Analyse de données textuelles

Svetlana Zhuk

1. Introduction

En 1982, la Loi constitutionnelle de 1982 est adoptée par le Canada sans le consentement du Québec. Brian Mulroney, député du Parti progressiste-conservateur, promet au Québec une stratégie visant à l'intégrer dans l'Acte constitutionnel et à reconnaître son caractère distinct. En 1984, Brian Mulroney remporte les élections et forme le 33e Parlement. Notre objectif est d'analyser le premier mois de travail de la 33e législature et d'identifier les sujets les plus discutés. Notre question de recherche est donc la suivante :

- Quelles ont été les principales discussions menées par Brian Mulroney au cours du premier mois de la 33e législature ?

Pour répondre à cette question, nous utilisons des ensembles de données parlementaires contenant des transcriptions de débats fournies par LIPAD. Le mois qui nous intéresse est novembre 1984.

2. Les données et nettoyage

Les données utilisées sont des débats transcrits du Parlement du Canada, ce qui nous permet d'effectuer une analyse des données textuelles. Pour mieux comprendre les discussions politiques, nous utiliserons le lexicoder dictionary.

Ce dictionnaire facilite l'analyse automatisée des textes et est une source ouverte et gratuite. Pour permettre une analyse plus nuancée qui tienne compte du contexte, nous avons également développé la catégorie "Québec" afin d'examiner la fréquence des discussions liées à la situation du Québec au Parlement.

Dans les sections suivantes, nous détaillons les processus de nettoyage des données, de fusion des ensembles de données, d'intégration des dictionnaires et de représentation graphique.

```
Package version: 3.3.1
Unicode version: 14.0
ICU version: 71.1
```

```
Parallel computing: 8 of 8 threads used.

See https://quanteda.io for tutorials and examples.

Skipping install of 'clessnverse' from a github remote, the SHA1 (aa697496) has not changed s
  Use `force = TRUE` to force installation

DISCLAIMER: As of July 2023, `clessnverse` is no longer under active development.
To avoid breaking dependencies, the package remains available "as is" with no warranty of any

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x ggplot2::%+%()  masks crayon::%+%()
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
#dictionary

lexicoder_en <- dictionary(file = "~/Dropbox/fas_1001_Zhuk/_tp/_tp3/dictionary/policy_agen
```

```r
#new_dictionary_Québec

quebec_dictionary <- list(quebec = c("sign", "patriation", "referendum", "veto", "federali
  dictionary()
```

```r
#fusionner deux dictionnaires

lexicoder_en_m <- lexicoder_en |> stack()
quebec_dictionary_m <- quebec_dictionary |> stack()
new_dictionary <- bind_rows(lexicoder_en_m, quebec_dictionary_m) |>
  unstack(values ~ ind) |>
  dictionary()
```

```
Rows: 258 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 23 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (11): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (2): subsubtopic, speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 316 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 450 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 329 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
```

```
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 389 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 263 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 397 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 285 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
```

```
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 355 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 367 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 365 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 425 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
```

```
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 297 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 386 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 444 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 328 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 485 Columns: 15
-- Column specification --------------------------------------------------------
```

```
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 467 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (12): hid, pid, opid, speakeroldname, speakerposition, maintopic, subto...
dbl   (1): basepk
lgl   (1): speakerriding
date  (1): speechdate

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
#fusion des bases de donnes

data_parl_merged <- bind_rows(
  data_parl_1, data_parl_2, data_parl_3, data_parl_4, data_parl_5, data_parl_6, data_parl_
                              data_parl_9, data_parl_10, data_parl_11, data_parl_12, data_
                              data_parl_15, data_parl_16, data_parl_17, data_parl_18, data
```

```r
data_parl_clean <- data_parl_merged |>
  select(speechdate, speechtext, speakername) |>
  mutate(speechtext = tolower(speechtext)) |>  #lowercase
  na.omit() # remove missing variables
```

