Introduction aux mégadonnées en sciences sociales (FAS 1001)

Cours 6: Analyses textuelles automatisées partie 2

Nadjim Fréchet

13 février 2024

Retour sur le TP2 📄

- Pas de panique tout le monde!
- On me prévient dans les temps si on a un problème (code ou `GitHub`). Rendez-vous, si vous pensez que le problème est plus sérieux.
- Pour le code, souvent, la plupart des solutions à vos problèmes se touvent également en ligne. **D'autres codeurs ont eu les mêmes problèmes que vous**.
- `GitHub` fait également partie de l'apprentissage…

Invité surprise pour le dernier cours (26 mars)

- Nael Shiab: Journaliste de données à CBC/Radio-Canada

Invité surprise pour le dernier cours (26 mars)

- Je vous invite à consulter son **portfolio**
- Ou son **GitHub**…

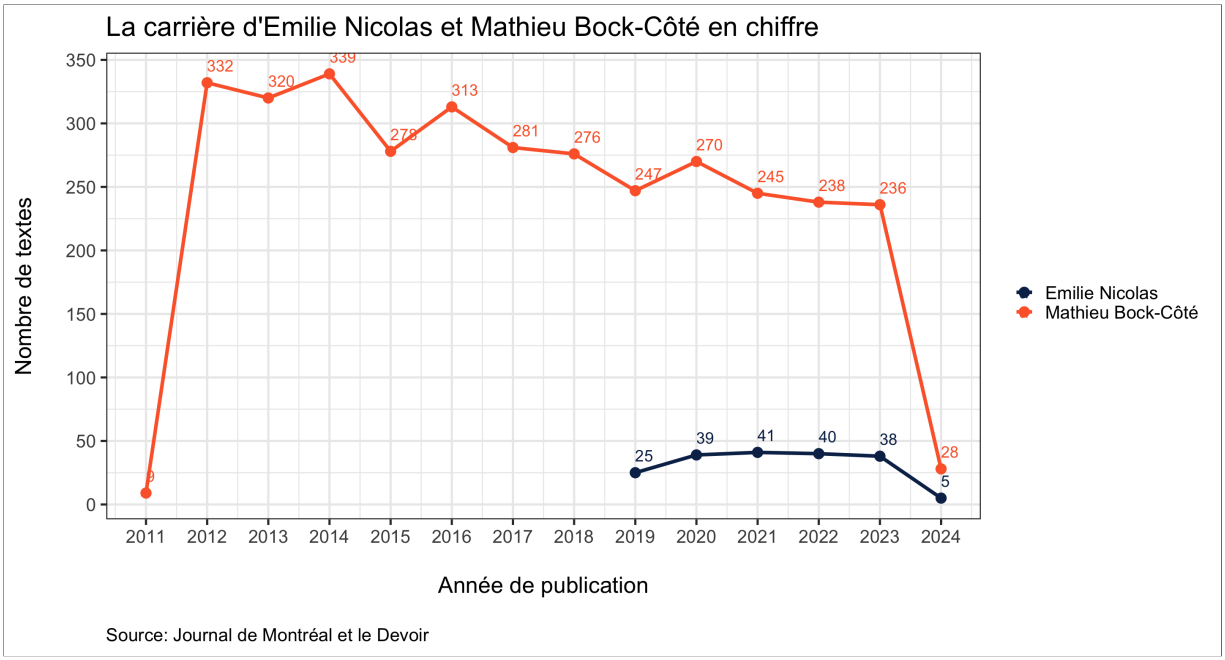
Aucune lecture cette semaine 📚

- Avez-vous profité de cette pause pour lire 😅? Notamment Grimmer et Stewart (**2013**)?
- J'ai vu que certains d'entre vous se sont attaqués aux exercices 😮
- Comment trouvez-vous les exercices?

