Master Thesis Proposal

Manon Costa, Sébastien Gadat, Christophe Gaillac

September 2024

Student profile: Master 2 student in Applied Probability, Statistics, Data Science. Programming skills in Python will be appreciated.

Keywords: Machine Learning, Applied Probability and Statistics, Text Analysis

Advisor and contact:

- Manon Costa, manon.costa@math.univ-toulouse.fr
- Sébastien Gadat, sebastien.gadat@tse-fr.eu

Working place: Toulouse School of Economics & Institut de Mathématiques de Toulouse

Length: 5 to 6 months, starting in February/March 2025

Remuneration: Standard internship grant : 550 euros/month.

Context This internship proposal is part of the research project funded by the Tiris grant "*Text-Ecop*" (Scaling-up Science program), a pluri-disciplinary research project between Toulouse School of Economics and Université Paul Sabatier. The project deals with the use of applied mathematics, computer science and A.I. on textual data analysis for applications in various fields, including political economy as a specific focus in *Text-Ecop*.

Objectives and Contents of the Internship: The goal of the project is to develop relevant probability models of texts for document classification, to build some new statistical/econometric dissymmetry indicators, and to study the statistical robustness associated to these indicators when observing some empirical datasets.

The starting point is related to [1], which primarily considers both word count X and party affiliation Y of parliamentary speeches recorded in the *United States Congressional Record*, from the 43rd Congress to the 114th Congress. The authors introduce a multinomial distribution for the random variable $X|Y$ that highly depends on a choice probability that behaves differently according to Y, with the help of a logistic generalized linear model. During this internship, we are mainly interested in:

- Define some polarization indicators from the multinomial model, built on the posterior distributions of $Y|X$. These indicators may be derived from the law Y —X instead, where X might be a more general text representation (embeddings, see Chapter 6 of [3]), using transport distances among others, and some other divergences related to distributions. These tools will allow to perform inference on the evolution of the selected notion of polarization while being more credible. We also refer to [2] for complementary considerations on distances among documents.
- Use some different probability choices functions that will generalize the one introduced in [1], using for example extended GLM and more sophisticated A.I. methods.
- Implement some statistical and machine learning methods on the dataset of [1], `https://home.heinonline.org/content/U-S--Congressional-Documents/` using Python. The data have already been cleaned up (no data engineering is planned during this internship) and is easily accessible, which allows to start both theoretical and practical investigations rapidly.

Associated to the internship project, a PhD grant may be foreseen according to the evolution of the research project, the different opportunities of fundings in Toulouse and the motivation of the student.

Any student interested in this research project shall contact us to plan a short zoom meeting interview.

References

[1] Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. Econometrica, 87(4), 1307-1340.

[2] Matt Kusner, Yu Sun, Nicholas Kolkin, Kilian Weinberger (2015). From Word Embeddings To Document Distances. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:957-966.

[3] Jurafsky, D. and Martin, J. H. (2019). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed.). Draft available at https://web.stanford.edu/~jurafsky/slp3/