Big Data Processing
Project Proposal

Issam Falih

April 7, 2025

Goal: Design a recommender system

Data Description

- **movies_metadata.csv**: The main Movies Metadata file. Contains information on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.
- **keywords.csv**: Contains the movie plot keywords for our MovieLens movies. Available in the form of a stringified JSON Object.
- **credits.csv**: Consists of Cast and Crew Information for all our movies. Available in the form of a JSON Object.
- **links.csv**: The file that contains the TMDB and IMDB IDs of all the movies featured in the Full MovieLens dataset.
- **ratings.csv**: This file contains 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

Questions

1. Make a good analysis of the dataset.
2. Build regression models to predict movie revenue and vote averages.
3. Use collaborative filtering to build a movie recommendation system with two functions:
   - 3.1 Suggest top N movies similar to a given movie title.
   - 3.2 Predict user rating for the movies they have not rated for. You may use a test set to test your prediction accuracy, in which the test ratings can be regarded as not rated during training .

Data can be found here
https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

Context & problematic

Recommender systems *Recommender systems represent user preferences for the purpose of suggesting items to purchase or examine. They have become fundamental applications in electronic commerce and information access, providing suggestions that effectively prune large information spaces so that users are directed toward those items that best meet their needs and preferences. [?].*

Recommender systems

- Content-based.
- Collaborative filtering.
- Hybrid.

Recommender systems

- Content-based :

Look at the content of the items and try to retrieve features specific for a certain type of items. Textual features are usually used as features for content-based systems..

- Collaborative filtering
- Hybrid.

Recommender systems

- Content-based.
- Collaborative filtering :

Can be divided into user-based and item-based collaborative filtering. In the user-based case, similarity between two users is computed while in the item-based case, similarity between two items is computed..

- Hybrid.

Recommender systems

- Content-based.
- Collaborative filtering
- Hybrid :

Combine the outputs of collaborative filtering and content-based methods using for example linear combinations of predicted ratings or voting schemes. Latent factor models were also added to collaborative filtering and content-based to improve the representation of user preferences.

Context & problematic

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	1	-	4	-	-	2	2	-	3	-
u_2	-	3	1	-	2	-	-	5	4	-
u_3	4	-	2	-	-	5	-	5	-	5
u_4	1	-	4	-	2	2	2	-	2	-
u_5	-	3	-	5	-	5	-	-	3	3

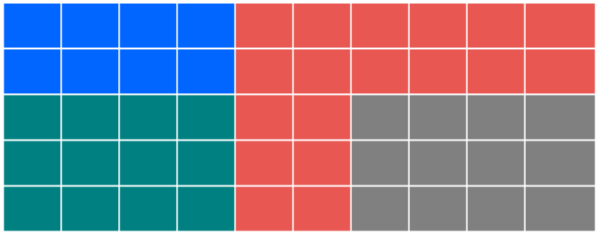
Table: Users-Items matrix

Recommendation system

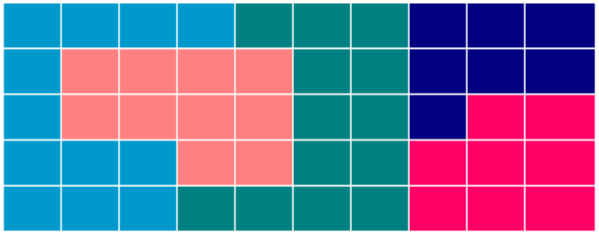
Learning system : clustering on both users and items

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	1	-	4	-	-	2	2	-	3	-
u_2	-	3	1	-	2	-	-	5	4	-
u_3	4	-	2	-	-	5	-	5	-	5
u_4	1	-	4	-	2	2	2	-	2	-
u_5	-	3	-	5	-	5	-	-	3	3

Users-Items matrix



Clustering users



Clustering items

Recommendation system

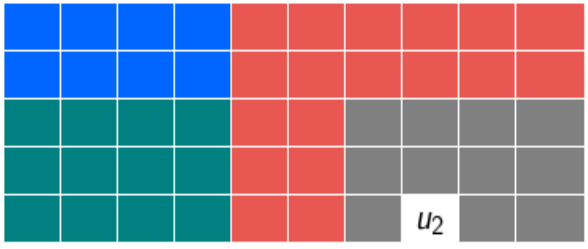
Prediction system (Recommendation) : find the rate value given by the user u2 for the item i7

Step 1 : find the corresponding clusters of user u2 and item i7

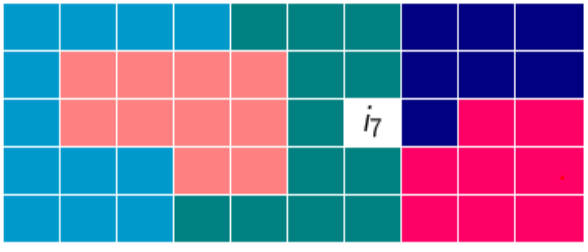
	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	1	-	4	-	-	2	2	-	3	-
u_2	-	3	1	-	2	-	?	5	4	-
u_3	4	-	2	-	-	5	-	5	-	5
u_4	1	-	4	-	2	2	2	-	2	-
u_5	-	3	-	5	-	5	-	-	3	3

Users-Items matrix

Find the corresponding cluster for user u2



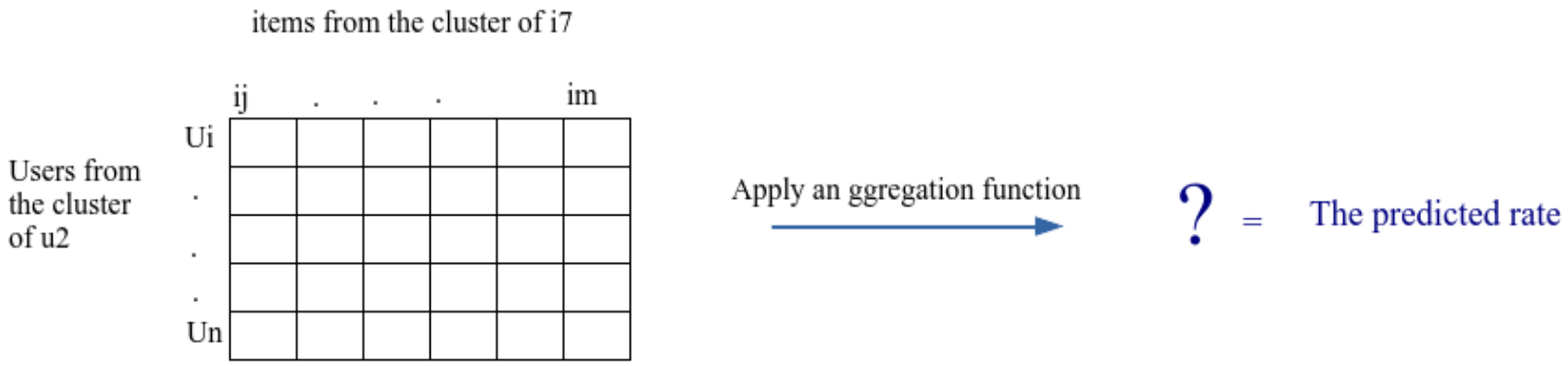
Find the corresponding cluster for item i7



Recommendation system

Prediction system (Recommendation) : find the rate value given by the user u2 for the item i7

Step 2 : aggregate the ratings from the corresponding clusters



Aggregation function : statistical measure i.e. mode, median, ...

Deposit & Practical information

Due Date

The due date is on **April 30, 2025**. All materials including your report should be dropped on Moodle (boostcamp) by **all** members of the team using this format: `LastName1_LastName2.zip`