BT4222 Source Code & Datasets

Recommendation System 7



| **Name** | **Matriculation Number** |
| --- | --- |
| Aarushi Mundra | A0246207J |
| Dylan Wo Qiying | A0233634J |
| Kaavya Senthil Kumar | A0246227E |
| Kenny Low Sheng Wei | A0233869N |
| Ng Wei Han | A0234338E |
| Tasneem Benazir | A0238339W |

**Datasets used**

The [dataset](https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset) we chose consists of the following files:

1. [movies\_metadata](https://drive.google.com/file/d/1fiKyla5QlCGq1ERb5IT7GlDVxehkLK0P/view?usp=drive_link).csv: The main Movies Metadata file. Contains 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.
2. [keywords.csv](https://drive.google.com/file/d/1pwuGPB-f8sX0HBZm4lXIjwVNxpgBM6EE/view?usp=drive_link): Contains the movie plot keywords, in stringified JSON Object form.
3. [credits.csv](https://drive.google.com/file/d/1OgAdH8h-8gBTh3IvZAKkfilgJ7NP_PYJ/view?usp=drive_link): Consists of cast and crew Information for all our movies, in stringified JSON Object form.
4. [ratings\_small.csv](https://drive.google.com/file/d/1rVEKuncfTQUeS8eTzT_TxThskkCqeppV/view?usp=drive_link): The subset of 100,000 ratings from 700 users on 9,000 movies. Contains user ID, movie ID, rating and timestamp.
5. [links\_small.csv](https://drive.google.com/file/d/1CeVsfXl2MHqunJXJsPhGKg9EqT5rWNhq/view?usp=drive_link): Contains IMDB and TMDB IDs of all movies featured in the ratings\_small.csv file (About 9000 movies)

For CBF, we need sufficient information about a movie’s content, which can be extracted from movies\_metadata.csv, keywords.csv and credits.csv. We will also need data about each user’s preferences, which can be deduced from their rating and timestamp from ratings\_small.csv.

Likewise, for both user-based and item-based CF to work, we will also need information about a movie’s content and all user’s preferences.

We have also enlisted a CNN model that requires these files to run.

1. `saved\_model.pb` is the file where the models parameters are saved.
2. `variables` folder stores the variables of the processed posters utilized by the model.
3. `posters` folder stores the posters required to make recommendations.

**Source Code**

Entire Drive link - <https://drive.google.com/drive/folders/1CwjOOM9oVBcRKeLBOz3yxg-b5LereLny?usp=drive_link>

Step by Step notebook links -

[BT4222\_step1\_Data\_Preprocessing.ipynb](https://colab.research.google.com/drive/1SZaT4Y7AmMuONCRJTKzLwJLipFJFJl8S?usp=drive_link)

[BT4222\_step2\_Demographic Filtering](https://colab.research.google.com/drive/1rlRhagDkcgkRbzpO1pR7KJW7cT1yw6ym?ouid=115158940571709552379&usp=drive_link)

[BT4222\_step3\_Collaborative Filtering.ipynb](https://colab.research.google.com/drive/1hCAQ699tklUtYFj5weBAjfa1lOiH_-b7?ouid=115158940571709552379&usp=drive_link)

[BT4222\_step4\_CBF\_Hybrid.ipynb](https://drive.google.com/file/d/12qWP0twe9CivZFAbxm02a37CclCxaZMs/view?usp=drive_link)