MODULE TEXT MINING: TEXT PRE-PROCESSING

EXERCICES
MASTER DS 2020, S3
J.ZAHIR

Le corpus webtext de NLTK sera utilisé dans ce TP:

- `nltk.download('webtext')`,
- `from nltk.corpus import webtext`

1. Enoncé

(1) Stocker dans un dataframe de quatre colonnes, le nom du document, le nombre de stop words, de tokens et de tokens en Majuscule pour chaque document. Afficher les 5 premières et les 5 dernières lignes du dataframe.
(2) Eliminer les stops words et transformer toutes les majuscules en minuscules.
(3) Trouver les 10 tokens les plus fréquents en utilisant la fonction most_common du Counter
(4) Trouver les 10 tokens les plus rares et les éliminer.
(5) Ecrire une fonction qui élimine les chiffres et la ponctuation dans un document textuel.
(6) Ecrire une fonction qui convertit les chiffres dans un document textuel en mots ($1 => un$).
(7) Ecrire une fonction pour le comptage d'emojis dans un texte
(8) Appliquer un stemming de porter (`from nltk.stem import PorterStemmer`).
(9) Appliquer un stemming de Lancaster et comparer avec le résultat précèdent
    - `from nltk.stem.lancaster import LancasterStemmer`.
(10) Reprendre le text initial et appliquer la lemmatization (`from nltk.stem import WordNetLemmatizer`). Qu'est ce que WordNet ? Quelles langues sont supportées ?