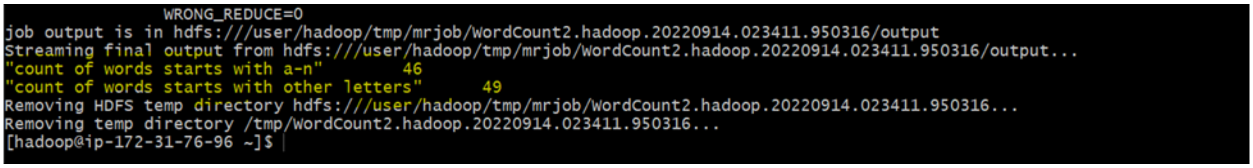
Assignment 3

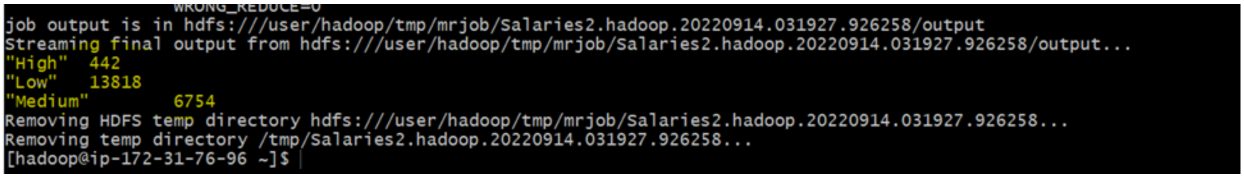
Created	@September 14, 2022 12:28 PM
Class	CSP 524- Big Data Technologies
Type	Assignment
Materials	

```
from mrjob.job import MRJob
import re

WORD_RE = re.compile(r"[\w']+")
alphabet = {'a','b','c','d','e','f','g','h','i','j','k','l','m','n'}
a_n_count = 0
other_count = 0

class MRWordCount(MRJob):

    def mapper(self, _, line):
        for word in WORD_RE.findall(line):
            if any(word.startswith(firstletter) for firstletter in alphabet):
                yield 'count of words starts with a-n', 1
            else:
                yield 'count of words starts with other letters', 1

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

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


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



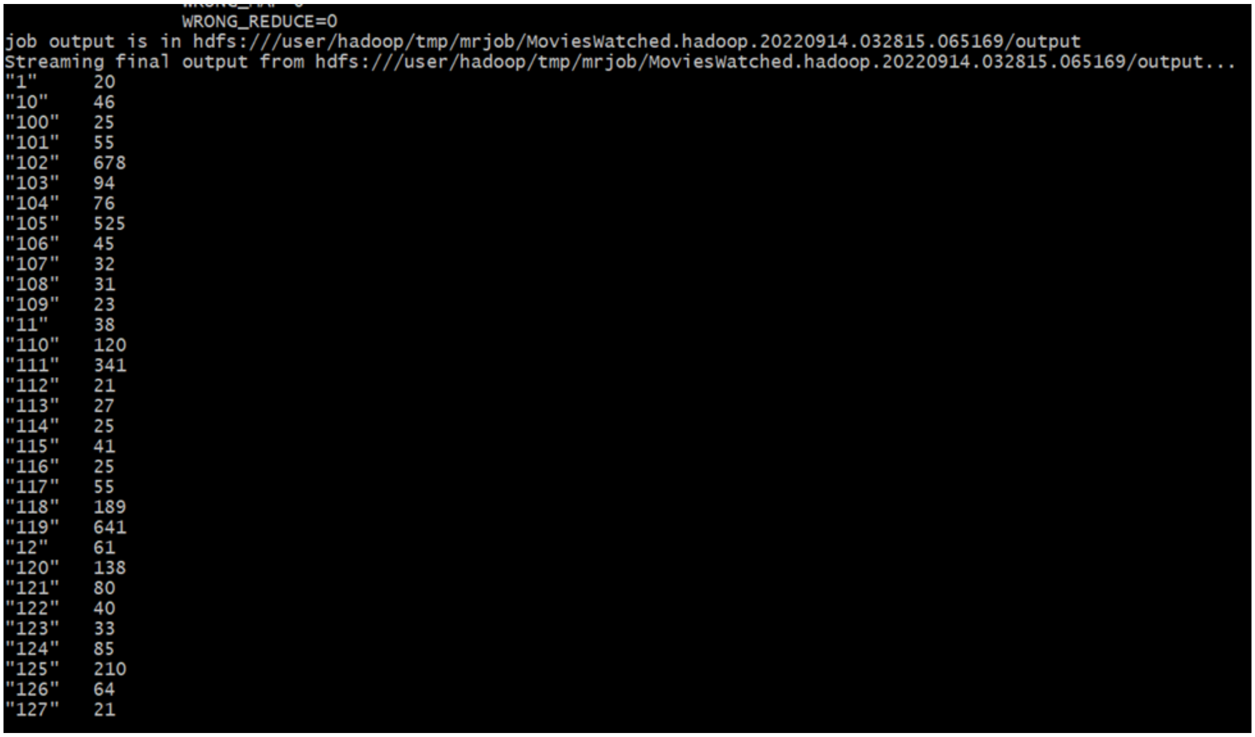
```
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220914.023411.950316/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220914.023411.950316/output...
"count of words starts with a-n"        46
"count of words starts with other letters"      49
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220914.023411.950316...
Removing temp directory /tmp/WordCount2.hadoop.20220914.023411.950316...
[hadoop@ip-172-31-76-96 ~]$
```

```
from mrjob.job import MRJob

class MRSalaries(MRJob):

    def mapper(self, _, line):
        (name,jobTitle,agencyID,agency,hireDate,annualSalary,grossPay) = line.split('\t')
        actualAnnualSalary = float(annualSalary)
        if actualAnnualSalary >= 100000:
            yield 'High', 1
        if actualAnnualSalary >= 50000:
            yield 'Medium', 1
        if actualAnnualSalary >= 0:
            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()
```



```
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220914.031927.926258/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220914.031927.926258/output...
"High"  442
"Low"   13818
"Medium"        6754
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220914.031927.926258...
Removing temp directory /tmp/Salaries2.hadoop.20220914.031927.926258...
[hadoop@ip-172-31-76-96 ~]$
```

```
from mrjob.job import MRJob

class MRRatings(MRJob):

    def mapper(self, _, line):
        (userid, movieid, rating, dateTimeValue) = line.split(',')
        yield userid, 1

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

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


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



```
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/MoviesWatched.hadoop.20220914.032815.065169/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/MoviesWatched.hadoop.20220914.032815.065169/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
```