Etape 6 : Pré-entrainement des modèles d'apprentissage automatique

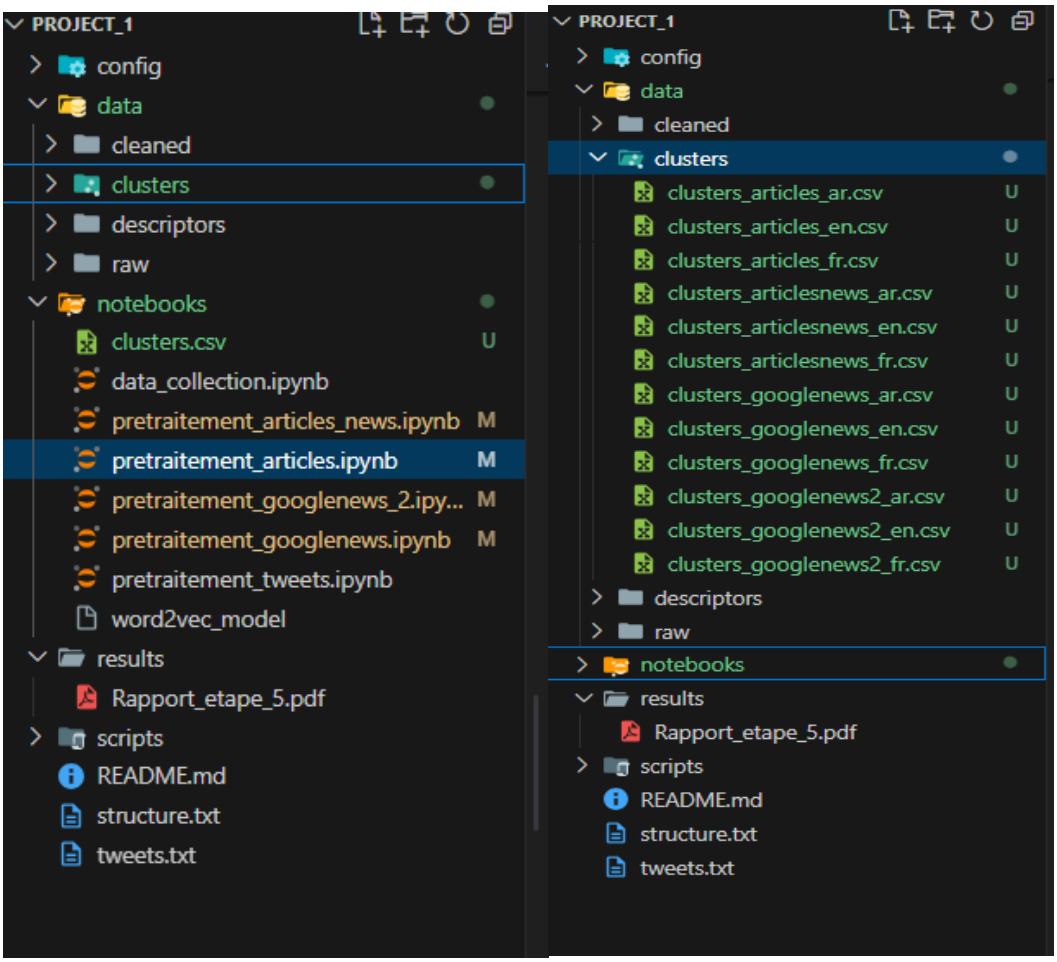
1. Introduction

Le présent projet a pour objectif l'analyse de documents textuels issus de différentes sources telles que des articles, des blogs, ou des publications sur les réseaux sociaux, afin d'identifier des patterns de désinformation médicale. Nous avons choisi d'adopter une approche de **clustering** non supervisé, en raison de l'absence d'annotations explicites des documents. L'objectif est d'organiser ces documents en fonction de leurs similarités thématiques et de détecter les sujets de désinformation médicale qui émergent, en utilisant des descripteurs sélectionnés à partir du contenu textuel.

2. Méthodologie

1. **Préparation des données** :
   La première étape consiste à **nettoyer et prétraiter les données textuelles**. Nous appliquons des techniques telles que la suppression des stopwords, la gestion des caractères spéciaux et la lemmatisation des mots pour uniformiser le texte.
   Ensuite, nous utilisons la méthode **TF-IDF** (Term Frequency-Inverse Document Frequency) pour transformer les textes en représentations numériques qui reflètent la pertinence des mots dans chaque document.
2. **Application de LDA** :
   Nous appliquons **LDA** pour extraire les topics sous-jacents du corpus. LDA permet d'identifier les topics principaux, chacun représentant un groupe de mots fréquemment co-occurrents. Chaque document est alors représenté par un vecteur indiquant la proportion de chaque topic qu'il contient. Ce processus permet de détecter les sujets latents qui peuvent être liés à des formes de désinformation.
3. **Clustering des documents** :
   Sur la base des représentations obtenues grâce à LDA, nous appliquons trois algorithmes de clustering différents :
   - **K-means**
   - **DBSCAN**
   - **Agglomerative Clustering**
     Ces algorithmes permettent de regrouper les documents selon leur affinité thématique, ce qui facilite la détection des documents traitant des mêmes sujets liés à la désinformation médicale.
4. **Analyse des clusters** :
   Après avoir formé les clusters, nous analysons les **termes dominants** dans chaque cluster pour interpréter les sujets. Par exemple, certains clusters peuvent être dominés par des termes tels que "remèdes miracles", "théories du complot sur les vaccins", ou encore "médicaments non prouvés".

Sauvegarde des résultats



3. Conclusion

Les méthodes de **clustering** basées sur LDA ont permis de découvrir des groupes de documents traitant des sujets similaires en lien avec la désinformation médicale. Les clusters ont révélé plusieurs thèmes clés de désinformation, comme les fausses informations sur les vaccins et les remèdes non prouvés. Cette approche peut être utilisée comme base pour la création d'outils automatiques de détection de la désinformation médicale, facilitant ainsi la gestion et la lutte contre ce phénomène.
Les résultats obtenus peuvent également être utilisés pour sensibiliser le public à la désinformation médicale, en mettant en lumière les thèmes récurrents et leurs variations au sein du corpus étudié.