Projet TAL - RI

Thomas Tremble Luca Scimone

May 2021

Sommaire

1 Introduction

L'objectif de ce projet est l'étude d'une base de données contenant une collection de débats de l'assemblée nationale en France. Ce projet se décompose en deux étapes.

2 Étape 1

2.1 Description

Le jeu de données sur lequel nous allons travailler comporte des tours de parole des députés dans les débats de l'Assemblée nationale en France. Ces tours de paroles sont fourni avec une date et le parti politique des députés ayant pris la parole. Voici un extrait de la base de données fournit: .

	Unnamed: 0	dateSeance	texte	groupe_politique
0	0	20190115150000000	Prochaine séance, ce soir, à vingt et une heur...	LaREM
1	1	20190115150000000	\n \nLe Directeur du service du com...	NaN
2	2	20190115150000000	Monsieur le Premier ministre, comme vous le sa...	NaN
3	3	20190115150000000	Il appartiendra donc aux parlementaires ou, pl...	NaN
4	4	20190115150000000	Dans cet esprit, pourquoi avoir refusé un chan...	SOC

L'objectif est de notre projet prédire si le tour de parole est celui d'un député de gauche ou de droite.
Pour obtenir un nombre satisfaisant d'exemples par groupe politique, on considérera uniquement deux groupes : "droite" et "gauche". Nous considérerons que le parti LR fait partie de la classe "droite", et que les partis SOC, GDR et FI font partie de la classe "gauche". Les autres groupe politique seront ignorés.

2.2 Pré-traitement

Dans un premier temps nous avons réétiqueté les discours selon si le partis était considéré comme de gauche ou comme de droite et ignorés les partis non concerné dans notre étude.

Nous avons remarqué que les prises de paroles courtes ne contenaient pas assez d'information pour la construction de notre modèle. Ainsi nous avons décidé de conserver les allocutions de plus de 20 items. En faisant cela nous avons grandement diminué notre jeu de données. Cependant ont conservé un peu près 2000 discours ce qui était largement assez pour nos algorithmes.

Dès lors après cette opération nous obtenons un jeu de données avec un discours rattaché à une orientation politique.

	texte	orientation_politique
0	Dans cet esprit, pourquoi avoir refusé un chan...	gauche
1	Le pompier qui est dans le coma, ce n'est pas ...	gauche
2	Le groupe La France insoumise ne désespère pas...	gauche
3	Le groupe de la Gauche démocrate et républicai...	gauche
4	Est-ce donc cela, votre vision de l'impartiali...	droite

Ensuite nous avons tokenizé et nettoyé nos données en éliminant les stopwords dans les allocutions. Pour faire cela nous avons utilisé nltk qui propose des méthode de tokenisation et une liste de stopwords pour la langue française.

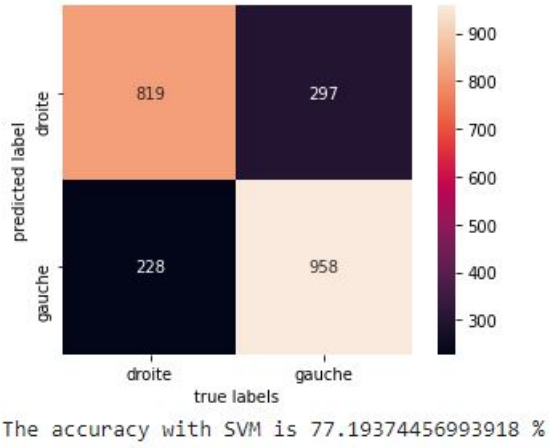
Ensuite nous avons vectorisé nos données texte avec la méthode TFidf. TFidf est une méthode de pondération très utilisé dans la fouille de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus.
Ainsi avec cette vectorisation nous pourrons appliquer nos différent algorithme de prédictions.

2.3 Résultats

La phase de pré-traitement étant effectué nous allons pouvoir entraîner nos différents algorithme et voir celui qui fonctionne le mieux. Voici les différents algorithmes testé :

- MultinomialNB : Le classificateur multinomial Naive Bayes est génarelement utilisé pour la classification de texte avec des caractéristiques discrètes.
- SVM : Support vector machine, SVM sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression. Les SVM sont une généralisation des classifieurs linéaires
- RandomForestClassifier : Une forêt aléatoire est un méta-estimateur qui ajuste un certain nombre de classificateurs à arbre de décision sur divers sous-échantillons de l'ensemble de données et utilise la moyenne pour améliorer la précision prédictive et contrôler l'ajustement excessif.

Parmi c'est trois algorithme celui qui a donné le meilleur résultat est SVM avec une précision de prédiction de 77 % (67 % pour MultinomialNB et 74 % pour RandomForestClassifier).
Ci-dessous la matrice de confusion obtenue avec l'algorithme SVM :




```
The accuracy with SVM is 77.19374456993918 %
```

La précision obtenue est assez satisfaisante pour ce genre de données.

3 Étape 2

3.1 Description

Lors de la deuxième étape, nous avons créé à partir du fichier débats.csv Debats2.csv, qui contient les allocutions labellisés de la gauche et de la droite.

