MODULE TEXT MINING: TD1 - TEXT VECTORIZATION

EXERCICES
MASTER DS 2020, S3
J.ZAHIR

Pour réaliser ce TP, vous aurez besoin d'une installation Python et des packages TextBlob, Sklearn, NLTK et pandas, éventuellement. L'objectif est de répondre aux questions de la section 1 en effectuant les tests sur les documents présentés dans la section 2.

1. Enoncé

(1) Ecrire une fonction TF, qui reçoit un mot w et un document d et calcule $TF(w, d)$.
(2) Ecrire une fonction IDF, qui reçoit un mot w et une collection de documents D et calcule $IDF(w, D)$.
(3) Ecrire une fonction $TF\text{-}IDF$ qui prend en entrée un mot w, un document d et une collection D et calcule $TF\text{-}IDF(w, d, D)$. Utiliser **TextBlob** pour la tokenization.
(4) Quelles sont les valeurs TF, IDF et $TF\text{-}IDF$ pour le mot « **boy** » du document 1 ?
(5) Créer et afficher la matrice *term-document*, pour les 4 documents ci- dessous, en utilisant la fonction $TF\text{-}IDF(w, d, D)$ pour les poids et les *mots* pour les attributs. La matrice ne doit contenir aucune valeur « NaN ». Convertir la matrice obtenue au format *document-term*.
(6) Créer et afficher la matrice *term-document* pour les même documents, mais cette fois-ci en utilisant **TfidfVectorizer** de **sklearn.feature_extraction.text**
(7) Comparer la matrice obtenue dans la question 6 avec celle obtenue dans la question 5. Sont elles identiques ? Si non, pourquoi ?
(8) Importer le corpus shakespeare de **NLTK** et créer la matrice *term-document* en utilisant $TF\text{-}IDF$ pour le poids et des bigrams pour les attributs.

2. Documents

(1) "You are trying to code TF-IDF all by yourself like a big girl/boy."
(2) "So this is a tinny doc."
(3) "And another tinny doc to test few stuff."
(4) "So in total, we are four documents, have fun ;)."