HW5

Deadline: May. 2nd 2:59 P.M. (before class)

Data

MovieLens is a dataset that is collected by the GroupLens Research Project at the University of Minnesota and made available rating data sets from the MovieLens web site. Download and unzip the MovieLens 100K Dataset (ml-100k.zip).

http://grouplens.org/datasets/movielens/

u.data is the dataset for this assignment:

The full u data set, 100000 ratings by 943 users on 1682 items. Each user has rated at least 20 movies. Users and items are numbered consecutively from 1. The data is randomly ordered. This is a tab separated list of

user_id<tab>item_id<tab>rating<tab>timestamp.

The time stamps are unix seconds since 1/1/1970 UTC

TODO

Perform the recommender system required in Option with Spark.
A tutorial of spark can be found at
https://spark.apache.org/docs/1.6.0/mllib-collaborative-filtering.html

Again, you need to output top 10 movies for all the users with the following format

userID<\tab>itemID1,itemID2,itemID3 ...,itemID10
…

Submission

1. Spark program
2. Output file with the required format.