

REPUBLIQUE DU CAMEROUN

Paix-Travail-Patrie

\*\*\*\*\*\*\*\*\*\*

UNIVERSITE DE YAOUNDE I

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

ECOLE NATIONALE SUPERIEUR POLYTECHNIQUE DE YAOUNDE

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

DEPARTEMENT DU GENIE INFORMATIQUE

REPUBLIC OF CAMEROON

Peace-Work-Fatherland

\*\*\*\*\*\*\*\*\*\*\*\*

UNIVERSITY OF YAOUNDE I

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

NATIONAL ADVANCED SCHOOL

OF ENGINEERING OF YAOUNDE

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

DEPARTEMENT OF COMPUTER SCIENCE ENGINEERING

**AGENTS INTELLIGENTS**

Rapport de Projet :

Détection des Langues

Etudiant

NGUAZONG AUREL 20P001

Sous la supervision : M. Bitha

Année académique :

2023/2024

Table des matières

[INTRODUCTION 3](#_Toc166846102)

[I. Méthodologie de conception 4](#_Toc166846103)

[1. Importation et exploitation de la dataset 4](#_Toc166846104)

[2. Analyse univariée 4](#_Toc166846105)

[3. Prétraitement des données 5](#_Toc166846106)

[4. Séparation des données d’entraînement et de test 6](#_Toc166846107)

[5. Vectorisation des données 7](#_Toc166846108)

[6. Entraînement par régression logistique 7](#_Toc166846109)

[7. Evaluation sur les données test 8](#_Toc166846110)

[8. Démonstration 9](#_Toc166846111)

[II. Remarques et commentaires 10](#_Toc166846112)

# INTRODUCTION

Intitulé du devoir :

Projet AGENTS INTELLIGENTS

A partir du jeu de données Languages basé sur la reconnaissance de la langue en fichier joint, il vous est demandé de mettre en place une solution capable de pouvoir reconnaitre la langue utilisée lors de la saisie d'un texte. Le jeu de données présente plusieurs langues parmi lesquelles :

1) Anglais

2) Malayalam

3) Hindi

4) Tamoul

5) Kannada

6) Français

7) Espagnol

8) Portugais

9) Italien

10) Russe

11) Suédois

12) Néerlandais

13) Arabe

14) Turc

15)Allemand

16) Danois

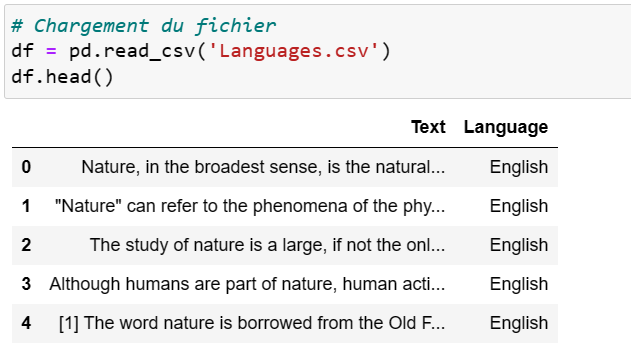
17) grec

Vous utiliserez à votre convenance des algorithmes d'apprentissage automatique ou en profondeur pour la résolution de la problématique susmentionnée.

# Méthodologie de conception

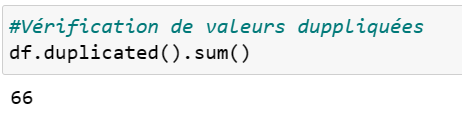
## Importation et exploitation de la dataset

La dataset est importée via pandas et l’on vérifie ces paramètres.

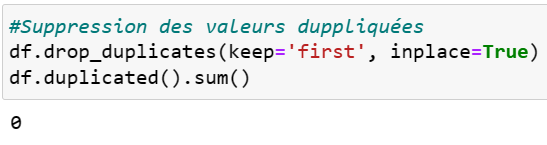


Par la suite l’on exploite ces données pour pouvoir mieux les manipuler via un processus simple :

