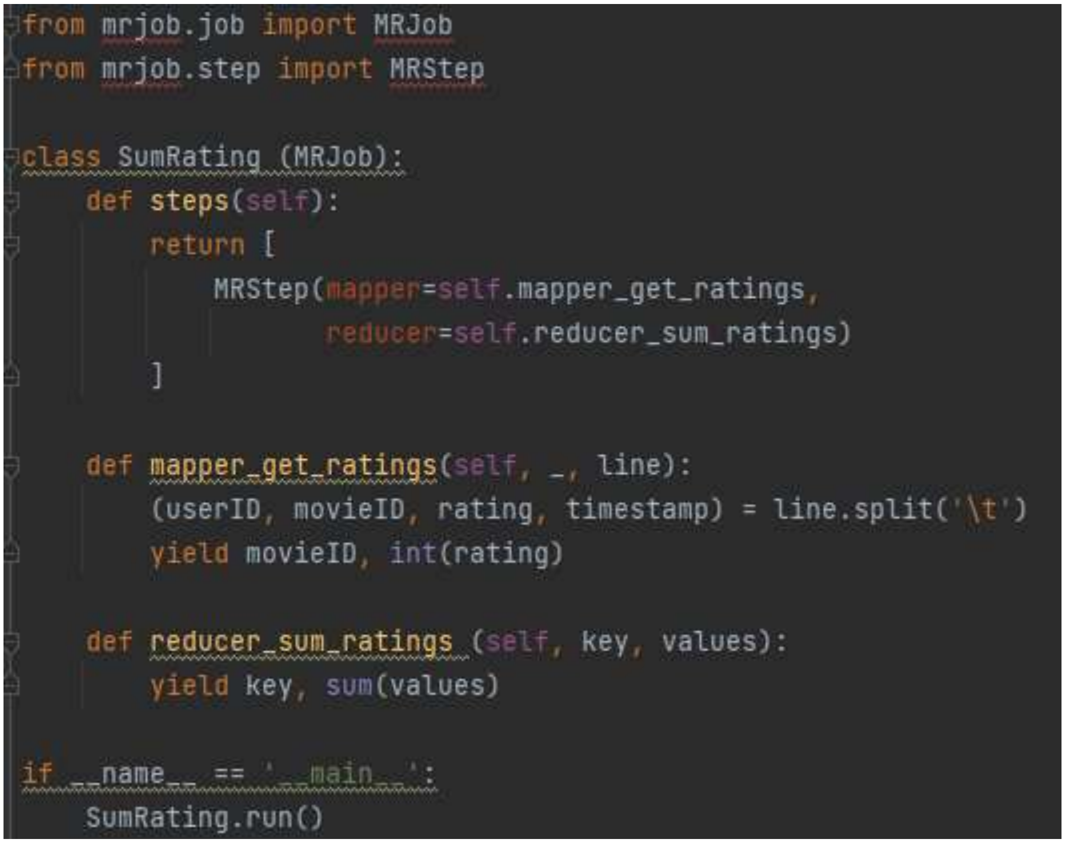
Abdullah Ghayumi

634072

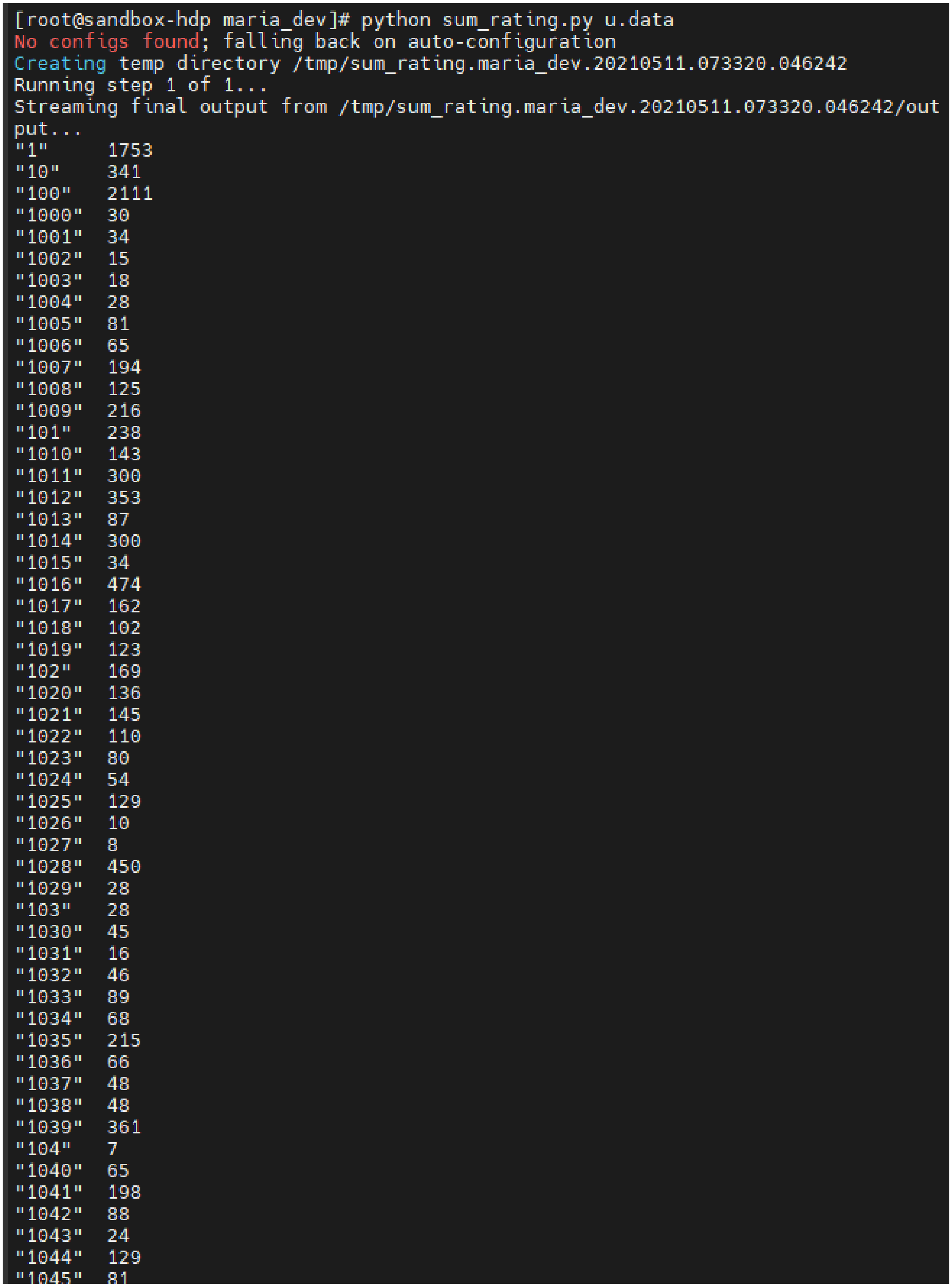
Github URL: https://github.com/Aras53/Hadoop

After struggling to get everything installed, I finally got it working. Afterwards I used the default script that was presented during the class to create the sum ratings script for each movie.



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

class SumRating (MRJob):
    def steps(self):
        return [
            MRStep(mapper=self.mapper_get_ratings,
                   reducer=self.reducer_sum_ratings)
        ]

    def mapper_get_ratings(self, _, line):
        (userID, movieID, rating, timestamp) = line.split('\t')
        yield movieID, int(rating)

    def reducer_sum_ratings (self, key, values):
        yield key, sum(values)

