Movies Recommendation System

Reem Ibrahim Amer

1

About Our Dataset

Movies Dataset

- These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset.
- The dataset consists of movies released on or before July 2017.
- Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages.
- This dataset also has files containing 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.

Movies Dataset cont.

This dataset consists of the following files:

- **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 stringified JSON Object.
- **links.csv:** The file that contains the TMDB and IMDB IDs of all the movies featured in the Full MovieLens dataset.
- **links_small.csv:** Contains the TMDB and IMDB IDs of a small subset of 9,000 movies of the Full Dataset.
- **ratings.csv:** Contains about 26 million of users ratings.

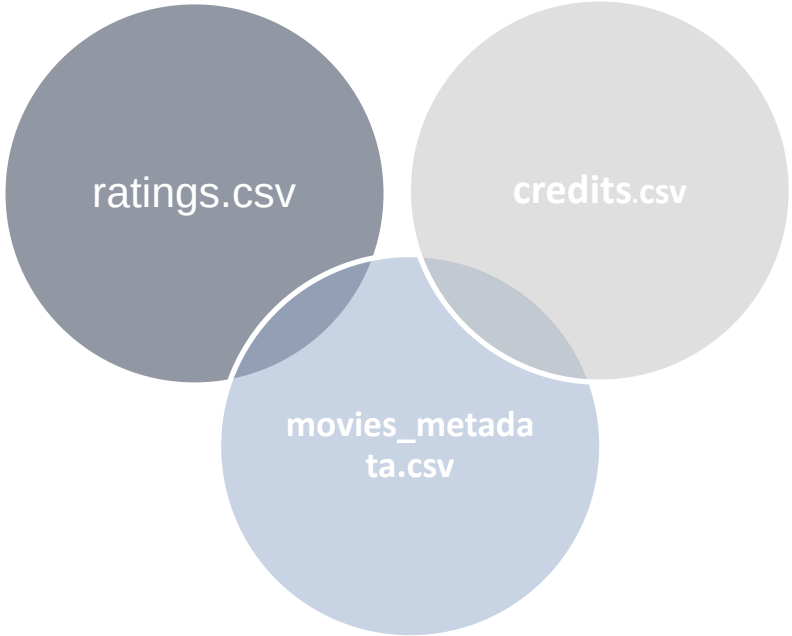
Movies Dataset cont.

This dataset consists of the following files:

- **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.
- **credits.csv:** Consists of Cast and Crew Information for all our movies. Available in the form of a stringified JSON Object.
- **ratings.csv:** Contains about 26 million of users ratings.

Movies Dataset cont.

We will be using the following files only:



ratings.csv

	userId	movieId	rating	timestamp
0	1	110	1.0	1425941529
1	1	147	4.5	1425942435
2	1	858	5.0	1425941523
3	1	1221	5.0	1425941546
4	1	1246	5.0	1425941556

- Consists of 4 columns and 26024289 rows
- Contains the ratings of each user for multiple movies.

Credits.csv

	cast	crew	id
0	[{'cast_id': 14, 'character': 'Woody (voice)',...	[{'credit_id': '52fe4284c3a36847f8024f49', 'de...	862
1	[{'cast_id': 1, 'character': 'Alan Parrish', '...	[{'credit_id': '52fe44bfc3a36847f80a7cd1', 'de...	8844
2	[{'cast_id': 2, 'character': 'Max Goldman', 'c...	[{'credit_id': '52fe466a9251416c75077a89', 'de...	15602
3	[{'cast_id': 1, 'character': "Savannah 'Vannah...	[{'credit_id': '52fe44779251416c91011acb', 'de...	31357
4	[{'cast_id': 1, 'character': 'George Banks', '...	[{'credit_id': '52fe44959251416c75039ed7', 'de...	11862

- Consists of 3 columns and 45476 rows
- Cast and crew are in json format.

Movies_metadata.csv

Column	Non-Null Count	Dtype
adult	45466 non-null	object
belongs_to_collection	4494 non-null	object
budget	45466 non-null	object
genres	45466 non-null	object
homepage	7782 non-null	object
id	45466 non-null	object
imdb_id	45449 non-null	object
original_language	45455 non-null	object
original_title	45466 non-null	object
overview	44512 non-null	object
popularity	45461 non-null	object
poster_path	45080 non-null	object
production_companies	45463 non-null	object
production_countries	45463 non-null	object
release_date	45379 non-null	object
revenue	45460 non-null	float64
runtime	45203 non-null	float64
spoken_languages	45460 non-null	object
status	45379 non-null	object
tagline	20412 non-null	object
title	45460 non-null	object
video	45460 non-null	object
vote_average	45460 non-null	float64
vote count	45460 non-null	float64

- Consists of 24 columns and 45466 rows
- Missing values in multiple columns

Project Idea

Project Idea

1. Build recommendation system using deep learning and avoid using traditional ways in machine learning such as matrix factorization.
2. build our system by using implicit feedback rather than explicit feedback.
   - **Explicit feedback:** explicit feedback are direct and quantitative data collected from users
   - **Implicit feedback:** are collected indirectly from user interactions.
3. Use leave-one-out methodology as a strategy for train-test split.
4. Use Ranking Evaluation Metrics for Recommender Systems such as HR(Hit Ratio)

Thank You