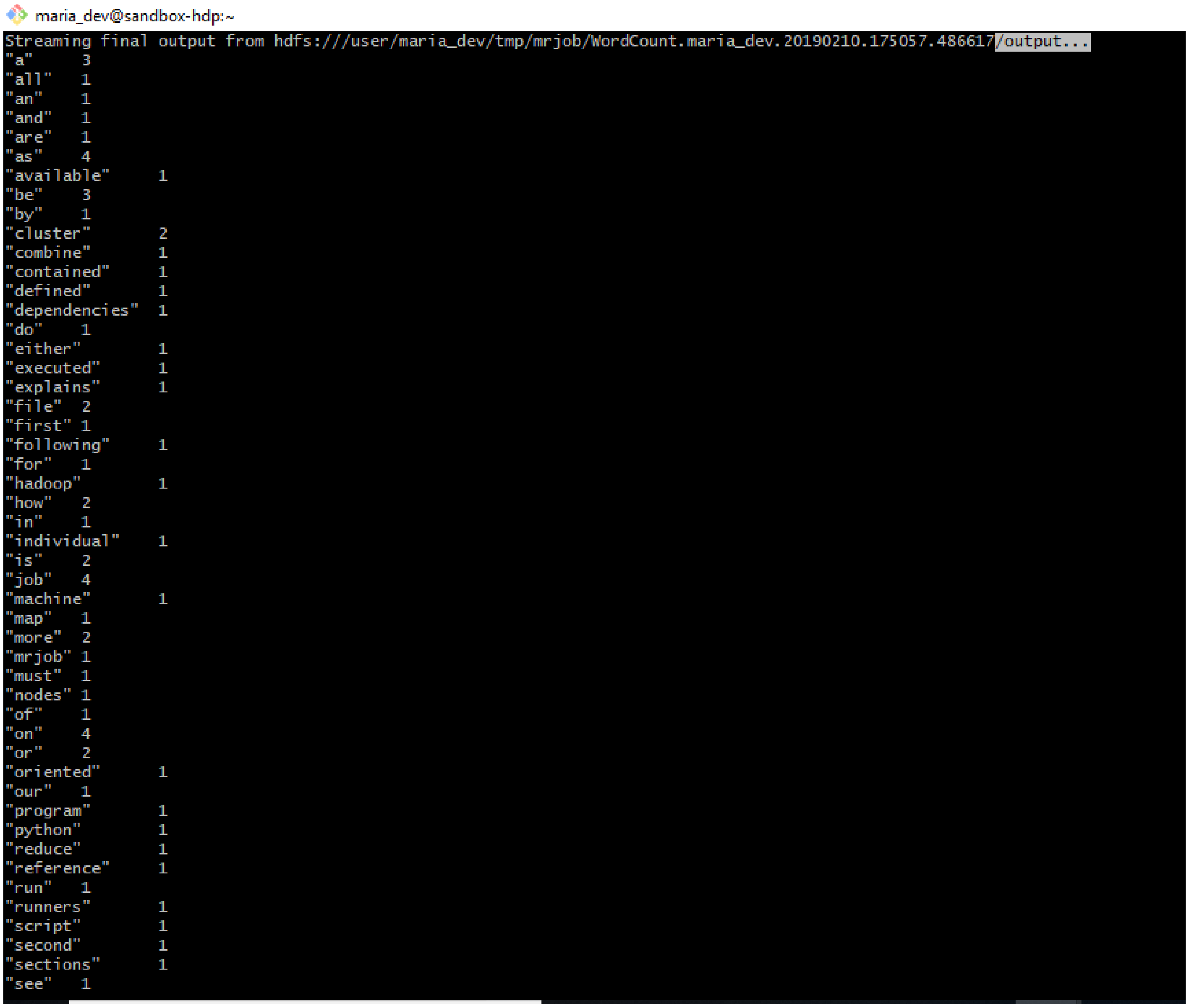
Q1. Submit a copy of this modified program and a screen shot of the results of the program's execution as the output of your assignment.

