Assignment 3

ANALYZING THE MOVIELENS DATASET USING SPARK
VERSION 1.0

DUE DATE: Wednesday, November 10th, 2021 @ 5:00 pm

OBJECTIVE

The objective of this assignment is to gain experience in developing Spark programs. As part of this assignment, you will be working with the MovieLens datasets that describe ratings and free-text tagging activities from MovieLens, a movie recommendation service. This dataset was created by GroupLens and primarily hosted at Kaggle. You will be using Apache Spark (version 3.1.2) to implement this assignment.

This assignment may be modified to clarify any questions (and the version number incremented), but the crux of the assignment and the distribution of points will not change.

Grading: This assignment will account for **15 points** towards your cumulative course grade. There are a few components to this assignment, and the points-breakdown is listed in the remainder of the text. This is a team assignment to be done in a group of two. The scoring process will involve a one-to-one interview session of approximately 20 minutes where you will execute the programs in Spark and demonstrate the approaches for the questions based on the inputs that will be provided to you. The slots for these interview sessions will be posted a few days prior to the submission deadline.

1 Cluster setup [3 points]

As part of this assignment you are responsible for setting up your own HDFS and Spark cluster with Spark running on every node.

Refer these documents and videos:

Hadoop Setup – http://www.cs.colostate.edu/~cs555/HadoopInstallationGuide.pdf
(all links are provided in the above pdf)

Infospaces video - https://infospaces.cs.colostate.edu/watch.php?id=257

Spark Setup - http://www.cs.colostate.edu/~cs555/Apache-Spark.pdf
(all links are provided in the above pdf)

Infospaces videos -
Spark Setup - https://infospaces.cs.colostate.edu/watch.php?id=184
Running jobs using spark shell for debugging : https://infospaces.cs.colostate.edu/watch.php?id=186
Compiling and creating jar using SBT and submitting job on Spark standalone cluster: https://infospaces.cs.colostate.edu/watch.php?id=185

2 Analysis of MovieLens Dataset

You should develop Spark programs that leverage the DataFrame construct to process the main *and* supplementary datasets to answer the following questions. Questions Q1-Q5 account for **1 point each while Q6-Q7 account for 2 points each. Cumulatively, Q1-Q7 account for 9 points.**

Q1. How many movies were released for every year within the dataset?
The title column of movies.csv includes the year each movie was published. Some movies might not have the year, in such cases you can ignore those movies.

Q2. What is the average number of genres for movies within this dataset?

Q3. Rank the genres in the order of their ratings? Again, a movie may span multiple genres; such a movie should be counted in all the genres.

Q4. What are the top-3 combinations of genres that have the highest ratings?

Q5. How many movies have been tagged as "comedy"? Ignore the "case" information (i.e. both "Comedy" and "comedy" should be considered).

Q6. What are the different genres within this dataset? How many movies were released within different genres? A movie may span multiple genres; in such cases, that movie should be counted in all the genres?

Q7. Be creative and come up with your own data analytics question and answer for MovieLens dataset

[3 points]
You should include a PDF report that substantiates the results from your analysis. Please only include a PDF document (no Word or OpenOffice or Google Docs please).

2.1.1 Dataset Description

For this assignment you will be using the MovieLens 20M Dataset.

The datasets describe ratings and free-text tagging activities from MovieLens, a movie recommendation service. It contains 20,000,263 ratings and 465,564 tag applications across 27,278 movies. These data were created by 138,493 users between January 09, 1995 and March 31, 2015. The dataset was generated on October 17, 2016.

Users were selected at random for inclusion. All selected users had rated at least 20 movies. No demographic information is included. Each user is represented by an id, and no other information is provided.

The data are contained in six files.

- tags.csv that contains tags applied to movies by users:
  - userId
  - movieId
  - tag
  - timestamp

- ratings.csv that contains ratings of movies by users:
  - userId
  - movieId
  - rating
  - timestamp

- movies.csv that contains movie information:
  - movieId
  - title
  - genres

- links.csv that contains identifiers that can be used to link to other sources:
  - movieId
  - imdbId
  - tmbdId

- genome-scores.csv that contains movie-tag relevance data:
  - movieId
  - tagId
  - relevance

- genome-tags.csv that contains tag descriptions:
  - tagId
  - tag

References:

1. https://www.kaggle.com/grouplens/movielens-20m-dataset

2. F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages.

3 Provided Resources

You can download the dataset here (http://www.cs.colostate.edu/~cs555/ml-20m.zip)

Submission deadline:

Please submit a zip file containing the source codes and report (pdf) for your project by 5:00 pm on the due date using **Canvas**. We will rely on the honor system: please do not make any modifications to the codebase after the submission deadline has elapsed. There will be steep deductions for making modifications to the source code after you have submitted it.

> *Nota Bene*: Please do not e-mail the source codes to the Professor or the GTA – there will be a 2 **point** deduction for doing this.

4 Change History

This section will reflect any changes that were made to a particular version of the assignment. Generally, these changes are made to better clarify the spirit of the assignment.

Version	Date	Change
1.0	10/6/2021	First release of the assignment