2eme année Second Cycle Option : IASD
Module : Natural Language Processing

Fiche TP N° 01 : LE PRETRAITEMENT EN NLP

Objectif :

➔ Connaitre les notions de base des prétraitements en NLP

Librairies nécessaires

Nltk

SpaCy

WordCloud

Pywaffle

A. Préparation de données

1. Importer le jeu de données **spooky.csv** à partir de l'URL https://github.com/GU4243-ADS/spring2018-project1-ginnyqg/raw/master/data/spooky.csv en utilisant pandas et afficher les **10** premiers échantillons.

B. Nettoyage d'un texte

1. Gérer les caractères répétitifs (par exemple « cooooool » → « cool »)
2. Manipuler des homoglyphes (par exemple « $tupide » → « stupide »)
3. Transformer les entrées spéciales telles que les URL, les adresses e-mail et les balises HTML à une forme canonique.
4. Mettre tous les caractères en minuscule.
5. Supprimer la ponctuation
6. Supprimer les mots-vide.

C. Segmentation

1. Segmenter chaque phrase sur les espaces / la ponctuation
2. Segmenter chaque phrase avec un algorithme de segmentation basé sur des règles
3. Segmenter chaque phrase avec un algorithme de segmentation en sous mots (Subword Tokenization)

D. Reconnaissance d'entité nommée

1. Pour chaque phrase représentait les entités nommées (avec NLTK ou SpaCy).

E. Réduction des formes

1. Avec NLTK utiliser la lemmatisation et la racinisation.

Optionnel: Effectuer les mêmes tâches avec SpaCy

F. Analyse des fréquences

1. Compter le nombre de phrases, pour chaque auteur, où apparaît le mot **Great**.

2. Utiliser **pywaffle** pour obtenir un graphique qui résume de manière synthétique le nombre d'occurrences du mot "

great ” par auteur.

3. Refaire l’analyse avec le mot “impossible ”.

4. En utilisant la fonction **wordCloud**, faire trois nuages de mots pour représenter les mots les plus utilisés par chaque auteur.

5. En utilisant la fonction **wordCloud**, montrer les 100 meilleurs mots positifs et négatifs utilisés par les auteurs.