(a) Output for WordCount.py



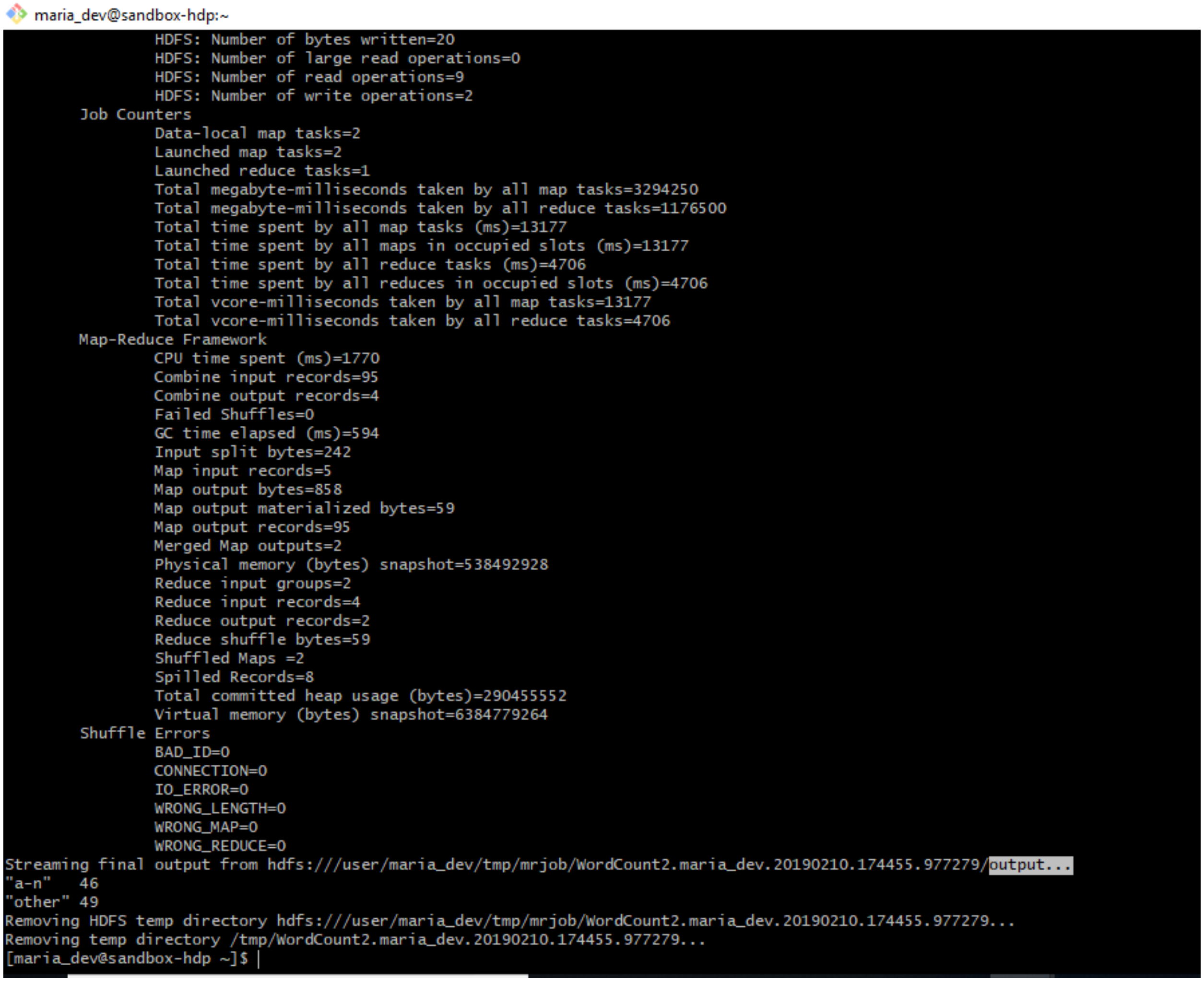
maria_dev@sandbox-hdp:~

```
Streaming final output from hdfs:///user/maria_dev/tmp/mrjob/WordCount.maria_dev.20190210.175057.486617/output...
