Introduction à l'analyse textuelle avec R

Marc-Antoine Martel
Université de Montréal
Janvier 2024

Plan de la présentation

- L’analyse de données textuelles comme méthode de recherche
- Avantages/limites
- Quelques exemples de recherches
- Démonstration dans R (librairie Quanteda)

L’analyse de contenu

- Méthode très utilisée en sciences sociales
- Permet d’analyser le comportement des acteurs politiques
  - Traitement de l’information par les médias
  - Stratégies de communication déployées par les partis politiques
  - Circulation de la désinformation sur les médias sociaux
- Analyse manuelle & analyse automatisée

Les analyses automatisées

- **Caractéristiques d'un corpus de textes**
  - Longueur des textes
  - Occurrence des mots
  - Registres de langue
- **Dictionnaires**
  - Lexiques associant une série de mots à des thèmes ou des attributs (enjeux politiques, tonalité, cadrages)
- **Combiner des dictionnaires**
  - Exemple: tonalité et enjeux.

L’analyse textuelle: quelques étapes à prévoir

1. Collecte des données
2. Structuration du corpus
3. Construction du dictionnaire (optionnel)
4. Production des résultats
5. Visualisation des résultats
6. Analyse des résultats
    a. Description des tendances
    b. Intégration des variables à des modèles de régression
7. Rédaction

Avantages

- Permet l’analyse d’un grand volume de données
- Offre la possibilité de reproduire les recherches facilement
- Représente une option économique

Limites

- **À considérer**
  - autobus
  - tabac
  - enjeux polysémiques
- Métaphores, sarcasme…
- Créer un dictionnaire exhaustif peut être chronophage
- N’est pas un remplacement à l’analyse qualitative. Attention à l’interprétation. Importance de la validation. Attrait des méthodes mixtes.

Données à analyser

...tous les écrits disponibles au format numérique!

- Articles de journaux
- Plateformes électorales
- Communiqués de presse
- Publications sur les médias sociaux (Facebook, Twitter, YouTube, etc.)
- Pages web, blogs, etc.
- Transcriptions (période des questions à l'Assemblée nationale)

Construction d'un dictionnaire

- Critères de sélection des mots à intégrer au dictionnaire
- Désuétude
- Méthodes supervisées & non-supervisées (topic-models)

Exemples de recherches

- Niveau de politisation des nouvelles sur la COVID-19 en fonction de la visibilité des acteurs

Science Communication
Volume 42, Issue 5, October 2020, Pages 679-697

https://doi.org/10.1177/1075547020950735

Research Note

Politicization and Polarization in COVID-19 News Coverage

P. Sol Hart [1], Sedona Chinn [2], and Stuart Soroka [1]

Abstract

This study examines the level of politicization and polarization in COVID-19 news in U.S. newspapers and televised network news from March to May 2020. Using multiple computer-assisted content analytic approaches, we find that newspaper coverage is highly politicized, network news coverage somewhat less so, and both newspaper and network news coverage are highly polarized. We find that politicians appear in newspaper coverage more frequently than scientists, whereas politicians and scientists are more equally featured in network news. We suggest that the high degree of politicization and polarization in initial COVID-19 coverage may have contributed to polarization in U.S. COVID-19 attitudes.

Keywords

COVID-19, politicization, polarization, media coverage, risk communication

Exemples de recherches

- Transcription des interventions de citoyens lors d'assemblées de village.
- Capacité à influencer la conversation
  - Comparaison du sujet abordé par une personnes aux sujets abordés lors des interventions subséquentes.

American Political Science Review (2019) 113, 3, 623–640
doi:10.1017/S0003055419000182

Deliberative Democracy in an Unequal World: A *Text-As-Data* Study of South India's Village Assemblies

RAMYA PARTHASARATHY *Stanford University*
VIJAYENDRA RAO *World Bank*
NETHRA PALANISWAMY *World Bank*

This paper opens the "black box" of real-world deliberation by using text-as-data *methods on a corpus of transcripts from the constitutionally mandated* gram sabhas, *or village assemblies, of rural India. Drawing on normative theories of deliberation, we identify empirical standards for "good" deliberation based on one's ability both to speak and to be heard, and use natural language processing methods to generate these measures. We first show that, even in the rural Indian context, these assemblies are not mere "talking shops," but rather provide opportunities for citizens to challenge their elected officials, demand transparency, and provide information about local development needs. Second, we find that women are at a disadvantage relative to men; they are less likely to speak, set the agenda, and receive a relevant response from state officials. And finally, we show that quotas for women for village presidencies improve the likelihood that female citizens are heard.*

Exemples de recherches

Revue canadienne de science politique (2023), **56**, 26–48
doi:10.1017/S0008423922000907

CAMBRIDGE UNIVERSITY PRESS

ÉTUDE ORIGINALE / RESEARCH ARTICLE

Couverture médiatique des enjeux électoraux au Québec de 1994 à 2018

Marc-Antoine Martel[1]* et Richard Nadeau[2]

- Visibilité médiatique des enjeux en campagne électorale.