if __name__ == '__main__':
    SumRating.run()
```

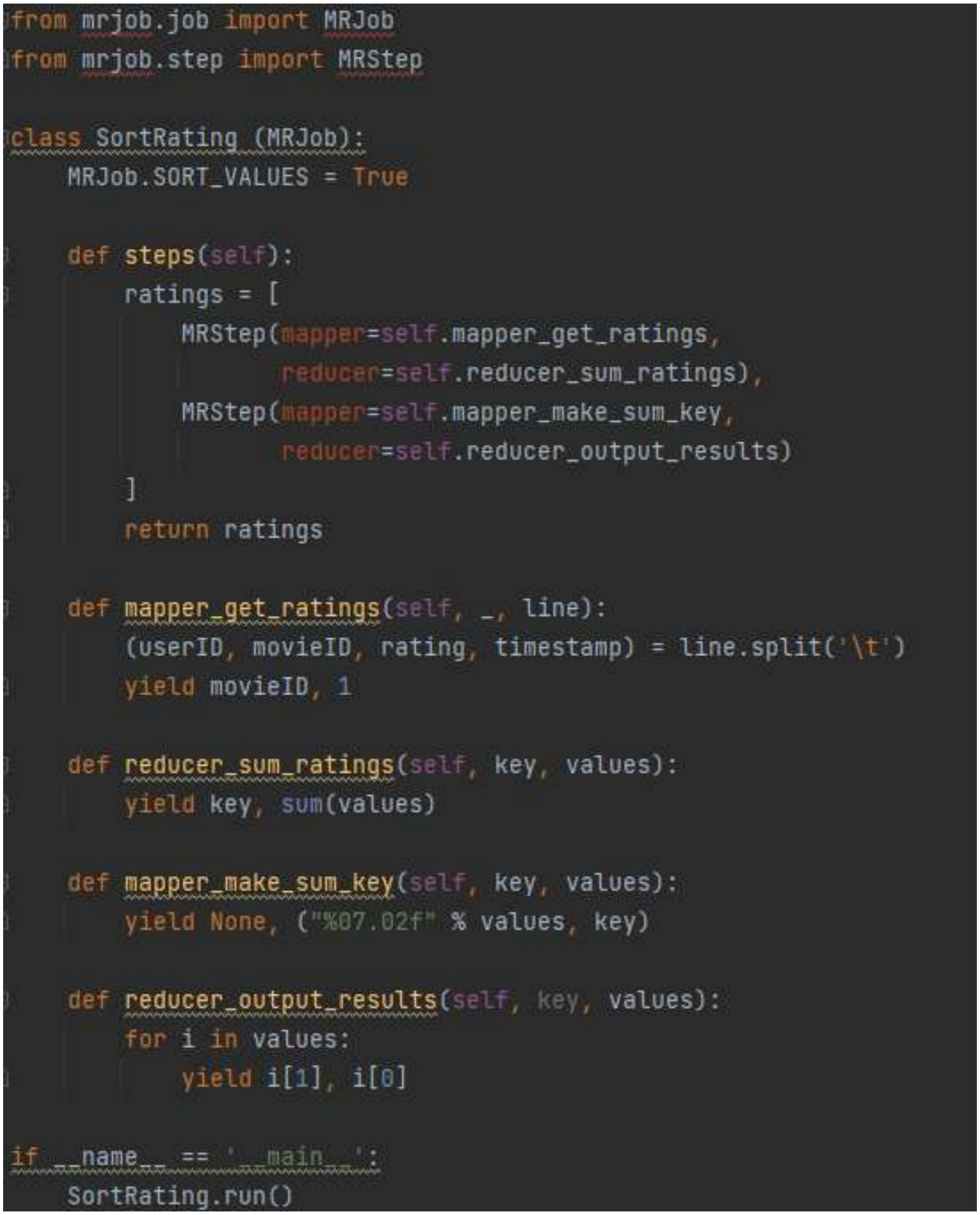
The mapper_get_ratings function will get the various columns and values from the data file. Then it will only yield the movieID and the rating that I had casted to int. I casted it to int to be able to use it in the reducer_sum_ratings function. Here I take the key, that is the movieID and sum all the ratings that the movie got. With that done, I used the following command `python sum_rating.py u.data` in the terminal to get the following results:



