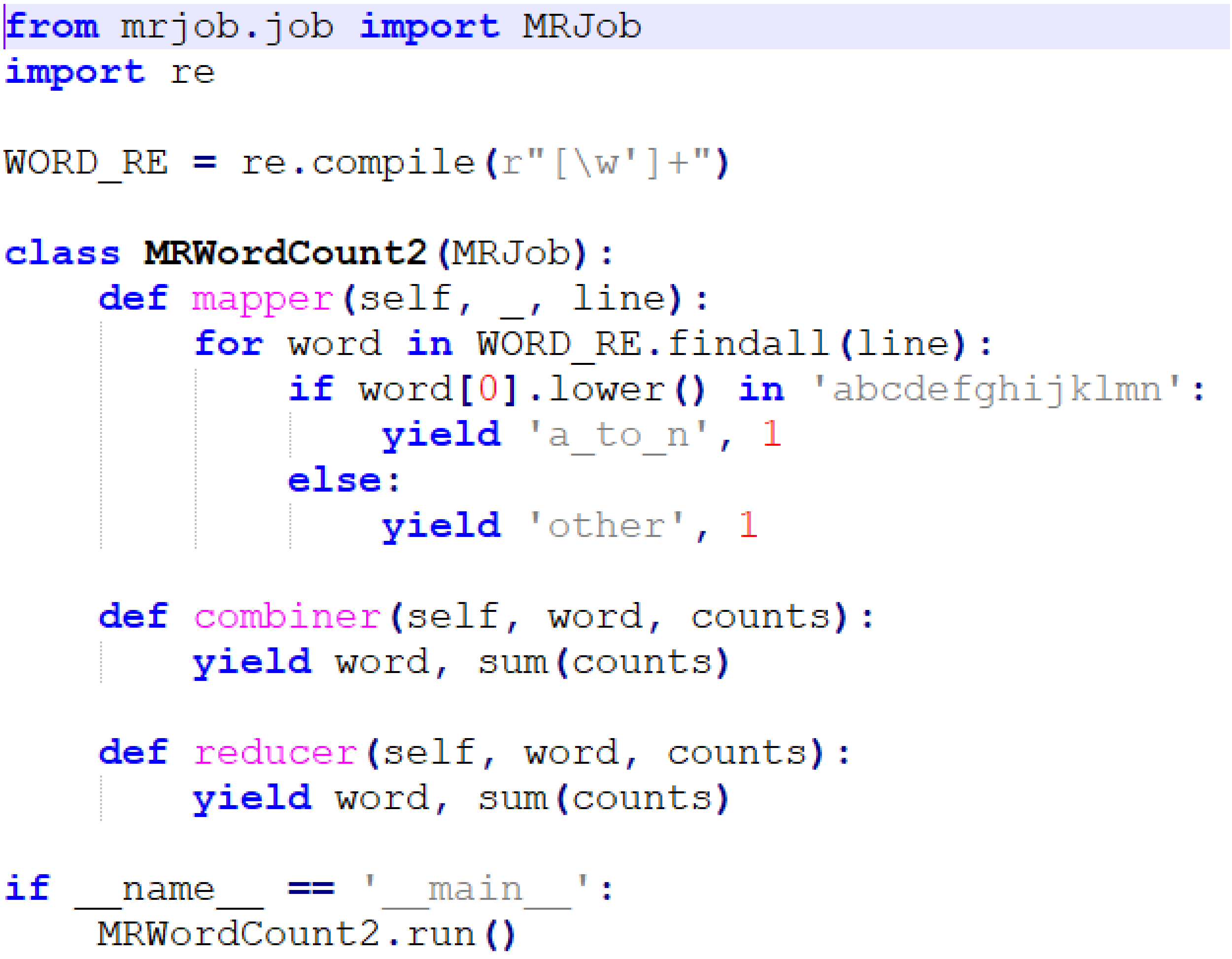
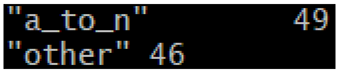
CSP554—Big Data Technologies

Assignment #3 (Modules 03a & 03b, 15 points)

6) (5 points) 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.