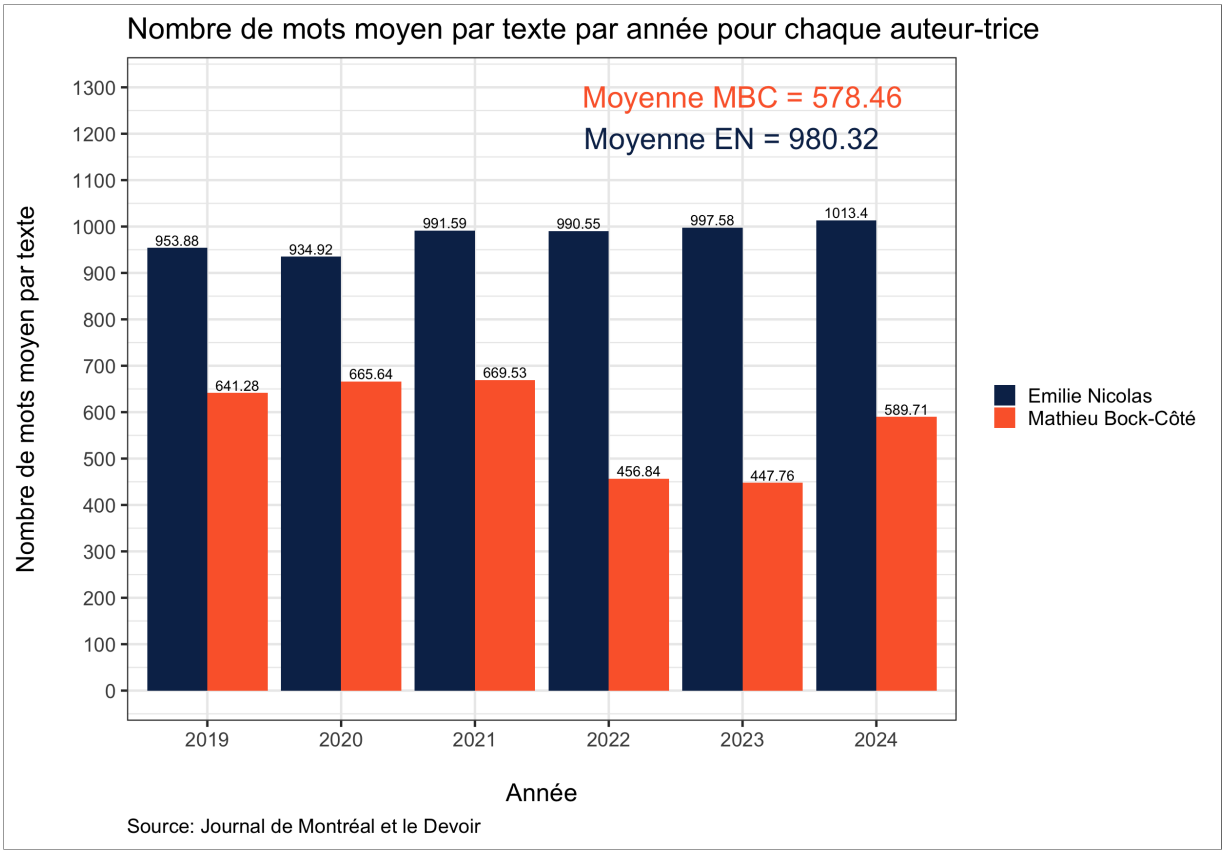
Retour sur l'analyse du dictionnaire

- Exemple à partir de données scrapées du ***Devoir*** et du ***Journal de Montréal***.
- Plus précisément, deux chroniqueur chroniqueuse associé-es à la droite et à la gauche au Québec…

Chroniqueur et chroniqueuse aux carrières différentes



Longueur moyenne des textes



D'abord le dictionnaire: 1ère façon

```
# 0.1 - Package ----

library(quanteda)    # Package pour le lire
library(clessnverse) # Package pour utiliser plus facilement un dic

# 0.2 - Dictionnaire ----

lexicoder_en <- dictionary(file = "_dictionnaires/policy_agendas_en

lexicoder_en
```

```
Dictionary object with 28 key entries.
- [macroeconomics]:
  - aggregate demand, aggregate supply, business cycle, demand shock,
demand side, demand-side, econom, employment rate, full employment,
food price, industr, keynes, bank of canada, bank of england, bear
market, bretton woods, budget, bull market, changing demographic,
coinage [ ... and 62 more ]
- [civil_rights]:
  - civil right, ableism, abortion, access to info, african american,
anti-choice, anti-semit, bill 101, charter of the french language,
biphobi, bisexual, charter of rights, civil libert, disabilit,
discriminat, diversity, equal employm, equal opportunit, equal right,
equalit [ ... and 65 more ]
- [healthcare]:
  - aids, alcoholism, allerg, anaesthesiolog, anesthesiolog, cancer,
cardiolog, cardiothoracic, cardiovascular, cigarette, dermatolog,
```