* L’on vérifie les lignes dupliquées

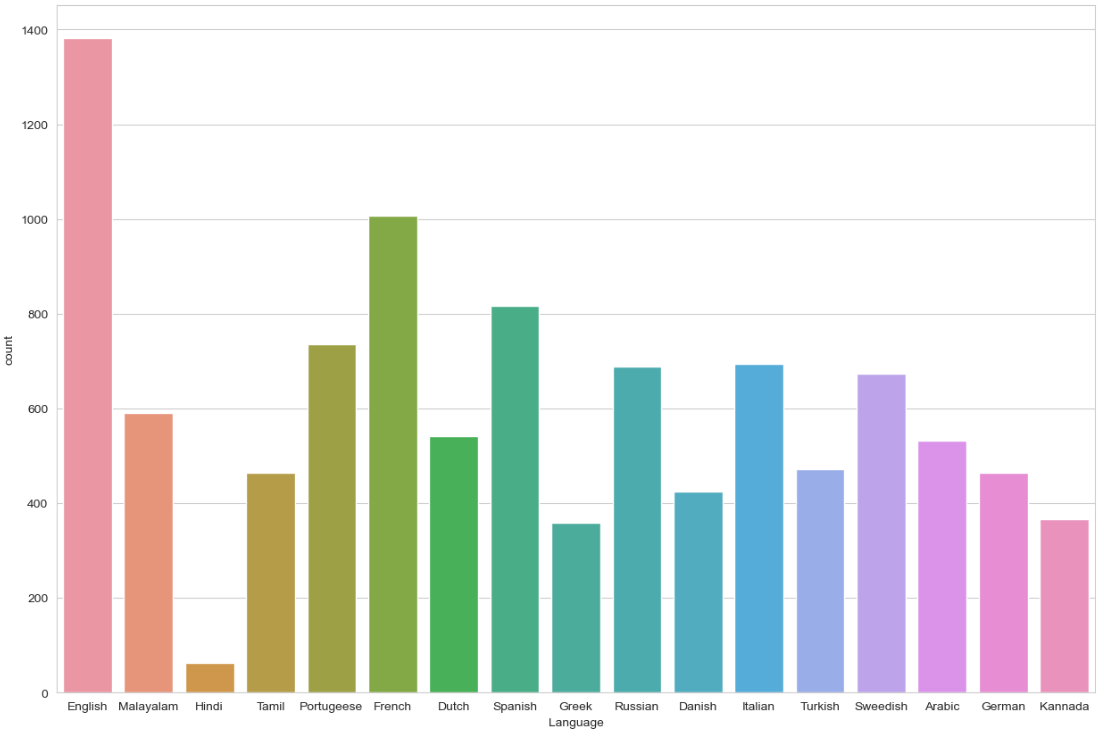


* Puis l’on supprime ces valeurs



## Analyse univariée

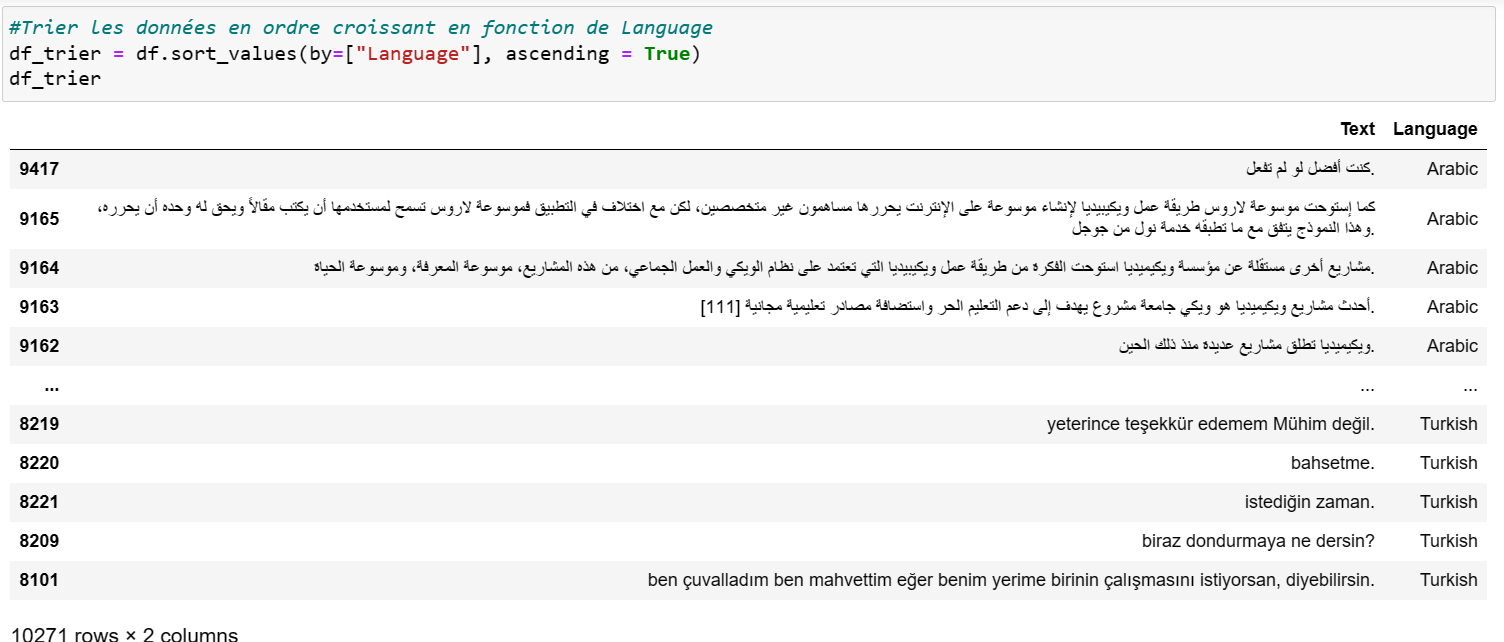
Concernant l’analyse univariée de la dataset, l’on remarque que les langues dominantes tel que l’anglais, le français ou l’espagnol auront plus de chance d’être prédite à cause du nombre extrême de features dans la dataset, contrairement à l’Hindi ou le grec qui seront moins performant lors de la prédiction.



