**AYDIN ADNAN MENDERES UNIVERSITY**

**Engineering Faculty**

**Computer Engineering**

**CSE424 BIG DATA ANALYSIS**

**LAB 03**

**Recommendation System with Spark**

**Prepared by**

**Emre BİLAL - 151805029**

**Introduction**

* I selected movie lens small latest dataset(**ratings.csv)** from kaggle.
* <https://www.kaggle.com/shubhammehta21/movie-lens-small-latest-dataset>

**Summary**

This dataset (ml-latest-small) describes 5-star rating and free-text tagging activity from [MovieLens](http://movielens.org/), a movie recommendation service. It contains 100836 ratings and 3683 tag applications across 9742 movies. These data were created by 610 users between March 29, 1996 and September 24, 2018. This dataset was generated on September 26, 2018.

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 the files links.csv, movies.csv, ratings.csv and tags.csv.

**Content and Use of Files**

**User Ids:** MovieLens users were selected at random for inclusion. Their ids have been anonymized. User ids are consistent between `ratings.csv` and `tags.csv` (i.e., the same id refers to the same user across the two files).

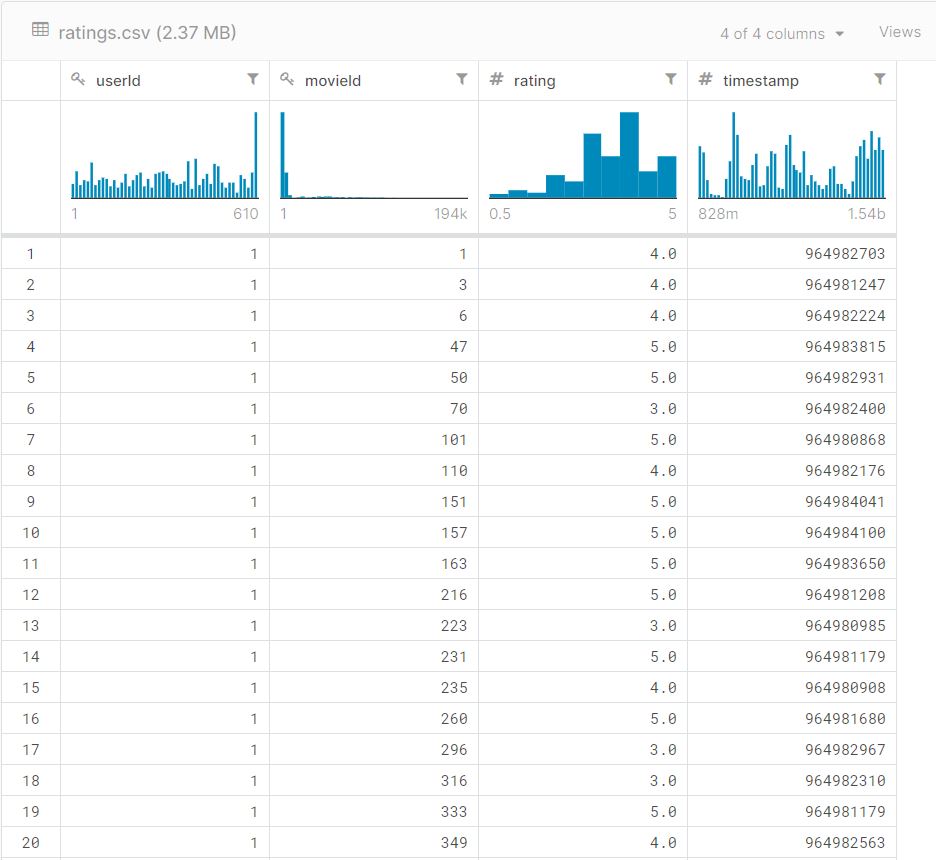
**Movie Ids:** Only movies with at least one rating or tag are included in the dataset. These movie ids are consistent with those used on the MovieLens web site (e.g., id `1` corresponds to the URL <https://movielens.org/movies/1>). Movie ids are consistent between `ratings.csv`, `tags.csv`, `movies.csv`, and `links.csv` (i.e., the same id refers to the same movie across these four data files).

