

# CSP554-BIG DATA TECHNOLOGIES

## HOMEWORK-3

BAIRI ROHITH REDDY

A20526972

7. Modify WordCount.py to WordCount2.py to count words that begin with the small letters a-n and how many start with anything else.

```
hadoop@ip-172-31-49-146:~
GNU nano 2.9.8 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 re.match(r'^[a-n]\w*', word):
                yield 'a-n', 1
            else:
                yield 'other', 1

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

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

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

Output:-

```
hadoop@ip-172-31-49-146:~
Job Counters
  Data-local map tasks=4
  Killed map tasks=1
  Launched map tasks=4
  Launched reduce tasks=1
  Total megabyte-milliseconds taken by all map tasks=69539328
  Total megabyte-milliseconds taken by all reduce tasks=13658112
  Total time spent by all map tasks (ms)=45273
  Total time spent by all maps in occupied slots (ms)=2173104
  Total time spent by all reduce tasks (ms)=4446
  Total time spent by all reduces in occupied slots (ms)=426816
  Total vcore-milliseconds taken by all map tasks=45273
  Total vcore-milliseconds taken by all reduce tasks=4446

Map-Reduce Framework
  CPU time spent (ms)=5260
  Combine input records=95
  Combine output records=6
  Failed shuffles=0
  GC time elapsed (ms)=1027
  Input split bytes=448
  Map input records=6
  Map output bytes=858
  Map output materialized bytes=135
  Map output records=95
  Merged Map outputs=4
  Physical memory (bytes) snapshot=2042712064
  Reduce input groups=2
  Reduce input records=6
  Reduce output records=2
  Reduce shuffle bytes=135
  Shuffled Maps=4
  Spilled Records=12
  Total committed heap usage (bytes)=1576009728
  Virtual memory (bytes) snapshot=17866485760

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0

job output is in hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230225.190722.357745/output
Streaming output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230225.190722.357745/output...
"a-n" 46
"other" 49
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230225.190722.357745...
Removing temp directory /tmp/wordcount2.hadoop.20230225.190722.357745...
[hadoop@ip-172-31-49-146 ~]$
```

## 10. Edit the Salaries.py to Salaries2.py to output the employees with high low and medium salaries.

```
hadoop@ip-172-31-49-146:~$ nano salaries2.py
GNU nano 2.9.8 salaries2.py

from mrjob.job import MRJob

class MRSalaries(MRJob):

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

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

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

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

## Output:-

```
hadoop@ip-172-31-49-146:~$
Data-local map tasks=4
Killed map tasks=1
Launched map tasks=4
Launched reduce tasks=1
Total megabyte-milliseconds taken by all map tasks=66573312
Total megabyte-milliseconds taken by all reduce tasks=13246464
Total time spent by all map tasks (ms)=43342
Total time spent by all maps in occupied slots (ms)=2080416
Total time spent by all reduce tasks (ms)=4312
Total time spent by all reduces in occupied slots (ms)=413952
Total vcore-milliseconds taken by all map tasks=43342
Total vcore-milliseconds taken by all reduce tasks=4312
Map-Reduce Framework
CPU time spent (ms)=6350
Combine input records=13818
Combine output records=12
Failed Shuffles=0
GC time elapsed (ms)=982
Input split bytes=472
Map input records=13818
Map output bytes=129922
Map output materialized bytes=231
Map output records=13818
Merged Map outputs=4
Physical memory (bytes) snapshot=2056171520
Reduce input groups=3
Reduce input records=12
Reduce output records=3
Reduce shuffle bytes=231
Shuffled Maps =4
Spilled Records=24
Total committed heap usage (bytes)=1639972864
Virtual memory (bytes) snapshot=17866338304
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230225.191214.458541/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230225.191214.458541/output...
High 442
Low 7064
Medium 6312
Removing HDFS temporary directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230225.191214.458541...
Removing temp directory /tmp/Salaries2.hadoop.20230225.191214.458541...
[hadoop@ip-172-31-49-146 ~]$
```

### 13. Number of movies reviewed by each user.

```
hadoop@ip-172-31-49-146:~$ nano MoviesCount.py
GNU nano 2.9.8

from mrjob.job import MRJob

class MRMovieReviews(MRJob):

    def mapper(self, _, line):
        userId, movieId, rating, timestamp = 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__':
    MRMovieReviews.run()
```

Read 18 lines

Get Help Write Out Where Is Cut Text Justify Cur Pos M-U Undo  
Exit Read File Replace Uncut Text To Linter Go To Line M-E Redo

34°F Mostly cloudy 1:19 PM 2/25/2023

Output:-

```
hadoop@ip-172-31-49-146:~$
WRONG_MAP=0
WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/MoviesCount.hadoop.20230225.191555.956856/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/MoviesCount.hadoop.20230225.191555.956856/output...
"100:" 25
"101:" 55
"102:" 678
"103:" 94
"104:" 76
"105:" 525
"106:" 45
"107:" 32
"108:" 31
"109:" 23
"110:" 46
"111:" 120
"112:" 341
"113:" 21
"114:" 27
"115:" 25
"116:" 41
"117:" 25
"118:" 55
"119:" 189
"120:" 641
"121:" 38
"122:" 138
"123:" 80
"124:" 40
"125:" 33
"126:" 85
"127:" 210
"128:" 64
"129:" 21
"130:" 323
"131:" 26
"132:" 61
"133:" 375
"134:" 44
"135:" 94
"136:" 178
"137:" 311
"138:" 22
"139:" 50
"140:" 80
"141:" 81
"142:" 68
"143:" 53
```

34°F Mostly cloudy 1:17 PM 2/25/2023