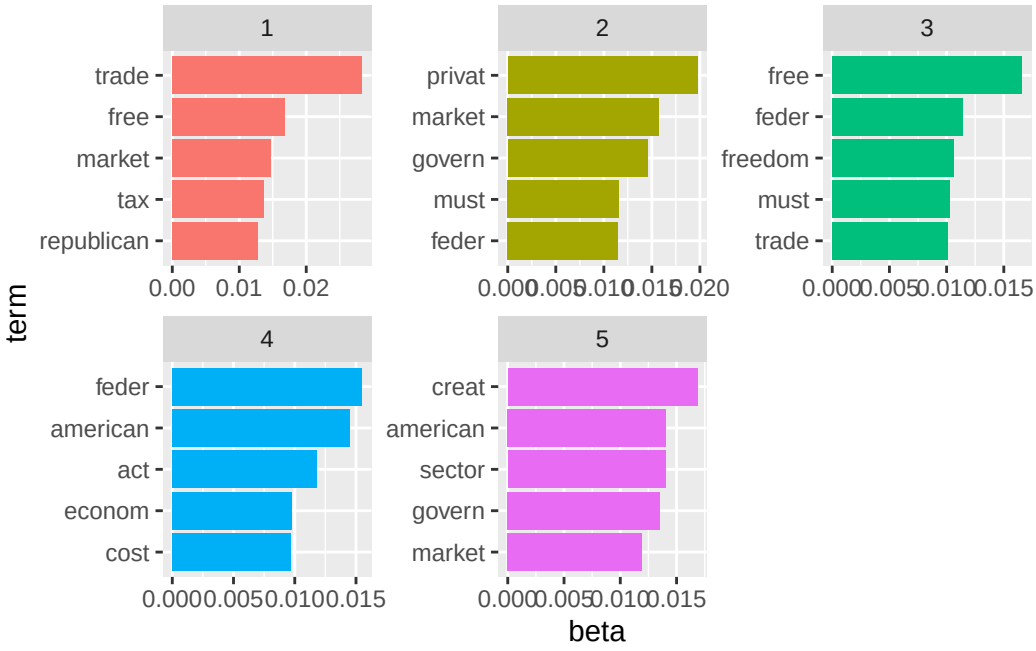
```r
print(dim(lda@beta))
```

```
[1]    5 1397
```

```r
topic_words <- tidy(lda, matrix = "beta") |>
  group_by(topic) |>
  slice_max(beta, n = 5) |>
  ungroup() |>
  arrange(topic, -beta)
topic_words
```

```
# A tibble: 25 x 3
   topic term         beta
   <int> <chr>       <dbl>
 1     1 trade      0.0283
 2     1 free       0.0168
 3     1 market     0.0146
 4     1 tax        0.0137
 5     1 republican 0.0128
 6     2 privat     0.0197
 7     2 market     0.0157
 8     2 govern     0.0145
 9     2 must       0.0116
10     2 feder      0.0115
# i 15 more rows
```

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




```
ggsave(glue::glue("img/top_terms_{n_topics}.png"), width = 12, height = 8)
```

```
# manifesto_topics <- tidy(lda, matrix = "gamma") |>
#   left_join(text_tidy, by = c("document" = "manifest_id"))
#
# yearly_topics <- manifesto_topics |>
#   #filter(country == "United States") |>
#   mutate(year = year(date)) |>
#   group_by(year, party, topic) |>
#   summarise(gamma = sum(gamma)) |>
#   group_by(year, party) |>
#   mutate(year_share = gamma/sum(gamma)) |>
#   ungroup() |>
#   mutate(topic = factor(topic))
#
# yearly_topics
#
# yearly_topics |>
#   filter(topic %in% 1:5) |>
#   ggplot(aes(x = year, y = year_share, group = topic, colour = topic, fill = topic)) +
#   geom_line() +
```

```
#   facet_wrap(vars(party), ncol = 1) +
#   theme_minimal()
#
# # save plot
# ggsave(glue::glue("img/topic_groups_{n_topics}.png"), width = 12, height = 8)
# #
# library(LDAvis)
# topicmodels2LDAvis <- function(x, ...){
#  post <- topicmodels::posterior(x)
#  if (ncol(post[["topics"]]) < 3) stop("The model must contain > 2 topics")
#  mat <- x@wordassignments
#  json <- LDAvis::createJSON(
#    phi = post[["terms"]],
#    theta = post[["topics"]],
#    vocab = colnames(post[["terms"]]),
#    doc.length = slam::row_sums(mat, na.rm = TRUE),
#    term.frequency = slam::col_sums(mat, na.rm = TRUE)
#  )
#  return(json)
# }
# json <- topicmodels2LDAvis(lda)
# serVis(json)
```

```
#The definition of a "good" topic is not universal but task dependent.

#A good topic model is one that helps us to answer the research question we have in mind,

#Sometimes we want the big picture (few topics), sometimes we want fine-grained detail (ma

#What we should ensure is that any results we present are not an artefact of an arbitrary
```

```
# evaluate performance
#A quick heuristic for naming topics is to concatenate the top 3 terms for each topic.
#If we want to use our model then we should give meaningful names to topics by inspecting

#Loss/heldout-likelihood based measures work by comparing the topic predictions of words i

#Coherence based measures work by assessing whether the words in a topic are similar. If w
```

4. Topic model description

Describe the topic model. What topics does it contain? How are these distributed across the data?

5. Answering your research question

Use your topic model to answer your research question by showing plots or statistical results. Discuss the implications of what you find, and any limitations inherent in your approach. Discuss how the work could be improved upon in future research.

Sources

- Lehmann, Pola / Franzmann, Simon / Burst, Tobias / Regel, Sven / Riethmüller, Felicia / Volkens, Andrea / Weßels, Bernhard / Zehnter, Lisa (2023): The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2023a. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB) / Göttingen: Institut für Demokratieforschung (IfDem). https://doi.org/10.25522/manifesto.mpds.2023a

Resources

- `manifestoR` vignette
- https://manifesto-project.wzb.eu/tutorials/primer
- https://manifesto-project.wzb.eu/tutorials/main-dataset
- https://manifesto-project.wzb.eu/tutorials/firststepsmanifestoR