Élection	La Presse	Le Devoir
1994	476	286
1998	445	287
2003	495	237
2007	578	296
2008	401	257
2012	396	350
2014	652	303
2018	379	229

Exemples de recherches

Revue canadienne de science politique (2023), **56**, 26–48
doi:10.1017/S0008423922000907

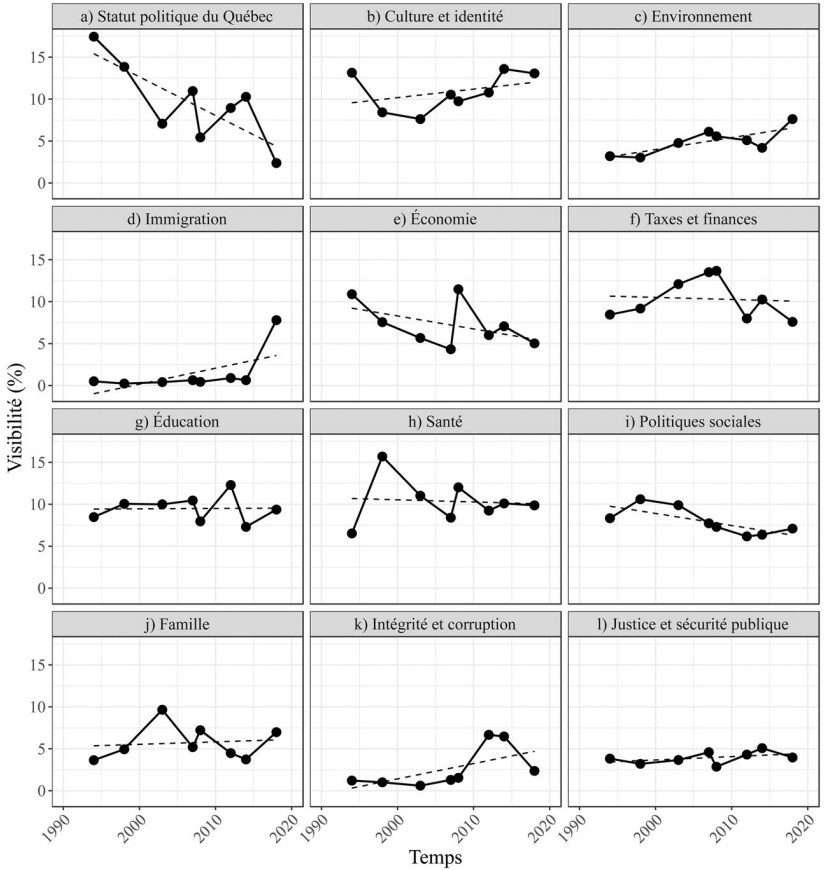
CAMBRIDGE UNIVERSITY PRESS

ÉTUDE ORIGINALE / RESEARCH ARTICLE

Couverture médiatique des enjeux électoraux au Québec de 1994 à 2018

Marc-Antoine Martel[1]* et Richard Nadeau[2]

- Visibilité médiatique des enjeux en campagne électorale.



Exemples de recherches

Revue canadienne de science politique (2023), **56**, 26–48
doi:10.1017/S0008423922000907