```
[root@sandbox-hdp maria_dev]# python sum_rating.py u.data
No configs found; falling back on auto-configuration
Creating temp directory /tmp/sum_rating.maria_dev.20210511.073320.046242
Running step 1 of 1...
Streaming final output from /tmp/sum_rating.maria_dev.20210511.073320.046242/out
put...
"1"     1753
"10"    341
"100"   2111
"1000"  30
"1001"  34
"1002"  15
"1003"  18
"1004"  28
"1005"  81
"1006"  65
"1007"  194
"1008"  125
"1009"  216
"101"   238
"1010"  143
"1011"  300
"1012"  353
"1013"  87
"1014"  300
"1015"  34
"1016"  474
"1017"  162
"1018"  102
"1019"  123
"102"   169
"1020"  136
"1021"  145
"1022"  110
"1023"  80
"1024"  54
"1025"  129
"1026"  10
"1027"  8
"1028"  450
"1029"  28
"103"   28
"1030"  45
"1031"  16
"1032"  46
"1033"  89
"1034"  68
"1035"  215
"1036"  66
"1037"  48
"1038"  48
"1039"  361
"104"   7
"1040"  65
"1041"  198
"1042"  88
"1043"  24
"1044"  129
"1045"  81
```

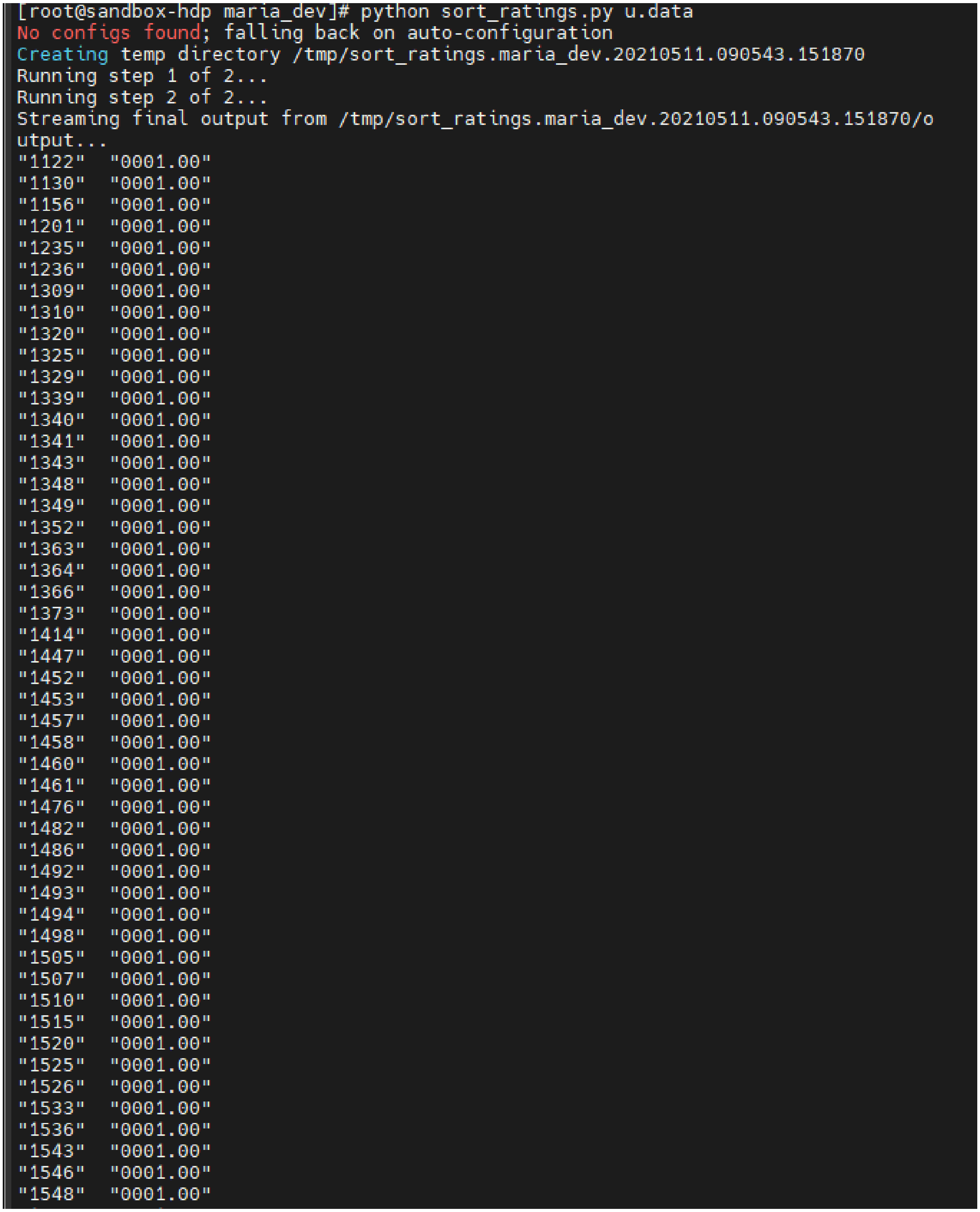
For the sorting assignment I used a chaining of MRSteps. First, I got a list with movieID and the total ratings they got. With that list, I created a new list on values that adds all the keys and values in the values list. Besides that, I made the total ratings a string so that the sorting would work, with ints it gave a result like this: 1, 10, 100 etc.

I tried doing this with sorted() and .sort(), but was not successful so this was my final solution.

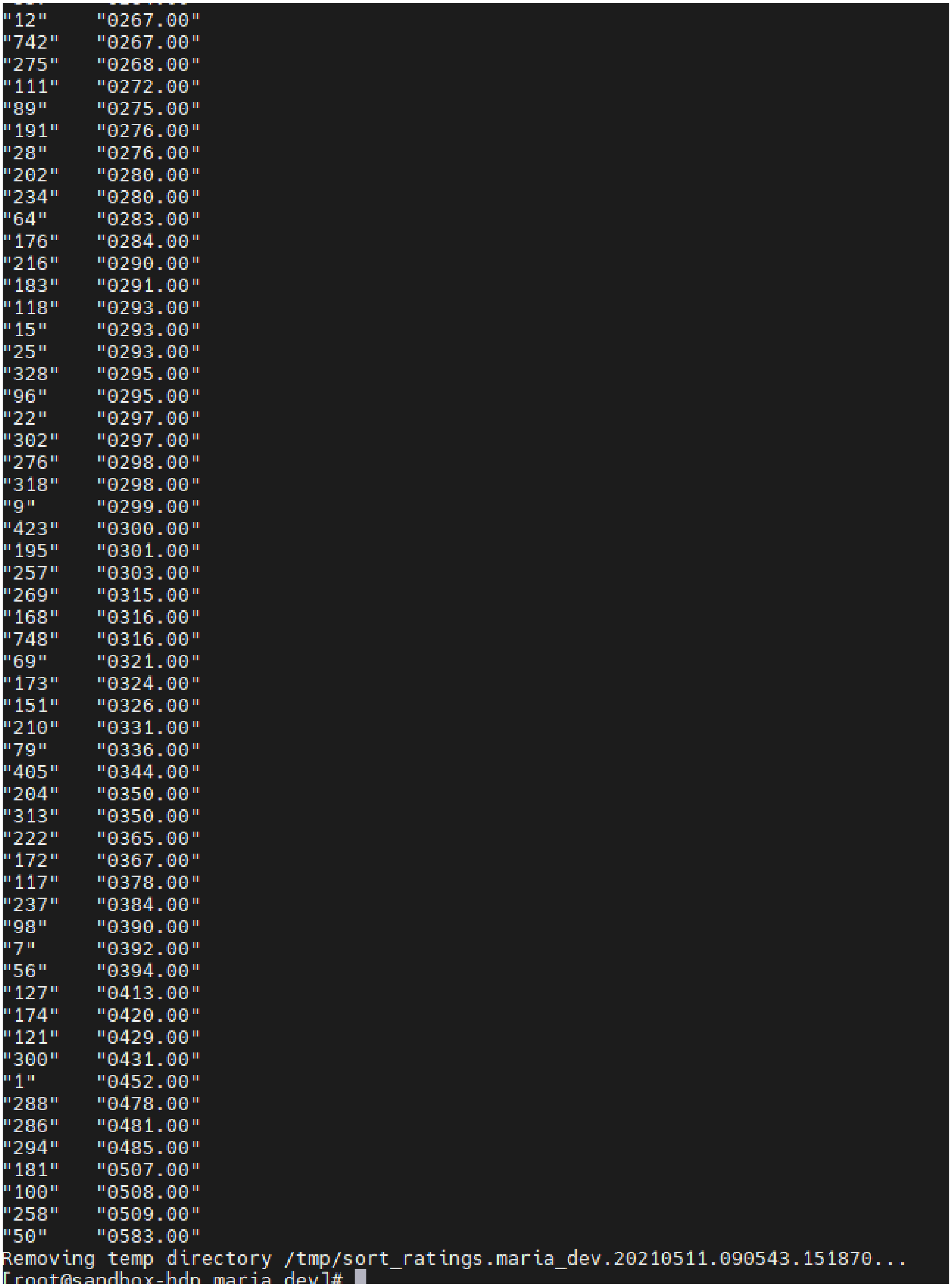


