DATA CHALLENGE

Link prediction for the French Web

Palaiseau – 10 Janvier 2020

Oualid EL HAJOUJI
Othman GAIZI
Jad SAADANI HASSANI

Table of contents

I. *Introduction*

II. *Feature engineering*

III. *Parameter tuning*

IV. *Conclusion*

Introduction

Data : - Directed graph: enumeration of edges

Problem : - Given a couple of node Ids, predict presence of a directed link

Example :

- 23	56	→	1
56	23	→	0
340	23	→	0

Feature engineering

Graph structure

- 1-hop features

  Ex : $CN(u,v) = \mid \Gamma(u) \cap \Gamma(v) \mid$

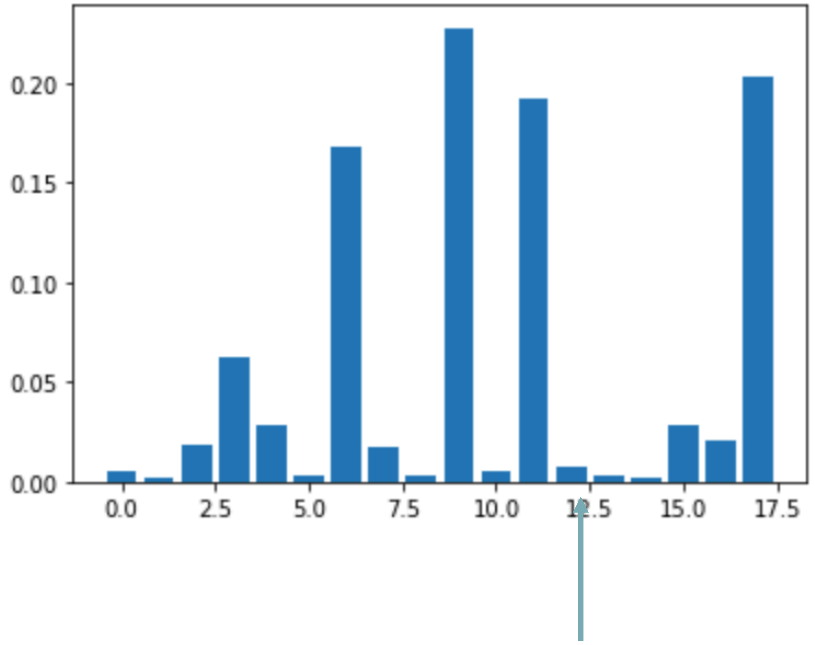
  $JC(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$

  $UD = \mid \Gamma(u) \mid$

  $VD = \mid \Gamma(v) \mid$

- Multi-hops features

  - Number of paths of length 2/3 (Katz)
  - Shortest path length
  - Community (Louvain, Infomap)
  - Node2Vec