Voici un aperçu de ce fichier :

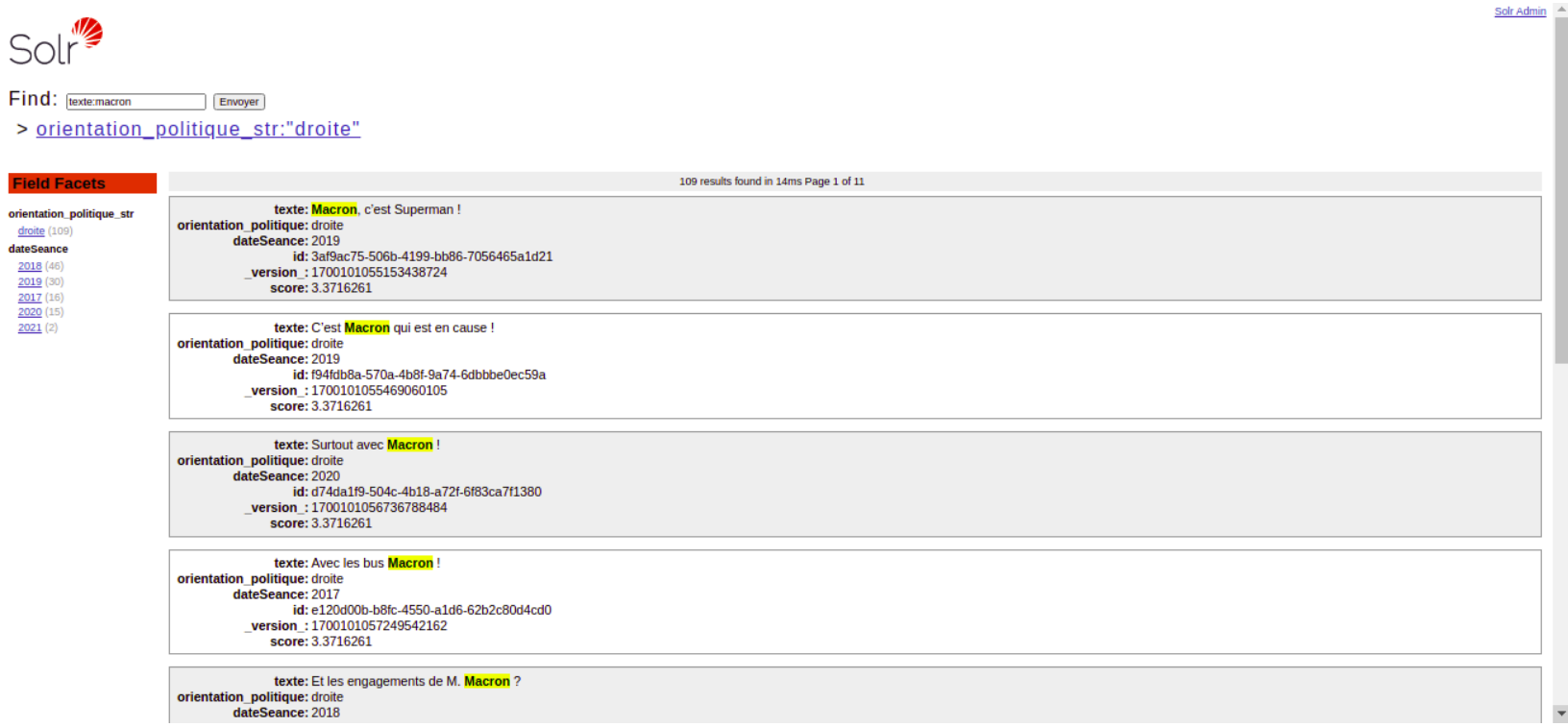
	texte	orientation_politique	dateSeance
1	Dans cet esprit, pourquoi avoir refusé un chantier o...	gauche	2019
2	Le pompier qui est dans le coma, ce n'est pas des c...	gauche	2019
3	Vous en savez quelque chose !	gauche	2019
4	Le groupe La France insoumise ne désespère pas de...	gauche	2019
5	Le groupe de la Gauche démocrate et républicaine ...	gauche	2019
6	Très juste !	droite	2019
7	Quelle est la question ?	droite	2019
8	Elle vient de dire que nous y contribuerons !	droite	2019
9	C'est pourtant ce qui se passe depuis plusieurs mois !	gauche	2019
10	La décision est prise de fait !	gauche	2019
11	Est-ce donc cela, votre vision de l'impartialité et de ...	droite	2019
12	Rappelez-nous le rôle de Macron !	droite	2019

3.2 Configuration

Nous avons essayé de faire en sorte que la zone de recherche par défaut soit le texte des allocutions malheureusement cela n'a pas fonctionné.
Ainsi pour trouver un mot dans les différentes allocution il faut taper dans la barre de recherche texte:{mot} pour trouver l'occurrence de ce mot dans les discours de la base de données.
On peut aussi appliquer des filtres à notre recherche comme l'orientation politique (gauche ou droite) et l'année.

3.3 Résultats

Voici un exemple avec la recherche du mot macron avec le filtre "droite" qui permet de voir uniquement les allocutions faites par les partis de droite.



4 Conclusion

Lors de ce projet, nous avons été amené à la création d'un modèle de catégorisation de textes. Nous avons également appris à utiliser le serveur de recherche Solr. Ces deux étapes nous ont permis d'apprendre catégoriser des textes et à les analyser au sein d'un moteur de recherche.

5 Répartition du travail

Afin de travailler sur ce projet à deux, nous avons utilisé les outils suivant:

- Google Colab : notebook jupyter sur le cloud qui permet de fusionner nos travaux.
- Overleaf : Une plateforme en ligne gratuite permettant d'éditer du texte en LATEX sans aucun téléchargement d'application et de partager en "live" l'édition du fichier.
- Discord : Outil de discussion en ligne et de screen sharing.

6 Références

- StackOverflow
- solr.apache.org
- wikipedia
- cdata.com
- tutorialspoint.com
- scikit-learn.org
- https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34