TP n°3 : Présentation vectorielle du texte

L'objectif de ce TP est de comprendre les concepts (N-grams et TF-IDF) expliqués précédemment.

Exercice 1

Dans cet exercice, nous allons utiliser le dataset suivant : **tweets.csv**

1. Appliquer les opérations nécessaires de prétraitement à ce dataset.
2. Donner la représentation en utilisant la méthode des n-grammes (avec n =2,3,4).

Exercice 2

Dans cet exercice, nous allons utiliser le dataset suivant :

Natural_Language_Processing_Text.txt

1. Appliquer les opérations nécessaires de prétraitement à ce dataset en plus d'une tokenisation du texte sous forme de phrase (token = sentence).
2. Donner la représentation en utilisant la méthode des n-grammes (avec n =2,3,4).
3. Effectuer la vectorisation de texte à l'aide de TF-IDF.

Exercice 3

Dans cet exercice nous allons utiliser un dataset (**train_rel_2.tsv**) qui comprend les réponses des élèves à dix ensembles différents de questions courte et les scores attribués par deux évaluateurs humains. L'ensemble de données est disponible ici sous forme de fichier (TSV). L'ensemble de données se compose des variables suivantes :

- **Id** : Un identifiant unique pour chaque tentative d'étudiant.
- **EssaySet** : un identifiant pour chaque ensemble d'essais (de 1 à 10).
- **Score 1** : score de l'évaluateur 1 (allant de 0 à 2).
- **Score2** : score de l'évaluateur2 (allant de 0 à 2).
- **EssayText** : réponse de l'étudiant (données textuelles).

Chaque document comprend un ensemble de mots contribuant au sens de la phrase, ainsi que des mots vides (par exemple, articles, prépositions, pronoms et conjonctions) qui n'ajoutent pas beaucoup d'informations au texte. Étant donné que les mots vides sont très courants et pourtant ils ne fournissent que des informations de bas niveau, les supprimer du texte peut nous aider à mettre en évidence les mots les plus importants pour chaque document. De plus, les textes en minuscules-majuscules et la lemmatisation sont d'autres facteurs qui peuvent avoir un impact sur la vectorisation du texte.

1. Appliquer un processus de prétraitement au texte qui implique la suppression des mots vides, la conversion des lettres majuscules en lettres minuscules et la lemmatisation.
2. Effectuer la vectorisation de texte à l'aide de TF-IDF.

Exercice 4

Dans cet exercice nous allons utiliser le dataset (**News_Category_Dataset.json**). Ce dataset est constitué de deux colonnes : « Text » et « Category ».

- Le texte représente « news article »
- La catégorie peut être l'une de ces 4 : 'BUSINESS', 'SPORTS', 'CRIME', 'SCIENCE'.

1. Appliquer les opérations nécessaires de prétraitement à ce dataset.
2. Donner la représentation vectorielle en utilisant les techniques ci-dessous :
    a. Bag of words
    b. N-grams
    c. Tf-IDF