2020 Massive Data Analysis Term Project

Recommendation System

109062623 林鎰鋒 Group51

A movie recommendation system which describes the details as following.
In the first part, I will build an "Item(movie)-Based Collaborative Filtering" to find the similarity of movie using "Cosine Similarity", and in the second part I will build a simple recommender system for "Rating Predictions" to predict the movie rating for each user by the top 10 similar movies' rating from first part.
The whole program is implement by MapReduce on PySpark.

I. Dataset

From MovieLens: https://grouplens.org/datasets/movielens/
This dataset (***ml-latest-small***) describes 5-star rating from MovieLens, a movie recommendation service.
It contains **100836 ratings** across **9742 movies** which created by **610 users.**
All ratings are contained in the file ***ratings.csv***. Each line of this file after the header row represents one rating of one movie by one user, and has the following format:

userId,movieId,rating,timestamp

II. Item-item Collaborative Filtering

Calculate the similarity for each movie pairs

- Cosine similarity:

$$sim(x,y) = \cos(r_x, r_y) = \frac{r_x \cdot r_y}{\|r_x\| \cdot \|r_y\|}$$

*** The mapper and reducer implement detail please see the *ipynb* file**

Output Result: https://www.dropbox.com/s/ao53hl6rfkbwl4p/similarity.out?dl=0

III. Rating Predictions

Select top 10 similarity to calculate the movie rating for each user

- Rating Predictions:

$$r_{xi} = \frac{\sum_{j \in N(i;x)} s_{ij} \cdot r_{xj}}{\sum_{j \in N(i;x)} s_{ij}}$$

*** The mapper and reducer implement detail please see the *ipynb file***

Output Result: https://www.dropbox.com/s/wtcogl0lv1fxx7n/predict.out?dl=0