Deuxième façon

```
lexicoder_fr <- read_csv("_dictionnaires/lexicoder_merged.csv") |> 
  select(categorie, traductionGoogle) |>
  unstack(traductionGoogle~categorie) |>  # On transforme la catégo
  dictionary() # On transforme le tout en dictionnaire avec la fonc

lexicoder_fr
```

```
Dictionary object with 28 key entries.
- [aboriginal]:
  - aborigène, algonquin, amérindien, anishinaabe, assiniboine,
beothuk, cayuga, chipewyan, comos, cowichan, cri, dunneze, première
nation, premières personnes, gwich'in, haida, huron, acte indien,
affaire indienne, innu [ ... and 33 more ]
- [agriculture]:
  - agricole, bétail, cultiver, grain, blé, orge, du boeuf, porc, la
volaille, tracteur, battre, verger, nourriture inspecter, ferme,
importation de nourriture, aquacole, pied et bouche, surgir,
agroalimentaire, pesticide [ ... and 8 more ]
- [civil_rights]:
  - droit civil, le capacitisme, avortement, accès à l'information,
afro-américain, anti-choix, antisémite, facture 101, charte de la
langue française, biphobi, bisexuel, charte des droits, liberté
civile, handicapé, discriminer, la diversité, travail égal, égalité
```

Troisième façon

```
cultural_value_dictionary <- list(racisme = c("racisme* systémique*
                                  nationalisme = c("souveraineté du
  dictionary() # On transforme le tout en dictionnaire avec la fonc

cultural_value_dictionary
```

```
Dictionary object with 2 key entries.
- [racisme]:
  - racisme* systémique*, personne* racisée*, personne* de couleur,
ethni*, discrimination raciale*, préjugé* racia*, islamophobie*,
antisémitisme*, colonialisme*, xénophobie*, islam*, musulman*
- [nationalisme]:
  - souveraineté du québec, nationalisme*, patriotisme*,
chauvinisme*, autodétermination*, séparatisme*, souverainisme*
```

Fusionner deux dictionnaires

```
# 0.3 - Fusionner deux dictionnaires ----

cultural_value_dictionary_m <- cultural_value_dictionary |> stack()

lexicoder_fr_m <- lexicoder_fr |> stack() # Même chose ici

new_dictionary <- bind_rows(cultural_value_dictionary_m, lexicoder_
  unstack(values~ind) |>  # Tansformation en list()
  dictionary() # Puis en dictionnaire

new_dictionary
```

```
Dictionary object with 30 key entries.
- [racisme]:
  - racisme* systémique*, personne* racisée*, personne* de couleur,
ethni*, discrimination raciale*, préjugé* racia*, islamophobie*,
antisémitisme*, colonialisme*, xénophobie*, islam*, musulman*
- [nationalisme]:
  - souveraineté du québec, nationalisme*, patriotisme*,
chauvinisme*, autodétermination*, séparatisme*, souverainisme*
- [aboriginal]:
  - aborigène, algonquin, amérindien, anishinaabe, assiniboine,
beothuk, cayuga, chipewyan, comos, cowichan, cri, dunneze, première
nation, premières personnes, gwich'in, haida, huron, acte indien,
affaire indienne, innu [ ... and 33 more ]
- [agriculture]:
  - agricole, bétail, cultiver, grain, blé, orge, du boeuf, porc, la
volaille, tracteur, battre, verger, nourriture inspecter, ferme,
```

On essaye le tout

```r
# 1.1 - Fusion des données ----

Chr_data <- bind_rows(Mbc_data, Nicolas_data) |>
  filter(year >= 2019) |>       # Filtre pour 2019 et plus
  select(author, text) |> # On garde les variables pertinentes
  mutate(text = tolower(text))  # minuscules

glimpse(Chr_data)
```

```
Rows: 1,452
Columns: 2
$ author <chr> "Mathieu Bock-Côté", "Mathieu Bock-Côté", "Mathieu
Bock-Côté", …
$ text   <chr> "je croyais l’indépendance absolument nécessaire, mais
j’avais …
```

