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

Fiche TP N° 02: Feature extraction and embeddings

Objectif :

➔ Familiariser avec les méthodes de vectorisation.

A. Préparation de données

Importer le jeu de données **spooky.csv** que vous avez déjà prétraité (la version finale après le prétraitement).

B. Encodage de la variable à prédire (facultatif)

Encoder les labels en utilisant l'encodage OneHot.

C. Construction des bases d'entraînement et de test

1. Diviser le dataset en deux parties : entrainement et test, en utilisant train_test_split, la taille du test 30% et random_state=0.
2. Le dataset est déséquilibré (Imbalanced dataset), stratifier les échantillons de manière à obtenir une répartition similaire dans chaque classe du dataset.

D. Méthodes de vectorisation

1. Utiliser la méthode de fréquence lexicale et one-hot encoding pour vectoriser le dataset d'entrainement et du test.
2. Entrainer un modèle de vectorisation TF-IDF sur la partie d'entrainement et vectorisez-le.
3. En utilisant le même modèle, vectoriser la partie du test.

E. Entrainement

1. Créer trois modèles du type MLPClassifier. (Vous pouvez changer l'algorithme d'apprentissage : utiliser les autres algorithmes de scikit-learn)
2. Entrainer ces trois modèles sur les trois représentations vectorielles.
3. Prédire les classes en appliquant les trois modèles sur les trois représentations d'entrainement.
4. Afficher le rapport de classification en utilisant les mesures de performance (accuracy, precision, recall…).

F. Test

1. Prédire les classes en appliquant les trois modèles sur les trois représentations de test.
2. Afficher le rapport de classification en utilisant les mesures de performance (accuracy, precision, recall…)
3. Calculer le temps de prédiction pour chaque modèle

G. Vectorisations basées sur les *embeddings* de mots

1. Utiliser les techniques de représentation vectorielle basées sur les prolongements de mots (Word embedding) :
    a. Word2Vec (CBOW et Skip gram)
    b. Glove
    c. FastText.

H. Entrainement / Test

1. Les mêmes questions des sections E et F mais avec les nouvelles représentations vectorielles présentées dans la section G.
2. Comparer tous les modèles réalisés dans ce TP.
3. Analyser les différentes méthodes de vectorisations et déterminer quelle méthodes est la meilleure et pourquoi ? Justifier votre réponse.