2020 Massive Data Analysis Term Project

Deadline: 2021/1/1(五) 23:59

Recommendation System: Item-item Collaborative Filtering

From lecture **Recommendation Systems P.27 ~ 30**

Similarity: cosine similarity with subtract mean

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

output format:

(item, item), similarity

以上為基本分

Rating Predictions

Select top 10 similarity to calculate the movie rating for each user.
i.e., N = 10

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

output format:

(user, item), rating

以上為更高分

Input Data

From MovieLens: https://grouplens.org/datasets/movielens/

It contains 610 users and 9742 movies.

format:

(userID, movieID, rating)

Other Topics

Select one of the algorithms from Chapter 10

You can get your dataset from any source on the Internet, ex: Facebook, Instagram, YouTube, PTT, or any other dataset on the Internet etc.

Remember to describe the dataset you used in your report.

Filename:

Term_Project_Groupnumber.java or

Term_Project_Groupnumber.ipynb

Report:

Term_Project_Groupnumber.pdf

Describe what you done in your term project, includes your code and some explanation. Also, describe your dataset.

Pack the above files in Term_Project_Groupnumber.zip and upload to ilms.

Demo:

From 2021/1/4 ~ 2021/1/8 at **資電館 831**

Caution:

Except web crawler, you have to complete Term Project with Hadoop or PySpark. i.e., you have to use MapReduce to complete this Term Project.