Rapport Projet DATA732

Objectifs du projet

L'objectif de ce projet était de mener des études et des analyses de données sur différents corpus mis à notre disposition.

Pour ce faire, nous devions utiliser différentes méthodes telles que la vectorisation, par exemple. Nous pouvions utiliser le bag of words, les Word Embeddings ou encore les cooccurrences.

Pour réaliser ces analyses, nous devions utiliser des corpus représentant des articles avec les mots, les personnes ou encore les localisations présents dans ceux-ci.

Démarches

Pour mener à bien ce projet j’ai tout d’abord commencer par me renseigner sur les différentes technologies et méthodes que nous allons devoir utiliser afin de mieux comprendre lesquelles seraient les plus pertinentes à utiliser en fonction des objectifs que je donnerais à mes dashboard.

J’ai donc commencé par me renseigner sur les méthodes de vectorisation courantes comme le Term Frequency, le Term Frequency * Inverse Document Frequency ou encore la méthode de Word2Vec ou comment faire des matrices de concurrences.

Par la suite j’ai regardé quelles librairies je devrais utiliser pour réaliser ces différentes méthodes.

J’ai aussi regardé comment étaient faits les différents corpus afin de mieux pouvoir les parcourir lorsque je les analyserais, notamment où étaient placés les différentes informations importantes comme le nombre d’occurrences des mots, les descriptions

d’articles ou encore le fait que certaines parties regroupées certaines données pour être traités plus facilement et plus rapidement pour les programmes que j’utiliserais.

Une fois ces analyses terminées j’ai commencé à élaborer des programmes simples pour traiter certains corpus, notamment les moins gros pour que cela aille plus vite. J’ai principalement commencé par récupérer les données des mots les plus utilisés, en utilisant des méthodes de TF ou de TF*IDF. Je les ai enregistrés dans des json afin de les utiliser par la suite plus facilement et plus rapidement.

Par la suite j’ai utilisé le fonctionnement de la méthode Word2Vec afin de voir les similarités entre les différents mots des textes et de savoir lesquels sont les plus proches et lesquels sont les plus éloignés.

J’ai ensuite créé de nouveaux csv avec des données comme les pays, les personnes ou les keywords, afin de réaliser plus rapidement des graphiques, des cartes, des hitmaps, des histogrammes, etc… dans mon futur dashboard.

Pour finir j’ai réalisé le dashboard final regroupant l’ensemble des données que j’ai créé auparavant, avec les csv et les json. Dans celui-ci mon objectif était de pouvoir voir les différences entre les différents corpus proposés et pour chacun de voir les analyses des pays, des keywords et des personnes avec différentes méthodes que j’expliquerai par la suite.

Technologies et méthodes utilisées

Pour l’ensemble de mon analyse j’ai donc utilisées les technologies suivantes :

- langage : python (simple fichier et jupyter notebook)
- librairies :
  - pandas
  - dash
  - plotly

    - gensim
    - matplotlib
    - wordcloud
    - …

Pour l’ensemble de mon analyse j’ai donc utilisées les méthodes suivantes :

- Term Frequency
- Term Frequency * Inverse Document Frequency
- Word2Vec
- Matrice de coocurrence
- wordcloud

Problèmes rencontrés

Le principal problème que j’ai rencontré lors de mes différentes analyses de ces corpus est un problème de taille de corpus et de temps d’exécution du parcours de ces corpus. En effet pour certaines analyses et certains calculs je devais parcourir plusieurs fois les corpus et comme certains sont très gros le temps de calcul peut très vite être très long.

Un autre problème que j’ai rencontré a été de trouver des idées d’analyses à faire, notamment quel graphique utiliser pour quels documents, quoi comparer, etc…