Spark Movielens RDD Assignments

Skills Required: Apache Spark

Description:

This case study is used to analyze Movie lens Data set which has details about users, ratings, movies, genre using Apache Spark RDD API

Input Data Set:

Sample Data:

movies.csv

```
movieId,title,genres
1,Toy Story (1995),Adventure|Animation|Children|Comedy|Fantasy
2,Jumanji (1995),Adventure|Children|Fantasy
3,Grumpier Old Men (1995),Comedy|Romance
4,Waiting to Exhale (1995),Comedy|Drama|Romance
```

ratings.csv

```
userId,movieId,rating,timestamp
1,1,4.0,964982703
1,3,4.0,964981247
1,6,4.0,964982224
1,47,5.0,964983815
```

Tags.csv

```
userId,movieId,tag,timestamp
2,60756,funny,1445714994
2,60756,Highly quotable,1445714996
2,60756,will ferrell,1445714992
```

Source Data: http://files.grouplens.org/datasets/movielens/ml-latest-small.zip

Requirement

Use the movielens dataset as input and perform the following analysis:

a. Find the list of movies which have "**funny**" word in the tag and save it to a file called funnyMovies.txt
b. List top 5 movies based on maximum ratings received (where rating >= 3)
c. Get a list of Top 5 movies where genre = "Children" based on maximum ratings received (where rating >= 3)