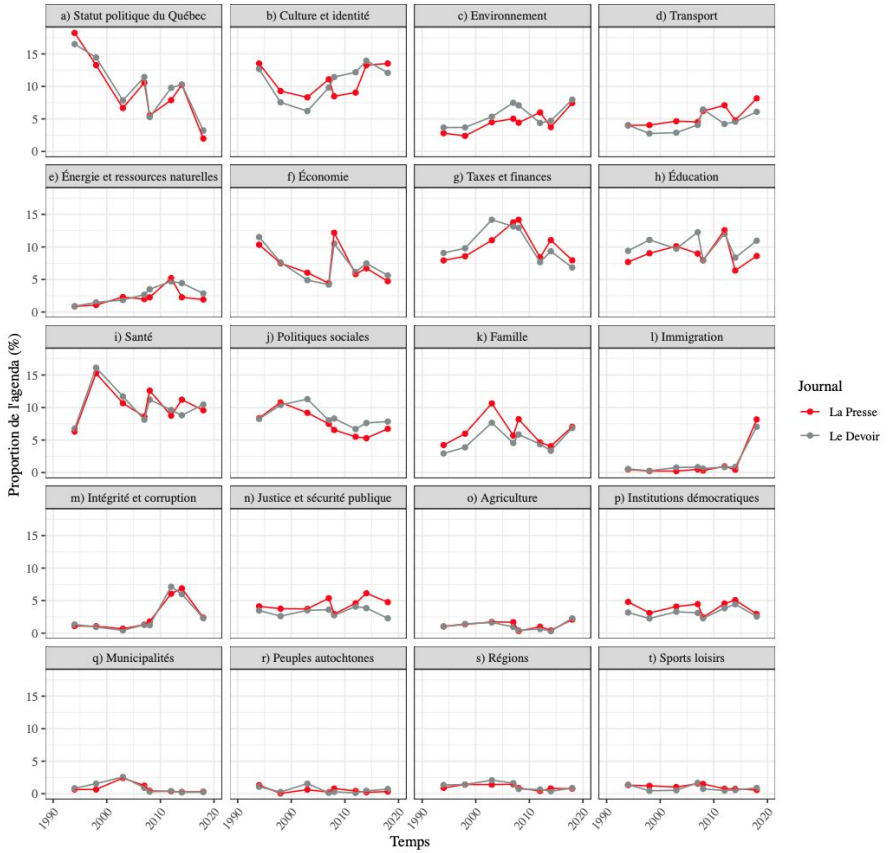
CAMBRIDGE UNIVERSITY PRESS

ÉTUDE ORIGINALE / RESEARCH ARTICLE

Couverture médiatique des enjeux électoraux au Québec de 1994 à 2018

Marc-Antoine Martel[1]* et Richard Nadeau[2]

- Visibilité médiatique des enjeux en campagne électorale.

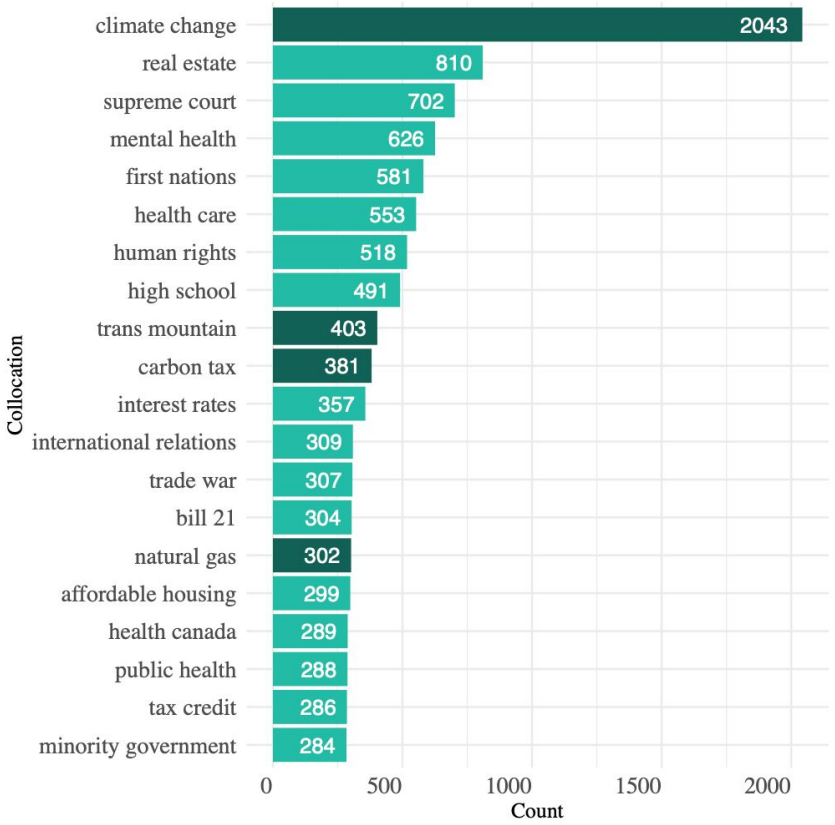


Exemples de recherches

What is a climate election? The case of the 2019 Federal Election in Canada (Nadeau, Lachapelle, Martel)

Top 20 Collocations Associated with Political Issues in Media Campaign Coverage

N = 11 653 articles de journaux



Ressources

Collecte de données

- X/Twitter: https://developer.twitter.com/en/docs/twitter-api
- Facebook: https://www.crowdtangle.com/
- Articles de journaux: http://eureka.cc/fr/
- Archives de sites web: https://archive.org/web/
- Collection de l'Université Laval: https://www.poltext.org/fr

R

- Web scraping: rvest, Selector Gadget, ParseHub.
- Packages: tm, tidytext, Quanteda (https://tutorials.quanteda.io/)

Recommandations de lecture

- Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. Text as data: A new framework for machine learning and the social sciences. Princeton University Press, 2022.
- 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.
- Young, Lori, et Stuart Soroka. 2012. « Affective News: The Automated Coding of Sentiment in Political Texts ». Political Communication 29 (2) : 205-31. https://doi.org/10.1080/10584609.2012.671234.

Atelier

- [Cliquez ici] pour installer R et Rstudio
- [Cliquez ici] pour accéder aux documents prévus pour l'atelier.
- Sources
  - https://quanteda.io/
  - Duval, D. and Pétry, F. (2016) "L'analyse automatisée du ton médiatique : construction et utilisation de la version française du Lexicoder Sentiment Dictionary", Revue canadienne de science politique, 49(2), pp. 197–220.
- Partage d'écran

Merci!

- N’hésitez pas à me contacter pour discuter!
- X: https://twitter.com/_MAMartel
- Courriel: marc-antoine.martel@umontreal.ca