TP TIW - Natural Language Processing

Ce TP consiste à prédire si un commentaire sur une vidéo youtube est un spam ou un ham.

Jeu de données à télécharger :
https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection

1. Importer les données en un dataframe contenant tous les commentaires (CONTENT) et leur label (CLASS). La classe est 1 pour les spam et 0 pour les ham.

2. Installer la librairie *nltk* et réaliser les 4 étapes de prétraitements (normalisation, tokénisation, suppression stopwords et lemmatization) sur le commentaire de votre choix.

3. Créer une fonction qui effectue ces 4 étapes à partir du commentaire passé en paramètre et appliquez-la sur la totalité des commentaires du dataframe contenant les données.

4. À l'aide de librairie *sklearn*, transformer votre liste de mots en matrice de nombre avec TFIDF pour pouvoir effectuer une classification par la suite.

5. Appliquer un SVM ou un autre algorithme de classification sur ce nouveau format de données et regarder la qualité des prédictions avec la matrice de confusion et l'accuracy.