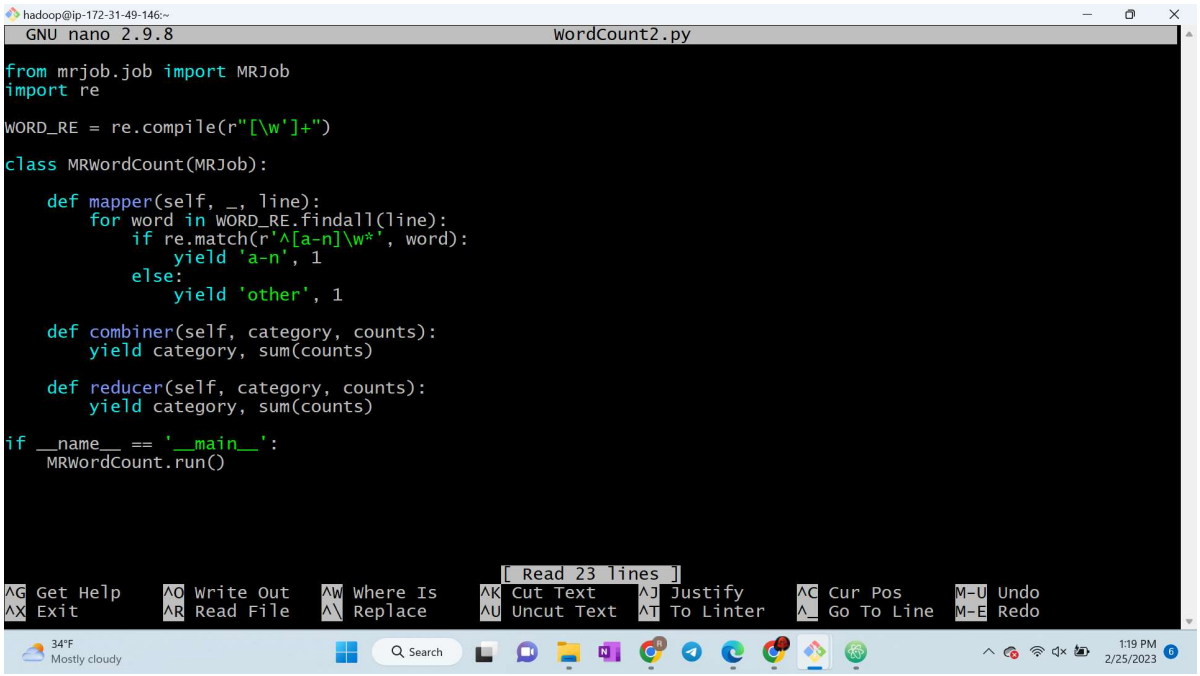
CSP554-BIG DATA TECHNOLOGIES
HOMEWORK-3

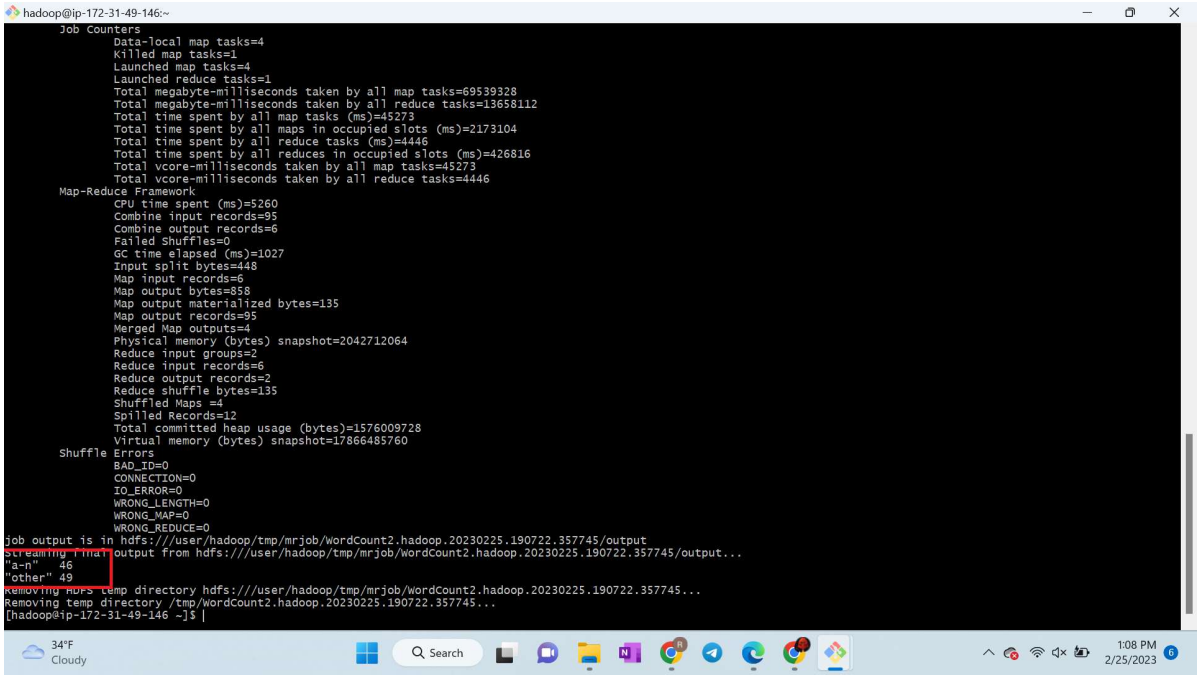
BAIRI ROHITH REDDY **A20526972**

---

