Python 大作业描述

# About Dataset:

* Links.csv

|  |  |  |
| --- | --- | --- |
| **movieId** | **imdbId** | **tmdbId** |

* Movies.csv

|  |  |  |
| --- | --- | --- |
| **movieId** | **title** | **genres** |

* Rating.csv

|  |  |  |  |
| --- | --- | --- | --- |
| **userId** | **movieId** | **rating** | **timestamp** |

* Tags.csv

|  |  |  |  |
| --- | --- | --- | --- |
| **userId** | **movieId** | **tag** | **timestamp** |

**Summary:**

There are a total of **9742 Movie's** in our dataset with **100836 Users Rating** the movie and **3683 Tags**.

**userId**: Unique Id provided for each User

***userId*** were selected at random for inclusion. Their ids have been anonymized. User ids are consistent between ratings.csv and tags.csv (i.e., the same id refers to the same user across the two files).

**movieId**: Unique Id provided for each Movie

Only movies with at least one rating or tag are included in the dataset. These movie ids are consistent with those used on the MovieLens web site (e.g., id 1 corresponds to the URL [**Movie Lens**](https://movielens.org/movies/1). Movie ids are consistent between ratings.csv, tags.csv, movies.csv, and links.csv (i.e., the same id refers to the same movie across these four data files).

**rating** (rating.csv): Ratings are made on a 5-star scale, with half-star increments (0.5 stars - 5.0 stars).

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

*Mean* show the average rating is **3.2-Star**

**genres**: Genres are a pipe-separated list, and are selected from the following:

Action

Adventure

Animation

Children's

Comedy

* + Crime
  + Documentary
  + Drama
  + Fantasy
  + Film-Noir
  + Horror
  + Musical
  + Mystery
  + Romance
  + Sci-Fi
  + Thriller
  + War
  + Western
  + (no genres listed)

# 基本内容

**Task**: Predict movie star ratings based on movie attributes

• Randomly divide 9742 movies into training set (data\_train) and validation set (data\_val)

• Use the training set to train the model and predict the labels of the validation set movies

• Repeat "training set/validation set division, training, and testing" three times, and report the average accuracy of the three times

**数据集划分：**

* 可组队
* 不多于3人
* 提交时间
* 2023.12.31

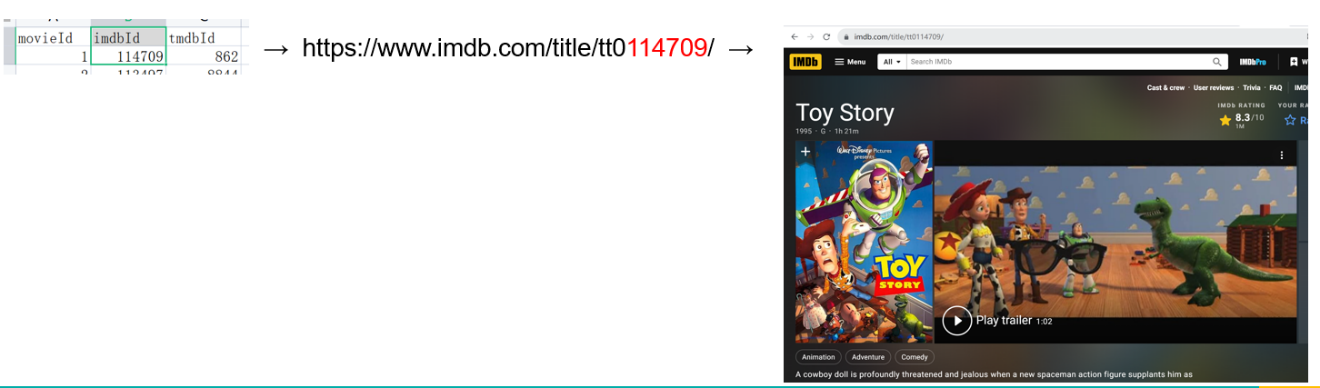
**提交内容**

* 可运行复现的pytorch代码，运行说明，作业报告（pdf）
* 打分依据
* 平均准确率 + 算法实验报告
* 同组同分

# Task tips

1. **Data reading**: table operation
2. **Data exploration** --- the relationship between each attribute and the star rating, how to display it, etc.
3. **Data preprocessing** --- missing values, multiple users, etc.
4. **Data modeling**: In addition to the ready-made information in csv, is additional information such as movie posters related to star ratings?

Encourage the use of link.csv link crawlers to obtain additional information

**For example:**