

Movie Lens Data Analysis

Objective

The objective of the project is to analyze movies dataset and solve below mentioned KPIs. The dataset contains following files:

movies:

MovieID::Title::Genres

- Titles are identical to titles provided by the IMDB (including year of release)
- Genres are pipe-separated and are selected from the following genres:
 - Action
 - Adventure
 - Animation
 - Children's
 - Comedy
 - Crime
 - Documentary
 - Drama
 - Fantasy
 - Film-Noir
 - Horror
 - Musical
 - Mystery
 - Romance
 - Sci-Fi
 - Thriller
 - War
 - Western
- Some MovieIDs do not correspond to a movie due to accidental duplicate entries and/or test entries
- Movies are mostly entered by hand, so errors and inconsistencies may exist

ratings:

UserID::MovieID::Rating::Timestamp

- UserIDs range between 1 and 6040
- MovieIDs range between 1 and 3952
- Ratings are made on a 5-star scale (whole-star ratings only)
- Timestamp is represented in seconds since the epoch as returned by time(2)

users:

UserID::Gender::Age::Occupation::Zip-code

- Gender is denoted by a "M" for male and "F" for female
- Age is chosen from the following ranges:
 - 1: "Under 18"
 - 18: "18-24"
 - 25: "25-34"
 - 35: "35-44"
 - 45: "45-49"
 - 50: "50-55"
 - 56: "56+"
- Occupation is chosen from the following choices:
 - 0: "other" or not specified
 - 1: "academic/educator"
 - 2: "artist"
 - 3: "clerical/admin"
 - 4: "college/grad student"
 - 5: "customer service"
 - 6: "doctor/health care"
 - 7: "executive/managerial"
 - 8: "farmer"
 - 9: "homemaker"
 - 10: "K-12 student"
 - 11: "lawyer"
 - 12: "programmer"
 - 13: "retired"
 - 14: "sales/marketing"
 - 15: "scientist"
 - 16: "self-employed"
 - 17: "technician/engineer"
 - 18: "tradesman/craftsman"
 - 19: "unemployed"
 - 20: "writer"

Data Download Link

<https://drive.google.com/file/d/0B6rxRECSt4WdamFMVldOVjZMZ3M/view?usp=sharing>

Develop MapReduce programs to solve following KPIs:

1. Top ten most viewed movies with their movies Name (Ascending or Descending order)
2. Top twenty rated movies (Condition: The movie should be rated/viewed by at least 40 users)
3. We wish to know how have the genres ranked by Average Rating, for each profession and age group. The age groups to be considered are: 18-35, 36-50 and 50+.