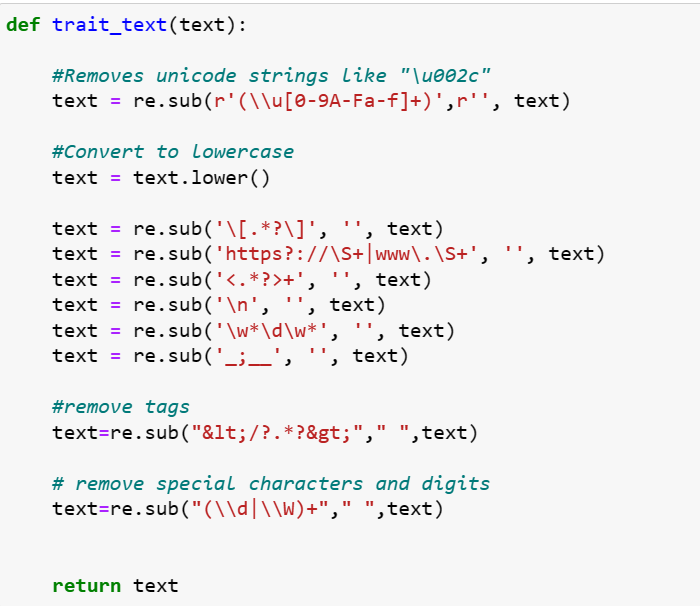
## Prétraitement des données

Après l’analyse, viens le prétraitement de données où l’on doit pouvoir :

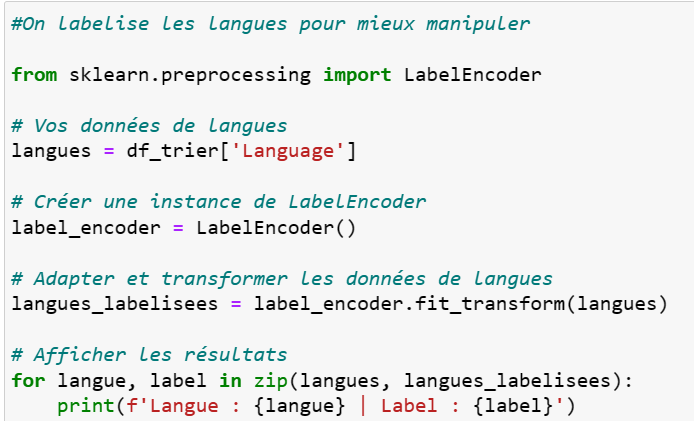
* Trier les langues en ordre alphabétique :



