CentraleSupélec

Documentation technique

Identification des thèmes d'un livre par NLP

<u>Client :</u> Timothée BOHÉ & Nicolas GAUDEMET - Onepoint

Amine CHERIF HAOUAT
Tom LABIAUSSE
Cyrine NABI
Pierre OLLIVIER

I Description du projet

L'objectif de ce projet est de prévoir les ventes de livres à partir de leur quatrième de couverture. Plus précisément, notre travail s'inscrit dans un projet plus global, qui se divise en deux parties : une dans laquelle les livres sont classifiés grâce à leur quatrième de couverture (c'est la partie dont nous nous sommes occupés) et une autre dans laquelle chaque classe de livres est associée à un nombre de ventes.

II Vue d'ensemble de l'architecture

L'architecture de notre projet est assez simple : nous disposons de fichiers contenant les données (au format .xslx), stockés sur un Drive et d'un fichier de code pour faire les calculs, qui se connecte au Drive pour récupérer les données. Schématiquement, notre architecture est donc la suivante :

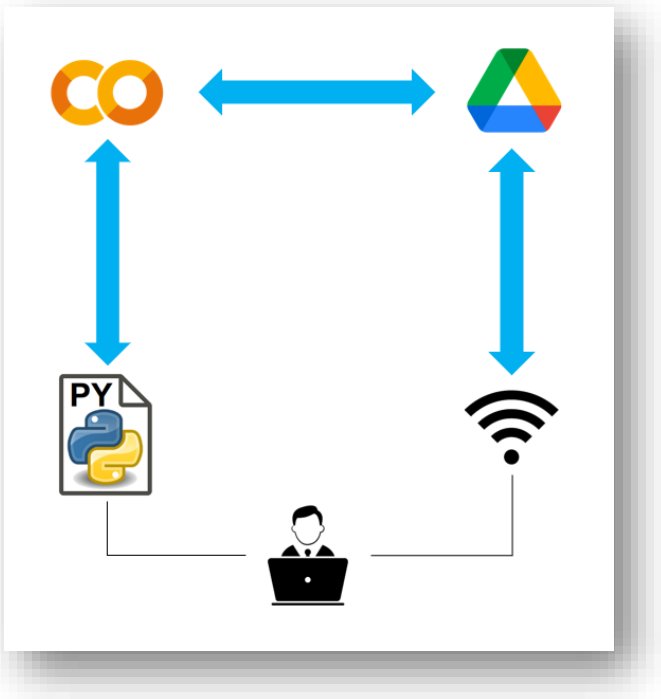

Figure 1 : Architecture du projet.

L'architecture de notre projet peut finalement se découper en deux parties :

- La conception du code par l'utilisateur. Le code est écrit en langage Python, pour tirer profit des bibliothèques de machine learning qui sont particulièrement développées dans ce langage.

Le code est stocké en ligne, sur Google Colaboratory, pour pouvoir se connecter au Drive et rendre le projet accessible à plusieurs utilisateurs.

- Le code et les données. Une fois authentifié, l'utilisateur peut accéder au Drive sur lequel sont présentes les données. Google Colaboratory interagit avec ce Drive.

Pour les données, nous avons choisi de les stocker sur un Drive pour qu'il n'y ait pas besoin de les importer manuellement à chaque fois : c'est plus pratique et ça permet d'exécuter le code même sans disposer des données en local.

Bien sûr, cette architecture a un inconvénient : elle nécessite une connexion internet. Il est cependant possible d'exécuter le code en local moyennant quelques modifications.

III Architecture détaillée

Les deux notebooks de ce projet ***Topic_Modeling_LDA.ipynb*** et ***Topic_Modeling_BERTopic.ipynb*** utilisent plusieurs packages python. Certains nécessitent une version particulière qui peut être installée avec une commande du type :

```
pip install package==version
```

Les détails des versions précises à utiliser tout comme une présentation des fonctionnalités des notebooks sont disponibles dans les fichiers ***README_LDA.txt*** et ***README_BERT.txt***. Enfin, l'utilisation de notebooks présente l'avantage de pouvoir mêler parties de codes et d'explications. Ainsi, la majorité des indications concernant le code sont directement accessibles via les notebooks.