Code Summary: Content-Based and Collaborative Filtering for Movie Recommendations

Harsh Anand

May 13, 2024

1 Content-Based Filtering

1. **Data Preparation**:
   - Load movie data from a CSV file (`movies.csv`) containing movie IDs, titles, and genres.
   - Clean and preprocess the data by replacing '—' with spaces in the 'genres' column.
2. **TF-IDF Vectorization**:
   - Utilize `TfidfVectorizer` from scikit-learn to compute TF-IDF vectors for movie genres.
   - Reduce dimensionality using `TruncatedSVD` to limit the size of the TF-IDF vectors.
3. **User Preferences Representation**:
   - Load user tag data from another CSV file (`tags.csv`), containing user IDs and their tagged genres.
   - Group user preferences by user ID and aggregate tags.
4. **TF-IDF Vectorization for Users**:
   - Similar to movies, compute TF-IDF vectors for user preferences, reducing dimensionality using `TruncatedSVD`.
5. **Calculating Similarities**:
   - Calculate cosine similarity between each user's TF-IDF vector and all movie TF-IDF vectors.
   - Recommend the top movies with the highest similarity scores for each user.

2 Collaborative Filtering (Attempted)

1. **Data Preparation**:
   - Load user ratings data from a CSV file (`ratings.csv`) containing user IDs, movie IDs, and ratings.
2. **Creating User-Item Matrix**:
   - Convert the ratings data into a sparse user-item matrix.
3. **Memory Optimization**:
   - Attempt to split the user-item matrix into smaller subsets due to memory constraints.
4. **Recommendation Generation**:
   - Utilize the `surprise` library to implement KNN-based collaborative filtering.
   - Define functions to generate recommendations for each user segment based on similar users' ratings.

3 Challenges Encountered

- Memory errors due to the large size of the dataset, especially for collaborative filtering.
- Difficulty in proceeding with collaborative filtering due to dataset size and memory constraints.

4 Conclusion

- Successfully implemented content-based filtering for movie recommendations.
- Attempted collaborative filtering but faced challenges due to dataset size and memory constraints.