Analyse du dictionnaire

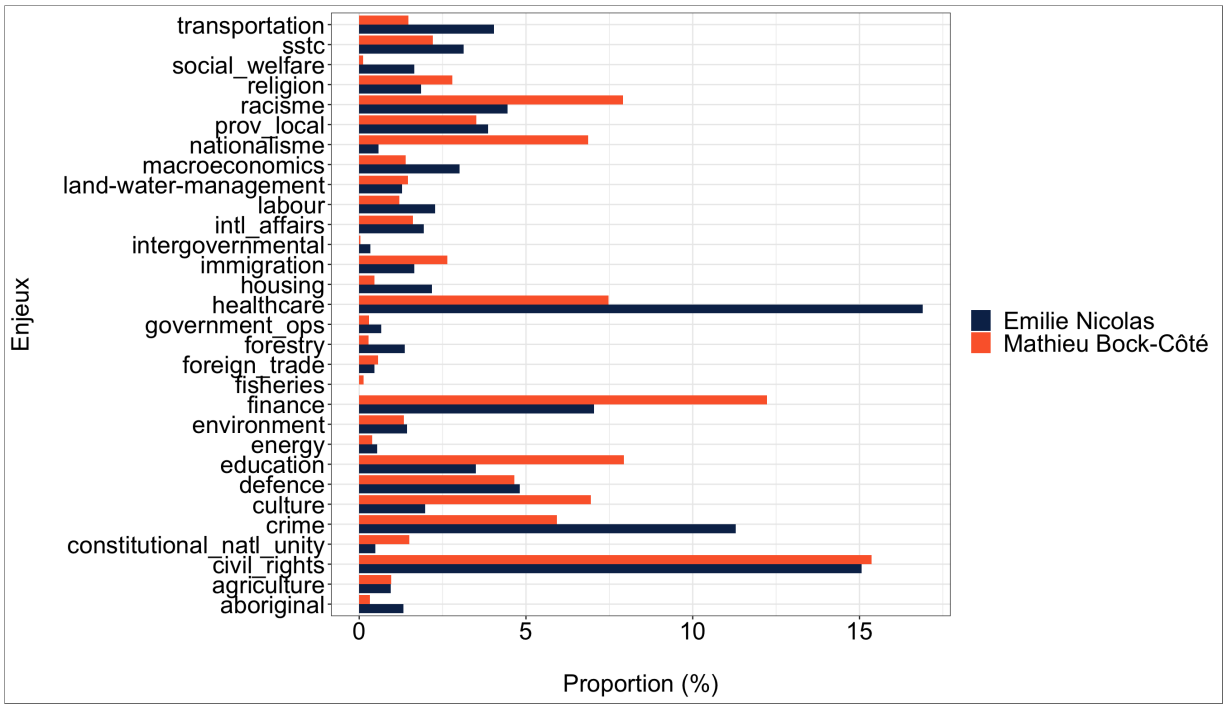
```
# 1.2 - Analyse du dictionnaire ----

Data_dictionary <- run_dictionary(data = Chr_data, text = text, dic
```

```
0.953 sec elapsed
```

```
# 1.3 - On fusionne de nouveau les bases de données

Data_dictionary_clean <- Data_dictionary |>
  bind_cols(Chr_data) |>
  select(-c(doc_id, text)) |>  # On retire les variables qu'on a pl
  pivot_longer(!c(author), names_to = "categorie", values_to="n") |
  ungroup() |>
  group_by(author, categorie) |>
  summarise(n=sum(n)) |>
  mutate(prop = round(n/sum(n),4)*100)

glimpse(Data_dictionary_clean)
```

```
Rows: 60
Columns: 4
Groups: author [2]
$ author    <chr> "Emilie Nicolas", "Emilie Nicolas", "Emilie
Nicolas", "Emili…
$ categorie <chr> "aboriginal", "agriculture", "civil_rights",
"constitutional…
$ n         <dbl> 32, 23, 366, 12, 274, 48, 117, 85, 13, 35, 171, 0,
11, 33, 1…
$ prop      <dbl> 1.32, 0.95, 15.07, 0.49, 11.29, 1.98, 4.82, 3.50,
0.54, 1.44…
```

Graphique


transportation
sstc
social_welfare
religion
racisme
prov_local
nationalisme
macroeconomics
land-water-management
labour
intl_affairs
intergovernmental
immigration
housing
healthcare
government_ops
forestry
foreign_trade
fisheries
finance
environment
energy
education
defence
culture
crime
constitutional_natl_unity
civil_rights
agriculture
aboriginal
Enjeux
0
5
10
15
Proportion (%)
Emilie Nicolas
Mathieu Bock-Côté

Pourquoi utiliser des analyses plus automatisées? 📈

- Limite du dictionnaire: contexte, oui. Mais aussi la possibilité de se tromper.
- Dictionnaire peu pertinent pour le corpus de textes en question.
- Pourquoi ne pas laisser le corpus de textes parler pour nous?
- Analyses non-supervisées: **topic modeling (LDA, STM, autres)** (voir **Blei, Ng, et Jordan 2003**; **Roberts et al. 2014**), **wordfish** (voir **Slapin et Proksch 2008**) et ***Word embeddings***, autres.

