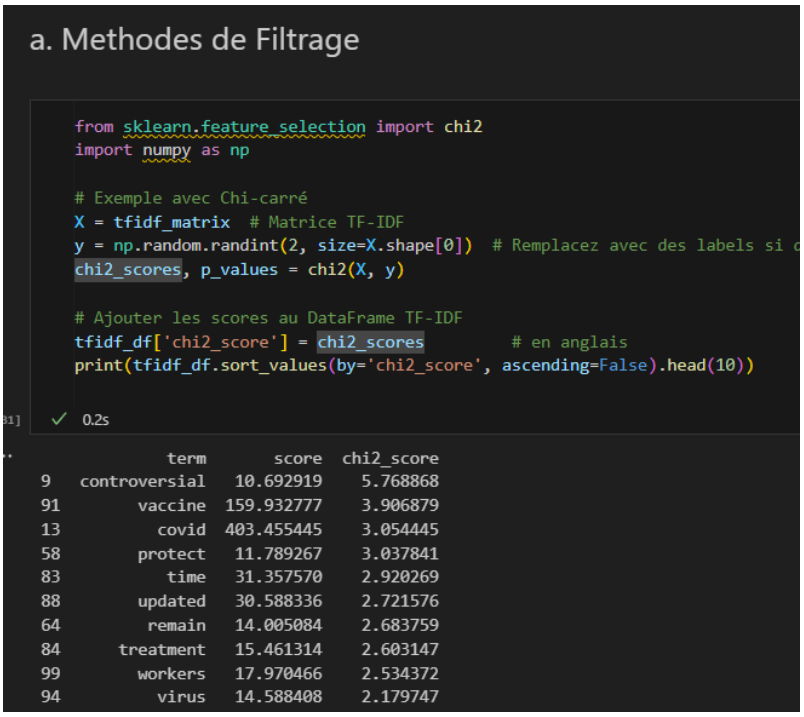
Etape 5 : Analyse des Descripteurs Textuels et Désinformation Médicale

1. Introduction

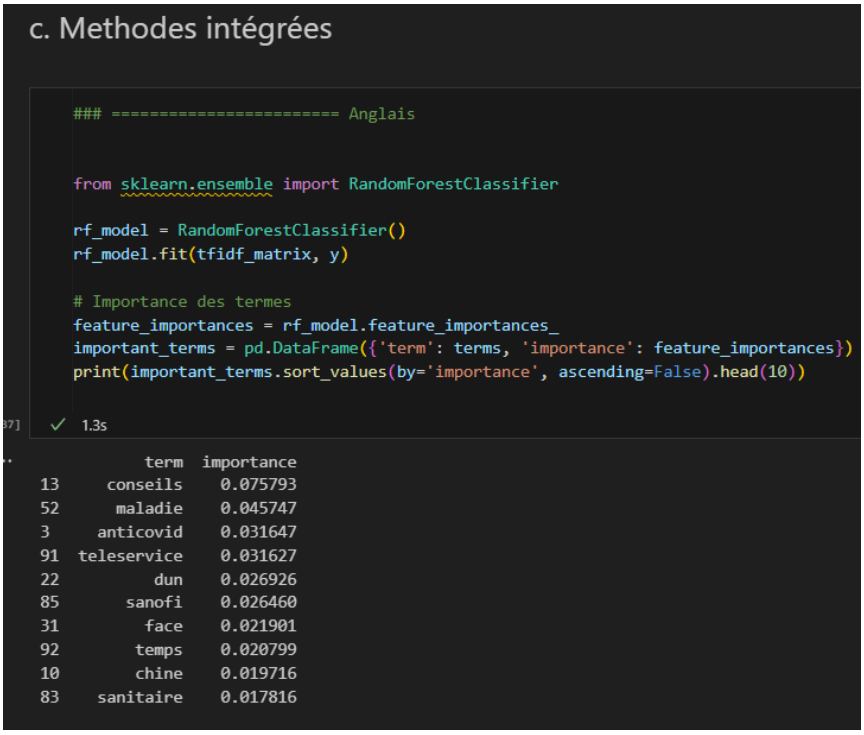
Ce projet vise à explorer et analyser les données textuelles issues de différentes sources (articles, tweets, etc.) dans le but d'identifier des descripteurs pertinents liés à la désinformation médicale. Les méthodes employées permettent de sélectionner les termes les plus représentatifs tout en mettant en évidence leur lien avec des indices de désinformation médicale, tels que des expressions sensationnalistes ou pseudoscientifiques.

