Project Documentation: Hybrid Recommendation System

1. Introduction

This project focuses on building a hybrid recommendation system, combining collaborative filtering and content-based filtering. The objective was to provide personalized recommendations for users, based on their past interactions and product similarities.

2. Problem Statement

With the rapid increase in online shopping platforms and media consumption, providing personalized recommendations becomes critical for user engagement and satisfaction. The goal was to improve the existing recommendation system by combining two approaches: collaborative filtering (user-item interaction) and content-based filtering (item similarity).

3. Dataset Overview

The project involved two primary datasets:

- Product Dataset: Contained information about products such as product_id, product_name, brand_name, and product features.
- Review Dataset: Contained user interactions with the products including author_id, product_id, and ratings.

We faced several challenges in terms of data preparation, feature extraction, and handling sparsity within these datasets.

4. Recommendation Techniques Used

1. Collaborative Filtering:
- We used Singular Value Decomposition (SVD) to predict ratings based on user-item interactions.
- Collaborative filtering recommends items based on similar users or items the user interacted with previously.

2. Content-Based Filtering:
- Cosine similarity was calculated between products based on their features (e.g., product descriptions, brand, etc.).
- Content-based filtering suggests products similar to those the user has interacted with.

3. Hybrid Method:
- Both collaborative and content-based recommendations were combined using a weighted average to provide final recommendations.

5. Challenges and Solutions

1. Data Filtering and Merging Issues:

- Problem: Initially, we had separate datasets for products and reviews, with inconsistencies in product names and IDs.
  Solution: We performed inner joins on the product_id to align both datasets.
  Additionally, we removed duplicates and handled missing values in the product dataset, ensuring only relevant products were considered in the recommendations.

2. Sparsity of the User-Item Matrix:

- Problem: With over 99% sparsity, the user-item matrix was mostly empty, which negatively impacted collaborative filtering.
  Solution: We filtered the dataset, keeping only users and products with enough interactions to reduce the sparsity, though sparsity remained a challenge. We relied on SVD, which helps in dealing with sparse data by reducing dimensionality.

3. Using the Whole Dataset as Training Set:

- Problem: Due to data sparsity and limited interactions for many users and products, splitting the data into training and test sets wasn't feasible.
  Solution: We used the entire dataset as the training set for collaborative filtering to ensure the model had enough information to learn from user interactions.

4. Dealing with Inconsistent Entries:

- Problem: There were inconsistencies in product names and product_ids across datasets, leading to mismatches during recommendation generation.
  Solution: We standardized the datasets by ensuring product names and IDs were consistent across the different stages. Cleaning and reindexing were essential to ensure correct alignment between datasets.

6. Evaluation Metrics

To evaluate the hybrid recommendation system, the following metrics were used:
1. Root Mean Squared Error (RMSE): Evaluated the accuracy of predicted ratings against the true ratings.
2. Mean Absolute Error (MAE): Provided another measure of prediction accuracy.

7. Code Structure

The main components of the project are as follows:

- Data Preprocessing: Merging product and review datasets. Handling missing and inconsistent entries.
- Recommendation System: Collaborative filtering using SVD. Content-based filtering based on cosine similarity. Hybrid system combining both approaches.
- Evaluation: Calculation of RMSE, MAE to assess system performance.

8. Conclusion

This project successfully implemented a hybrid recommendation system that combines collaborative filtering and content-based filtering. By overcoming challenges related to data sparsity, inconsistencies, and limited training data, we were able to generate relevant product recommendations for users. The hybrid approach allowed for personalized suggestions by leveraging both user interaction history and product similarities.

9. Future Work

Future improvements can focus on:
- Enhancing model performance by experimenting with different collaborative filtering algorithms (e.g., ALS, NMF).
- Incorporating additional product features (e.g., product categories, user demographics) to improve content-based filtering.
- Further reducing data sparsity by employing matrix factorization techniques or imputation strategies.