Codebook for FCP Dataset
ISQS 6338 (D20 & D21) – Summer II – 2022

ratings.csv

This file contains user ratings for movies as comma-separated-values (CSV), and includes user data, movie data, rating data, and a timestamp when a specific rating was given. It also includes the genre(s) of each movie (if any).

tagged.csv

This file contains free-text movie-tagging activities of users for movies as comma-separated-values (CSV), and includes user data, movie data, free-text tagging data, and a timestamp when a movie was tagged by a user. It also includes the genre(s) of each movie (if any).

genome-scores.csv

This file contains data about how strongly movies exhibit particular properties as represented by tags. Each movie is given a relevance score for each of 1,128 tags.

Below you will find a codebook contain details about each column in the various CSV files:

- *userId* – the system-assigned identifier of a user
- *birthdate* – the birth date of a given user
- *gender* – gender of the user, if know. A value of "u" is assigned if unknown/ not specified
- *zip* – U.S.-based 5-digit zip code of the user's location
    - Note: If a zip value is less than five digits, then the preceding zero(es) have been omitted by the program and would need to be added back to the value.
- *occupation* – occupation of user at time of account sign-up
- *movieId* – system-assigned identifier of a movie
- *yearReleased* – year the movie was released
- *genres* – assigned movie genre(s) (if any)
    - Note: Multiple genres for a movie are separated by a pipe (the | symbol)
- *imdbId* – identifier for movies used by imdb.com. e.g., http://www.imdb.com/title/tt0114709/
- *tmdbId* – identifier for movies used by themoviedb.org. e.g., https://www.themoviedb.org/movie/862
- *rating* – the user rating for a movie on a 5-star scale (0.5 stars - 5.0 stars)
- *tagId* – the system-assigned identifier for tags contained within the genome-scores.csv file
- *tag* – tag for a given movie
    - Note: The tags in the tagged.csv file are free-text/ user-generated, whereas the tags in the genome-scores.csv file have been generated by a machine learning algorithm. Hence, you should not assume that the tags in each respective CSV contain the same list of values. While there may be some overlap, there are certainly discrepancies.
- *timestamp* – date and time when a given movie was rated/ tagged by a user