Machine Learning Challenge

Movie lens is a dataset developed by the University of Minnesota's Group Lens Research group. This dataset contains all the data related to thousands of movies and rankings given to those movies by thousands of people (https://movielens.org)

In this exercise we are focusing on the following datasets that are *slightly modified* version of original dataset:

movies.dat
This file contains the movie information such as title and genre(s) of the movie and,
it organized as follows:

```
MovieID::Title::Genres
```

The genres are list of genres for that movie that are separated by `|`

ratings.dat
This file represents rating of movies by users,
and has the following format:

```
UserID::MovieID::Rating::Timestamp
```

Ratings should be on a 5-star scale, with half-star increments. Timestamp is the unix timestamp a.k.a `Coordinated Universal Time (UTC)` (i.e. seconds since midnight of January 1st, 1970) that shows the date and time of rating.

Recommender Systems seeks to predict the "rating" or "preference" a user would give to an item. The goal of this assignment is to develop a recommender system that generates movie recommendations for a specific user.

Details of Approach

The code should be developed in **python** or **scala**. We encourage you to build your solution as scalable as possible (Spark, Dask, . . .). You should provide your source file(s) and/or jupyter notebooks that have covered all the steps and questions below.

First, read the data from the following files (movies.dat, ratings.dat) and *reorganize/clean* the data to be used in the model.
Now answer the following questions:

- (Q1) What are the titles of top 5 most popular movies i.e. have the most ranking in the whole dataset?
- (Q2) What are the top 5 ranked movie genres on average in the whole dataset?
- (Q3) How many movies have been ranked the most consecutive days?

Second, split the data into test and training sets and create a recommender system.
Now answer the following questions:

- (Q4) What are the top 5 recommended movies made to one user, e.g. , `UserID = 122` (any user can be selected)

- (Q5) What are the top 5 movies that are most frequently recommended by your model? (use training set)
- (Q6) Calculate the RMSE of your model for your test set.

What are we looking for?

We want to see how you handle:

- Machine Learning best practices
- Code quality
- Innovative technologies and frameworks
- Messy (i.e. real) data

Please make sure answers are clear and complete with as much commentary as possible and are mentioned in your code/notebook.