Analyzing the French *Journal Officiel*

Project Proposal - NGSA 2018

Adib Baziz
CentraleSupélec
3 Rue Joliot Curie
Gif-sur-Yvette 91190
adib.baziz@student.ecp.fr

Reuben Dorent
CentraleSupélec
3 Rue Joliot Curie
Gif-sur-Yvette 91190
reuben.dorent@student.ecp.fr

Samuel Joutard
CentraleSupélec
3 Rue Joliot Curie
Gif-sur-Yvette 91190
samuel.joutard@student.ecp.fr

Joël Seytre
CentraleSupélec
3 Rue Joliot Curie
Gif-sur-Yvette 91190
joel.seytre@student.ecp.fr

ABSTRACT

This proposal proposes the outline of our suggested project for the class of Network Science Analysis, that focuses on the following: **given the French Journal Officiel scrapped from the web explore different graph construction methods and community detection algorithms to detect significant articles as well as political trends and the tendencies of different political administrations.**
Our proposal is focused on exploiting this newly gathered dataset.

CCS CONCEPTS

•Mathematics of computing → Graph Theory;

KEYWORDS

Journal Officiel, Network Science Analysis, Graphs, French Politics, Community Detection

ACM Reference format:
Adib Baziz, Reuben Dorent, Samuel Joutard, and Joël Seytre. 2018. Analyzing the French *Journal Officiel* . In *Proceedings of CentraleSupélec '17-'18, Paris-Saclay, France, March 2018,* 2 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION: THE FRENCH JOURNAL OFFICIEL AND MOTIVATION

The French *Journal Officiel* (JO) contains additions and amendments to the French law and is published almost every day. These modifications are reported in articles which sometimes refer to one another and therefore inherently present a graph structure where each article can be seen as a node and references are analogous to edges.
Almost every day, a new publication of the French Journal Officiel is published on http://www.journal-officiel.gouv.fr/. One of our group member's previous work consisted in setting up a proper scrapping of the website and storing the data cleanly in an ElasticSearch database. The relevant GitHub repository is public: https://github.com/alexis-thual/parsing-journal-officiel.
Every JO publication contains on average 100 articles, which in a year approximately sums up to 20,000 pages of text. The work done previously makes it possible to search the ElasticSearch database, through it's built-in search function (which operates using Term-Frequency Inverse-Document Frequency [TF-IDF]).
Once the data has been gathered, it is crucial to compare the different ways of building a graph for this dataset and also evaluate the relative importance of the articles.

CentraleSupélec '17-'18, Paris-Saclay, France

DOI: 10.1145/nnnnnnn.nnnnnnn

2 PROBLEM DEFINITION

Given the scrapping tool that was built, the problem becomes the interpretation and the insights that can be derived from the very large amount of text data, structured into articles.
To that end, the main goal will be to compare different ways of building a JO-article graph, be it on "natural links" (see Section 3) or links based on the content of the articles. The problem can be formulated as follows: **what insights can we derive from different graph constructions of the JO data? Can we identify different groups of articles? Different types of articles?**

3 METHODOLOGY

We will explore different ways of building a graph based on either natural or a textual similarity function.
We identified the following ways of building the graph:

- The natural links: JO articles refer to other articles and that link is stored during the scrapping. A given article generally refers to ≈ 10 articles;
- Links based on the content (text) of the articles: using TF-IDf, Word2Vec or other derived methods (Paragraph Vector, etc), it is possible to extract embeddings of a JO article and / or establish a similarity function between two articles;
- a potential mixture of the two where we treat natural links differently: is it a reference to a code of law, is it a reference

to a similar previous article, etc?

Once the graph is constructed, we will exploit its structure to draw conclusions on JO articles. We will mainly focus on three elements we would like to highlight:

- First, find fundamental articles which appear as important hubs in the graph. We also would like to take into account the temporal dimension in this task to detect trend effects;
- Then, find outliers that seem to stand out of the graph stream over time;
- Finally, identify tendencies in the natural links through community detection. We expect to part the graph into homogeneous subsets of the corresponding ministries, specific topics or the same reform depending of our partitioning scale.

Other tasks that seem relevant and that we will consider in our exploitation are:

- Identification of decrees that follow each other (e.g 'nomination of current Prime Minister' ⇒ 'nomination of former Prime Minister' ⇒ etc).
- Application of Topic Modeling algorithms, such as *Latent Dirichlet Allocation*, to identify the underlying topics appearing in the articles

4 EVALUATION

We will evaluate the different ways of building the graph by comparing the insights that can be derived through the methods mentioned in Section 3.

5 REFERENCES

- *Distributed Representations of Sentences and Documents*, Q. Le, T. Mikolov - https://arxiv.org/abs/1405.4053
- *The Pursuit of Hubbiness: Analysis of Hubs in Large Multidimensional Networks*, M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, D. Pedreschi - http://www.michelecoscia.com/wp-content/uploads/2012/08/socoms.pdf
- *Fast unfolding of communities in large networks*, - http://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008/meta
- *Outlier Detection in Graph Streams*, - http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.297.8978&rep=rep1&type=pdf
- *Decentralized Topic Modelling with Latent Dirichlet Allocation*, I. Colin, C. Dupuy - https://arxiv.org/pdf/1610.01417.pdf