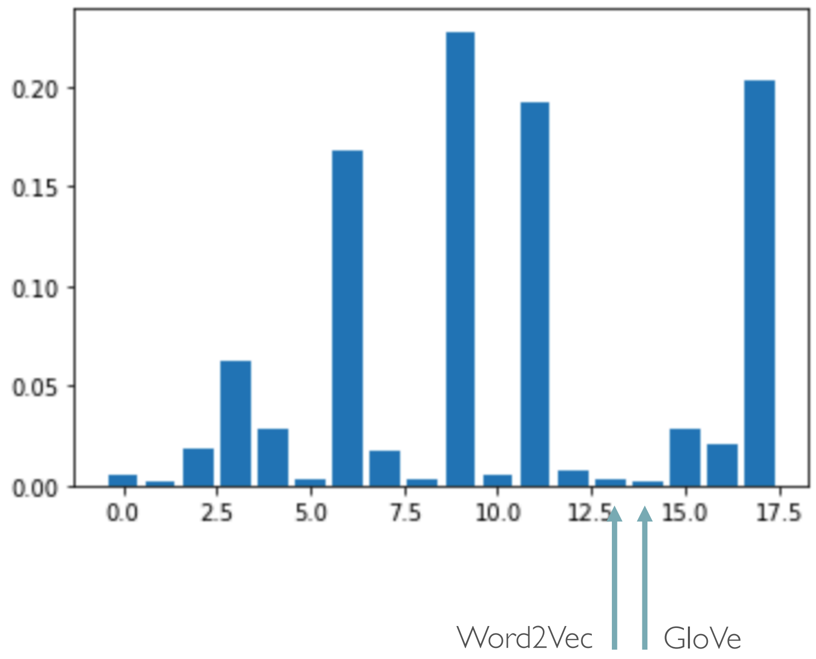
Node2vec

- Engineering all features considering directionality …

Feature engineering

Nodes text

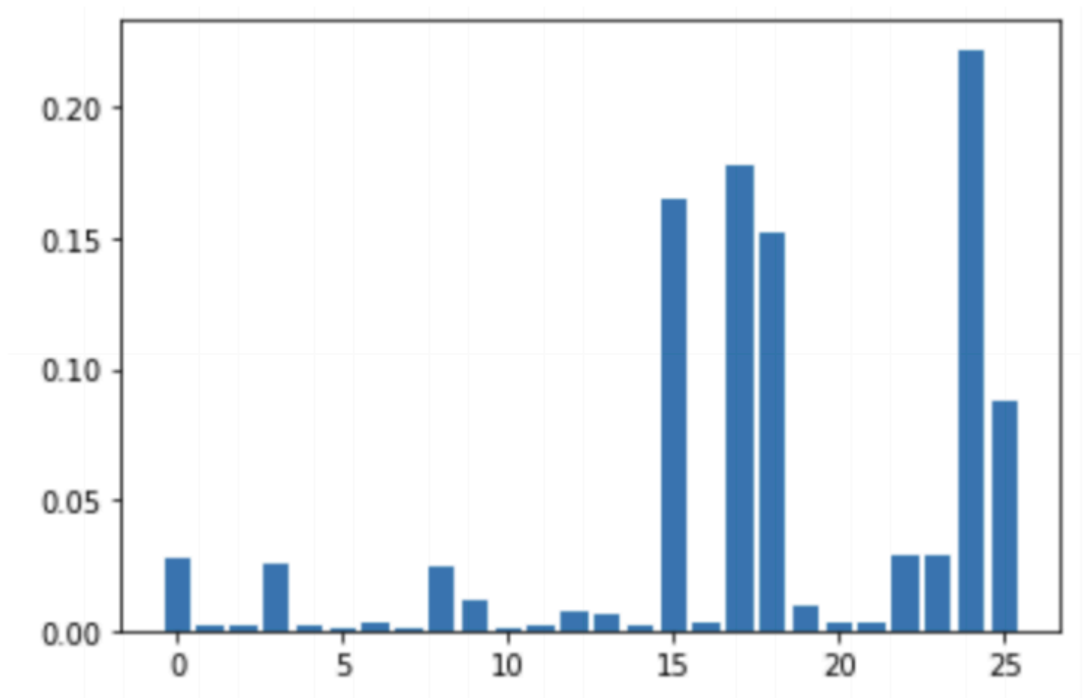
- TF-IDF
- Keywords extraction with PageRank
- Word2Vec/GloVe (pretrained models)



- Recomputing all features after cleaning the data with nltk library

Parameter tuning

- Model used: XGBoost



Features importance of our final model. 0-21: basic neighborhood features, 22: TF-IDF, 23: shortest path length, 24: community, 25: number of paths of length 3.
It is this tool that helped us throughout our experiments to evaluate our model and features selection.

Parameter tuning

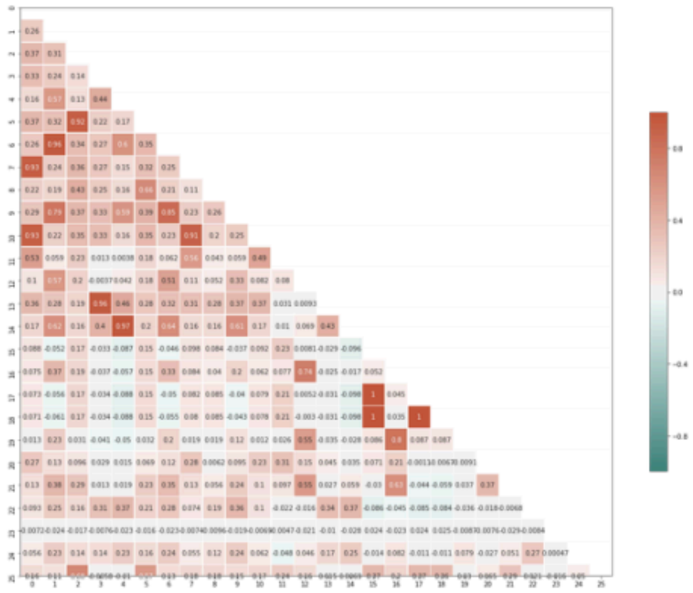
- Xgboost caveats : - large number of parameter (depth complexity, learning rate, number of trees, regularization)
  - Overfitting

- Constraint : 5 submissions / day

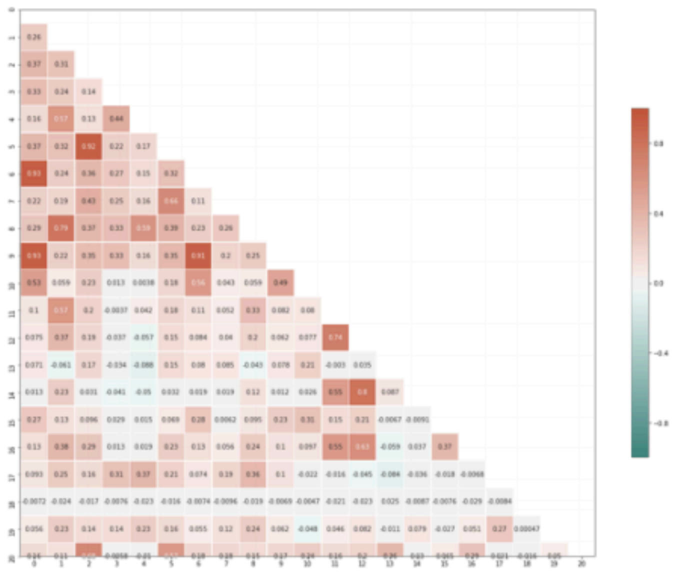
- Strategy :
  - Dividing data into training/test set for local evaluation
  - Parameters grid search with cross validation

Final model

- Features correlation



All features correlation



Filtered features correlation

- Comparison and voting classifier

- Logistic Regressor,
- Decision Tree Classifier,
- SVM,
- Naive bayes classifier
- Linear Discriminant Analysis,
- Quadratic Discriminant Analysis,
- Random Forest Classifier,
- KNN Classifier …