CREST - GENES
Cours doctoraux 2023 – 2024

NATURAL LANGUAGE PROCESSING

Julien Boelaert
CERAPS, Université de Lille

SCHEDULE	Monday	13th November 2023 20th November 2023	From 13:00 to 16:15	TBA
	Thursday	16th November 2023 23th November 2023	From 13:00 to 16:15	TBA

Aims and objectives

The aim of this course is to provide an introduction to the main contemporary methods for natural language processing, and to illustrate them with recent uses of *text as data* in social sciences.

Natural language processing has made giant steps during the last decade, as illustrated in 2023 by the resounding popularity of chatGPT. In addition, text corpora have become increasingly available for exploitation by social scientists, be it through digitization of originally paper sources (*eg.* Parliamentary sessions transcripts, printed newspapers, books, historical sources, ...) or audio sources (through automatic transcription), or through the advent of natively digital sources (from social media, online newspapers, …).

The course will start with the standard (aka pre-neural) methods of the late 20th century, based on large document-feature matrices. We will then cover more recent developments: word embeddings (for improved NLP, or studies about bias in text corpora), topic modeling with Latent Dirichlet Allocation (unsupervised detection of topics), and Transformer models (current state of the art, BERT- and GPT-like models). Each session will comprise a theoretical lecture, and applied examples on R or python.

Outline

1. From text to large matrices: the DTM and its uses (3hrs).
   - Introduction
   - Pre-processing steps for an efficient DTM
   - Weighting, distance between words or documents
   - Uses: description, exploration, inference

2. Word embeddings: words as dense vectors (3hrs).
   - Motivation, history
   - Main models: word2vec, GloVe, fastText
   - Uses: better NLP, analysis of embeddings for bias studies

3. Topic modeling (3hrs).
   - Motivation, history
   - Core method: latent Dirichlet allocation
   - Variants: structural topic models, seeded LDA

- Applications in social science

4. Transformer-based models (3hrs).
    - Crash course in neural networks
    - The transformer model
    - BERT: state-of-the-art NLP
    - GPT: text generation, and how to use it

Pre-requisites

Knowledge of basic statistical modeling (regressions) and linear algebra are assumed, as well as some experience in a scripting/programming language for statistical analysis (R or python).

Some related literature

* D. Jurafsky and J. H. Martin (2023), *Speech and Language Processing* (3rd ed. draft), esp. chapters I.6, I.7 and I.9-12 Draft available at https://web.stanford.edu/~jurafsky/slp3

* D. M. Blei, A. Y. Ng, and M. I. Jordan (2003), "Latent dirichlet allocation" *Journal of Machine Learning Research*, 3, 993-1022.

* P. DiMaggio, M. Nag, and D. Blei (2013), "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding", *Poetics*, 41(6), 570-606.

* N. Fligstein, J. Stuart Brundage, and M. Schultz (2017), "Seeing like the Fed: Culture, cognition, and framing in the failure to anticipate the financial crisis of 2008", *American Sociological Review*, 82(5), 879-909.

* T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013), "Distributed representations of words and phrases and their compositionality", *NeurIPS* (https://arxiv.org/abs/1310.4546)

* P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov (2017), "Enriching word vectors with subword information", *Transactions of the Association for Computational Linguistics*, 5:135–146. (https://arxiv.org/abs/1607.04606)

* D. Stoltz and M. Taylor (2021), "Cultural cartography with word embeddings", *Poetics*, Vol. 88

* J. J. Jones, M. Ruhul Amin, J. Kim, and S. Skiena, (2019), "Stereotypical Gender Associations in Language Have Decreased Over Time.", *Sociological Science*, 7: 1-35

* Vaswani *et al* (2017), "Attention is all you need", *Proceedings of the 31st Conference on Neural Information Processing Systems*

* S. Do, É. Ollion, and R. Shen, (2022), "The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy", *Sociological Methods & Research*