Advanced Text-As-Data: Word Embeddings, Deep Learning, and Large Language Models

Winter School - Iesp UERJ

Tiago Ventura and Sebastian Vallejo Vera

Course Description

In recent years, the surge in the availability of textual data, ranging from the digitalization of archival documents, political speeches, social media posts, and online news articles, has led to a growing demand for statistical analysis using large corpora. Once dominated by sparse bag-of-words models, the field of Natural Language Processing (NLP) is now increasingly driven by dense vector representations, deep learning, and the rapid evolution of Large Language Models (LLMs). This course offers an introduction to this new generation of models, serving as hands-on approach to this new landscape of computational text analysis with a focus on political science research and applications.

The class will cover four broad topics. We start with an overview of how to represent text as data, from a sparse representation via bag-of-words models, to a dense representation using word embeddings. We then discuss the use of deep learning models for text representation and downstream classification tasks. From here, we will discuss the foundation of the state-of-art machine learning models for text analysis: transformer models. Lastly, we will discuss several applications of Large Language Models in social science tasks.

The course will consist of lectures and hands-on coding in class. The lecture will be conducted in English, but students are free to ask questions in Portuguese. Students will have time in the afternoon to practice the code seen in class, and we will suggest additional coding exercises. We assume students attending this class have taken, at a minimum, an introductory course in statistics and have basic knowledge of probability distributions, calculus, hypothesis testing, and linear models. The course will use a mix of R and Python, two computational languages students should be familiar with. That being said, students should be able to follow the course even if they are just starting with any of the two programming languages.

Instructors

Tiago Ventura

- Assistant Professor in Computational Social Science, Georgetown University
- Pronouns: He/Him
- Email: tv186@georgetown.edu

Sebastian Vallejo

- Assistant Professor in the Department of Political Science at the University of Western Ontario
- Pronouns: He/Him
- Email: sebastian.vallejo@uwo.ca

Required Materials

Readings: We will rely primarily on the following textbook for this course. The textbook is freely available online

- Daniel Jurafsky and James H. Martin. Speech and Language Processing, 3nd Edition. - [SLP]

The weekly articles are listed in the syllabus

Schedule & Readings

Day 1: Text Representation: Sparse & Dense Vectors. Deep Learning Models for Text Analysis

- **Readings**:
  - SLP: Chapters 7
  - Lin, Gechun, and Christopher Lucas. "An Introduction to Neural Networks for the Social Sciences." (2023)

Day 2: Word Embeddings: Theory and Applications

- **Theory Papers**:
  - SLP Chapter 6, Vector Semantics and Embeddings
  - Meyer, David. How Exactly Does Word2Vec Work?
  - Spirling and Rodriguez, Word embeddings: What works, what doesn't, and how to tell the difference for applied research
- **Applied Papers**:
  - Austin C. Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." American Sociological Review 84, no. 5: 905–49. https://doi.org/10.1177/0003122419877135
  - Garg, Nikhil, Londa Schiebinger, Dan Jurafsky and James Zou. 2018. "Word embeddings quantify 100 years of gender and ethnic stereotypes." Proceedings of the National Academy of Sciences 115(16):E3635–E3644.
  - Rodman, E., 2020. A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. Political Analysis, 28(1), pp.87-111
  - Rheault, Ludovic, and Christopher Cochrane. "Word embeddings for the analysis of ideological placement in parliamentary corpora." Political Analysis 28, no. 1 (2020): 112-133.
  - Gennaro, Gloria, and Elliott Ash. "Emotion and reason in political language." The Economic Journal 132, no. 643 (2022): 1037-1059.

Day 3: Transformes: Theory and Fine-tuning a Transformers-based model

- **Theory Papers**
  - [SLP] - Chapter 10.
  - Jay Alammar. 2018. "The Illustrated Transformer."
  - Vaswani, A., et al. (2017). Attention is all you need. Advances in neural information processing systems, 30.
  - Timoneda and Vallejo Vera (2025). BERT, RoBERTa or DeBERTa? Comparing Performance Across Transformer Models in Political Science Text. Journal of Politics.
- **Applied Papers**

- Vallejo Vera et al. (2025). Semi-Supervised Classification of Overt and Covert Racism in Text. Working Paper.

Day 4: Large Language Models: Social Science Applications

- **Applied Papers**
  - Wu, Patrick Y., Joshua A. Tucker, Jonathan Nagler, and Solomon Messing. "Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting." arXiv preprint arXiv:2303.12057 (2023).
  - Rathje, Steve, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire Robertson, and Jay J. Van Bavel. GPT is an effective tool for multilingual psychological text analysis. (2023).
  - Davidson, Thomas: Start Generating: Harnessing Generative Artificial Intelligence for Sociological Research, PrePrint
  - Bisbee, James, Joshua Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer Larson. 2023. "Synthetic Replacements for Human Survey Data? the Perils of Large Language Models." SocArXiv. May 4.
  - Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C. and Wingate, D., 2023. Out of one, many: Using language models to simulate human samples. Political Analysis, 31(3), pp.337-351.
  - Walker, C., & Timoneda, J. C. (2024). Identifying the sources of ideological bias in GPT models through linguistic variation in output. arXiv preprint arXiv:2409.06043.
  - Timoneda, Joan C., and Sebastián Vallejo Vera. "Memory Is All You Need: Testing How Model Memory Affects LLM Performance in Annotation Tasks." arXiv preprint arXiv:2503.04874 (2025).