Bogdankov Nikita

Introduction

The movie industry has witnessed a significant shift in recent years, with movie recommendations playing a pivotal role in enhancing user experience. This report focuses on an extensive exploration of movie-related data.

Data analysis

User data distribution:

1. Gender distribution pie chart was visualized, from which we gained that males give rating to movies more often.
2. Also histogram of the ages was explored, and from this graph we can see that the average age of those who give rating is between 25 and 40.

Movie data distribution:

1. The most popular genre is drama, then comes comedy, and the least popular are fantasy, film-noir and animation (not considering unknown).
2. The distribution of movie ratings was visualized using a histogram from which I received that the most common rating is 4.
3. The release year distribution of movies was explored. The movies which have the largest number of reviews are those who released in 1990-1996

Model Implementation

The solution focuses on implementing an Implicit Alternating Least Squares (ALS) Matrix Factorization model for movie recommendations.

Data Preprocessing:
The 'ua.base' file was used to create a sparse user-item matrix. The data was then fed into the ALS model.

ALS Model:
The ALS model was configured with 10 factors, regularization of 0.02, and 600 iterations. The model was trained on the sparse matrix, enabling it to learn latent factors for users and items.

Recommendations:
The model generated personalized movie recommendations for a given user, showcasing movie IDs and associated scores.

Model Advantages and Disadvantages

Advantages:

- Implicit Matrix Factorization enables capturing latent factors for both users and items.
- The model is capable of providing personalized recommendations based on user behavior.

Disadvantages:

- The model's performance heavily relies on the quality and quantity of user-item interactions in the training data.
- Tuning hyperparameters might be required for optimal results.

Training Process

The ALS model was trained using the 'ua.base' data, allowing it to learn the underlying patterns in user-item interactions.

Evaluation

The model's performance was evaluated using the 'ua.test' dataset. Precision at K (P@K) was computed for each user, and the average P@K was calculated to assess the overall effectiveness of the model. The received precision is more than 10%.

Results

The precision at K values for each user and the average precision at K across all users were calculated, providing insights into the model's overall performance (10.4% is the precision).

In conclusion, this report outlines a comprehensive exploration of movie-related data and the implementation of an Implicit Matrix Factorization model for movie recommendations.