Projet 2 : Analyse des Sentiments sur les Critiques de Films avec Naive Bayes

Contexte du projet :

L'objectif de ce projet est de classifier des critiques de films en sentiments positifs ou négatifs en utilisant des techniques de traitement du langage naturel (NLP) et un modèle Naive Bayes. Nous utiliserons un jeu de données contenant des critiques de films et appliquerons des méthodes comme la lemmatisation, la vectorisation (Bag of Words et TF-IDF) et la classification probabiliste.

Données :

Le jeu de données contient les informations suivantes :

- **Critique** : Le texte de la critique du film.
- **Sentiment** : La classification du sentiment (positif ou négatif).

Étapes du projet :

1. Récupération du Jeu de Données

- Utiliser un jeu de données préexistant, comme **IMDB Reviews Dataset** disponible sur Kaggle.
- Charger les données et les inspecter pour comprendre leur structure.

2. Prétraitement du Texte

- **Nettoyage du texte** :
    - Suppression de la ponctuation et des caractères spéciaux.
    - Conversion en minuscules.
    - Suppression des mots vides (**stopwords**).
- **Lemmatisation** : Réduction des mots à leur forme racine (ég. "running" devient "run").

3. Vectorisation du Texte

- **Bag of Words (BoW)** : Création d'une matrice de comptage des mots.
- **TF-IDF (Term Frequency - Inverse Document Frequency)** : Conversion du texte en une représentation numérique en pondérant l'importance des mots.

4. Séparation des Données

- Diviser les données en **ensemble d'entraînement** (80%) et **ensemble de test** (20%).

5. Entraînement du Modèle Naive Bayes

- Utilisation de **Multinomial Naive Bayes**, bien adapté à la classification de texte.
- Entraînement du modèle sur les données vectorisées.

6. Évaluation du Modèle

- Calcul des métriques de performance :
    - **Précision (accuracy)**
    - **Rappel (recall)**
    - **Score F1**
    - **Matrice de confusion** pour visualiser les erreurs de classification.

7. Analyse des Résultats

- Identifier les critiques mal classifiées.
- Observer les mots les plus influents pour chaque classe (positif/négatif).
- Tester l'impact des différentes techniques de vectorisation sur la précision du modèle.