TP : Classification des Emails en Spam ou Ham avec Naive Bayes

Objectif :

L'objectif de ce TP est de construire un modèle de classification des emails en utilisant l'algorithme Naive Bayes. Nous allons explorer les étapes nécessaires pour préparer les données, entraîner le modèle, l'évaluer et l'utiliser pour prédire de nouvelles étiquettes pour les emails.

Étapes du TP :

1. **Importer les données :**
   Commencez par importer les bibliothèques nécessaires, telles que scikit-learn, pandas, numpy, etc.
2. **Prétraitement des données :**

En plus du nettoyage du texte, envisagez d'autres techniques de prétraitement :

- Suppression des doublons : Éliminez les doublons pour éviter d'inclure plusieurs fois le même email.
- Extraction de fonctionnalités : Identifiez des caractéristiques pertinentes dans les emails, telles que la longueur du texte, la présence de liens, etc.
- Analyse des pièces jointes : Si vos emails contiennent des pièces jointes, explorez également leur contenu.

3. **Exploration des données :**

- Utilisez des visualisations pour comprendre la distribution des étiquettes (spam vs. ham).
- Analysez les caractéristiques des emails (longueur, fréquence de mots, etc.).

4. **Séparation des ensembles d'entraînement et de test :**

- Divisez vos données en ensembles d'entraînement et de test .
- Assurez-vous que les proportions de spam et de ham sont similaires dans les deux ensembles.

5. **Entraînement du modèle Naive Bayes :**

- Utilisez la bibliothèque scikit-learn pour entraîner un classificateur Naive Bayes.
- Il est basé sur le théorème de Bayes et suppose que les caractéristiques sont indépendantes.
- Les paramètres incluent le type de modèle (Multinomial, Bernoulli, etc.) et les hyperparamètres (lissage, etc.).

6. **Évaluation du modèle :**

- Utilisez des mesures telles que la précision, le rappel, le score F1 et la matrice de confusion pour évaluer les performances du modèle.
- Verifier est-ce-que ces emails sont-ils des spam ?
  - Email1: 'Hey mohan, can we get together to watch footbal game tomorrow?',
  - Email2: 'Upto 20% discount on parking, exclusive offer just for you. Dont miss this reward!'