Topic modeling

- **On assume**: on estime un nombre K de sujet(s) dans notre corpus de texte. Chaque document est fait par une personne qui décide quel sujet il ou elle aborde dans le texte (les mots et termes ne sont pas là par hasard).

```
library(stm)
library(tidytext)
library(quanteda)

# 2.1 - Un nettoyage plus poussé est nécessaire ----

stopwords_new <- sprintf("\\b(%s)\\b", paste0(c(quanteda::stopwords

Mbc_topics <- Mbc_data |>
  filter(year == 2023) |> # Regardons les sujets pour l'année 2023
  mutate(text = tolower(text),
         text = str_remove_all(text, stopwords_new),
         text = str_squish(str_remove_all(text, "[[:punct:]]"))) |>
  corpus() |>
  tokens() |>
  dfm() |>
  dfm_trim(min_termfreq = 2, min_docfreq = 2) # On va être picky on

Mbc_topics
```

```
Document-feature matrix of: 236 documents, 5,053 features (97.21%
sparse) and 8 docvars.
       features
docs    jetons abord œil europe vieux monde semble jamais emporté
décadence
  text1      1     1   1      2     1     1      1      1       1
1
  text2      0     1   0      1     0     1      1      0       0
0
  text3      0     0   0      0     0     5      0      0       0
0
  text4      0     0   0      0     0     1      0      0       0
0
  text5      0     0   0      0     1     0      0      2       0
0
  text6      0     0   0      0     0     0      0      0       0
```

Topic modeling

```
# 2.2 - Topic models ----

set.seed(123)

topic_model <- stm(Mbc_topics, K = 6, verbose = F)

summary(topic_model) # on visualise le tout
```

```
A topic model with 6 topics, 236 documents and a 5053 word
dictionary.

Topic 1 Top Words:
     Highest Prob: monde, pays, guerre, histoire, vie, politique,
fois
     FREX: sanitaire, conflit, américains, indiana, jones, paix,
russie
     Lift: admirable, attaché, climatique, confinement, confinements,
covid, écolos
     Score: jones, sanitaire, russie, conflit, passeport, russes,
poutine
Topic 2 Top Words:
     Highest Prob: pays, vie, français, france, droit, islamisme,
dire
     FREX: islamisme, terrorisme, islamistes, islamiste, roi, juifs,
verre
     Lift: boire, compagne, françaises, french, historien, livreur,
no
```

Et Emilie Nicolas?

```
A topic model with 6 topics, 38 documents and a 2294 word dictionary.

Topic 1 Top Words:
     Highest Prob: femmes, hommes, féminisme, monde, femme,
politique, plusieurs
     FREX: féminisme, féministes, femmes, hommes, femme, féministe,
mouvements
     Lift: emblée, étrangères, féminine, féminisme, féministe,
prévention, sexisme
     Score: intersectionnelle, féminisme, féministes, femmes,
féministe, hommes, mouvements
Topic 2 Top Words:
     Highest Prob: politique, canada, pays, médias, monde, années,
depuis
     FREX: médias, pays, démocratie, canadien, entreprises, paroles,
information
     Lift: chine, creuse, doigt, égard, empire, géants, opinions
     Score: paroles, démocratie, médias, entreprises, canadien,
```

