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 HD_rating.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar u.data
or can be run locally with:
python HD_rating.py u.data

```
  GNU nano 2.3.1                                              File: HD

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

class RatingCount(MRJob):
    def steps(self):
        return [
            MRStep(
                mapper=self.mapper,
                reducer=self.reducer
            ),
            MRStep(
                reducer=self.sorting
            )
        ]

    def mapper(self, _, line):
        (user_id, movie_id, rating, rating_time) = line.split('\t')
        yield int(movie_id), int(rating)

    def reducer(self, movie_id, ratings):
        yield None, (len(list(ratings)), movie_id)

    def sorting(self, _, countPairs):
        for count, movie_id in sorted(countPairs, reverse=True):
            yield movie_id, count

if __name__ == '__main__':
    RatingCount.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

HD_rating.py	This is the code to run the file
Assignment 1.docx	This is the documentation
Result.jpg	This is the result image of the code

Result:

```
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/HD_rating.maria_dev.20210611.125433.954698/output
Streaming final output from /tmp/HD_rating.maria_dev.20210611.125433.954698/output...
50      583
258     509
100     508
181     507
294     485
286     481
288     478
1       452
300     431
121     429
174     420
127     413
56      394
7       392
98      390
237     384
117     378
172     367
222     365
313     350
204     350
405     344
79      336
210     331
151     326
173     324
69      321
748     316
168     316
269     315
257     303
195     301
423     300
9       299
318     298
276     298
302     297
22      297
328     295
96      295
118     293
25      293
15      293
183     291
216     290
176     284
64      283
234     280
202     280
191     276
28      276
89      275
111     272
275     268
742     267
```