"a"     3
"all"   1
"an"    1
"and"   1
"are"   1
"as"    4
"available"     1
"be"    3
"by"    1
"cluster"       2
"combine"       1
"contained"     1
"defined"       1
"dependencies"  1
"do"    1
"either"        1
"executed"      1
"explains"      1
"file"  2
"first" 1
"following"     1
"for"   1
"hadoop"        1
"how"   2
"in"    1
"individual"    1
"is"    2
"job"   4
"machine"       1
"map"   1
"more"  2
"mrjob" 1
"must"  1
"nodes" 1
"of"    1
"on"    4
"or"    2
"oriented"      1
"our"   1
"program"       1
"python"        1
"reduce"        1
"reference"     1
"run"   1
"runners"       1
"script"        1
"second"        1
"sections"      1
"see"   1
```

Updated Code: WordCount2.py

```
from mrjob.job import MRJob
import re
WORD_RE = re.compile(r"[\w']+")
class MRWordCount(MRJob):
    def mapper(self, _, line):
        for word in WORD_RE.findall(line):
            if(word[0]>="a" and word[0]<="n"):
                yield "a-n",1
            else:
                yield "other",1
    def combiner(self, word, counts):
        yield word, sum(counts)
    def reducer(self, word, counts):
        yield word, sum(counts)
if __name__ == '__main__':
    MRWordCount.run()
```

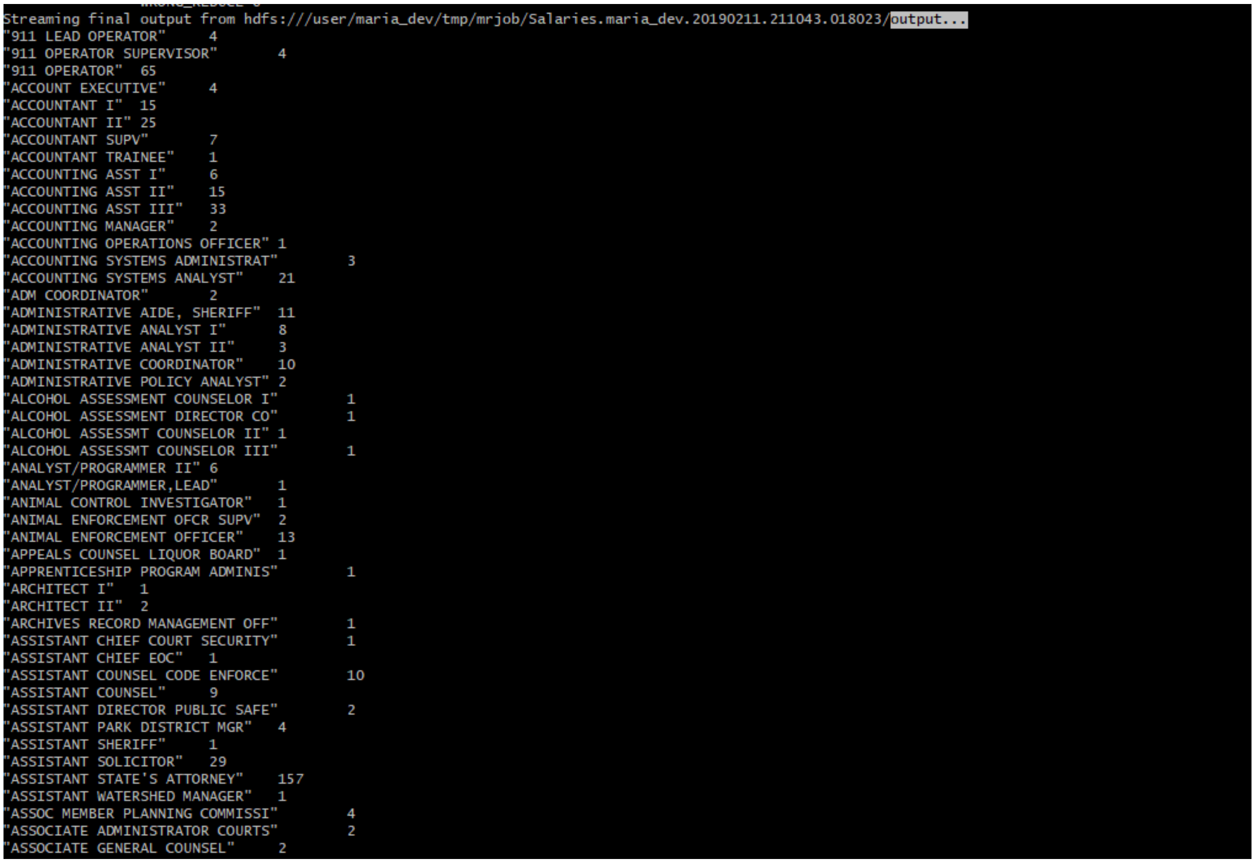
(b) WordCount2.py



```
                HDFS: Number of bytes written=20
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Data-local map tasks=2
                Launched map tasks=2
                Launched reduce tasks=1
                Total megabyte-milliseconds taken by all map tasks=3294250
                Total megabyte-milliseconds taken by all reduce tasks=1176500
                Total time spent by all map tasks (ms)=13177
                Total time spent by all maps in occupied slots (ms)=13177
                Total time spent by all reduce tasks (ms)=4706
                Total time spent by all reduces in occupied slots (ms)=4706
                Total vcore-milliseconds taken by all map tasks=13177
                Total vcore-milliseconds taken by all reduce tasks=4706
        Map-Reduce Framework
                CPU time spent (ms)=1770
                Combine input records=95
                Combine output records=4
                Failed Shuffles=0
                GC time elapsed (ms)=594
                Input split bytes=242
                Map input records=5
                Map output bytes=858
                Map output materialized bytes=59
                Map output records=95
                Merged Map outputs=2
                Physical memory (bytes) snapshot=538492928
                Reduce input groups=2
                Reduce input records=4
                Reduce output records=2
                Reduce shuffle bytes=59
                Shuffled Maps =2
                Spilled Records=8
                Total committed heap usage (bytes)=290455552
                Virtual memory (bytes) snapshot=6384779264
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/maria_dev/tmp/mrjob/WordCount2.maria_dev.20190210.174455.977279/output...
