

Catalogue de données pour le projet d'analyse et recommandation des films dans la plateforme Netflix

Description des données :

L'ensemble de données sur les évaluations de films comprend une collection complète de 100 000 évaluations de films fournies par 943 utilisateurs pour 1 682 éléments distincts. Chaque utilisateur a consciencieusement évalué au moins 20 films, et les utilisateurs ainsi que les éléments sont identifiés de manière unique par des identifiants numériques attribués consécutivement. Les données sont organisées dans un format séparé par des tabulations.

1. Movies (**u_item.txt**) :

Contient des informations sur chaque film, notamment son titre, sa date de sortie et ses genres.

Fields :

- Movie ID
- Title
- Release Date
- Genres (Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western)

2. Users (**u_user.txt**) :

Contient des informations démographiques sur chaque utilisateur, telles que l'âge, le sexe, la profession et le code postal.

Fields :

- User ID
- Age
- Gender
- Occupation
- Zip Code

3. Ratings (**u_data.txt**) :

Contient les notes attribuées par les utilisateurs aux films, ainsi que les horodatages.

Fields :

- User ID
- Movie ID
- Rating
- Timestamp

4. Genres (u_genre.txt) :

Contient une liste de genres.

Fields :

- Genre ID
- Genre Name

5. Occupations (u_occupation.txt) :

Contient une liste de professions.

Fields :

- Occupation ID
- Occupation Name

Données utilisées dans le projet :

u_data :

Un ensemble de données fondamentales contenant 100 000 évaluations par 943 utilisateurs pour 1 682 films. Les évaluations des utilisateurs, accompagnées d'horodatages, fournissent une vue dynamique des préférences des utilisateurs au fil du temps.

u_item :

Informations sur les films, y compris les titres, les dates de sortie et les classifications par genre. Les genres sont représentés sous forme d'indicateurs binaires, ce qui permet une exploration détaillée des divers éléments thématiques présents dans chaque film.

u_genre :

Cet ensemble de données sert de référence pour les genres, en fournissant une liste de genres avec des identifiants uniques. Il constitue un lien essentiel pour comprendre et classer les films en fonction de leur contenu thématique.

Comment on a utilisé les données :

Exploration des évaluations de films et des préférences des utilisateurs :

Dataset: u_data

Objectif : Comprendre les préférences des utilisateurs et les modèles d'engagement.

Approche :

- Utilisez l'ensemble de données u_data pour explorer les évaluations de films fournies par les utilisateurs.
- Analysez les préférences des utilisateurs en fonction des classements, des genres et des modèles temporels.
- Identifiez les films et les genres populaires pour informer les stratégies de recommandation.

La curation de contenu basée sur le genre :

Datasets: u_genre, u_item

Objectif : Créer du contenu basé sur des genres spécifiques.

Approche :

- Exploitez u_genre et u_item pour comprendre le paysage des genres.
- Identifiez les genres en vogue et adaptez les recommandations de contenu en conséquence.
- Mettez en œuvre des stratégies de promotion de contenus alignés sur les préférences des utilisateurs.

Analyse des données en temps réel avec Kafka, Elasticsearch et Kibana :

Datasets: u_data

Objectif : Permettre le traitement des données en temps réel pour des recommandations dynamiques.

Approche :

- Configurez Kafka pour la diffusion en continu des données d'évaluation des films.
- Développez un module de transformation Python pour adapter les messages Kafka à Elasticsearch.
- Visualiser les données présentes dans Elasticsearch avec Kibana.

Date de publication : 05/03/2024

Signé : Sefdine Nassuf