Projet : Extraction et Recommandation de Films et Séries avec RDFLib et SPARQL

Auteurs :

- **Nima Nikookar**
- **Goumba Fate Amar**
- **Mamadou WADE**

Objectif

L'objectif principal de ce projet est de créer un système de recommandation de films et séries en utilisant des données RDF extraites de la base DBPedia. Ce système vise à fournir des recommandations précises basées sur des préférences utilisateurs et des algorithmes d'apprentissage automatique. Ce projet repose également sur l'utilisation de technologies modernes pour assurer une manipulation efficace des données complexes et une personnalisation optimale des recommandations.

1. Extraction des données

Description

- Source des données : DBPedia via des requêtes SPARQL.
- Langage utilisé : Python (modules SPARQLWrapper et RDFLib).

Données extraites

- Films :
  - Titre
  - Réalisateur
  - Année de sortie
  - Genre (e.g., action, comédie)
  - Résumé
  - Liste d'acteurs
  - Durée
- Réalisateurs : Nom, biographie et liste de films réalisés.
- Acteurs : Nom, biographie et liste de films joués.
- Genres : Nom et description.
- Utilisateurs : Identifiant et préférences cinématographiques (acteurs/genres préférés).

Processus

- Transformation des données extraites en triplets RDF.
- Ajout des triplets RDF dans un graphe existant en utilisant la méthode `g.add()` de **RDFLib.**
- Mise en place d'un pipeline automatisé pour assurer la cohérence et la fluidité des étapes d'extraction et de transformation des données.

2. Prétraitement des données

Nous avons effectué le nettoyage et la normalisation des triplets RDF pour assurer la qualité des données.

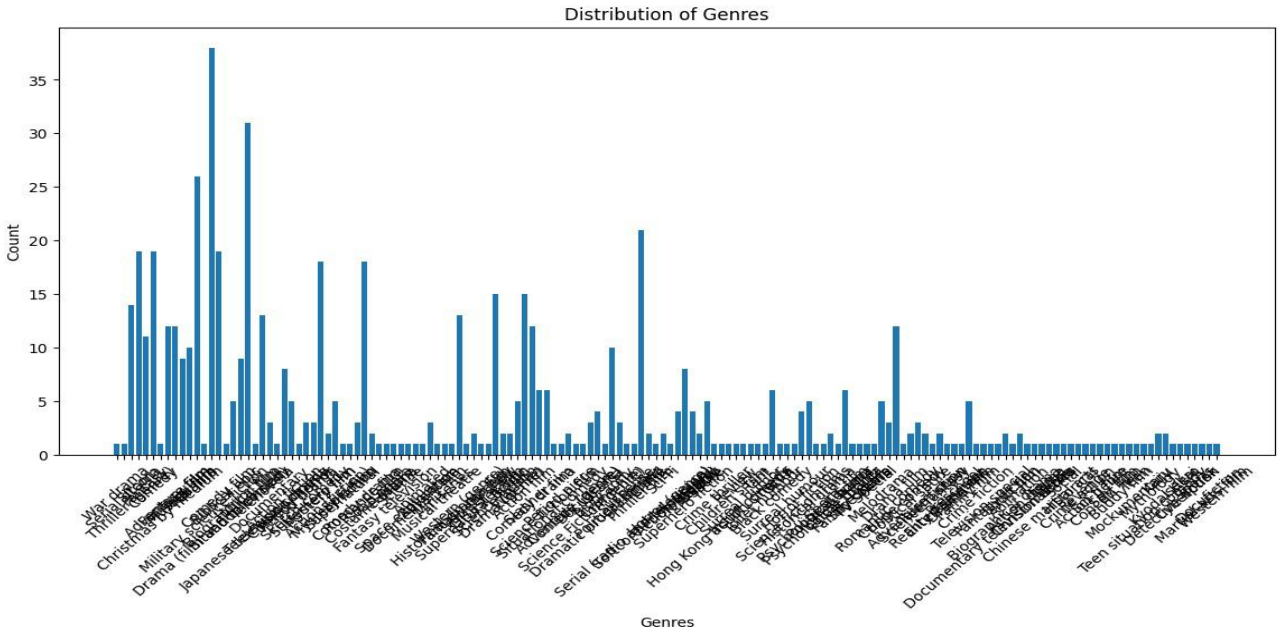
Actions réalisées

- Suppression des doublons : Identification des triplets dupliqués.
- Gestion des valeurs manquantes : Filtrage des littéraux vides.
- Normalisation : Conversion des littéraux en minuscules et suppression des espaces inutiles.
- Amélioration des références croisées : Association des entités connexes pour enrichir la structure du graphe RDF.