"a-n"   46
"other" 49
Removing HDFS temp directory hdfs:///user/maria_dev/tmp/mrjob/WordCount2.maria_dev.20190210.174455.977279...
Removing temp directory /tmp/WordCount2.maria_dev.20190210.174455.977279...
[maria_dev@sandbox-hdp ~]$ |
```

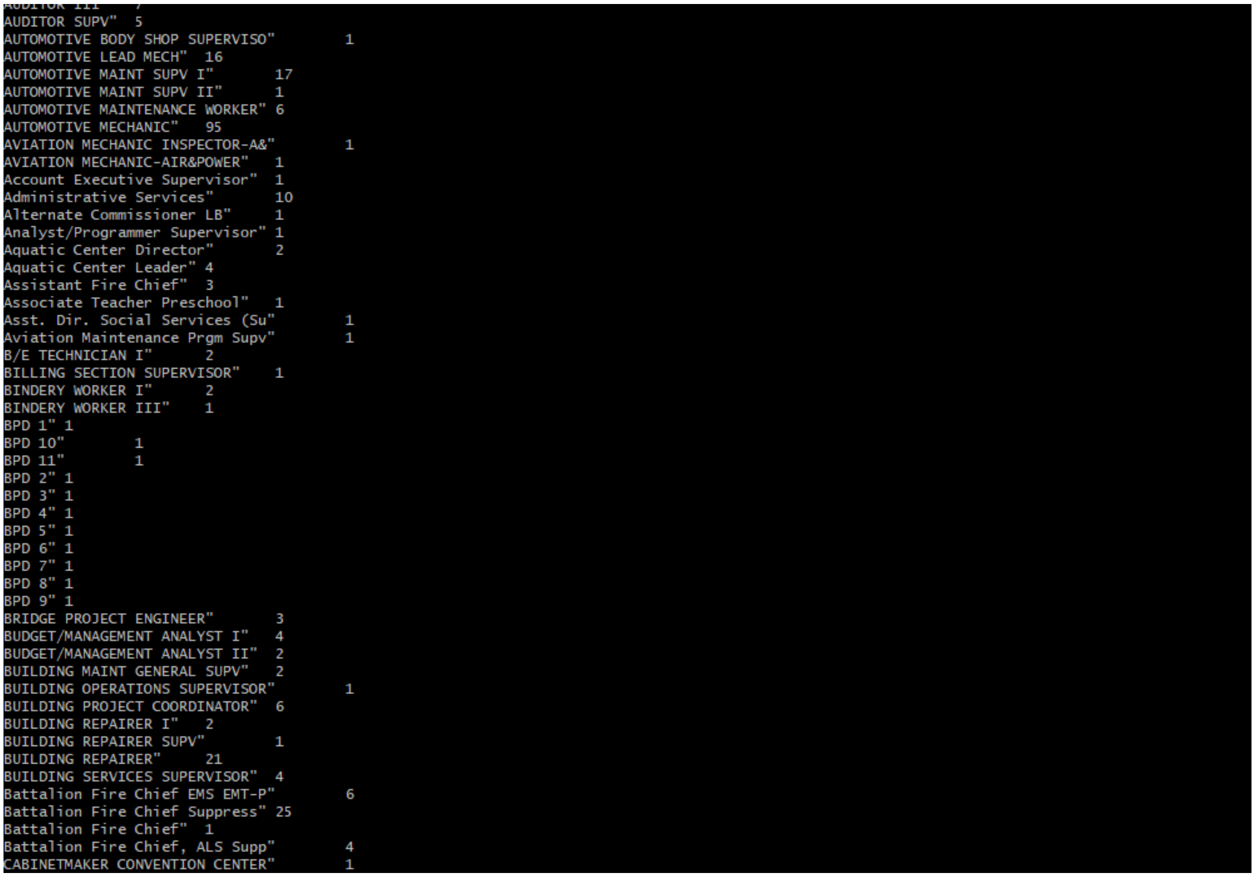
Q2. Submit a copy of this modified program and a screen shot of the results of the program's execution as the output of your assignment.

