TP Vectorisation

1 - Exploratory Data Analysis

Il vous est demandé de réaliser un premier notebook afin de **comprendre, explorer et nettoyer** les données. Vous devez notamment être capable de répondre aux questions suivantes :

- Quelle est la forme du DataFrame ?
- Y a-t-il des valeurs manquantes ou des valeurs dupliquées ?
- Quelles sont les colonnes qui vont nous intéresser ?
- Y a-t-il des données aberrantes ou des incohérences majeures ?
- Y a-t-il des tweets anormalement longs ou courts ? Peut-on les considérer comme des outliers ?
- Quel est le ratio de tweets parlant de **“catastrophes”** par rapport aux tweets normaux ?
- En regardant quelques tweets au hasard, peut-on deviner facilement la **"target"** ?
- Peut-on déjà détecter des **patterns** ou des **mots-clés** dans les tweets ?

2 - Text Processing

L'objectif est d'effectuer un premier **traitement des données textuelles** (colonne text). Il s’agira de **transformer les données textuelles en tokens** et de **réduire la dimensionnalité du corpus** en diminuant le nombre de tokens uniques.

Pour vous aider dans ce travail, voici quelques questions à explorer :

- Pouvez-vous écrire une fonction qui :
    - Tokenize un document ?
    - Supprime les stopwords ?
    - Supprime les tokens de moins de 3 lettres ?
- Comment peut-on **reconstituer le corpus** (c'est-à-dire un texte avec l’ensemble des documents) ?
- Une fois ce corpus constitué, combien de **tokens uniques** contient-il ? Ce nombre vous paraît-il faible, important ou gigantesque ?

- Comment peut-on **réduire la taille du vocabulaire** de ce corpus ?
- Combien de tokens sont présents **une seule fois** ? Ces tokens sont-ils utiles ?
- L'application d'une méthode de **stemmatisation** ou de **lemmatisation** peut-elle aider à réduire la dimensionnalité du corpus ?
- Comment visualiser graphiquement les tokens les plus présents (ex : **WordCloud**) ?
- Pouvez-vous appliquer tous ces traitements pour créer une **nouvelle colonne "text" plus pertinente** ?

3 - Appliquer les méthodes de vectorisation sur le dataset de tweets

4 - Entrainer 4 modèles par type de vectorisation

5 - Optimiser les hyper paramètres (CV etc)

6 - Comparer les différentes méthodes et choisir le meilleur tout en justifiant pourquoi.