MovieRecommender

Yograj Karki

5/8/2021

Introduction

In this project, I'll be looking at the movies database and create a recommender system. Entertainment services like Neflix, Hulu, YouTube have millions/billions of contents. It becomes challenging for the customer to select the right one. At this place, recommender systems come into the picture and help the user to find the right item by minimizing the options. The recommendation system will be based on collaborative filtering. The idea is if someone watches movies that you've watched and liked, you'll be recommended the other movies that user has watched. Therefore, recommending movies is dependent on creating a relationship of similarity between the two users. With the help of recommenderlab package, we can compute similarities using various operators like cosine, pearson etc.

Research questions

1. What are the top viewed movies?
2. What are the most rated movies?
3. What are the factors that most affect the movies recommendation?
4. Movies distribution on the basis of genres
5. What genre is most rated or liked?
6.

Approach

- Gather movies data and ratings data
- Prepare data
  - Deal with missing values, erroneous values, outliers
- Selecting useful data Selecting those movies which is watched and rated by certain number of people. Filtering out least watched and rated movies.
- Normalizing data In the case of some users, there can be high ratings or low ratings provided to all of the watched films. This will act as a bias while implementing the model. In order to remove this, data needs to be normalized.
- Binarizing the data Finally, binarizing the ratings as 1s and 0s. Ratings equal or above 3 will be 1 and below 3 will be 0.
- Collaborative Filtering system This is the part that needs most of the work. I'm still working on it.

Data

Data will be gathered from movielens platform of grouplens.org which has 20 million ratings and 465,000 tag applications applied to 27,000 movies by 138,000 users. https://grouplens.org/datasets/movielens/20m/

Required packages

1. ggplot2 - for charts and other visualizations
2. recommenderlab - to create a recommender algorithm
3. data.table - for data manipulation
4. reshape2 - data shape manipulation

Plots and Tables

1.
2. Heatmap of top users and movies
3. Distribution of average rating per user
4. Normalized ratings of the top users
5.

Question for the future steps

1. As there are different datasets available, which one to choose?
2. What scale range to set for the normalization of ratings?
3. Need to learn recommenderlab package