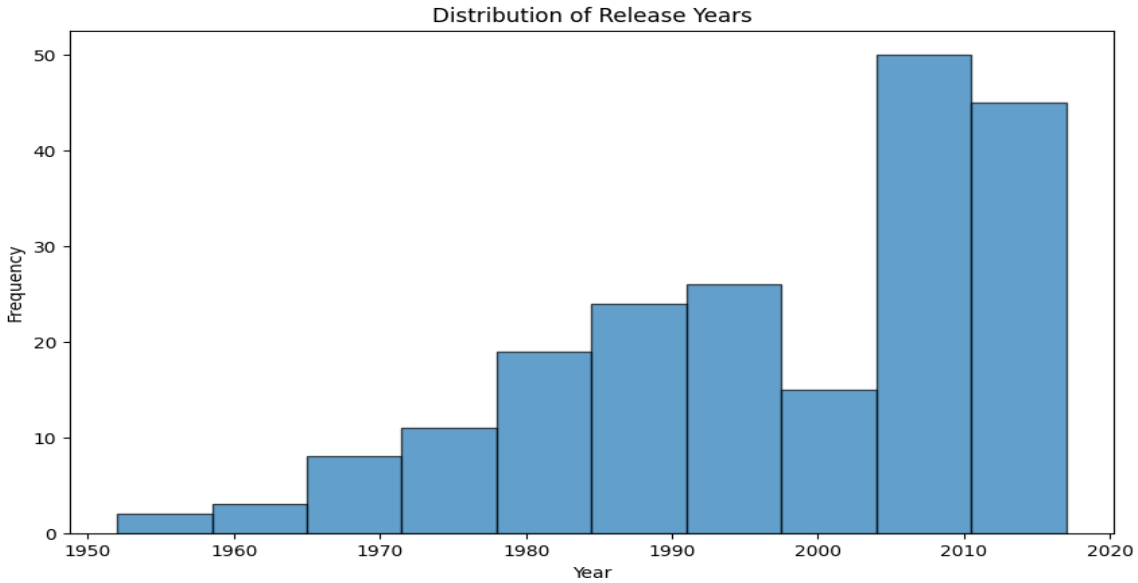
3. Analyse exploratoire des données (EDA)

Visualisations réalisées

1. Distribution des genres : Histogramme montrant le nombre de films par genre.



2. Distribution des années de sortie : Histogramme des années de sortie des films.



Observations

- Les genres les plus fréquents sont la comédie et l'action.
- La majorité des films répertoriés sont sortis entre 1990 et 2020.
- Des clusters distincts d'associations entre réalisateurs et genres ont été identifiés, ce qui pourrait guider les recommandations.

4. Système de Recommandation

Nous avons utilisé plusieurs approches :

1. **Basée sur le contenu** : Analyse des similarités entre les genres, réalisateurs et acteurs.
Pour cela le **machine Learning** est employé en utilisant de TF-IDF et de la similarité cosinus entre le vecteur de préférence de l'utilisateur et vecteur de chaque film pour générer des recommandations.
2. **Réseaux RDF enrichis** : Exploitation des relations entre entités pour affiner la personnalisation.

Recommandations Générées

Exemples de films recommandés (avec score de similarité) :

```
                          Label  Similarity
272               Call Me Claus    0.203589
289        Toy Story of Terror!    0.165601
293          Mazes and Monsters    0.161037
288                    Hung Out    0.157599
243          Glynis (TV series)    0.122028
207      Stir Crazy (TV series)    0.120555
3     The New Andy Griffith Show    0.118060
18     The New Bob Cummings Show    0.113192
213  Ricky Sprocket: Showbiz Boy    0.108581
202                   CarniK Con    0.108533
Evaluation Scores: {'MAP@k': 0.3, 'NDCG@k': np.float64(1.0)}
```

5. Évaluation des Recommandations

- **Méthodes**

Plusieurs méthodes ont été utilisées pour évaluer la qualité des recommandations. La mesure **MAP@k (Mean Average Precision at k)** permet de calculer la précision moyenne des recommandations jusqu'au rang k, tandis que **NDCG@k (Normalized Discounted Cumulative Gain at k)** prend en compte la pertinence des recommandations en fonction de leur position dans la liste. En complément, un feedback utilisateur simulé a été effectué en analysant qualitativement les recommandations à l'aide de cas tests spécifiques.

- **Résultats**

Pour la prédiction des préférences utilisateurs, les résultats montrent un score de **MAP@k égal à 0,3** et un score de **NDCG@k atteignant 1,0**. Le feedback simulé a révélé que les utilisateurs fictifs estimaient que les recommandations manquaient de diversité, suggérant des pistes d'amélioration.

- **Analyse**

Les scores relativement faibles indiquent que le système de recommandation est perfectible. Pour améliorer la personnalisation, l'intégration d'approches collaboratives pourrait être bénéfique. De

plus, une analyse plus fine des préférences des utilisateurs, en tenant compte des nuances dans leurs choix, permettrait d'augmenter significativement la pertinence des suggestions fournies par le système.

6. Conclusion et perspectives

- **Réalisations**

Ce projet a permis d'extraire avec succès des données RDF à partir de DBPedia en utilisant des requêtes SPARQL. Les données ont ensuite été intégrées dans un graphe RDF pour modéliser les relations entre les films, les réalisateurs, les acteurs et les utilisateurs. De plus, un système de recommandation basé sur le contenu a été implémenté, permettant de proposer des films en fonction des genres, des réalisateurs et des préférences des utilisateurs.

- **Améliorations futures**

Plusieurs améliorations pourraient être apportées pour optimiser ce système de recommandation. Tout d'abord, il serait bénéfique d'intégrer des méthodes de recommandation collaborative, qui exploiteraient les similitudes entre les utilisateurs pour affiner les suggestions. Ensuite, l'utilisation de jeux de données plus complets et diversifiés améliorerait la couverture des films et séries. Une évaluation approfondie avec des utilisateurs réels permettrait également de mieux mesurer l'efficacité des recommandations. Enfin, l'incorporation d'algorithmes d'IA avancés, tels que les modèles basés sur les graphes neuronaux, permettrait d'exploiter pleinement la structure des données RDF pour des recommandations plus précises et personnalisées.