(a) Salaries.py



```
Streaming final output from hdfs:///user/maria_dev/tmp/mrjob/Salaries.maria_dev.20190211.211043.018023/output...
"911 LEAD OPERATOR"     4
"911 OPERATOR SUPERVISOR"       4
"911 OPERATOR"  65
"ACCOUNT EXECUTIVE"     4
"ACCOUNTANT I"  15
"ACCOUNTANT II" 25
"ACCOUNTANT SUPV"       7
"ACCOUNTANT TRAINEE"    1
"ACCOUNTING ASST I"     6
"ACCOUNTING ASST II"    15
"ACCOUNTING ASST III"   33
"ACCOUNTING MANAGER"    2
"ACCOUNTING OPERATIONS OFFICER" 1
"ACCOUNTING SYSTEMS ADMINISTRAT"        3
"ACCOUNTING SYSTEMS ANALYST"    21
"ADM COORDINATOR"       2
"ADMINISTRATIVE AIDE, SHERIFF"  11
"ADMINISTRATIVE ANALYST I"      8
"ADMINISTRATIVE ANALYST II"     3
"ADMINISTRATIVE COORDINATOR"    10
"ADMINISTRATIVE POLICY ANALYST" 2
"ALCOHOL ASSESSMENT COUNSELOR I"        1
"ALCOHOL ASSESSMENT DIRECTOR CO"        1
"ALCOHOL ASSESSMT COUNSELOR II" 1
"ALCOHOL ASSESSMT COUNSELOR III"        1
"ANALYST/PROGRAMMER II" 6
"ANALYST/PROGRAMMER,LEAD"       1
"ANIMAL CONTROL INVESTIGATOR"   1
"ANIMAL ENFORCEMENT OFCR SUPV"  2
"ANIMAL ENFORCEMENT OFFICER"    13
"APPEALS COUNSEL LIQUOR BOARD"  1
"APPRENTICESHIP PROGRAM ADMINIS"        1
"ARCHITECT I"   1
"ARCHITECT II"  2
"ARCHIVES RECORD MANAGEMENT OFF"        1
"ASSISTANT CHIEF COURT SECURITY"        1
"ASSISTANT CHIEF EOC"   1
"ASSISTANT COUNSEL CODE ENFORCE"        10
"ASSISTANT COUNSEL"     9
"ASSISTANT DIRECTOR PUBLIC SAFE"        2
"ASSISTANT PARK DISTRICT MGR"   4
"ASSISTANT SHERIFF"     1
"ASSISTANT SOLICITOR"   29
