Movie Recommendation System using Spark

Information about the dataset:

This dataset (*ml-25m*) describes 5-star rating and free-text tagging activity from a movie recommendation service known as *MovieLens*. It contains 25000095 ratings across 62423 movies. These data were created by 162541 users between January 09, 1995 and November 21, 2019.

Users were selected at random for inclusion. All selected users had rated at least 20 movies. Furthermore, no demographic information is included. An ID represents each user, and no other information is provided.

The data are contained in the files *movies.csv, ratings.csv,* and *tags.csv*. The dataset files are written as CSV files with a single header row, and are encoded as UTF-8.

User IDs:

MovieLens users were selected at random for inclusion. User ids (anonymized) are consistent between ‘ratings.csv’ and ‘tags.csv’.

Movie IDs:

Only movies with at least one rating or tag are included in the dataset. Movie ids are consistent between ‘ratings.csv’, ‘tags.csv’, ‘movies.csv’, and ‘links.csv’.

Ratings Data File Structure (ratings.csv):

All ratings are contained in the file ‘ratings.csv’. Each line of this file after the header row represents one rating of one movie by one user, and has the following format:

userId, movieID, rating, timestamp

Ratings are made on a 5-star scale, with half-star increments (0.5 stars - 5.0 stars). Furthermore, timestamps represent seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.

Tags Data File Structure (tags.csv):

All tags are contained in the file ‘tags.csv’. Each line of this file after the header row represents one tag applied to one movie by one user, and has the following format:

userID, movieID, tag, timestamp

Tags are user-generated metadata about movies. Each tag is typically a single word or short phrase.

Movies Data File Structure (movies.csv):

Movie information is contained in the file ‘movies.csv’. Each line of this file after the header row represents one movie, and has the following format:

movieID, title, genres

Movie titles are entered manually or imported from https://www.themoviedb.org/, and include the year of release in parentheses.

Genres are a pipe-separated list, and are selected from the following:

- Action
- Adventure
- Animation
- Children's

- Comedy
- Crime
- Documentary
- Drama
- Fantasy
- Film-Noir
- Horror
- Musical
- Mystery
- Romance
- Sci-Fi
- Thriller
- War
- Western

The steps I have taken to preprocess this data:

1) Import *movies.csv, ratings.csv,* and *tags.csv* into Spark.
2) Create Spark dataframes for each of these CSV files. Let's call them *df_movies*, *df_ratings*, and *df_tags*.
3) Remove *timestamp* columns from *df_ratings* and *df_tags*, as they are not required for our project direction.