Assignment 1 – mrJob script

Name: Tim Gras

Studentnumber: 630259

Github URL => https://github.com/dantim1997/PDP-Assignments-630259/tree/main/Assignment%201

Setup

to run the code, install a VM with Hadoop.
Make sure to have MRJob, Pip and the u.data file installed.

execute

Log in as su root
Run the python file with the following command:
python assignment-1-Tim-Gras.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar u.data
or can be run locally with:
python assignment-1-Tim-Gras.py u.data

```
  GNU nano 2.3.1                                          File: assignment-1-Tim-Gras.py

from mrjob.job import MRJob
from mrjob.step import MRStep

class Ratings(MRJob):
    def steps(self):
                #Does the MRSteps
        return [
            MRStep(
                                mapper=self.mapper_get_all_movies,
                                combiner=self.combiner_get_count_ratings_by_movies,
                                reducer=self.reducer_sum_up_rating_counts_from_movies
                        ),
            MRStep(
                                reducer=self.reducer_sort_all_movies_by_ratings
                                )
        ]

    #Split the text on \t and get the movie_id
    def mapper_get_all_movies(self, _, line):
        (_, movie_id, _, _) = line.split('\t')
        yield movie_id, 1

    #This will combine the ratings count with their corresponding movie ids
    def combiner_get_count_ratings_by_movies(self, movie_id, ratings):
        yield movie_id, sum(ratings)

    #This will sum up the the count of the ratings corresponding to the movie_id
    def reducer_sum_up_rating_counts_from_movies(self, movie_id, ratings):
        yield None, (sum(ratings), movie_id)

    #This will sort the movies based on the ratings the movie has
    def reducer_sort_all_movies_by_ratings(self, _, movies):
        for count, movie_id in sorted(movies):
            yield (int(movie_id), int(count))

if __name__ == "__main__":
    Ratings.run()
```

Information

For this code to get every rating by the movies id. I made 2 steps, one to give every rating by the movie_id and the other one to order them from high to low.

So first it will come to the main when running the python application it will go to the function steps. There I declared 2 steps the one to get the data and one to order the data.

So first the mapper function: in this function it will split every line based on the tab split. This will create a table of the u.data and giving the movie_id and the rating back.

Then it will go the the reducer. This will count the amount of rating based on the key that is the movie_id. Now I have a list of movie_ids with the amount of rating each movie has.

For the second stap the sorter function will order the movies based on the amount of ratings each movie has. And I did this descending to get the most first.

Github

assignment-1-Tim-Gras.py	This is the code to run the file
Assignment 1.pdf	This is the documentation
Result.jpg	This is the result image of the code

Result:

```
[maria_dev@sandbox-hdp ~]$ python assignment-1-Tim-Gras.py u.data
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/assignment-1-Tim-Gras.maria_dev.20210722.085547.670378
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/assignment-1-Tim-Gras.maria_dev.20210722.085547.670378/output
Streaming final output from /tmp/assignment-1-Tim-Gras.maria_dev.20210722.085547.670378/output...
1122    1
1130    1
1156    1
1201    1
1235    1
1236    1
1309    1
1310    1
1320    1
1325    1
1329    1
1339    1
1340    1
1341    1
1343    1
1348    1
1349    1
1352    1
1363    1
1364    1
1366    1
1373    1
1414    1
1447    1
1452    1
1453    1
1457    1
1458    1
1460    1
1461    1
1476    1
1482    1
1486    1
1492    1
1493    1
1494    1
1498    1
1505    1
1507    1
1510    1
```

```
357     264
12      267
742     267
275     268
111     272
89      275
191     276
28      276
202     280
234     280
64      283
176     284
216     290
183     291
118     293
15      293
25      293
328     295
96      295
22      297
302     297
276     298
318     298
9       299
423     300
195     301
257     303
269     315
168     316
748     316
69      321
173     324
151     326
210     331
79      336
405     344
204     350
313     350
222     365
172     367
117     378
237     384
98      390
7       392
56      394
127     413
174     420
121     429
300     431
1       452
288     478
286     481
294     485
181     507
100     508
258     509
50      583
Removing temp directory /tmp/assignment-1-Tim-Gras.maria_dev.20210722.085547.670378...
[maria_dev@sandbox-hdp ~]$
```