MoviesRating-Spark
CPSC 531 Final Project

Prepare by: Yangling Cai, Kevin Huang

• Functionalities

1. Analyze movies' avg ratings and the total count of ratings by the users. Sorted by the total count of ratings.
2. Analyze selected user's rating and tags amount with movie title and time.
3. Analyze trends by user choice of genre.

• Architecture & Design

1. Main.java to take user input to choose which function to perform
2. MovieRankAnalyzer.java for functionalities 1. JavaRDD reads ratings.csv and movies.csv, then sets movieid as key with select info as value. Join both JavaPairRDD and reducebykey to group by movieid and calculate. Using TupleComparator.java to sort the RDD. A class in movies.jar. Save local files.
3. UserAnalyzer.java for functionalities 2. Take input as UserID to filter in the program. JavaRDD read ratings.csv and tags.csv, userid as key. Join movie.csv to get movies' titles. Sorted by Date/Time with recent action. A class in movies.jar. Save local files.
4. 4. GenreTrendAnalyzer.java for functionalities 2. Take input as genre to filter in the program. JavaRDD read movies.csv, movieId as key. Join the rating table for timestamps. Count for the genre of movies by years. A class in movies.jar. Save local files.
5. TupleComparator.java for tuple key sorting.
6. Movies.jar
7. Input files: movies.csv delimiter with ";", ratings.csv, tags.csv

• GitHub Location of Code

https://github.com/CC196/MoviesRating-Spark

• Deployment Instructions

- Require Apache Spark installed
- Clone or download from GitHub
- Unzip the files.zip, then copy all 3 csv files and paste back to the folder.
- Place csv files in the same folder with .jar file

• Steps to Run the Application

1. Open folder in terminal where .jar file located
2. Enter command spark-submit --class Main Movies.jar
3. Select analysis to run

• Test Results Sample

output/avg

(Forrest Gump (1994),(3518,4.0579877))
(Shawshank Redemption, The (1994),(3488,4.4360666))
(Pulp Fiction (1994),(3418,4.1824164))
(Silence of the Lambs, The (1991),(3227,4.1315465))
(Matrix, The (1999),(3064,4.1607375))
(Star Wars: Episode IV - A New Hope (1977),(2931,4.10116))
(Jurassic Park (1993),(2745,3.6874318))
(Schindler's List (1993),(2545,4.2760315))
(Fight Club (1999),(2530,4.2262845))
(Braveheart (1995),(2516,3.9974165))

output/user20
(Tue Aug 08 17:17:06 PDT 2006,(Star Wars: Episode V - The Empire Strikes Back (1980),5))
(Tue Aug 08 17:13:20 PDT 2006,(Alien (1979),5))
(Tue Aug 08 17:12:32 PDT 2006,(Sleepless in Seattle (1993),3))
(Tue Aug 08 17:12:24 PDT 2006,(Independence Day (a.k.a. ID4) (1996),5))
(Tue Aug 08 17:12:20 PDT 2006,(Die Hard (1988),5))
(Tue Aug 08 17:12:15 PDT 2006,(One Flew Over the Cuckoo's Nest (1975),5))
(Tue Aug 08 17:12:12 PDT 2006,(Aliens (1986),5))
(Tue Aug 08 17:12:08 PDT 2006,(Men in Black (a.k.a. MIB) (1997),4))
(Tue Aug 08 17:11:58 PDT 2006,(Ghostbusters (a.k.a. Ghost Busters) (1984),4.5))
(Tue Aug 08 17:11:22 PDT 2006,(Star Wars: Episode VI - Return of the Jedi (1983),bah))

output/genre_trend/action
(2007,278832)
(2008,317096)
(2009,237307)
(2010,238573)
(2011,200515)