Assignment #4

MapReduce Programming on MovieLens Data

Problem Description

An independent movie company is looking to invest in a new movie project. With limited finance, the company wants to analyze the reaction of audiences, particularly toward various movie genres, in order to identify beneficial movie project to focus on. The company relies on data collected from a publicly available recommendation service by MovieLens This dataset contains **24404096** ratings and **668953** tag applications across **40110** movies. These data were created by **247753** users between January 09, 1995 and January 29, 2016. This dataset was generated on October 17, 2016. From this dataset, several analyses are expected to derive information.

Data Set

The data set for this assignment is the data on /zfs/citi/movielens on Palmetto Cluster. You can copy the files to HDFS by using "hdfs dfs -put <localsrc> … <HDFS_dest_Path>" command and viewing them by "hdfs dfs -ls <HDFS_dest_Path>" command.

- Ratings Data File Structure

  All ratings are contained in the file ratings.csv. Each line of these files 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.

  Timestamps represent seconds since midnight Coordinated Universal Time (UTC) of January 09, 1995.

- Tags Data File Structure

  All tags are contained in the file tags.dat. Each line of this file represents one tag applied to one movie by one user, and has the following format:

  UserID::MovieID::Tag::Timestamp

  The lines within this file are ordered first by UserID, then, within user, by MovieID.

  Tags are user generated metadata about movies. Each tag is typically a single word, or short phrase. The meaning, value and purpose of a particular tag is determined by each user.

Timestamps represent seconds since midnight Coordinated Universal Time (UTC) of January 09, 1995.

- Movies Data File Structure

  Movie information is contained in the file movies.dat. Each line of this file represents one movie, and has the following format:

  MovieID::Title::Genres

  MovieID is the real MovieLens id.

  Movie titles, by policy, should be entered identically to those found in IMDB, including year of release. However, they are entered manually, so errors and inconsistencies may exist.

- 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

Programming Questions

Utilize MapReduce to find the answers to these questions.

1. Find the mean, median, and standard deviation of the ratings for each of the movie genres. For each statistic (mean, median, or standard deviation), only use a single MapReduce program (one pair of mapper and reducer).

2. Using a single MapReduce program (one pair of mapper and reducer), identify the user who provides the most rating. Which genre does this user watch the most?

Submission
Submit all electronic copies of the MapReduce programs that you implement AND also submit a printed electronic document (or screenshot) that provides the answers to the above questions.

Sample Test Results (Please be noted that your outputs may not necessarily be the same as the sample test results here. The sample here is only for format demonstration purpose instead of a real output from this assignment.):

- Create a 1000-lines test set from data file ratings.csv:
  head -1000 ratings.csv > sample.csv

  Output for sample mean calculation:
  Action 3.496598639455782
  Adventure 3.5662100456621006
  Animation 3.6923076923076925
  Children 3.324468085106383
  Comedy 3.5038560411311055
  Crime 3.6973684210526314
  Documentary 3.8333333333333335
  Drama 3.697674418604651
  Fantasy 3.699074074074074
  Film-Noir 4.2727272727272725
  Horror 3.3987341772151898
  Musical 3.688888888888889
  Mystery 3.775
  Romance 3.633720930232558
  Sci-Fi 3.5526315789473686
  Thriller 3.5848375451263537
  War 3.9903846153846154
  Western 3.5652173913043477

  Output for sample median calculation:
  Action 4.0
  Adventure 4.0
  Animation 4.0
  Children 3.5
  Comedy 3.5
  Crime 4.0
  Documentary 4.0
  Drama 4.0
  Fantasy 4.0
  Film-Noir 4.5
  Horror 3.5
  Musical 4.0

Mystery 4.0
Romance 4.0
Sci-Fi 4.0
Thriller 4.0
War 4.0
Western 4.0

Output for sample standard deviation calculation:
Action 1.0457178614293794
Adventure 1.0298954913019611
Animation 1.1895128533993933
Children 1.1547944617902528
Comedy 1.04677766775117
Crime 1.0701664509257338
Documentary 1.1055415967851334
Drama 0.9814532111235166
Fantasy 0.9928909688352056
Film-Noir 0.6863485850246136
Horror 1.1675456970306142
Musical 0.908532937633761
Mystery 0.9849915397267804
Romance 0.9099719475763407
Sci-Fi 1.0278328868124897
Thriller 0.9580672438010365
War 0.9119420560591924
Western 0.9005354424873034

Output for sample user identification:
12394 -- Total Rating Counts: 2 -- Most Rated Genre: Drama - 2
In the sample data set, user 12394 provided two ratings, and both movies contain the Drama genre.

- **Test Results on Full Data Set:**
  Mean
  Action 3.4213307400955033
  Adventure 3.4936211560747057
  Animation 3.5999880565273004
  Children 3.4184739602194236
  Comedy 3.436946311044954
  Crime 3.6656546597629625
  Documentary 3.7834593152512226
  Drama 3.6732628208935227
  Fantasy 3.502019481138221
  Film-Noir 4.012151194601495
  Horror 3.269243385726838
  Musical 3.56247843815335

Mystery 3.6776306613582186
Romance 3.553776441558182
Sci-Fi 3.396193464024223
Thriller 3.5071860967677653
War 3.7801731498090883
Western 3.555656921300586

Median
Action 3.5
Adventure 3.5
Animation 4.0
Children 3.5
Comedy 3.5
Crime 4.0
Documentary 4.0
Drama 4.0
Fantasy 3.5
Film-Noir 4.0
Horror 3.5
Musical 4.0
Mystery 4.0
Romance 4.0
Sci-Fi 3.5
Thriller 3.5
War 4.0
Western 4.0

Standard Deviation (using equation for sample, not population)
Action 1.0663429713964978
Adventure 1.052911290591791
Animation 1.0198106462057777
Children 1.0926580045999372
Comedy 1.0748705891497778
Crime 1.011868985282015
Documentary 1.0040662758800931
Drama 0.9954425330845901
Fantasy 1.065411123822357
Film-Noir 0.8864925130086234
Horror 1.150382535008721
Musical 1.0570693452090392
Mystery 0.9998890835526798
Romance 1.0304100915913417
Sci-Fi 1.092589221671842
Thriller 1.0310765443044845
War 1.01231127731633
Western 1.0237530124765744

User Identification
59269 -- Total Rating Counts: 7359 -- Most Rated Genre: Drama - 3657

Teamwork
You may form a team with up to 2 students (including yourself) in the class to work on this assignment together. Only one submission is needed from a team. Names of all team members have to be included in the submission.