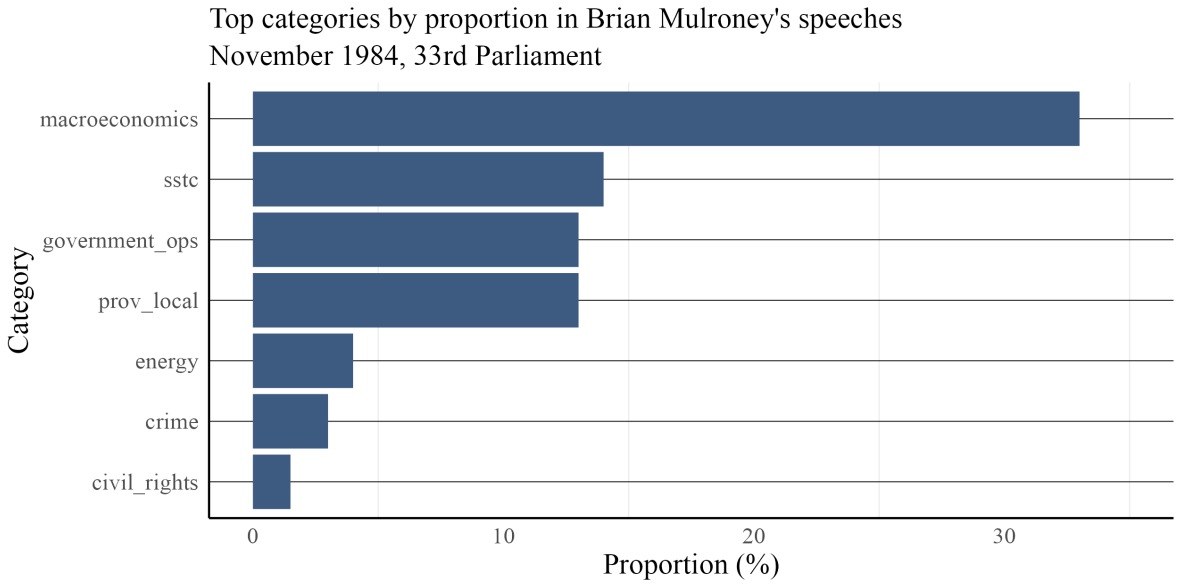
```r
#analyse du dictionnaire

nov_1984 <- run_dictionary(data = data_parl_clean,
               text = speechtext,
               dictionary = new_dictionary) |>
  bind_cols(data_parl_clean) |>
  select(-c(doc_id, speechtext)) |>
  pivot_longer(!c(speechdate, speakername),
               names_to = "category",
               values_to = "n") |>
```

```
    ungroup() |>
    na.omit() |>
    filter(speakername == "Martin Brian Mulroney") |>
    group_by(category) |>
    summarise(total = sum(n, na.rm = TRUE)) |>
    ungroup() |>
    mutate(prop = round(total/sum(total), 4)*100) |>
    slice_max(total, n = 10)
```

```
100% expressions/words found
```

```
1.117 sec elapsed
```

3. Résultats



Dans nos résultats, nous mettons en évidence les principales catégories discutées par Brian Mulroney au cours du premier mois de novembre, qui marque le début de la 33e législature après qu'il a formé un gouvernement. Le sujet le plus fréquemment abordé est la macroéconomie, probablement lié aux discussions sur les accords de libre-échange avec les États-Unis. Cela conduit à des discussions plus approfondies et à des analyses détaillées dans la catégorie "macroéconomie". Cependant, la catégorie "Québec", un autre domaine d'intérêt, n'est pas apparue dans les discussions les plus importantes. Il est probable que les mots-clés que nous avons sélectionnés pour la catégorie "Québec" n'ont pas permis de capturer de manière adéquate les discussions relatives à l'accommodement du Québec. Par conséquent, il est nécessaire d'affiner la catégorie "Québec" dans notre dictionnaire.

Bibliographie

- Beelen, K., Thijm, T. A., Cochrane, C., Halvemaan, K., Hirst, G., Kimmins, M., Lijbrink, S., Marx, M., Naderi, N., Rheault, L., Polyanovsky, R., and Whyte, T. (2017). "Digitization of the Canadian Parliamentary Debates." Canadian Journal of Political Science, 50(3), 849–864. https://doi.org/10.1017/S0008423916001165.
- Albugh, Quinn, Julie Sevenans and Stuart Soroka. 2013. Lexicoder Topic Dictionaries, June 2013 versions, McGill University, Montreal, Canada.