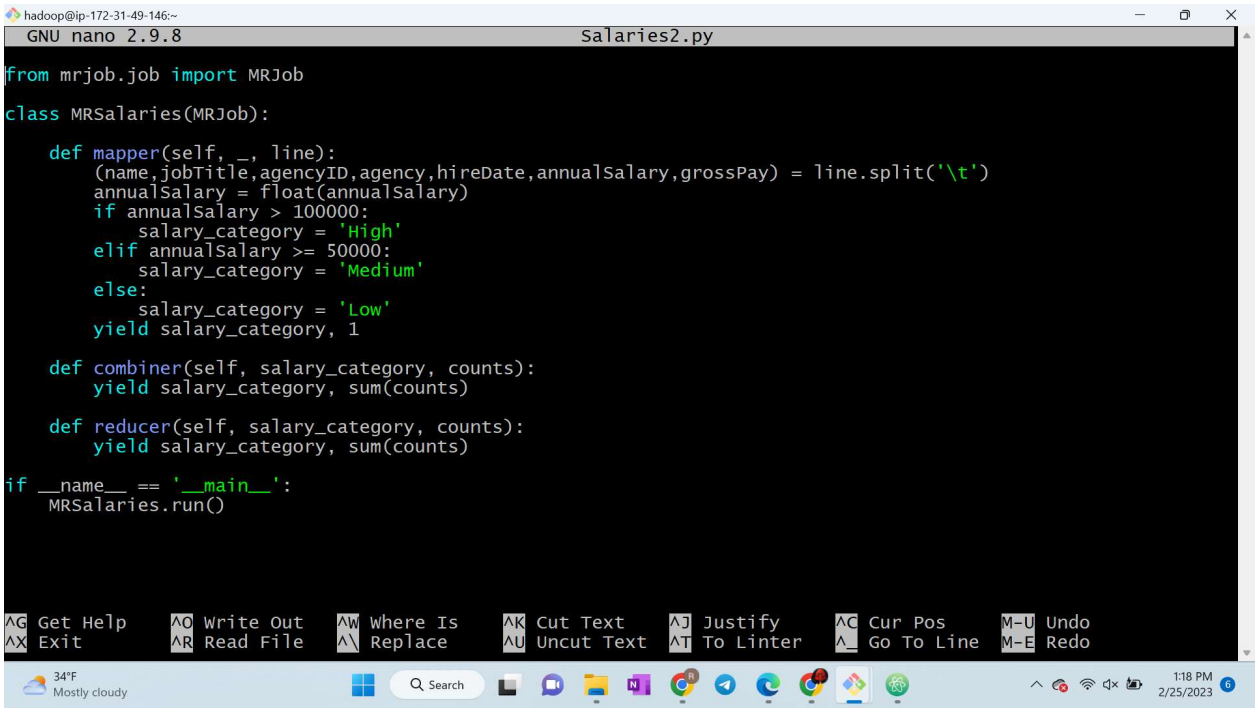
7. Modify WordCount.py to WordCount2.py to count words that begin with the small letters a-n and how many start with anything else.



