PDP assignment 1 – Chris Lips

Name: Christiaan Lips
Student number: 623434
Github: https://github.com/Lipsinator/Hadoop-fundamentals-Chris-Lips

Steps:

1. First of al a Hadoop vm needs to be started with VM WARE. Preferably a Hadoop 2.6.5 cluster. After this is all started up you can access this cluster with a tool like Putty or MobaXterm.
2. Secondly a series of commands needs to be run:
    - yum install python-pip
    - pip install google-api-python-client==1.6.4
    - pip install mrjob==0.5.11
    - yum install nano
3. A Third requirement for this assignment to run is to have the data for the assignment available on location. This can be achieved by running the following command on the cluster:
    - wget http://witan.nl/hadoop/u.data
4. Finally make sure the python file and the u.data file are in the directory in which the following command will be run:
    - Python assignment-1-Chris-Lips u.data

The result will be as follows if all went well:

```
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-Chris-Lips.maria_dev.20210722.090219.854711...
[root@sandbox-hdp ~]#
```

Code explanation:

1. First of all the base code for importing MRJOB is implemented and the class named Ratings is created
2. Secondly the mapper will map all the ids from the movies in the data file splited by tabs
3. After this the total count of ratings in the id of each movie is combined in combiner_count_ratings.
4. Then the total count of all the ratings is calculated.
5. And finally the movies are sorted by their rating to reduce amount of output and to create a better overview.