Assignment #1

IMDb

Movies.dat

- Each line has movie id, movie title(year), genre information for a movie.
- Each information in a line is separated by two consecutive colons (i.e., ::).
- Example)
  - 23::Assassins (1995)::Thriller
    - → Movie ID : 23, Title(year) : Assassins (1995), Genre : Thriller
  - 48::Pocahontas (1995)::Animation|Children's|Musical|Romance
    - → Movie ID : 48, Title (year) : Pocahontas (1995), Genre : Animation | Children's | Musical | Romance

Ratings.dat

- Each line has user id, movie id, movie title(year), rating, and XXX.
- Each information in a line is separated by two consecutive colons (i.e., ::).
- Example)
  - 1::1193::5::978300760
    - → USER ID : 1, Movie ID : 1193, Rating : 5

Q1. Top-k movies

- We need find k fantasy movies whose avg. rating is high
- Input file : movies.dat & ratings.dat
- Output : <movie name> <avg. rating>
    - Example)

    Jumanji (1995) 4.8

    …

- Submission
  - No document, no image
  - MapReduce source code
    - The file name for source code MUST BE **IMDBStudent<Your ID>.java**.
    - Example) If your student id is 20181047, your file for source code MUST BE **IMDBStudent20181047.java** .
    - If you don't follow the rule, your submission may be failed.
- When the due date is passed, the prof. will compile your code, run your mapreduce program, and check whether your program runs correctly or not.

- Your code MUST process three command line parameters.
  - The first parameter : input file path (hdfs)
  - The second parameter : output file path (hdfs)
  - The third parameter : k
- Your code may be executed through the following command:

  hadoop jar hadoop-project.jar IMDBStudent20181047 **movieinput movieoutput 2 (top-2)**
- The movieinput directory has two files (i.e., movies.dat and ratings.dat).

Youtube

youtube.dat

- Each line has some information such as category and rating.
- Each information in a line is separated by a single bar (|).
- The fourth and last substring : category and its average rating
- Example)
    - 47EWHY3E5AM|MgsTheFury404|1024|Entertainment|123|111371|4.77
        - This video falls into “Entertainment” and its rating is 4.77

Q2. Top-K category

- We need find k categories whose average rating is high.
- Input file : youtube.dat
- Output : category avg. rating
  - Example)

  Fantasy 4.8999

  …

- Submission
    - No document, no image
    - MapReduce source code
        - The file name for source code MUST BE **YouTubeStudent<Your ID>.java**.
        - Example) If your student id is 20181047, your file for source code MUST BE **YouTubeStudent20181047.java** .
        - If you don't follow the rule, your submission may be failed.
- When the due date is passed, the prof. will compile your code, run your mapreduce program, and check whether your program runs correctly or not.

- Your code MUST process three command line parameters.
    - The first parameter : input file path (hdfs)
    - The second parameter : output file path (hdfs)
    - The third parameter : k
- Your code may be executed through the following command:

```
hadoop jar hadoop-project.jar YouTubeStudent20181047
youtubeinput youtubeoutput 2 (top-2)
```