TP3

Sadou Ba

Introduction

Avec la montée du Parti Québécois dans les sondages depuis plusieurs semaines (Girard-Bossé, 2024), on pourrait se demander si un nouveau débat sur la question référendaire pourrait avoir lieu au sein du pays dans les prochaines années. Ainsi, après s'être posé la question, il serait intéressant de prévoir les éléments de langage des différents camps sur l'échiquier politique pour démystifier les discours. Ainsi, pour tenter d'anticiper les futurs discours, quoi de mieux que de se baser sur le passé. C'est en analysant les documents de la chambre des communes pendant la période référendaire de 1995 que l'on se familiarisera mieux avec les discours fédéralistes et indépendantistes. Nous présenterons la méthodologie utilisée ainsi que les résultats que nous obtiendrons. Et nous allons conclure sur l'efficacité de la méthode employé.

Méthodologie

Les données que nous avons utilisées proviennent du Lipad une initiative de digitalisation des retranscrit de l'ensemble des débats parlementaires de la chambre des communes effectué par des chercheurs en science politique de l'université de Toronto. Plus précisément, nous avons analysé les données concernant la retranscription des débats parlementaire du 27 octobre 1995 soit la dernière séance avant la date du référendum qui était le 30 octobre 1995. Les données recueillies ont été entièrement nettoyé et traités sur R studio. Après avoir importé la base de données, nous avons installé les paquetages nécessaires à l'analyse de textuelle soit le *quanteda* et le *stm.* Nous avons d'abord débuté en créant un corpus qui aura pour but de stocker nos données textuelles dans un environnement qui permettra par la suite de les convertir en DFM (Document feature matrix). Le DFM nous servira a séparé chaque mot dans une colonne distincte ce qui facilitera l'analyse textuelle non supervisé que l'on voudra faire. Pour créer le corpus, il faut indiquer une clé primaire qui est l'identifiant de chaque prise de parole en entrée ainsi qu'une variable textuelle dans ce cas-ci, il s'agit des discours des membres de la chambre. Ainsi, après avoir créer le corpus, la prochaine étape est de transformer notre corpus en DFM et en même temps que nous faisons cela, il est possible d'inclure dans le formatage du code un code pour nettoyer le texte. Pour ce faire, il sera question de mettre en minuscules tous les caractères, ensuite de se débarrasser de tous les symboles et de toutes les ponctuations

qui sont aussi considérés comme des caractères. Puis pour finir, il faut enlever l'ensemble des stopwords qui sont des mots du type préposition ou conjonction, soit des mots qui se répètent souvent, mais qui n'apporte pas de réel bénéfice à notre analyse textuelle, au contraire, cela pourrait même la fausser. Et puis finalement, nous avons indiqué que nous voulons que les mots présents dans le DFM soient au moins répétés 2 fois et dans 2 document différent. Ayant désormais un DFM il ne nous restait plus qu'à effectuer une analyse et dans ce cas-ci, il s'agit du topic modeling qui consiste à effectuer du machine learning ou le programme va définir la connexion entre les mots par rapport à leur répétition dans dans le DFM et ainsi définir une tendance et regrouper les mots qui vont ensemble sous différents topics. Dans notre cas, nous avons demandé 3 topics en raison du nombre de mots disponibles (Alexander, 2023).

Résultats

```
> labelTopics(topics_bef_ref)
Topic 1 Top Words:
        Highest Prob: small, business, businesses, loans, program, government, sbla
        FREX: small, business, businesses, loans, program, sbla, cost
        Lift: 12, 500, agencies, asset, centre, closure, companies
        Score: sbla, small, businesses, loans, business, cost, lenders
Topic 2 Top Words:
        Highest Prob: canada, mr, speaker, quebec, minister, people, canadians
        FREX: together, united, love, party, health, regional, national
        Lift: 18, affect, aspects, counsel, democracy, distinct, doubt
        Score: together, quebecers, proud, rally, french, environmental, english
Topic 3 Top Words:
        Highest Prob: mr, speaker, government, can, yes, house, quebec
        FREX: answer, democratic, position, unanimous, far, yes, report
        Lift: achieved, collective, council, domestic, far, life, listen
        Score: unanimous, parliamentary, democratic, quebecers, pursuant, p.m, hon
>
```

Conclusion

Comme nous pouvons le voir ci-dessus, le topic modeling que nous avons effectué a donné comme résultat trois topics en sortis. Dans le premier topic, nous pouvons voir que ce qui est abordé est le champ lexical de l'économie. Notamment en parlant d'entreprises, de prêt, de fermetures de coûts ou encore de programme. Ceci, pourrait être dû à la crainte de la fermeture d'entreprises dans le cas d'un référendum ou le oui l'emporterait ce qui explique le sentiment d'inquiétude face à l'économie.

Le deuxième topic est ce que l'on pourrait caractériser comme le champ lexical du camp du non. On le voit en observant la présence des mots ramenant à l'unité nationale soit rallye qui réfère au « love in » (Radio-Canada, 2010), unité, Canada et Québec.

Le troisième topic est le contraire du deuxième, soit une présence des mots faisant référence au oui. Notamment le « Yes », Québec, unanime, québécois.

En somme, nous voyons que trois sujets distincts ont pu être dégagés et que parfois certains mots se trouvaient dans plusieurs topics à l'instar de Québec, ce qui pourrait être expliqué par le fait que ces topics étaient des sous-sujet d'un sujet général qui était le référendum.

Pour conclure, nous ne pouvons nous empêcher de penser que nous aurions peut-être dû utiliser une autre analyse qui aurait été plus efficace pour la visualisation des données. Soit une analyse par dictionnaire ou par fréquence de mots.

Bibliographie

Rohan Alexander. (2023). *Telling stories with data*, https://rohanalexander.github.io/telling_stories-published/

Girard-Bossé, 2024, **Le Parti québécois conforte son avance**, https://www.lapresse.ca/actualites/politique/2024-02-06/sondage-leger/le-parti-quebecois-conforte-son-avance.php

Radio-Canada, 2010, https://ici.radio-canada.ca/nouvelle/457258/rdi-15-love-in

Annexe

```
library(quanteda)
install.packages("stm")
library(stm)
library(readr)
install.packages("wordcloud")
library(wordcloud)
before_referendum <- read_csv("_tp/_tp3/1995-10-27.csv")

before_referendum_clean <- before_referendum %>% select(speechtext
                                                        , speakerparty)


bef_ref_corpus <- corpus(before_referendum,
                         docid_field = "basepk",
                         text_field = "speechtext")

bef_ref_dfm <-
  bef_ref_corpus %>%
  tokens(
    remove_punct = TRUE,
    remove_symbols = TRUE
  ) %>%
  dfm(tolower = TRUE) %>%
  dfm_trim(min_termfreq = 2, min_docfreq = 2) %>%
  dfm_remove(stopwords(source = "snowball"))
```

```
# J'ai voulu faire 3 graphiques de wordcloud, mais je n'y suis pas parvenu.
# Etant donné que je ne voyais pas la visualisation pour le topic modeling, j'ai
# eu beaucoup de difficulté à trouver une bonne visualis

topics_bef_ref <- stm(documents = bef_ref_dfm, K = 3)

labelTopics(topics_bef_ref)

topic_1 <- c("bussiness", "loans", "program", "closure", "companies", "small",
             "lenders", "cost")
topic_2 <- c("Canada", "quebec", "regional", "national", "united", "rally", "french",
             "english", "together")
topic_3 <- c("yes", "unanimous", "far", "quebec", "government", "position", "achieved")

wordcloud(topic_1)
wordcloud(topic_2)
wordcloud(topic_3)
```