# Datasets

| Filename | Purpose/Content | Link |
|---|---|---|
| BT4222 Group 5 Movie Reco Project | Folder in which you can access ALL files, including subsidiary output | link |
| TMDB_movie_dataset_v11.csv | This is a full database of all movies with info like revenue, runtime, average rating, brief synopsis of movie, genre, production companies, updated till 2024. | link |
| ml-1m User ratings on movies dataset.zip<br>- users.csv<br>- ratings.csv<br>- movies.csv | Movielens dataset with user attributes such as gender, occupation and age as well as user ratings on the movies | link |
| Merged_df.csv | Merged & processed dataset with user attributes from movie lens and movie features from TMDB | link |
| top1movie-1.csv | Intermediary dataset - contains top 1 recommended movie from collab-model1.ipynb to be input for content-model1.ipynb to generate 5 more movie recommendations | link |

# Notebooks

| Filename | Purpose/Content | Link |
|---|---|---|
| Github repo in which files can be found | Overall Github Repo Link to all notebooks | link |
| data_processing.ipynb | This notebook is for our preprocessing before running the recommendation models. As we utilized data from 2 different sources, there is a need to merge the data from both sources together for a comprehensive analysis. This involved creation of a unique primary key for merging, string processing, and fuzzy wuzzy techniques to form a match on differing movie names which are actually referring to the same movie. | link |
| **Hybrid-model1 folder:**<br>collab-model1.ipynb<br>content-model1.ipynb | This notebook utilizes a 2-step approach. We first utilize the cosine similarity matrix approach to obtain the similarity matrix for user on user, based on their ratings of older movies and output a list of top 1 recommended movies for each user. Next we pass this list into the content based model which is based on movie to movie similarities, with a random probability of diversity, based on their features. We then recommend 5 newer movies (4 relevant movies + 1 random movie for diversity) for the same users based on the old movies' similarity to the newer ones. This method is a 2 step hybrid approach. | link |

| | | |
|---|---|---|
| hybrid-model2.ipynb | This notebook utilizes a neural collaborative filtering method to generate predictions for user ratings across all movies in the Movielens dataset using user attributes such as gender, occupation and age as features. The model is then extrapolated to newer movies from the Tmdb dataset such that it can recommend newer movies too. | link |
| hybrid-model3.ipynb | This notebook utilizes a deep learning approach, leveraging neural networks, and combines both content and collaborative filtering approaches to generate movie recommendations. This is the model which generates the best performance and is our main model to be utilized. It was also extrapolated to recommend newer movies from the TMDB dataset. Diversity was also addressed in this approach by recommending a movie that differed in genres (based on a weight) compared to the top 4 recommended movies. | link |