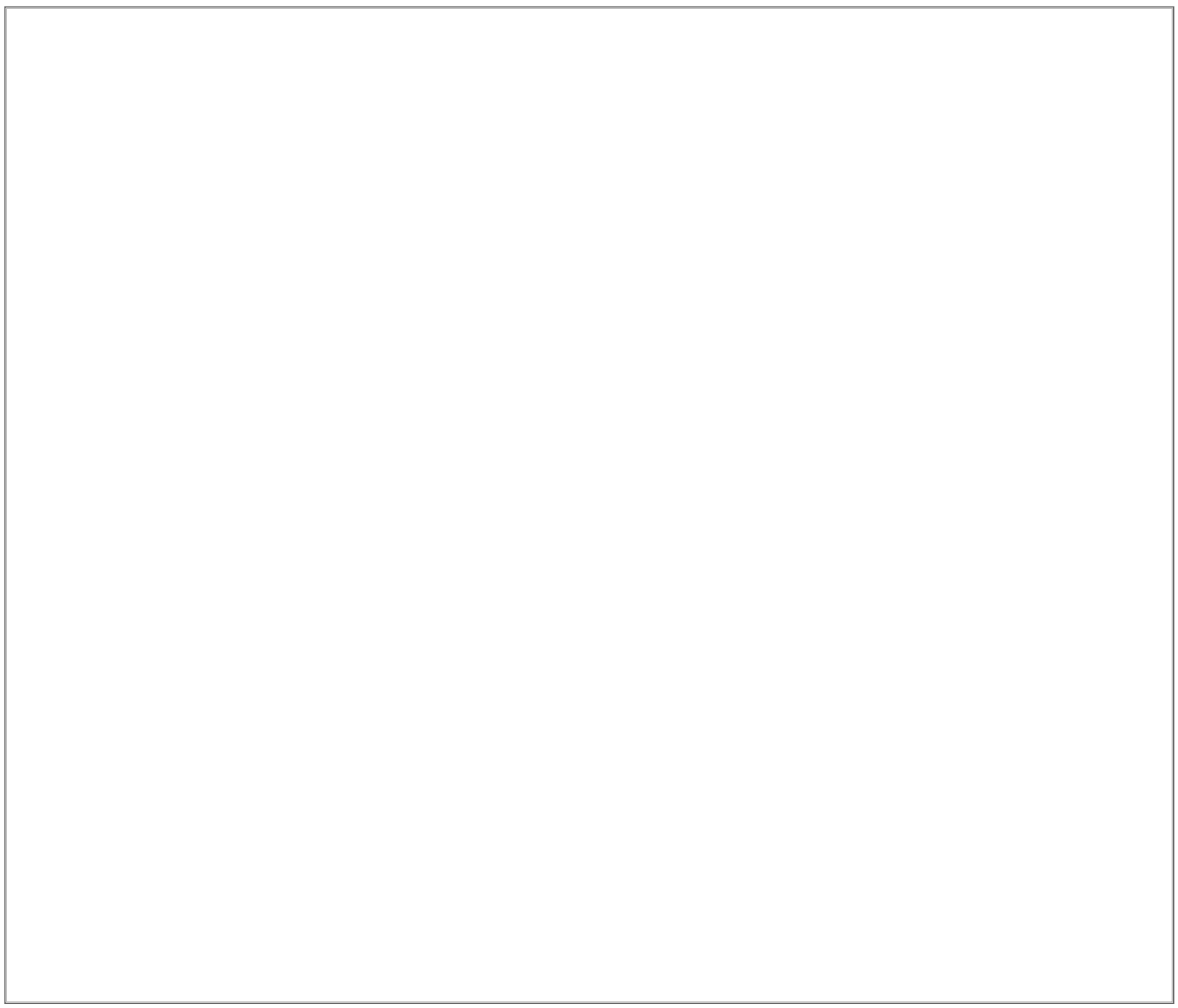
Exemple de représentation visuelle

Word embeddings

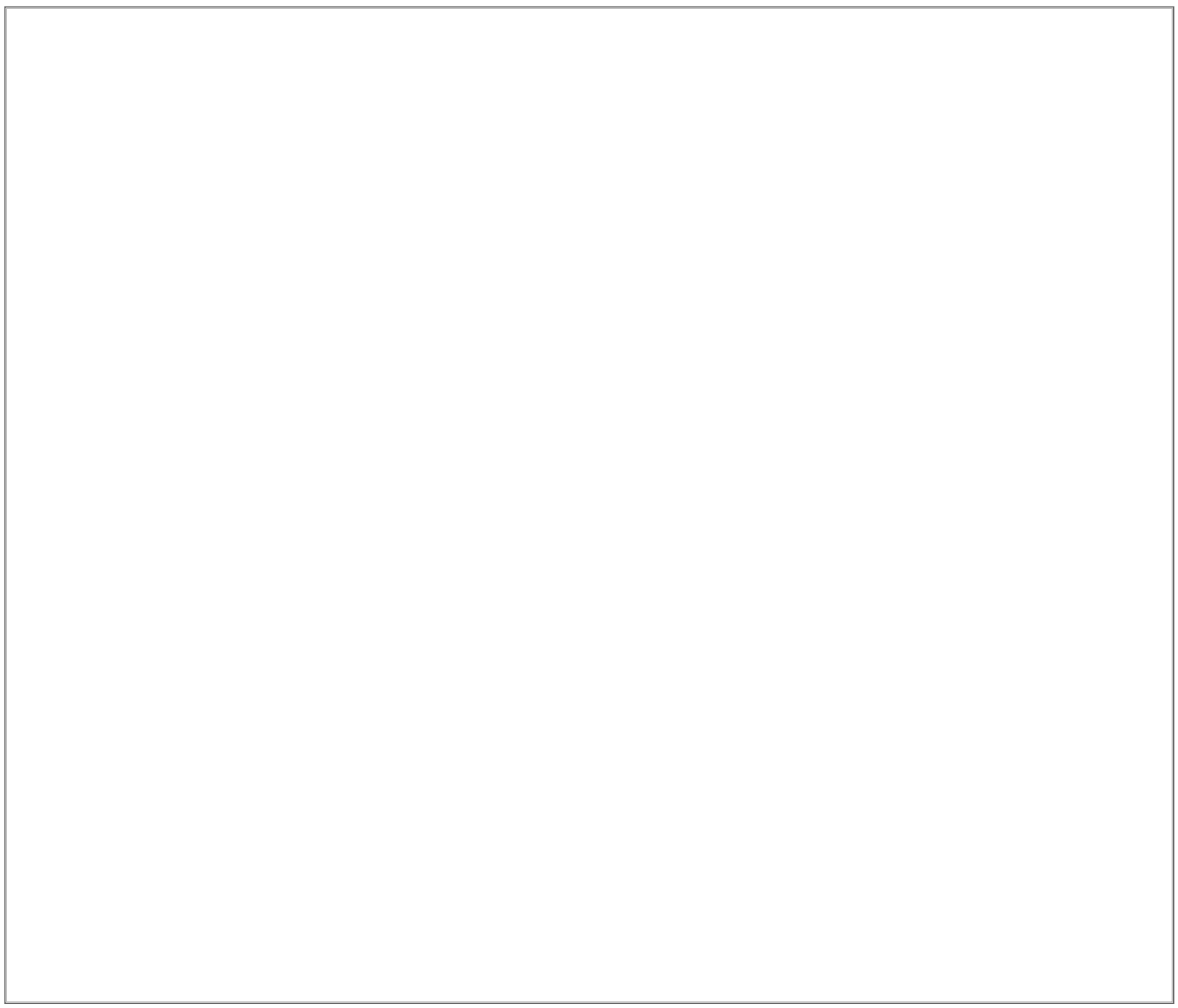
```
library(word2vec)
set.seed(123)

Test_mbc <- Mbc_data |>
    filter(year == 2023) |> # Regardons les sujets pour l'année 202
    select(text) |>
    mutate(text = tolower(text),
           text = str_squish(str_remove_all(text, stopwords_new)))
    corpus() |>
    corpus_reshape(to = "sentences") |>
    tokens(remove_punct = T, remove_symbols = T) |>
    as.list()

model_mbc_emb <- word2vec(Test_mbc, dim = 25, iter = 20, min_count
predict(model_mbc_emb, c("wokisme", "racisme", "québec"), type = "n
```

```
$wokisme
    term1         term2 similarity rank
1 wokisme aboutissement  0.8700056    1
2 wokisme       passion  0.8652310    2
3 wokisme       terrain  0.8559136    3
4 wokisme    socialisme  0.8543373    4
5 wokisme         cessé  0.8516707    5

$racisme
    term1      term2 similarity rank
1 racisme xénophobie  0.9131467    1
2 racisme systémique  0.8966184    2
3 racisme  antiblanc  0.8894600    3
4 racisme   surprise  0.8795394    4
5 racisme  courageux  0.8783245    5
```