2. Méthodes Utilisées

1. **Prétraitement des données textuelles :**
   - Nettoyage des données : suppression des stopwords, gestion des caractères spéciaux, lemmatisation.
   - Représentation des textes : transformation des corpus multilingues en représentations numériques via **TF-IDF** (Term Frequency-Inverse Document Frequency).
2. **Sélection des descripteurs :**
   - **Méthodes de filtrage** :
     - Utilisation du Chi-carré pour évaluer l'association statistique des termes avec les classes pertinentes.



- **Méthodes intégrées** :
  - Importance des caractéristiques calculée à l'aide de modèles tels que Random Forest ou régularisation L1 (Lasso).



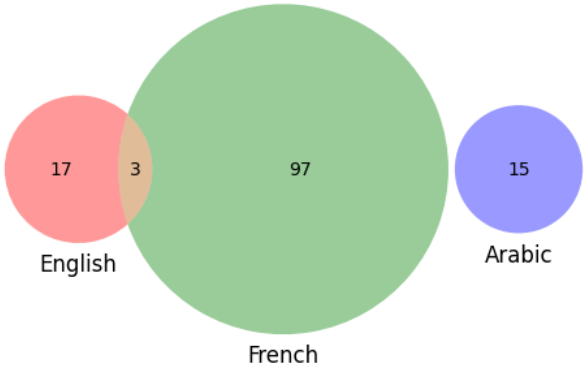
3. **Mise en contexte :**
   - Identification des termes liés à la désinformation médicale, par exemple :
     - Entités : COVID-19, vaccins, traitements, maladies.
     - Motifs linguistiques : phrases pseudoscientifiques ou sensationnalistes.

3. Analyse des Résultats

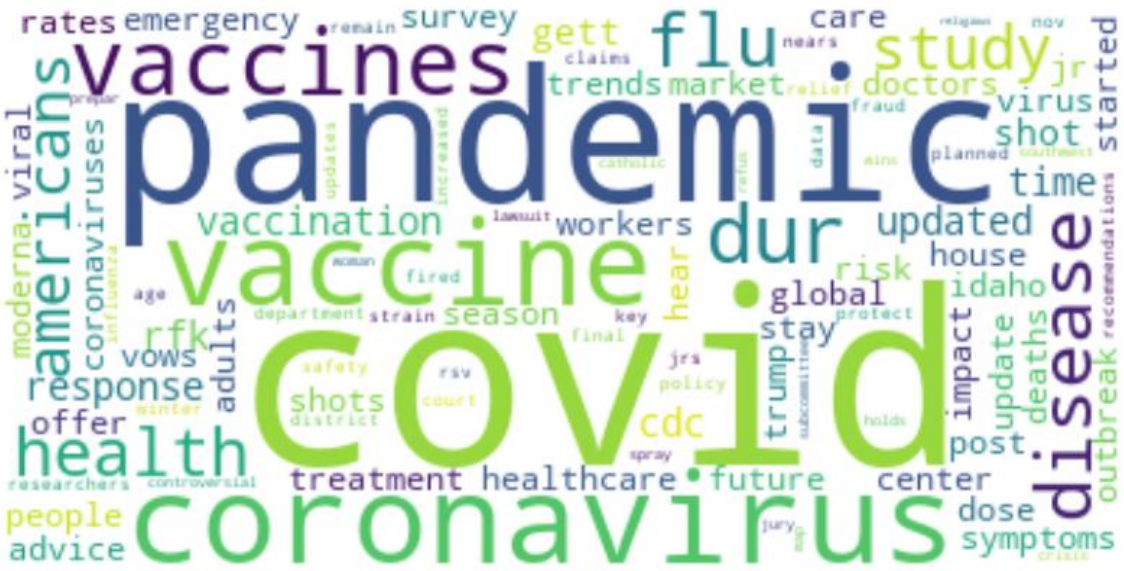
1. **Pertinence des termes identifiés :**
   - Des termes fréquents tels que "covid", "vaccine", et "disease" ont été identifiés comme des descripteurs clés.
   - Les termes avec un score élevé au Chi-carré indiquent un lien fort avec des indices de désinformation.
2. **Observations importantes :**
   - Certains termes sont universels (ex. "covid") tandis que d'autres sont spécifiques à une langue ou un contexte culturel.
   - Les méthodes de filtrage ont permis de réduire efficacement la dimensionnalité tout en conservant des termes pertinents.