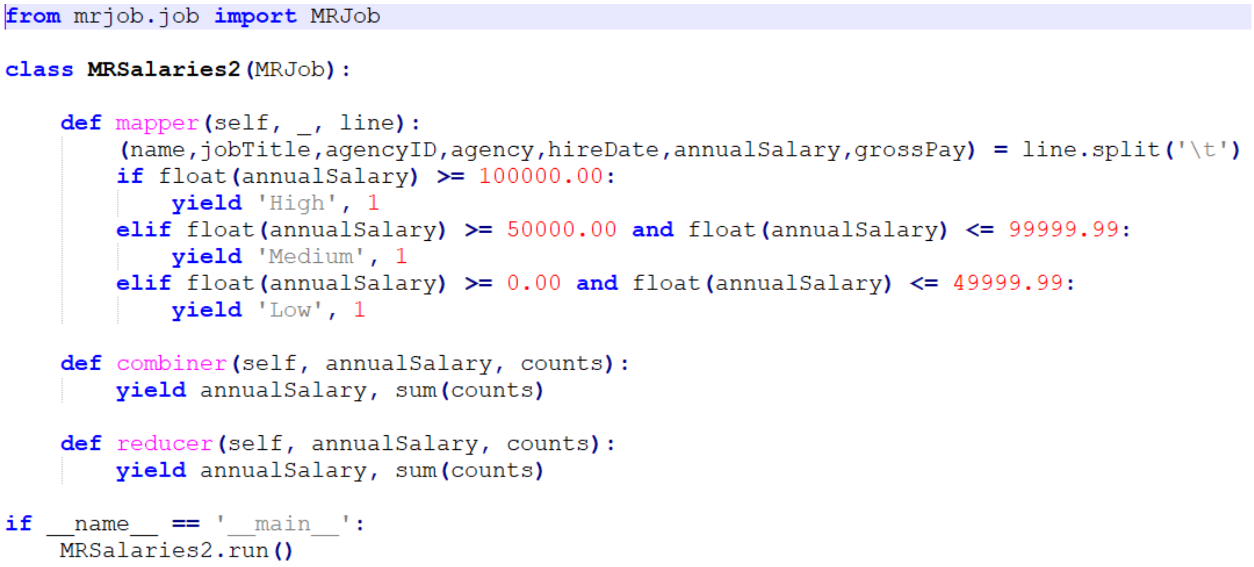


```
from mrjob.job import MRJob
import re

WORD_RE = re.compile(r"[\w']+")

class MRWordCount2(MRJob):
    def mapper(self, _, line):
        for word in WORD_RE.findall(line):
            if word[0].lower() in 'abcdefghijklmn':
                yield 'a_to_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__':
    MRWordCount2.run()
```

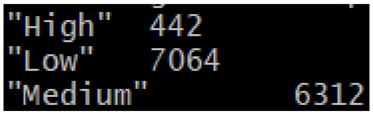


```
"a_to_n"        49
"other" 46
```

10) (5 points) 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.



```
from mrjob.job import MRJob

class MRSalaries2(MRJob):

    def mapper(self, _, line):
        (name,jobTitle,agencyID,agency,hireDate,annualSalary,grossPay) = line.split('\t')
        if float(annualSalary) >= 100000.00:
            yield 'High', 1
        elif float(annualSalary) >= 50000.00 and float(annualSalary) <= 99999.99:
            yield 'Medium', 1
        elif float(annualSalary) >= 0.00 and float(annualSalary) <= 49999.99:
            yield 'Low', 1

    def combiner(self, annualSalary, counts):
        yield annualSalary, sum(counts)

    def reducer(self, annualSalary, counts):
        yield annualSalary, sum(counts)

if __name__ == '__main__':
    MRSalaries2.run()
```



```
"High"   442
"Low"    7064
"Medium"        6312
```

12) (5 points) Review the slides 22-29 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.

