Preprocessing

Le TP se décompose en 2 parties :

1 - **Exploratory Data Analysis** : il vous est demandé de faire un premier notebook afin de **comprendre, d'explorer et d'effectuer un premier nettoyage** des 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 dans les données ?
- Y a t-il des tweets anormalement longs / courts ? Peut-on les considérer comme des outliers ?
- Quel est le ratio tweet qui parlent de "catastrophes" / tweet 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** : Il vous est demandé 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 réduisant le vocabulaire (le nombre de tokens différents).

Pour vous aider dans ce travail, essayez de répondre aux questions suivantes :

- 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 le constitue? Ce nombre vous apparaît-il faible, important, gigantesque ?
- Comment réduire ce nombre de tokens uniques, ou autrement dit “comment réduire la taille du vocabulaire” de ce corpus ?
- Combien de tokens sont présents une seule fois ? Ces tokens nous seront-ils utiles ?
- Appliquer une méthode de stemmatisation ou de lemmatisation peut-elle nous aider à réduire la dimensionnalité du corpus ?
- Comment visualiser graphiquement, par un WordCloud par exemple, les tokens les plus présents ?
- Pouvez vous appliquer tous les traitements évoqués afin de créer une nouvelle colonne “text” qui serait plus pertinente ?