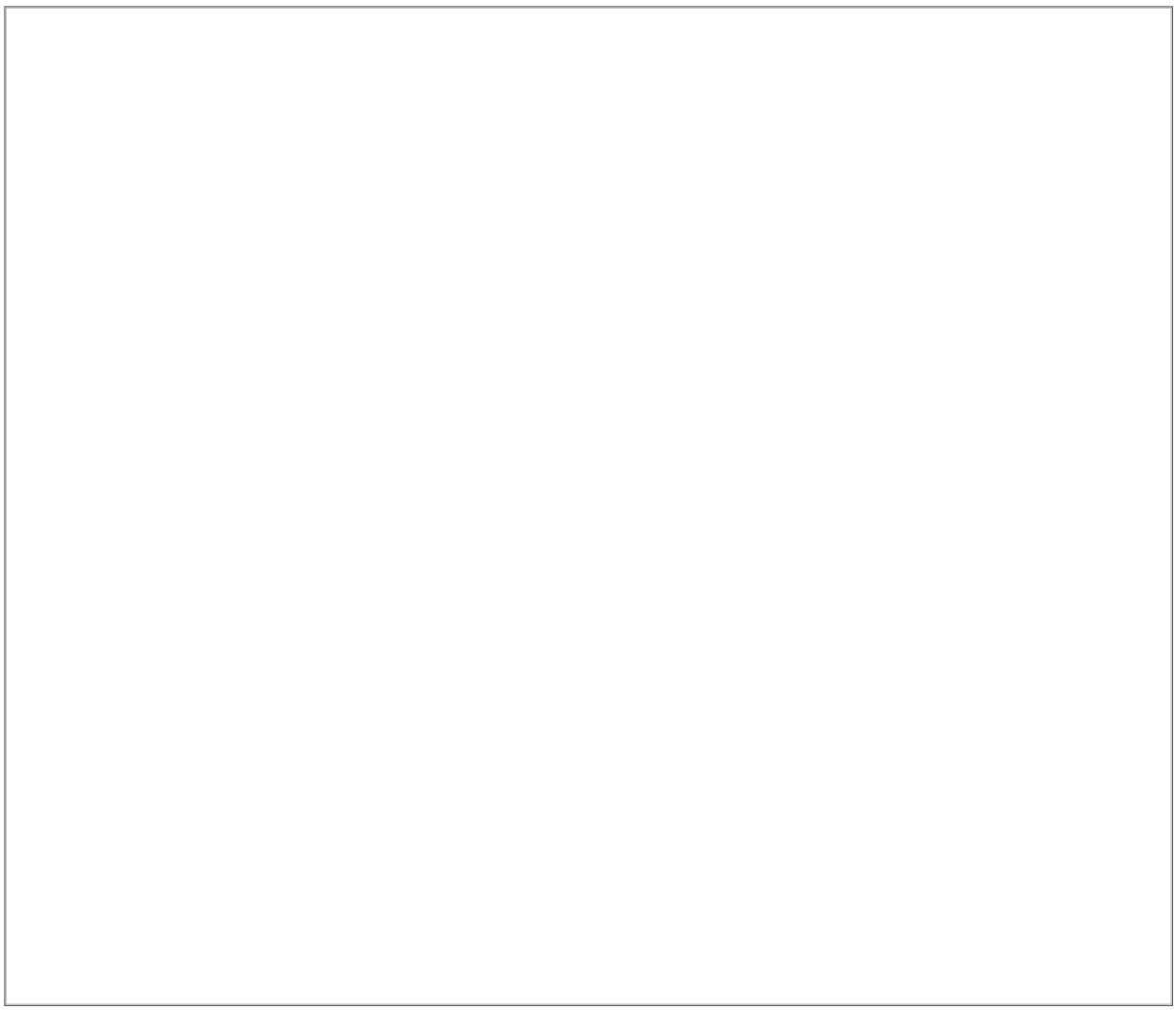
Exemple tiré de mes travaux



Échelle de pessimisme vs optimisme par rapport à la COVID-19



Algorithme de mise en échelle wordfish (question ouverte)



Questions + travail personnel 📚

Bibliographie

Blei, David M, Andrew Y Ng, et Michael I Jordan. 2003. « Latent Dirichlet Allocation ». *Journal of Machine Learning Research* 3 (Jan): 993-1022.

Grimmer, Justin, et Brandon M Stewart. 2013. « Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts ». *Political Analysis* 21 (3): 267-97.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, et David G Rand. 2014. « Structural Topic Models for Open-ended Survey Responses ». *American journal of political science* 58 (4): 1064-82.

Slapin, Jonathan, et Sven-Oliver Proksch. 2008. « A Scaling Model for Estimating Time-series Party Positions from Texts ». *American Journal of Political Science* 52 (3): 705-22.