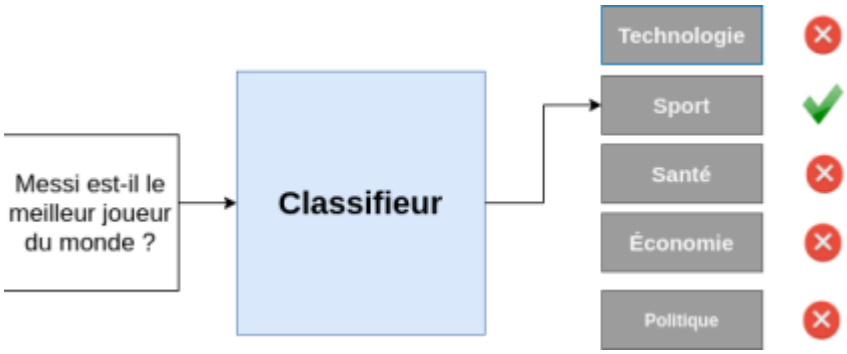
ÉCOLE NATIONALE SUPÉRIEURE D'ARTS ET MÉTIERS
UNIVERSITÉ HASSAN II DE CASABLANCA

Année Universitaire: 2021-2022

Projet de Machine Learning & Text Mining

Catégorisation des textes :

La classification (ou catégorisation) de textes est l'une des tâches de traitement du langage naturel (NLP : Natural Language Processing) les plus courantes. Elle consiste à associer un texte non-structuré à une étiquette, qui correspond à une classe bien précise. Si la catégorisation de textes nécessite beaucoup d'engouement, c'est à cause de ses nombreuses applications qui vont de l'analyse de sentiments à la détection de spams en passant par la détection de langue et les systèmes de recommandation.[1]



Dans ce projet, vous allez concevoir et développer un modèle intelligent pour classer les articles selon leur catégorie. Pour cela, 141 articles ont été extraits du site web Hespress. Les articles sont répartis en 5 catégories : sport, économie, culture, politique et société.

Travail à faire :

1. Vérifier les fichiers textes de chaque catégorie et regrouper les fichiers dans une seule data set. Nous signalons que les textes dans les fichiers actuels sont séparés par :
   ---------- numéro article ----------
2. Chercher le meilleur classifieur en menant une étude empirique comparative entre les techniques de machine learning (travaillées dans les séances de TP).
3. Les livrables sont : rapport + code source Python
   Dans le rapport,
   - il faut suivre une démarche adéquate avec un projet machine learning : collecte données, analyse et description du dataset…
   - N'hésiter pas à utiliser toutes les techniques permettant d'améliorer la performance du modèle (feature selection, optimisation,….)

[1] https://ledatascientist.com/

- L’évaluation est basée sur l’accuracy = Nombre articles bien classés/nombre total des articles
- La base de test est composée de 30% du dataset. La technique hold out tout en gardant la représentativité des catégories dans testing dataset.
- L’accuracy à utiliser est la moyenne de dix exécution du modèle
- La comparaison doit inclure aux moins 3 techniques vues en cours. Toutefois, vous pouvez utiliser d’autres techniques.