RAPPORT DE PROJET :
SYSTEME DE RECOMMANDATION DE FILMS BASE SUR LES RESUMES

Réalisé par :

Mariem Gmati

Oumayma Yakoubi

Natural language processing

5éme année informatique Groupe 3

[Polytechnique Sousse]

Introduction

Ce projet vise à développer un système de recommandation de films qui exploite la similarité textuelle des résumés de films pour suggérer des contenus similaires. Contrairement aux approches traditionnelles basées sur les évaluations ou les genres, ce système utilise des techniques de traitement du langage naturel (NLP) pour analyser les descriptions narratives des films.

1. Jeu de Données

- **Source** : Kaggle - IMDb Movies Dataset
- **Caractéristiques Clés** :
  - `Series_Title` : Titre du film.
  - `Overview` : Résumé du film.
  - `Genre`, `IMDB_Rating`, `Director`, etc.
- **Nettoyage** :
  - Suppression des doublons.
  - Gestion des valeurs manquantes.
  - Normalisation des textes (élimination des caractères spéciaux, mise en minuscule).
  - Conversion des colonnes numériques comme `Gross` en type approprié après traitement des chaînes de caractères.

2. Prétraitement des Données

- **Traitement des textes** :
  - Tokenisation des résumés.
  - Suppression des mots vides (stopwords).
  - Lemmatisation pour réduire les mots à leur forme canonique.
  - Application d'une pipeline de nettoyage pour homogénéiser les données textuelles.

3. Vectorisation

- **Méthode Utilisée** : TF-IDF (Term Frequency-Inverse Document Frequency)
  - Représentation des résumés sous forme de vecteurs pour capturer l'importance des mots.
  - Extraction des 5000 caractères les plus significatifs pour limiter la taille des vecteurs.

4. Calcul de la Similarité

- **Métrique** : Cosine Similarity
    - Utilisée pour mesurer la similarité entre les vecteurs TF-IDF des résumés.
    - Les films avec des scores de similarité élevés sont considérés comme pertinents.

5. Modèle de Recommandation

- **Logique** :
    - L'utilisateur entre un résumé.
    - Le système calcule les similarités entre le résumé entré et ceux des films dans la base de données.
    - Les films ayant les scores les plus élevés sont recommandés.
- **Implémentation** :
    - Le modèle s'appuie sur une base de données prétraitée pour rechercher des correspondances en temps réel.

6. Évaluation du Modèle

- **Métriques Utilisées** :
    - **Précision** : 0.60
    - **Rappel** : 1.00
    - **F1-score** : 0.75

Ces résultats montrent que le système identifie efficacement les films pertinents, bien qu'il inclue quelques recommandations non pertinentes.

- **Validation Croisée** : Une partie des données a été utilisée pour tester la qualité des recommandations.
- **Observations** :
    - Les films aux résumés similaires sont bien identifiés.
    - La performance varie en fonction de la qualité et de la longueur des résumés.

7. Interface Utilisateur

- **Technologies** : Flask, HTML/CSS
- **Fonctionnalités** :
    - Champ de saisie pour le résumé du film.
    - Bouton d'envoi pour déclencher la recherche.
    - Affichage des recommandations avec le titre et le résumé de chaque film.
- **Améliorations Apportées** :
    - Interface responsive pour une utilisation sur mobile et ordinateur.
    - Design simple et épuré pour une expérience utilisateur optimale.

8. Résultats et Discussion

- **Résultats** :
    - Le système fournit des recommandations pertinentes basées sur les similarités de contenu.
    - Les recommandations sont bien alignées avec les attentes pour les résumés bien écrits.

- **Points Forts** :
    - Exploitation efficace des résumés pour des suggestions précises.
    - Interface utilisateur intuitive.

- **Limites** :
    - Peut inclure des recommandations peu pertinentes si les résumés sont mal écrits.
    - Dépendance aux descriptions textuelles pour les recommandations.

9. Perspectives

- Intégration de méthodes avancées de NLP comme BERT ou GPT pour une meilleure compréhension contextuelle des résumés.
- Ajout de fonctionnalités basées sur les retours utilisateurs pour évaluer la qualité des recommandations.
- Élargissement du jeu de données pour inclure plus de films et diversifier les recommandations.
- Mise en place d'un système d'apprentissage continu pour améliorer les résultats au fil du temps.

Conclusion

Ce projet démontre comment les techniques de NLP et la similarité de contenu peuvent être utilisées pour créer un système de recommandation efficace. Avec des améliorations futures, ce système pourrait rivaliser avec des modèles plus complexes tout en restant compréhensible et simple à utiliser.