"ASSISTANT STATE'S ATTORNEY"    157
"ASSISTANT WATERSHED MANAGER"   1
"ASSOC MEMBER PLANNING COMMISSI"        4
"ASSOCIATE ADMINISTRATOR COURTS"        2
"ASSOCIATE GENERAL COUNSEL"     2
```

(a) Salaries.py (Rest output)

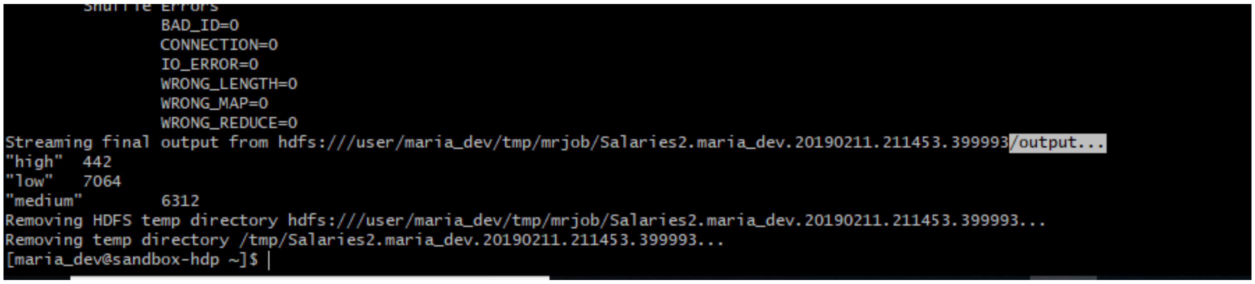


```
AUDITOR III"   7
AUDITOR SUPV"  5
AUTOMOTIVE BODY SHOP SUPERVISO"       1
AUTOMOTIVE LEAD MECH"  16
AUTOMOTIVE MAINT SUPV I"       17
AUTOMOTIVE MAINT SUPV II"      1
AUTOMOTIVE MAINTENANCE WORKER" 6
AUTOMOTIVE MECHANIC"   95
AVIATION MECHANIC INSPECTOR-A&"       1
AVIATION MECHANIC-AIR&POWER"   1
Account Executive Supervisor"  1
Administrative Services"       10
Alternate Commissioner LB"     1
Analyst/Programmer Supervisor" 1
Aquatic Center Director"       2
Aquatic Center Leader" 4
Assistant Fire Chief"  3
Associate Teacher Preschool"   1
Asst. Dir. Social Services (Su"       1
Aviation Maintenance Prgm Supv"       1
B/E TECHNICIAN I"      2
BILLING SECTION SUPERVISOR"    1
BINDERY WORKER I"      2
BINDERY WORKER III"    1
BPD 1" 1
BPD 10"        1
BPD 11"        1
BPD 2" 1
BPD 3" 1
BPD 4" 1
BPD 5" 1
BPD 6" 1
BPD 7" 1
BPD 8" 1
BPD 9" 1
BRIDGE PROJECT ENGINEER"       3
BUDGET/MANAGEMENT ANALYST I"   4
BUDGET/MANAGEMENT ANALYST II"  2
BUILDING MAINT GENERAL SUPV"   2
BUILDING OPERATIONS SUPERVISOR"       1
BUILDING PROJECT COORDINATOR"  6
BUILDING REPAIRER I"   2
BUILDING REPAIRER SUPV"        1
BUILDING REPAIRER"     21
BUILDING SERVICES SUPERVISOR"  4
Battalion Fire Chief EMS EMT-P"       6
Battalion Fire Chief Suppress" 25
