Université de Lorraine - IDMC
Masters SC & TAL – M1
2020-2021

Fiche de projet tutoré / Project form

Statistical modeling in textual data

Encadrement / Supervisors
1. Probability and Statistic team, Institut Elie Cartan de Lorraine
2. Marianne Clausel : marianne.clausel@univ-lorraine.fr
3. François Allard Huver; Anne Piponnier; Emmanuelle Simon

Description / Description
1. projet global/global project
This project will be part of a more global one between two laboratories : Institut Elie Cartan de Lorraine and Centre de Recherche sur les Médiations whose aim is to use quantitative tools to understand and to assess the links between the evolution with time of the discourse of institutional actors and that of traditional medias about this question of transparency in order to understand how a contreversy emerges.

The group of students shall focus on statistical tools allowing to study correlation and/or causality between the content of two corpora. The work will be divided in two parts
- Notion of causality in simple cases for random variables, the case of time series. In any cases we shall first understand the concept and then perform some classical tests on synthetic time series
- Application to textual data. Several corpora are available either at CREM, either Mazoyer et al.2020)

2. biblio. UE 705 (semestre 7)

The bibliographic part will focus on quantitative and statistical tools that will be useful to tackle the problem.
We shall focus on two main questions
- the notion of causality (Kalainathan, D., et al. 2020): case of random variable, case of time series. An introduction can be found here : https://blog.dominodatalab.com/understanding-causal-inference/
Use of the Python package causality (https://pypi.org/project/causality/)

- how can we summarize the content of a document or a corpus using toping modeling (Blei et al.2003) (implemented in NLTK and GenSim) and/or summarization technics and word embedding as BERT (Delvin et al. 2018) (using for example pretrained BERT version of

Pytoch).

3. réalisation. UE 805 (semestre 8)

We shall study time evolving corpora collected at CREM and explore causality between intitutional discourse and discourse in media on the contreversial question of Glyphosate

Informations diverses : matériel nécessaire, contexte de réalisation /
Various information: material, context of realization
- The corpus will be available. We shall use either corpus collected at CREM either a publicly available corpus as that of Mazoyer et al 2020
- Implementation will be done in Python

Livrables et échéancier / Deliverable and schedule
- Semester 1 : Report on linear causality, with possible applications to time series.
- Semester 2 : Use of Python tools to study causality inn discourse. Application to data collected at CREM

Bibliographie /References (max. 4-5)

Blei, David M., Andrew Y. Ng, & Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.

DEVLIN, Jacob, CHANG, Ming-Wei, LEE, Kenton, *et al*. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Kalainathan, D., Goudet, O., & Dutta, R. (2020). Causal Discovery Toolbox: Uncovering causal relationships in Python. *Journal of Machine Learning Research*, *21*(37), 1-5.

Mazoyer, B., Turenne, N., & Viaud, M.-L. (2017). **"Étude des influences réciproques entre médias sociaux et médias traditionnels"**. In "Amsaleg, L., Claveau, V. & Tannier, X. Actes de l'atelier Journalisme Computationnel 2017", 37–40