Analyse de Sentiments sur Twitter

Le projet **d'analyse de Tweet** de la base Sentiment140 vise à utiliser des techniques d'apprentissage automatique pour classer les tweets en fonction de leur polarité (positif, négatif ou neutre).

La base de données Sentiment140 contient 1,6 million de tweets étiquetés manuellement, ce qui en fait un jeu de données idéal pour entraîner et évaluer des modèles d'analyse de sentiments.

Les tweets ont été collectés via Twitter et comprennent des informations telles que le texte du tweet, le nom d'utilisateur, la date et l'heure de publication, ainsi que le nombre de retweets et de favoris.

Les modèles d'apprentissage automatique seront entraînés sur ces données pour prédire la polarité d'un tweet en fonction de son contenu textuel. Il sera intéressant de voir comment différentes variables, telles que la longueur du tweet, l'utilisation des émoticônes, les hashtags et les mentions, affectent la prédiction de polarité.

En outre, nous pourrons également explorer comment la polarité des tweets varie en fonction des heures et des jours de la semaine, ainsi que selon les sujets les plus discutés. Les résultats de ce projet pourraient être utilisés pour comprendre les opinions et les sentiments des gens sur différents sujets sur les réseaux sociaux. Cela pourrait être utile pour les entreprises qui cherchent à évaluer la réception de leurs produits ou campagnes publicitaires, ainsi que pour les chercheurs en sciences sociales qui étudient les tendances et les opinions sur les réseaux sociaux.
En somme, ce projet a pour objectif de comprendre comment les tweets sont utilisés pour exprimer les sentiments et les opinions sur les différents sujets, et comment ces sentiments évoluent au cours du temps.

La base de données de sentiment140 contient 1,6 millions de données d'entraînement. Ce qui est impossible à lire quand on appelle la méthode *read_csv* de pandas sans paramètres.
L'idée, pour vous, serait d'abord de réduire la taille des données afin d'avoir une bonne base de données pour entraîner les Tweets. Donc libre à vous de choisir la taille qui vous convient du moment où **le modèle que vous entraînez reste efficace** et capable de prédire n'importe quelle tweet avec une bonne accuracy ou une autre métrique.
Si par exemple, vous faites un premier extrait et que votre modèle **n'est pas efficace sur les données test**, pensez à augmenter votre base de données ou de trouver un moyen en machine learning permettant d'optimiser vos paramètres (GreadSearCV par exemple) pour avoir une bonne métrique.
On rappelle qu'on a en face de nous des vraies données et arriver à construire un modèle robuste capable de prédire un Tweet comme étant positif, négatif ou neutre serait pour vous une bonne prouesse.
Donc sortez le meilleur de vous et prenez du temps pour bien faire et structurer le projet et sachez que ce dernier ne se fera pas en 2 ou 3 jours mais sur une longue durée.

Il est attendu de vous un bon notebook bien détaillé avec beaucoup de commentaires et des arguments afin de défendre votre projet et dites-vous que vous allez le rendre à une personne qui ne connaît rien au NLP donc tâchez d'être le plus explicite possible.

<u>quelques informations sur les données:</u>

- Vous pouvez ouvrir les données d'entraînement (1,6 millions de lignes) grâce à des éditeurs comme Vscode, Notepad de window, sublim text etc . Donc libre à vous de trouver une manière de réduire la taille des données pour que ces derniers puissent être lus par pandas.
  Au lieu de celà aussi, vous pouvez utiliser une méthode

read_csv pour échantillonner des données et en choisir une partie **(regarder les paramètres de *read_csv*)**

- Les données test continent 3 étiquettes pour la variable cible : 0= négative, 2=neutre et 4= positive alors que les données d'entraînement contiennent que deux étiquettes pour la variable cible : 0=négative et 4=positive. Donc libre à vous de trouver un moyen pour traiter cette irrégularité mais sachez que ce sont les données du fichier de test qui doivent être utiliser comme donnée test, donc la méthode **traint_test_split** n'a pas lieu d'être dans votre code.
- Lien pour télécharger le fichier ; **https://www.kaggle.com/datasets/kazanova/sentiment140?resource=download**

Soyez Créatifs,
Bonne chance

Date limite de rendu du projet : 1er Novembre 2024