Battalion Fire Chief"  1
Battalion Fire Chief, ALS Supp"       4
CABINETMAKER CONVENTION CENTER"       1
```

Updated Code: Salaries2.py

```
from mrjob.job import MRJob

class MRSalaries(MRJob):

    def mapper(self, _, line):

        (name,jobTitle,agencyID,agency,hireDate,annualSalary,grossPay) = line.split('\t')

        if (float(annualSalary) >= 0.00   and float(annualSalary) <= 49999.99):

            yield "low", 1

        elif (float(annualSalary) >= 50000.00 and float(annualSalary) <= 99999.99):

            yield "medium", 1

        elif (float(annualSalary) >= 100000.00):

            yield "high", 1

    def combiner(self, salary, counts):

        yield salary, sum(counts)

    def reducer(self, salary, counts):

        yield salary, sum(counts)

if __name__ == '__main__':

    MRSalaries.run()
```

(b) Output for Salaries2.py



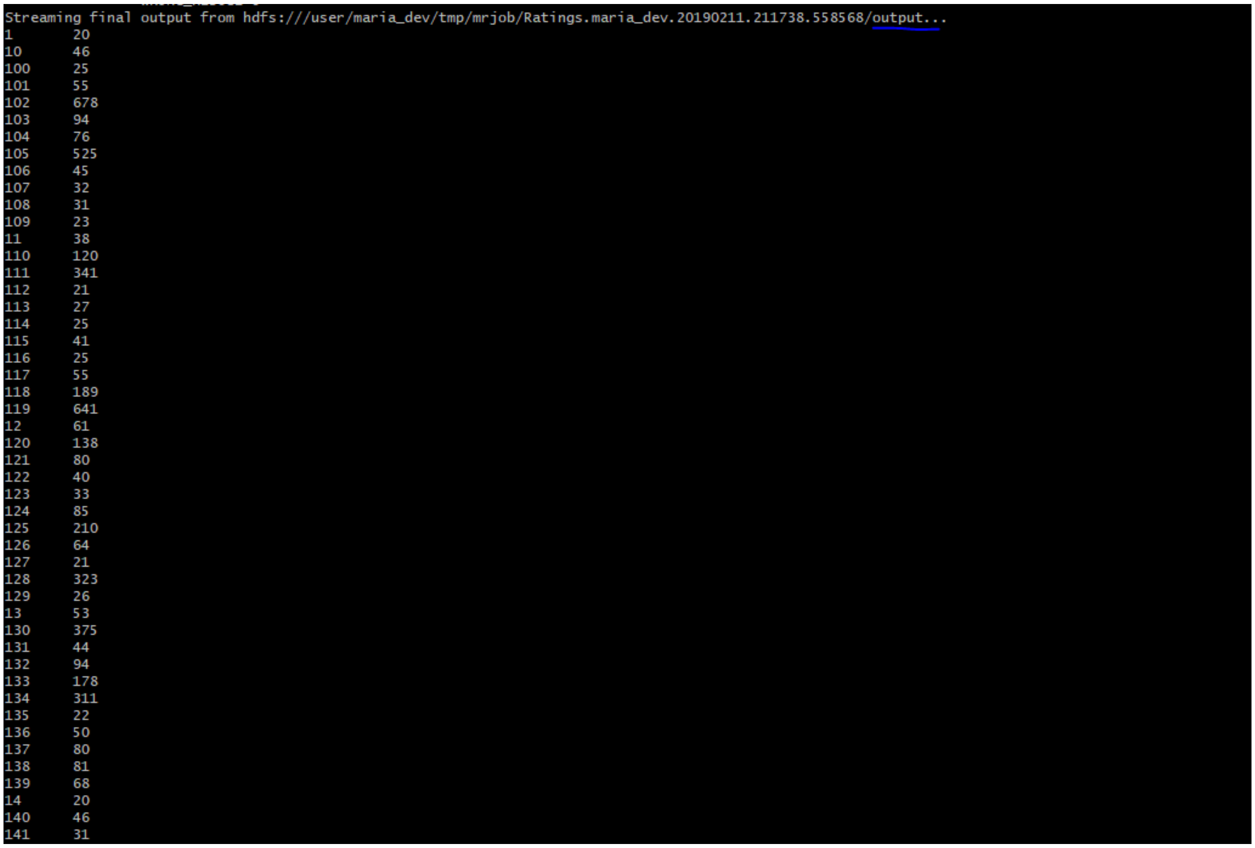
```
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/maria_dev/tmp/mrjob/Salaries2.maria_dev.20190211.211453.399993/output...
"high"  442
"low"   7064
"medium"        6312
Removing HDFS temp directory hdfs:///user/maria_dev/tmp/mrjob/Salaries2.maria_dev.20190211.211453.399993...
Removing temp directory /tmp/Salaries2.maria_dev.20190211.211453.399993...
[maria_dev@sandbox-hdp ~]$ |
```

Q3. Review the slides 17-22 in lecture notes Module 3b. Now write a program to perform the task of outputting a count of the number of movies each user (identified via their user id) reviewed.

