CS 210 Homework 1

Movie recommendation system

Overview

In this assignment, you'll implement a prototype movie recommendation program in Python according to the following guidelines.

You can work individually or in a group of up to 4 people.

What to submit

Write all the required functions described below in the given template file named `hw1.py`, and submit your completed `hw1.py`. If you are working in a group, only one person should submit, and your `hw1.py` file should have a comment that gives the names and netIDs of everyone in your group.

You can resubmit any number of times. The last submission will be graded.

You may also implement other helper functions as needed. Make sure you write all your test calls in the `main()` function. Do not write any code outside of any of the functions. We'll ignore your `main()` function when grading.

How to test your code

You can test your program by calling your functions in the `hw1.py` file. All test code must be in the `main()` function.

Make sure the test files are in the same folder as the program. You may develop and test your code in a Jupyter notebook, but for submission you will need to move your code over to `hw1.py` and execute it as above to make sure it works correctly.

You can run your tests on the given ratings and movies files, but testing on only these files may not be sufficient. You should make your own test files as well, to make sure that you cover the various paths of logic in your functions. You are not required to submit any of your test files.

You may assume that all parameter values given to your functions will be valid. Also, in any function that requires the returned values to be sorted or ranked, ties may be broken arbitrarily between equal values.

Data Input

- Ratings file: A text file that contains movie ratings. Each line has the name (with year) of a movie, its rating (range 0-5 inclusive), and the id of the user who rated the movie. A movie can have multiple ratings from different users. A user can rate multiple movies, but can rate a particular movie only once.

- Movies file: A text file that contains the genres of movies. Each line has a genre, a movie id, and the name (with year) of the movie. To keep it simple, each movie belongs to a single genre.

  Note: A movie name includes the year to disambiguate movies with the same title.

There are sample movies and ratings files provided on Canvas.

You may assume that input files will be correctly formatted, and data types will be as expected.

For all rating computations, do not round up (or otherwise modify) the rating unless otherwise specified.

Task 1: Reading Data

(8 points) Movie ratings

Write a function `read_ratings_data(f)` that takes in a ratings file name, and returns a dictionary. The dictionary should have movie as key, and the corresponding list of ratings as value.

For example:

```
{
    "The Lion King (2019)": [6.0, 7.5, 5.1],
    "Titanic (1997)": [7]
}
```

(8 points) Movie genres

Write a function `read_movie_genre(f)` that takes in a movies file name and returns a dictionary mapping from movie to genre.

For example:

```
{
    "Toy Story (1995)": "Adventure",
    "Golden Eye (1995)": "Action"
}
```

Watch out for leading and trailing whitespaces in movie name and genre name, and remove them when encountered.

Task 2: Processing Data

(8 points) Genre dictionary

Write a function `create_genre_dict` that takes as a parameter a movie-to-genre dictionary, of the kind created in Task 1.2. The function should return another dictionary in which a genre is mapped to all the movies in that genre.

For example:

```
{
    genre1: [m1, m2, m3],
    genre2: [m6, m7]
}
```

(where `genre1`, `m1`, etc. are placeholders for actual genre and movie names)

[8 points] Average rating

Write a function `calculate_average_rating` that takes as a parameter a ratings dictionary, of the kind created in Task 1.1. It should return a dictionary where the movie is mapped to its average rating computed from the ratings list.

For example, an input of

```
{
    "The Lion King (2019)": [6.0, 7.5, 5.1],
    "Spider-Man (2002)": [3, 2, 4, 5]
}
```

should generate

```
{
    "The Lion King (2019)": 6.2,
    "Spider-Man (2002)": 3.5
}
```

Task 3: Recommendation

(8 points) Popularity based

In services such as Netflix and Spotify, you often see recommendations with the heading "Popular movies" or "Trending top 10".

Write a function `get_popular_movies` that takes as parameters a dictionary of movie-to-average rating (as created in Task 2.2), and an integer n (default should be 10). The function should return a dictionary (`movie:average rating`, same structure as input dictionary) of top n movies based on the average ratings. If there are fewer than n movies, it should return all movies in order of top average ratings.

(8 points) Threshold rating

Write a function `filter_movies` that takes as parameters a dictionary of movie-to-average rating (same as for the popularity based function above), and a threshold rating with default value of 3. The function should filter movies based on the threshold rating, and return a dictionary with same structure as the input. For example, if the threshold rating is 3.5, the returned dictionary should have only those movies from the input whose average rating is equal to or greater than 3.5.

(8 points) Popularity & genre based

In most recommendation systems, genre of the movie/song/book plays an important role. Often features like popularity, genre, artist are combined to present recommendations to a user.

Write a function `get_popular_in_genre` that, given a genre, a genre-to-movies dictionary (as created in Task 2.1), a dictionary of movie:average rating (as created in Task 2.2), and an integer n (default 5), returns the top n most popular movies in that genre based on the average ratings. The return value should be a dictionary of movie-to-average rating of movies that make the cut. If there are fewer than n movies, it should return all movies in order of top average ratings.

Genres will be from those in the `movie:genre` dictionary created in Task 1.2. The genre name will exactly match one of the genres in the dictionary, so you do not need to do any upper or lower case conversion.

(8 points) Genre rating

One important analysis for the content platforms is to determine ratings by genre.

Write a function `get_genre_rating` that takes the same parameters as `get_popular_in_genre` above, except for n, and returns the average rating of the movies in the given genre.

(8 points) Genre popularity

Write a function `genre_popularity` that takes as parameters a genre-to-movies dictionary (as created in Task 2.1), a movie-to-average rating dictionary (as created in Task 2.2), and n (default 5), and returns the top n rated genres as a dictionary of `genre:average rating`. If there are fewer than n genres, it should return all genres in order of top average ratings.

Hint: Use the above `get_genre_rating` function as a helper.

Task 4: User focused

(8 points) User ratings

Read the ratings file to return a user-to-movies dictionary that maps user ID to the associated movies and the corresponding ratings. Write a function named `read_user_ratings` for this, with the ratings file as the parameter.

For example:

```
{
    u1: [(m1, r1), (m2, r2)],
    u2: [(m3, r3), (m8, r8)]
}
```

where `ui` is user ID, `mi` is movie, `ri` is corresponding rating. You can handle user ID as int or string type, but make sure you use the same type consistently everywhere in your code.

(10 points) User genre

Write a function `get_user_genre` that takes as parameters a user id, the user-to-movies dictionary (as created in Task 4.1 above), and the movie-to-genre dictionary (as created in Task 1.2), and returns the top genre that the user likes based on the user's ratings. Here, the top genre for the user will be determined by taking the average rating of the movies genre-wise that the user has rated. If multiple genres have the same highest ratings for the user, return any one of genres (arbitrarily) as the top genre.

(10 points) User recommendations

Recommend 3 most popular (highest average rating) movies from the user's top genre that the user has not yet rated. Write a function `recommend_movies` for this, that takes as parameters a user id, the user-to-movies dictionary (as created in Task 4.1 above), the movie-to-genre dictionary (as created in Task 1.2), and the movie-to-average rating dictionary (as created in Task 2.2). The function should return a dictionary of movie-to-average rating. If fewer than 3 movies make the cut, then return all the movies that make the cut in order of top average ratings.