

Assignment - 3

Bhasheyam Krishnan

A20380078

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
import string

WORD_RE = re.compile(r"[\w']+")
class MRWordCount(MRJob):

    def mapper(self, _, line):
        start = 'a'
        end = 'n'
        letters = string.ascii_lowercase
        allowed = letters[letters.index(start):letters.index(end)+1]
        for word in WORD_RE.findall(line):
            if any(word.lower().startswith(x) for x in allowed):
                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__':
    MRWordCount.run()
```

```
Streaming final output from hdfs:///user/maria_dev/tmp/mrjob/WordCount2.maria_dev.20180204.052132.754796...
"a to n" 49
"other" 46
Removing HDFS temp directory hdfs:///user/maria_dev/tmp/mrjob/WordCount2.maria_dev.20180204.052132.754796...
Removing temp directory /tmp/WordCount2.maria_dev.20180204.052132.754796...
```

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 MRSalaries(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):
            yield "Medium", 1
        else:
            yield "Low", 1

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

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

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

```
Streaming final output from hdfs:///user/aria_dev/tmp/mrjob/Salaries2.aria_dev.20180202.084453.331125/output...
"High" 442
"Low" 7064
"Medium" 6312
Removing HDFS temp directory hdfs:///user/aria_dev/tmp/mrjob/Salaries2.aria_dev.20180202.084453.331125...
Removing temp directory /tmp/Salaries2.aria_dev.20180202.084453.331125...
[aria_dev@sandbox ~]$
```

12) (5 points) 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.

```
from mrjob.job import MRJob

class Moviecount(MRJob):

    def mapper(self, _, line):
        (user,movieid, rate, timestamp) = line.split(',')
        yield user, 1

    def combiner(self, user, movieid):
        yield user,sum(movieid)

    def reducer(self, user, moviesum):
        yield user,sum(moviesum)

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

```
"666" 40
"667" 68
"668" 20
"669" 37
"67" 103
"670" 31
"671" 115
"68" 123
"69" 81
"7" 88
"70" 83
"71" 23
"72" 191
"73" 1610
"74" 49
"75" 145
"76" 20
"77" 315
"78" 263
"79" 55
"8" 116
"80" 37
"81" 160
"82" 39
"83" 161
"84" 116
"85" 107
"86" 190
"87" 31
"88" 255
"89" 66
"9" 45
"90" 50
"91" 150
"92" 123
"93" 159
"94" 196
"95" 299
"96" 76
"97" 128
"98" 71
"99" 188
Removing HDFS temp directory hdfs:///user/maria_dev/tmp/mrjob/Moviecount.
Removing temp directory /tmp/Moviecount.maria_dev.20180202.095545.885956.
[maria_dev@sandbox ~]$
```