```
from mrjob.job import MRJob
from mrjob.step import MRStep

class SortRating (MRJob):
    MRJob.SORT_VALUES = True

    def steps(self):
        ratings = [
            MRStep(mapper=self.mapper_get_ratings,
                   reducer=self.reducer_sum_ratings),
            MRStep(mapper=self.mapper_make_sum_key,
                   reducer=self.reducer_output_results)
        ]
        return ratings

    def mapper_get_ratings(self, _, line):
        (userID, movieID, rating, timestamp) = line.split('\t')
        yield movieID, 1

    def reducer_sum_ratings(self, key, values):
        yield key, sum(values)

    def mapper_make_sum_key(self, key, values):
        yield None, ("%07.02f" % values, key)

    def reducer_output_results(self, key, values):
        for i in values:
            yield i[1], i[0]

if __name__ == '__main__':
    SortRating.run()
```

The result is a list with the movieID on the left and the number of ratings sorted on the right in string format.



```
[root@sandbox-hdp maria_dev]# python sort_ratings.py u.data
No configs found; falling back on auto-configuration
Creating temp directory /tmp/sort_ratings.maria_dev.20210511.090543.151870
Running step 1 of 2...
Running step 2 of 2...
Streaming final output from /tmp/sort_ratings.maria_dev.20210511.090543.151870/o
utput...
"1122"  "0001.00"
"1130"  "0001.00"
"1156"  "0001.00"
"1201"  "0001.00"
"1235"  "0001.00"
"1236"  "0001.00"
"1309"  "0001.00"
"1310"  "0001.00"
"1320"  "0001.00"
"1325"  "0001.00"
"1329"  "0001.00"
"1339"  "0001.00"
"1340"  "0001.00"
"1341"  "0001.00"
"1343"  "0001.00"
"1348"  "0001.00"
"1349"  "0001.00"
"1352"  "0001.00"
"1363"  "0001.00"
"1364"  "0001.00"
"1366"  "0001.00"
"1373"  "0001.00"
"1414"  "0001.00"
"1447"  "0001.00"
"1452"  "0001.00"
"1453"  "0001.00"
"1457"  "0001.00"
"1458"  "0001.00"
"1460"  "0001.00"
"1461"  "0001.00"
"1476"  "0001.00"
"1482"  "0001.00"
"1486"  "0001.00"
"1492"  "0001.00"
"1493"  "0001.00"
"1494"  "0001.00"
"1498"  "0001.00"
"1505"  "0001.00"
"1507"  "0001.00"
"1510"  "0001.00"
"1515"  "0001.00"
"1520"  "0001.00"
"1525"  "0001.00"
"1526"  "0001.00"
"1533"  "0001.00"
"1536"  "0001.00"
"1543"  "0001.00"
"1546"  "0001.00"
"1548"  "0001.00"
```



```
"12"    "0267.00"
"742"   "0267.00"
"275"   "0268.00"
"111"   "0272.00"
"89"    "0275.00"
"191"   "0276.00"
"28"    "0276.00"
"202"   "0280.00"
"234"   "0280.00"
"64"    "0283.00"
"176"   "0284.00"
"216"   "0290.00"
"183"   "0291.00"
"118"   "0293.00"
"15"    "0293.00"
"25"    "0293.00"
"328"   "0295.00"
"96"    "0295.00"
"22"    "0297.00"
"302"   "0297.00"
"276"   "0298.00"
"318"   "0298.00"
"9"     "0299.00"
"423"   "0300.00"
"195"   "0301.00"
"257"   "0303.00"
"269"   "0315.00"
"168"   "0316.00"
"748"   "0316.00"
"69"    "0321.00"
"173"   "0324.00"
"151"   "0326.00"
"210"   "0331.00"
"79"    "0336.00"
"405"   "0344.00"
"204"   "0350.00"
"313"   "0350.00"
"222"   "0365.00"
"172"   "0367.00"
"117"   "0378.00"
"237"   "0384.00"
"98"    "0390.00"
"7"     "0392.00"
"56"    "0394.00"
"127"   "0413.00"
"174"   "0420.00"
"121"   "0429.00"
"300"   "0431.00"
"1"     "0452.00"
"288"   "0478.00"
"286"   "0481.00"
"294"   "0485.00"
"181"   "0507.00"
"100"   "0508.00"
"258"   "0509.00"
"50"    "0583.00"
Removing temp directory /tmp/sort_ratings.maria_dev.20210511.090543.151870...
[root@sandbox-hdp maria_dev]#
```