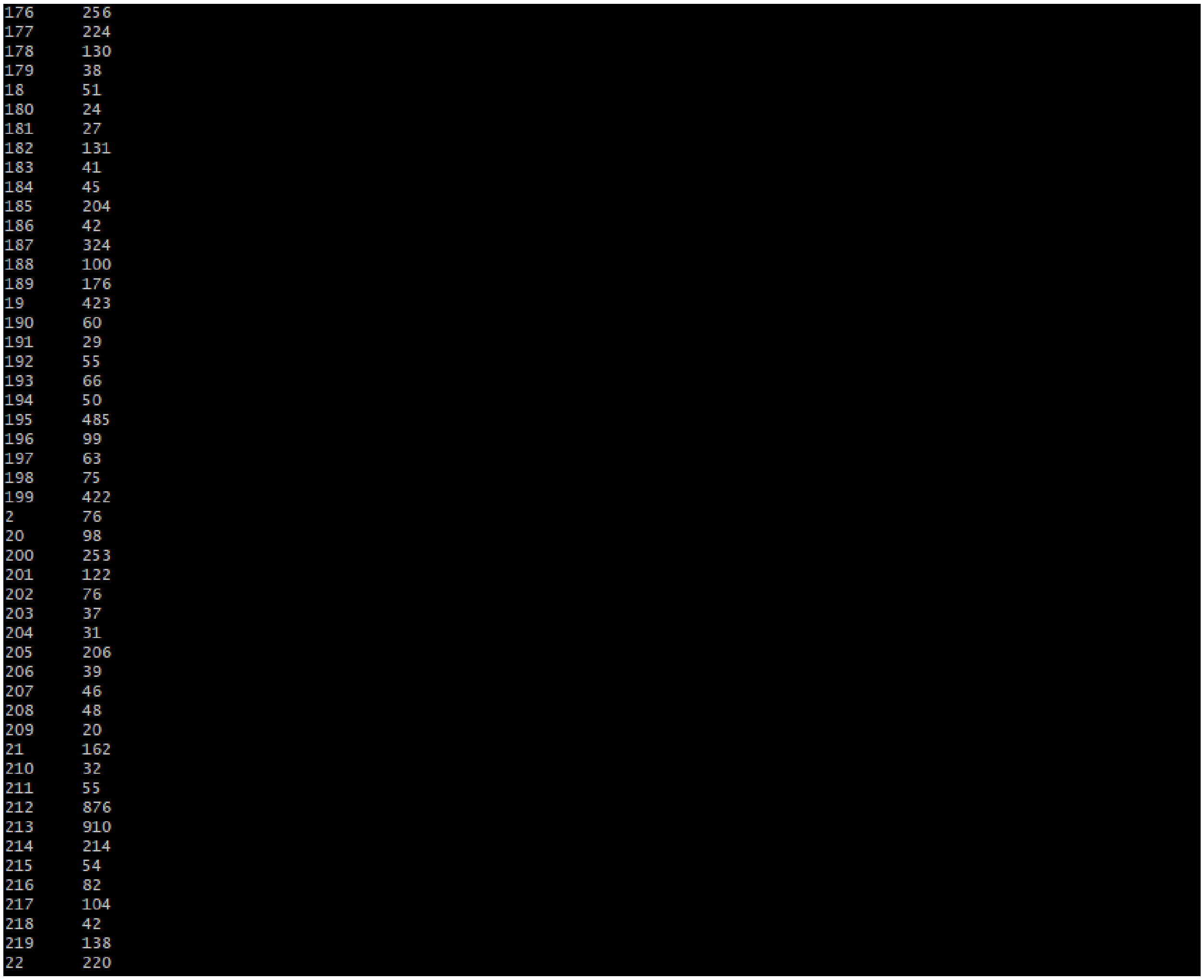
Code for Ratings

```
from mrjob.job import MRJob

from mrjob.step import MRStep

class Ratings(MRJob):

    def mapper(self, _, line):

        (userID, movieID, rating, timestamp) = line.split(',')

        yield int(userID), 1

    def combiner(self, userID, counts):

        yield int(userID), sum(counts)

    def reducer(self, userID, counts):

        yield int(userID), sum(counts)

if __name__ == '__main__':

    Ratings.run()
```

(a) Output for Ratings.py



```
Streaming final output from hdfs:///user/maria_dev/tmp/mrjob/Ratings.maria_dev.20190211.211738.558568/output...
1       20
10      46
100     25
101     55
102     678
103     94
104     76
105     525
106     45
107     32
108     31
109     23
11      38
110     120
111     341
112     21
113     27
114     25
115     41
116     25
117     55
118     189
119     641
12      61
120     138
121     80
122     40
123     33
124     85
125     210
126     64
127     21
128     323
129     26
13      53
130     375
131     44
132     94
133     178
134     311
135     22
136     50
137     80
138     81
139     68
14      20
140     46
141     31
```



```
176     256
177     224
178     130
179     38
18      51
180     24
181     27
182     131
183     41
184     45
185     204
186     42
187     324
188     100
189     176
19      423
190     60
191     29
192     55
193     66
194     50
195     485
196     99
197     63
198     75
199     422
2       76
20      98
200     253
201     122
202     76
203     37
204     31
205     206
206     39
207     46
208     48
209     20
21      162
210     32
211     55
212     876
213     910
214     214
215     54
216     82
217     104
218     42
219     138
22      220
```