* Supprimer les caractères spéciaux, les liens, les chiffres et mettre tout le texte en minuscule :

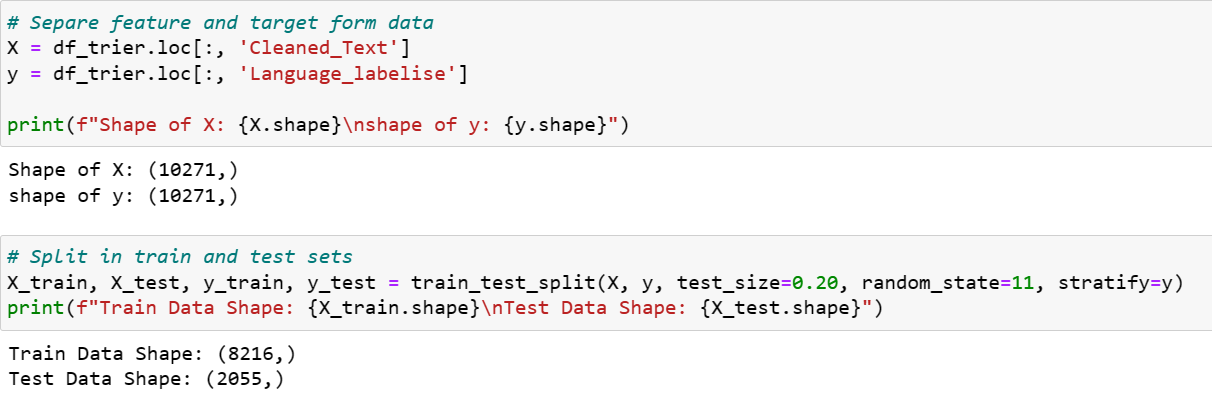


* Labéliser les langues pour l’entraînement de prédiction



## Séparation des données d’entraînement et de test

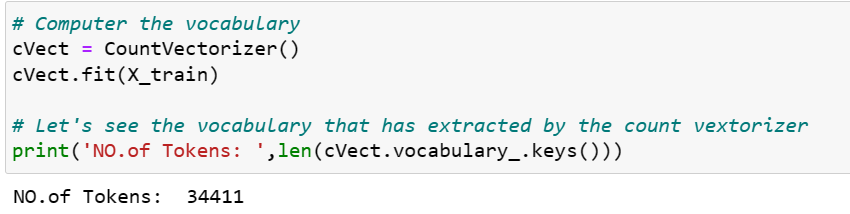
Pour l’entraînement, l’on sépare les données d’entraînement de test selon le ratio 8:2.



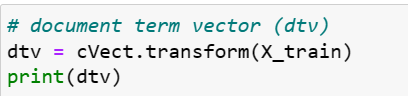
## Vectorisation des données

Pour la vectorisation, l’on doit :

* Dans un premier temps faire un countVectorizer pour dénombrer le vocabulaire à vectoriser



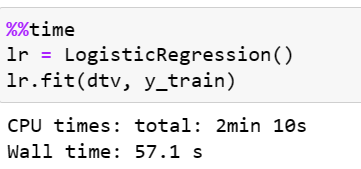
* Ensuite vectoriser chaque terme du vocabulaire



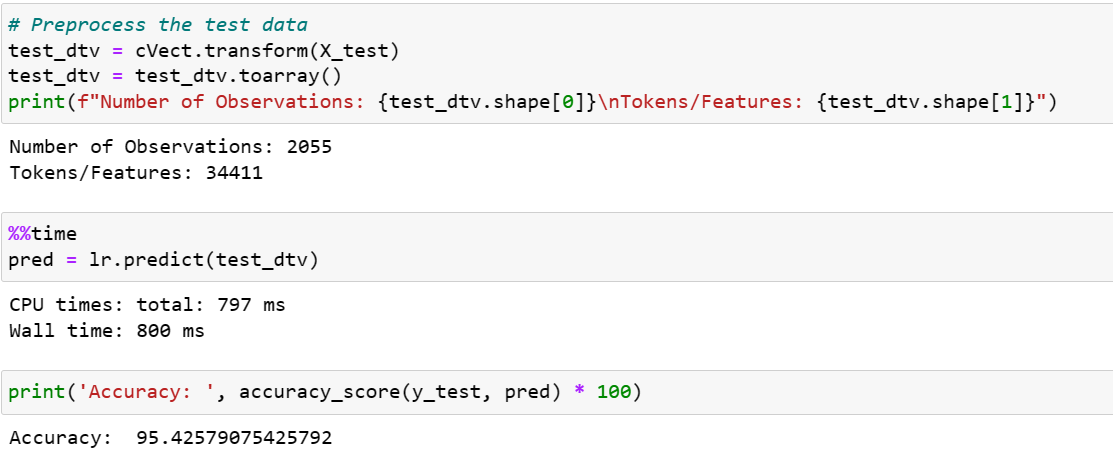
* Enfin transformer le tout en tableau pour qu’ils soient viables :