**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:

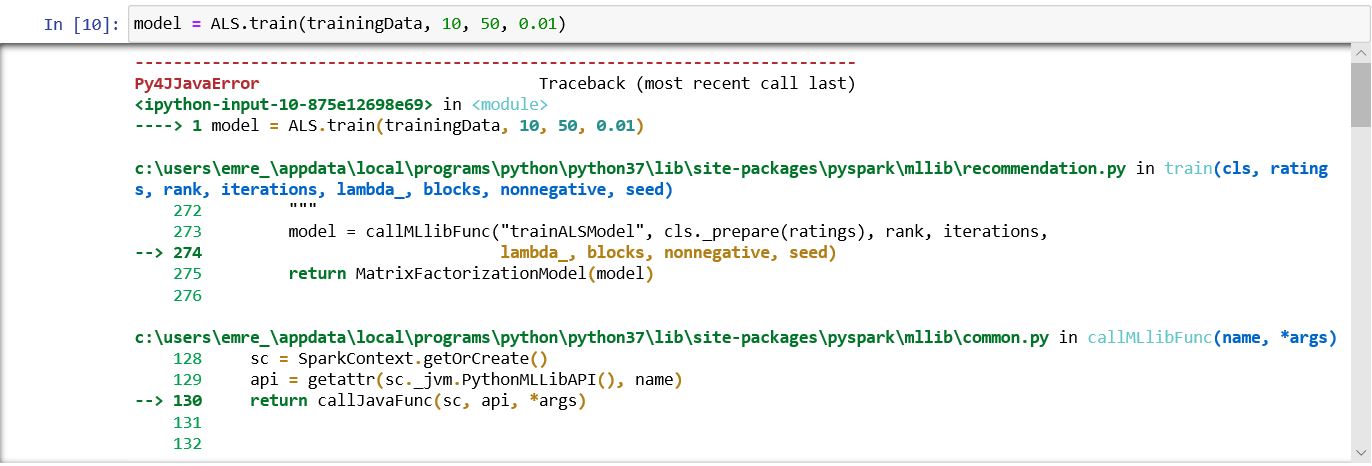
userId,movieId,rating,timestamp

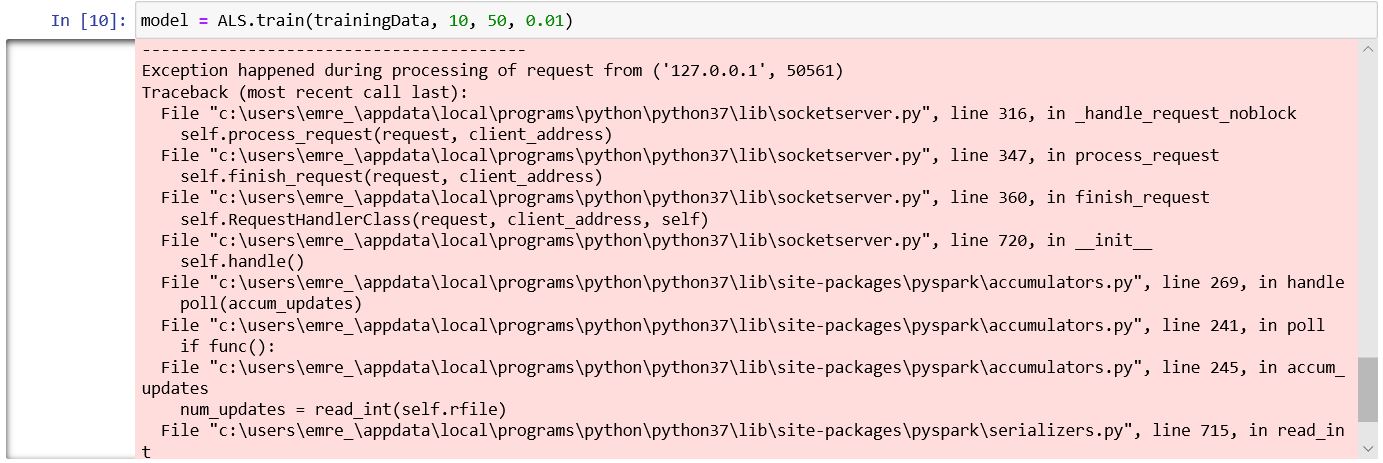
* The lines within this file are ordered first by userId, then, within user, by movieId.
* Ratings are made on a 5-star scale, with half-star increments (0.5 stars - 5.0 stars).
* Timestamps represent seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.



**Training Recommendation Model**

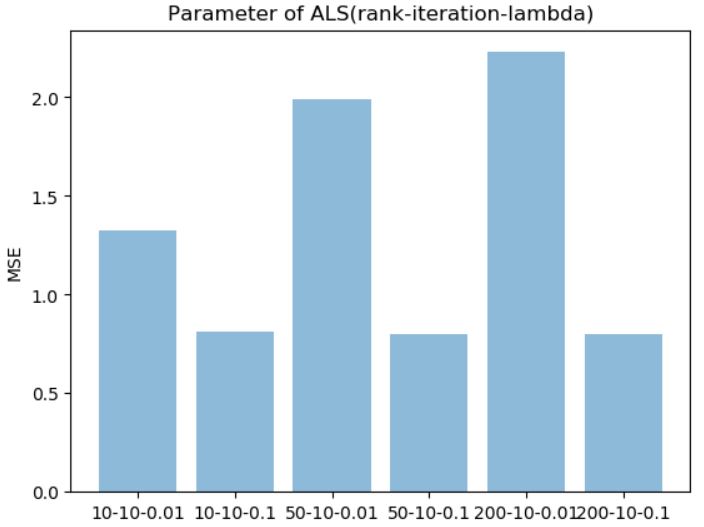
* I used ALS (Alternating Least Squares) for training recommendation model with last 4 digit of Emre’s student number as a “5029”.
* Also we changed the parameters of ALS re-run the algorithm for parameters “rank” (10, 50, 200), “iteration” (10, 50, 200) and “lambda” (0.01, 0.1).
* I tried all the combinations. The system failed 50 and 200 in iteration and did not work. So I continued with combinations other than them.





**Performance Evaluation**

* I tested the above combinations one by one to obtain Mean Squared Error values.



* According to these results, the values ​​of rank = 200, iteration = 10 and lambda = 0.1 are the smallest values. Accordingly, it is more performance than others.
* When the lambda value is 0.01 according to the table, the mse value increased when the rank value increased.
* When the lambda value is 0.1 according to the table, the mse value decreased when the rank value increased.