Global Objective

The main goal of the Capstone Project is to apply techniques and topics covered in the course on one of the standard recommender systems dataset. You will work in teams of three or four people.

Submissions will be divided into three parts - each part will be worth 40 points. Each of the assignments will have mandatory and optional tasks.

Datasets

1. MovieLens Dataset [1M version] - link
2. Book Crossing Dataset - link
3. Yelp Dataset - link
4. Jester Dataset - link
5. Last.fm Music Dataset - link

Submission #1: Setup and naïve implementations

Total Points: 40 (32 mandatory, 8 additional)
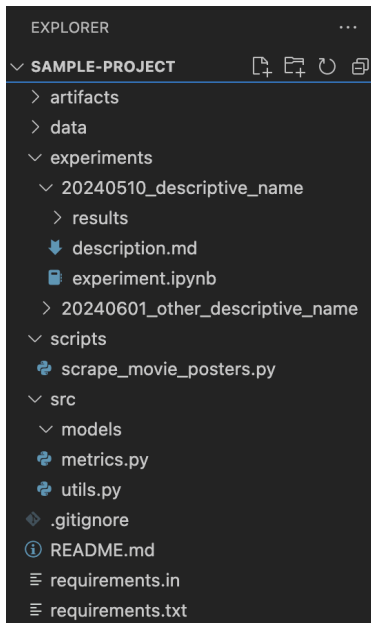Submission Structure:
- link to the repository
- branch and commit hash for evaluation
- pdf report [optional]

Mandatory Tasks (32 points)

Repository Setup And Exploratory Data Analysis(8 points)

Objective

Set up a repository with a clear and structured layout to support iterative experimentation, results tracking and code reusability. As a reference for the repository organization, you can use the attached screenshot, but you're not required to mimic it entirely.

EXPLORER …

∨ SAMPLE-PROJECT

> artifacts
> data
∨ experiments
  ∨ 20240510_descriptive_name
    > results
    description.md
    experiment.ipynb
  > 20240601_other_descriptive_name
∨ scripts
  scrape_movie_posters.py
∨ src
  ∨ models
  metrics.py
  utils.py
.gitignore
README.md
requirements.in
requirements.txt

Familiarize yourself with the dataset by conducting a comprehensive EDA to understand the dynamics of the selected dataset, identify patterns, and inform subsequent model development. EDA should contain statistical summaries, distribution visualizations, and any insights on data quality or peculiarity.

Deliverables

- GitHub repository
- README.md file with repository structure overview and setup instructions
- Jupyter Notebook with the EDA

Offline Evaluation Framework + Baseline Recommender (14 points)

Objective

Define an offline evaluation methodology for assessing recommendation quality. The methodology should cover both classical ML metrics for classification/regression and ranking specific approaches.
Offline evaluation approach should be demonstrated on the simple baseline recommender, such as a popularity-based or mean-rating recommender.

Deliverables

- Metrics implementation
- Description of evaluation methodology
- Baseline model implementation
- Jupyter notebook with the experiment that evaluates baseline recommender

Content-Based and Collaborative Filtering Implementation (12 points)

Objective

Implement content-based and collaborative filtering models.

- Content-based filtering model
- User-user and item-item collaborative filtering model
- Experiment with model evaluations and comparison to baseline recommender

Additional Task (8 points)

Graph-Based Analysis with PageRank (8 points)

Objective

Represent some part of the dataset as a connected graph and set up meaningful experiment with the PageRank algorithm. Provide analysis of the results, discussing potential impacts on recommender system strategies or insights into the dataset.

Deliverables

- PageRank implementation
- Jupyter Notebook with experiment setup and results interpretation