## Entraînement par régression logistique

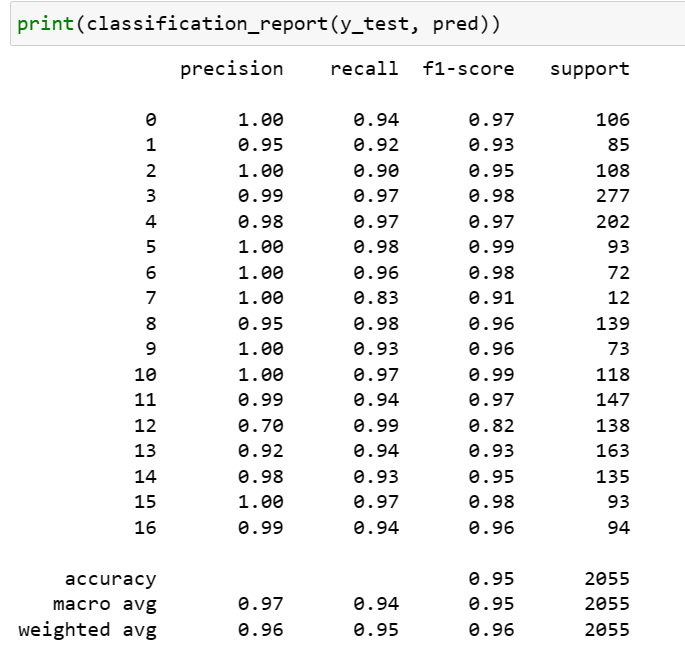


## Evaluation sur les données test



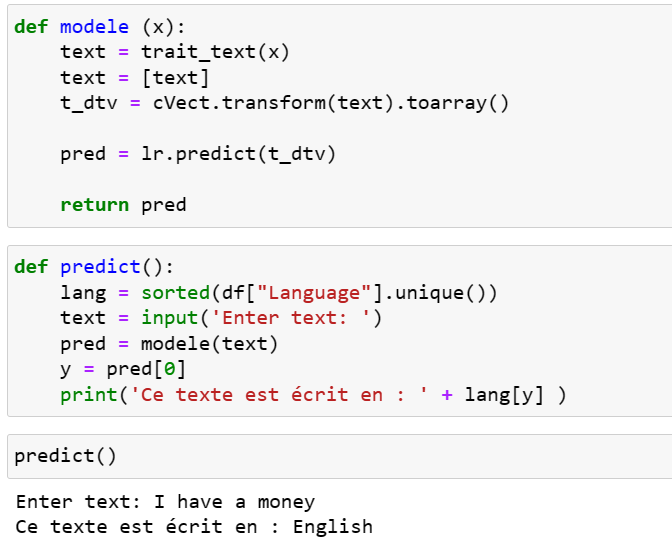
L’on remarque dans cette partie que le score au test est de 95,42% ce qui représente un bon score et ne nécessite pas forcement d’optimisation.

Concernant la précision pour chaque langue, elle varie assez mais reste stable avec une moyenne de 95%.



## Démonstration

Cette démo est juste la preuve que le code marche pour la plupart des langues et peut servir comme modèle de reconnaissance.



# Remarques et commentaires

Lors de la conception de ce modèle l’on a rencontré un certain nombre de problème et de remarque :

* Était-il nécessaire de supprimer les stopwords ? : J’ai jugé cela pas nécessaire à cause de la réflexion qui dit que même un stopword peut être une indication à la langue
* Était-il nécessaire d’optimiser les hyperparamètres du modèle ? J’ai jugé pas nécessaire à cause des 95% de précision.