Initial Technical Report

Data

* Source: <https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset>
  + Metadata on over 45,000 movies and 100k ratings
* EDA Notes
  + Movies Dataset
    - Top 3 Genres: Drama, Comedy, Thriller
    - Most common votes for a movie are within the range of 6-7 (out of 10)
    - Avg. Movie runtime is about an hour and half
  + Ratings Dataset
    - Most users average a rating between 3.5 – 4 (out of 5)
    - 12.5% of users are outliers in regard to rating movies (they have rated more than 285 unique movies)
* Data Handling
  + Filtered out – duplicates, non-null values, and non-English movies
  + Columns dropped:
    - Adult
    - Belongs\_to\_collection
    - Budget
    - Homepage
    - Imdb\_id
    - Original\_language (since the dataframe is filtered to only English movies)
    - Original\_title (duplicate column to “title”)
    - Production\_companies
    - Production\_countries
    - Revenue
    - Spoken\_language (filtered to all English movies)
    - Tagline
    - Video
    - Vote\_count
  + Converted JSON columns to Iterable lists, these columns include:
    - Genres
    - Keywords

Model Selection

* Recommendation Engine
  + Top 10 Movies (ranked by popularity)
    - If user provides no data/preferences
  + Content-Based Filtering
    - If the user provides preferred movie attributes but no additional movie ratings
  + Collaborative Filtering
    - If the user provides additional movie ratings but no preferred movie attributes
  + Hybrid Filtering
    - If the user provides all the required information (preferred movie attributes and additional movie ratings)

Performance Metrics

* MAP@K (Mean Average Precision at K)
* MAR@K (Mean Average Recall at K)
* Coverage
  + Percentage of movies that the recommendation will use
* Personalization
  + Does each user get different recommendations
* Intra-list Similarity
  + How similar are the movies that are recommended