4. Visualisations

1. **Diagramme de Venn :**
    - Illustre les chevauchements des descripteurs sélectionnés entre trois langues.
    - Exemples :
        - Termes communs : "vaccine", "covid".
        - Termes spécifiques : "miracle" (anglais), "remède" (français), "لقاح" (arabe).

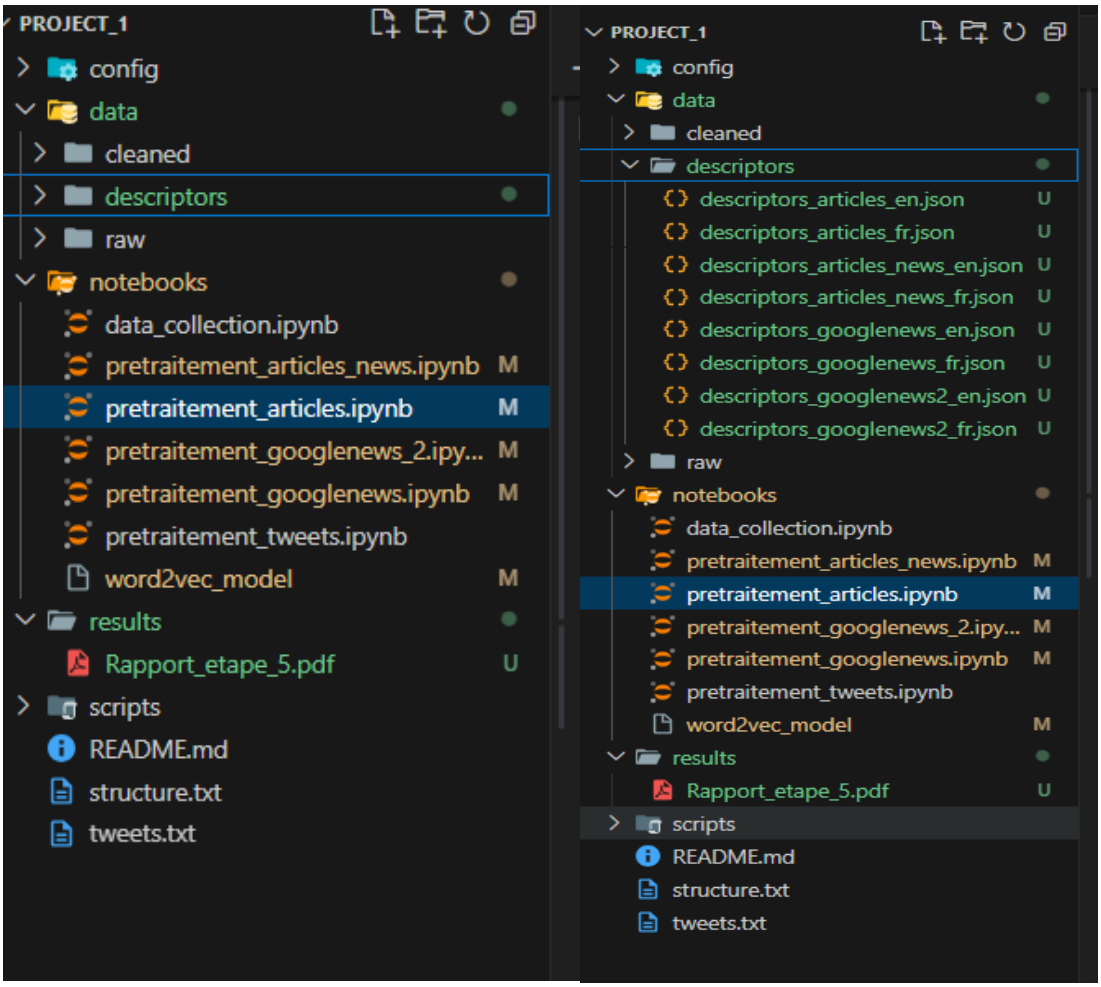


2. **Nuages de mots pondérés :**
    - Représentation visuelle des termes avec leur poids respectif (TF-IDF).
    - Les termes fortement associés à la désinformation (ex. "remède miracle") sont mis en avant.



Sauvegarde des résultats

Les descripteurs sélectionnés, accompagnés de leurs scores et méthodes associées, ont été sauvegardés dans le répertoire suivant :



5. Conclusion

Les méthodes employées permettent d'identifier efficacement les termes pertinents pour analyser la désinformation médicale. Les résultats mettent en évidence des similitudes et des spécificités entre les langues, ainsi que des termes particulièrement liés à la désinformation. Ces observations peuvent être utilisées pour développer des systèmes de détection automatique ou sensibiliser à la désinformation médicale.