janvier 2025

S6.C.01 Machine learning- deep learning

Une des problématiques que rencontrent les « vendeurs » de livres, de produits, …, tels que Amazon (Spotify, …) est d'évaluer la polarité des avis (commentaires) laissés par les acheteurs, les lecteurs, etc. Ces avis permettent à ces plateformes de mieux orienter leurs achats vers leurs potentiels consommateurs.

Dans le cas de la SAE, nous avons collecté des commentaires d'utilisateurs sur des livres. Ces livres sont décrits par des métadonnées « classiques » (titre, auteurs, éditeur, etc.), et on leur a associé les commentaires laissés par les utilisateurs.
Le dataset est composé de deux fichiers, le premier comporte la description des livres avec plusieurs métadonnées, et le second comporte les commentaires laissés par les lecteurs, en plus de ce commentaire les lecteurs ont également donné un score entre 1 et 5 (étoiles).

L'objectif de la SAE est de prédire la polarité et le score des commentaires en exploitant différentes représentations de textes.

Avant de se lancer dans cette phase de prédiction, il est demandé, d'effectuer quelques analyses afin de mieux comprendre la répartition des données de ce Dataset.

Quelques analyses de données à effectuer

- Distribution des livres sur le marché en fonction du genre
- Les mots les plus fréquents dans les revues ayant obtenues un score >5 (afficher sous forme de WordCloud)
- Les Livres les plus commentés (évalués)
- Les Livres les mieux notés
- Dans quels genres les lecteurs donnent-ils des évaluations positives et négatives ?
- Les 10 auteurs les mieux notés avec 5 étoiles
- Les 10 auteurs les mieux notés avec 1 étoile
- Existe t-il une corrélation entre l'éditeur et le commentaire (score)
- Existe-t-il une corrélation entre les auteurs et le commentaire (score).
- Distribution des sentiments négatifs, neutres et positifs dans l'ensemble du corpus
- Nombre le plus élevé de critiques négatives/positives/neutres sur les livres
- Existe-t-il une corrélation entre le prix et la revue (le score de la revue)

Ce qui est demandé :

- Prédiction de la polarité des commentaires : La première prédiction est assez simple, il s'agit de prédire la polarité du commentaire Positif, négatif ou neutre (on peut effectuer une réparation assez simple, si le score >3 alors le commentaire est Positif, score <3 polarité négative, 3 neutre) .

- Prédiction des scores des commentaires : construire un modèle capable de prédire le score que va attribuer le lecteur à partir de son commentaire).

Il est demandé d'exploiter et de comparer

- Différents modèles de représentation de textes :
  - Mots simples
  - Tf.idf
  - Embeddings BERT, modèle LLM ala GPT

- Différentes méthodes d'apprentissage :
  - Algorithmes classiques- algorithmes de Deep learning
  - Utilisation d'un modèle de type Transformer (BERT ?..) (Cf. la plate forme HuggingFace)

PS : une revue (commentaire) comporte différents champs utiles, en particulier, le résumé du commentaire, le texte complet du commentaire, Il est demandé d'évaluer l'impact de ces différents textes dans le modèle de prédiction.