FidelisProd

Catalogue des données

YONLI FIDELE

2023

Le catalogue des données est un référentiel centralisé répertoriant toutes les sources de données utilisées, leurs caractéristiques, leur origine et leur utilisation dans le projet de développement du "Système de Recommandation de Films".

Sources de Données

1. Ensemble de Données MovieLens :
   - Description : Données de notation de films collectées auprès des utilisateurs.
   - Type : Données tabulaires (CSV)
   - Origine : Téléchargées à partir du site MovieLens.
   - Colonnes Principales : userId, movieId, rating, timestamp.
2. Informations Utilisateur (u.user) :
   - Description : Profils des utilisateurs incluant l'âge, le genre et l'occupation.
   - Type : Données tabulaires (CSV)
   - Origine : Téléchargées à partir de MovieLens.
   - Colonnes Principales : userId, age, gender, occupation.
3. Informations Film (u.item) :
   - Description : Détails des films incluant le titre, la date de sortie et les genres.
   - Type : Données tabulaires (CSV)
   - Origine : Téléchargées à partir de MovieLens.
   - Colonnes Principales : movieId, title, release_date, genres.

Utilisation des Données

1. Entraînement du Modèle ALS : Les données de notation de films et les informations utilisateur sont utilisées pour entraîner le modèle de filtrage collaboratif.
2. Validation Croisée : Les ensembles d'entraînement et de test sont utilisés pour évaluer la performance du modèle.

Gestion des Données

1. Les données sont stockées localement dans des fichiers CSV pour une utilisation dans l'environnement Spark.
2. Des processus de nettoyage et de prétraitement sont appliqués pour garantir la qualité des données utilisées dans le modèle.