Output might look something like the following:

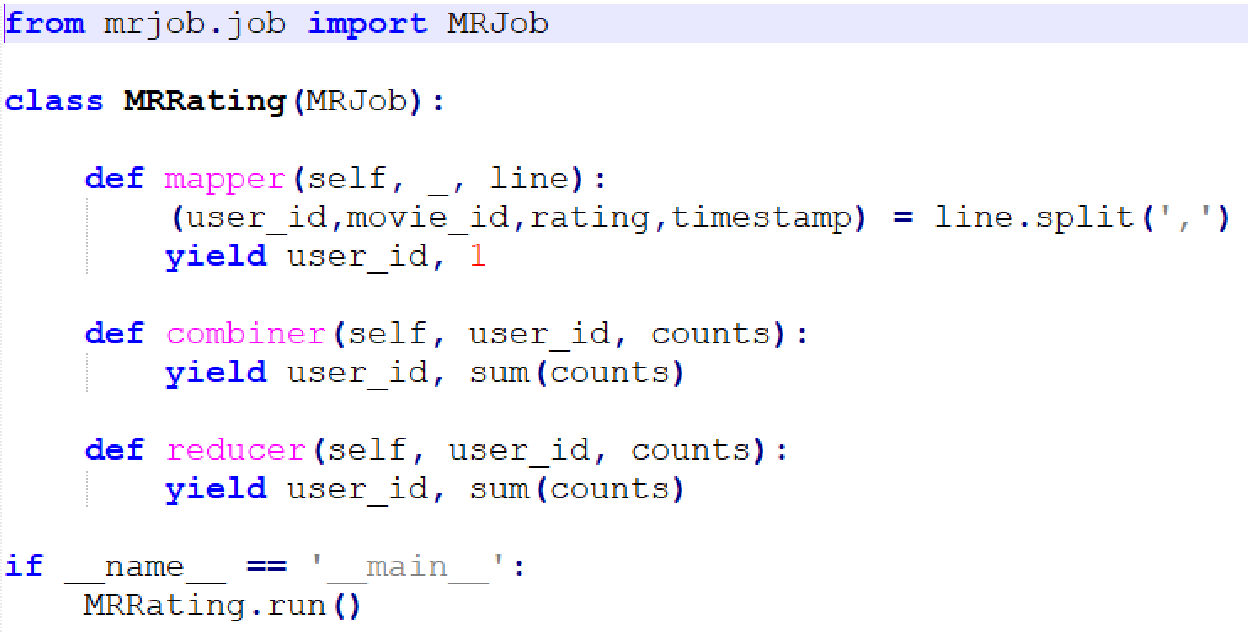
186: 2

192: 2

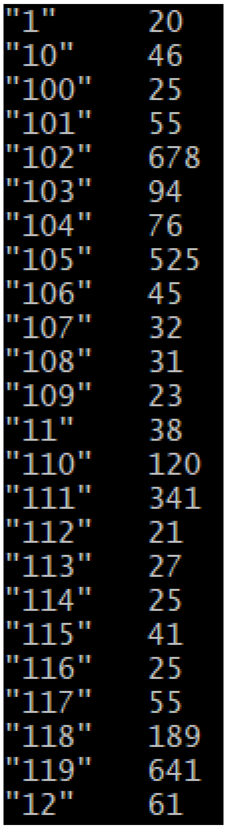
112: 1

etc.

Submit a copy of this program and a screen shot of the results of the program's execution (only 10 lines or so of the result) as the output of your assignment.



```
from mrjob.job import MRJob

class MRRating(MRJob):

    def mapper(self, _, line):
        (user_id,movie_id,rating,timestamp) = line.split(',')
        yield user_id, 1

    def combiner(self, user_id, counts):
        yield user_id, sum(counts)

    def reducer(self, user_id, counts):
        yield user_id, sum(counts)

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



```
"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
```