Output:-



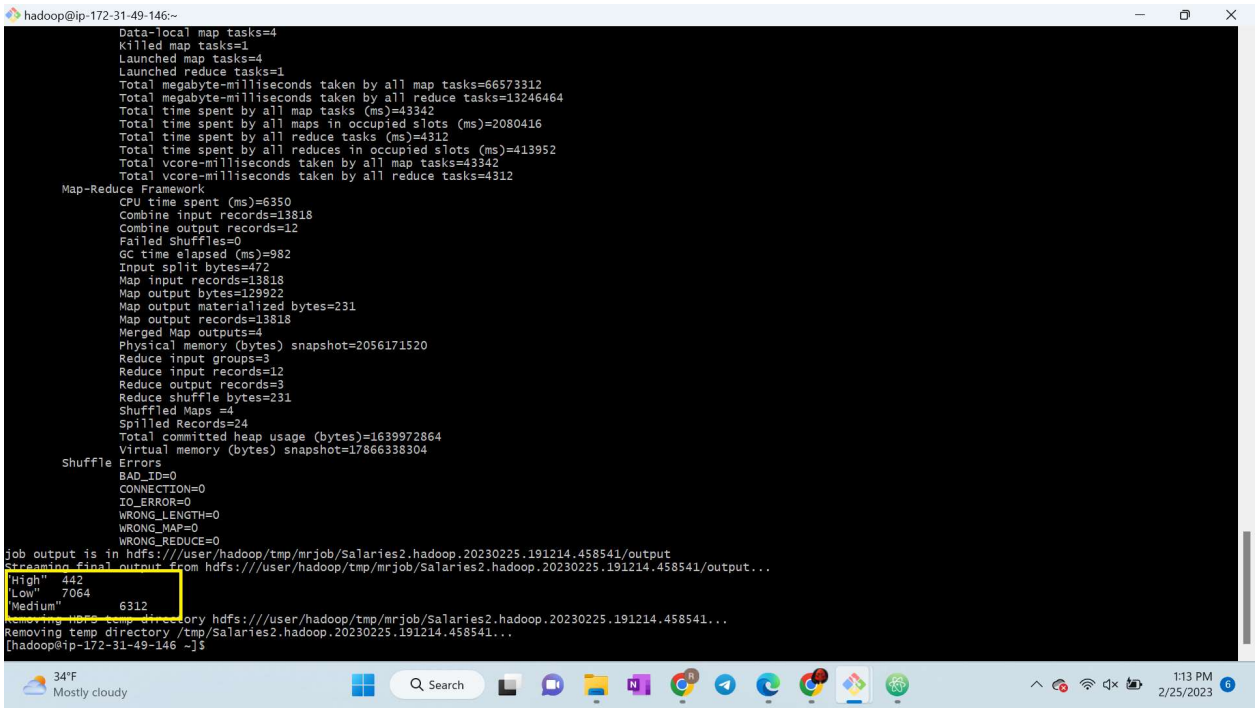
10. Edit the Salaries.py to Salaries2.py to output the employees with high low and medium salaries.



```
hadoop@ip-172-31-49-146:~
  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()

^G Get Help   ^O Write Out   ^W Where Is   ^K Cut Text   ^J Justify   ^C Cur Pos   M-U Undo
^X Exit       ^R Read File   ^\ Replace    ^U Uncut Text ^T To Linter  ^_ Go To Line M-E Redo
```

34°F Mostly cloudy Search 1:18 PM 2/25/2023

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 temp 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 ~]$
```

34°F Mostly cloudy Search 1:13 PM 2/25/2023

13. Number of movies reviewed by each user.

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

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 ]
^G Get Help    ^O Write Out   ^W Where Is    ^K Cut Text    ^J Justify     ^C Cur Pos     M-U Undo
^X Exit        ^R Read File   ^\ Replace     ^U Uncut Text  ^T To Linter   ^_ Go To Line  M-E Redo
```


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
"10:"   46
"110:"  120
"111:"  341
"112:"  21
"113:"  27
"114:"  25
"115:"  41
"116:"  25
"117:"  55
"118:"  189
"119:"  641
"11:"   38
"120:"  138
"121:"  80
"122:"  40
"123:"  33
"124:"  85
"125:"  210
"126:"  64
"127:"  21
"128:"  323
"129:"  26
"12:"   61
"130:"  375
"131:"  44
"132:"  94
"133:"  178
"134:"  311
"135:"  22
"136:"  50
"137:"  80
"138:"  81
"139:"  68
"13:"   53
```