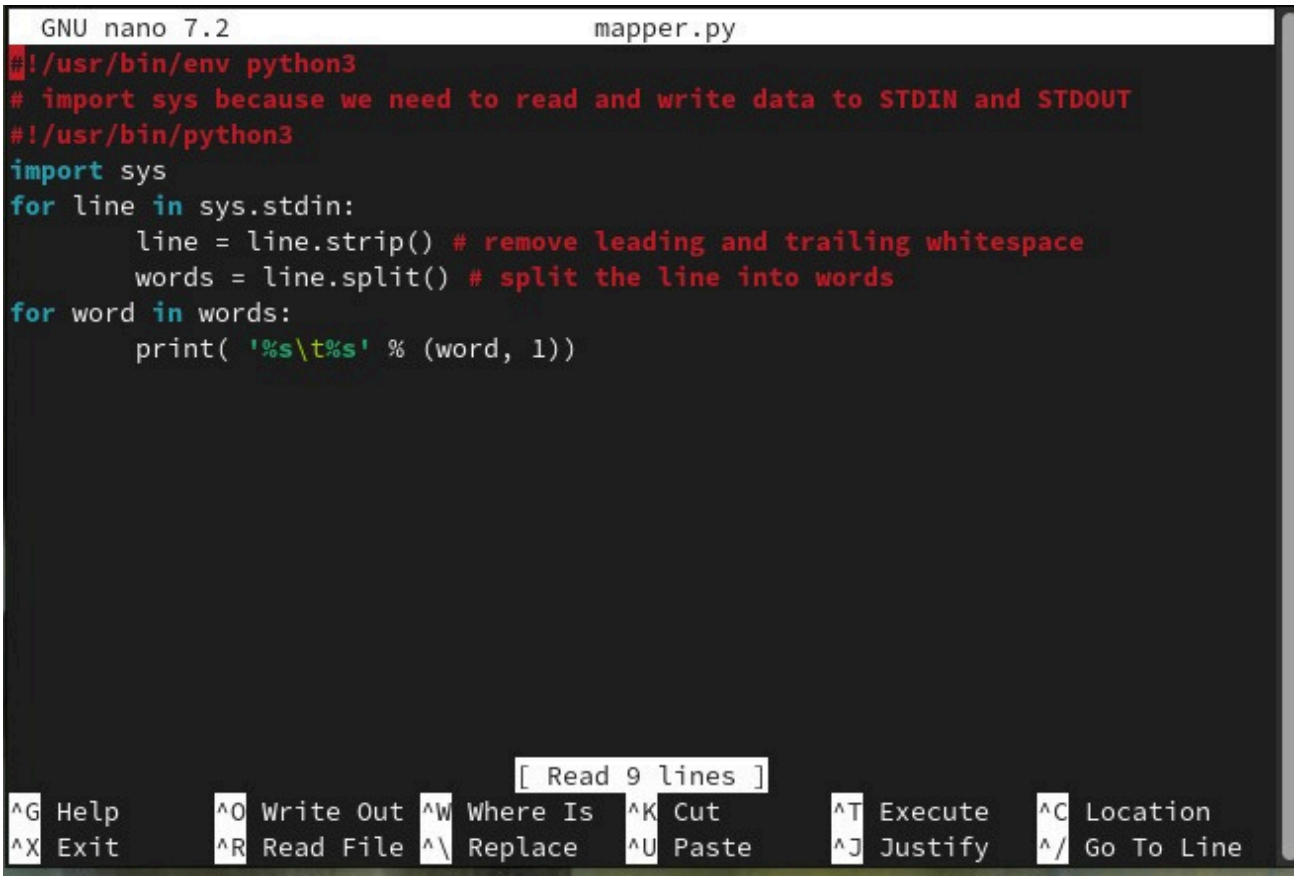
Exp. No : 2

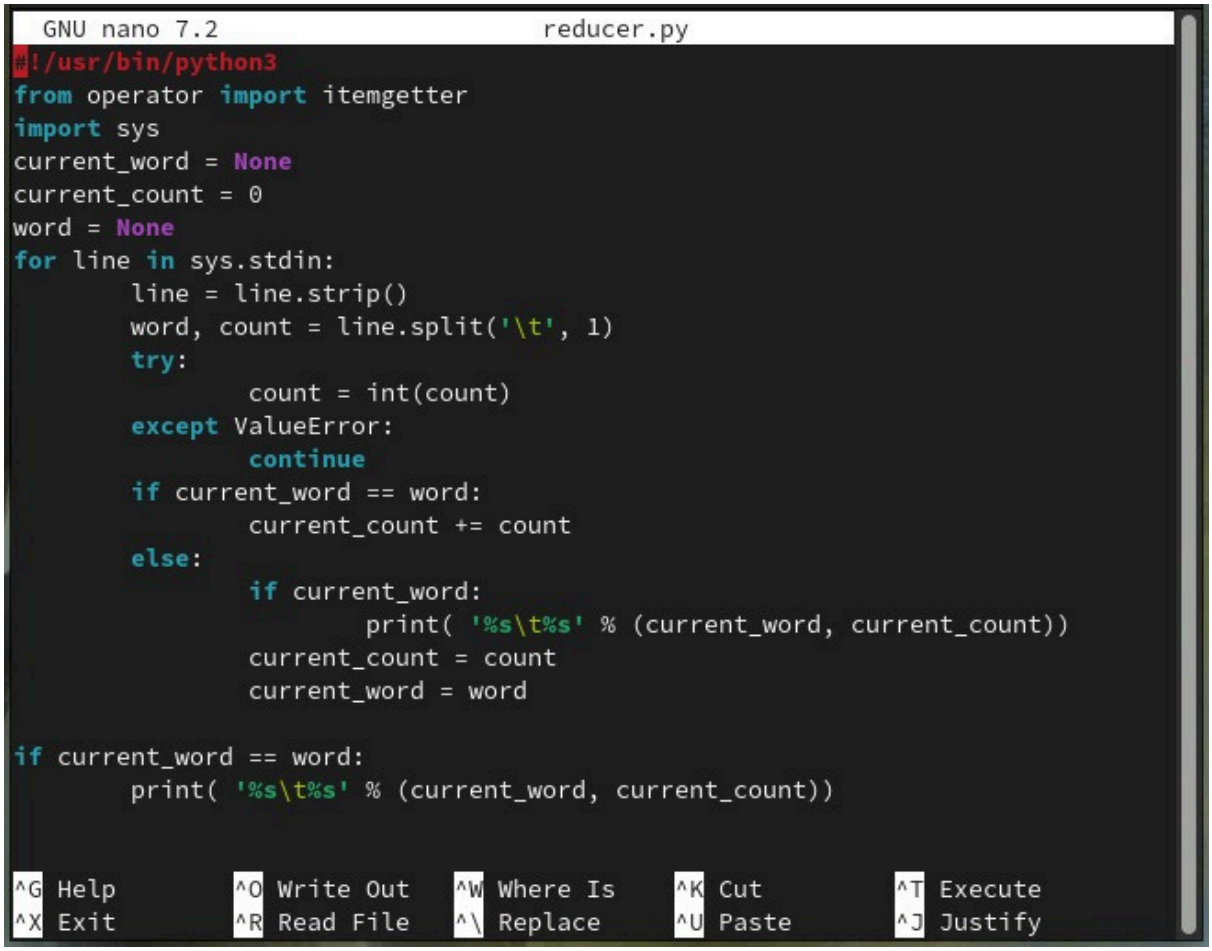
Word Count Map Reduce program

1. Create word_count.txt file

2. Create mapper.py program



3. Create reducer.py program.

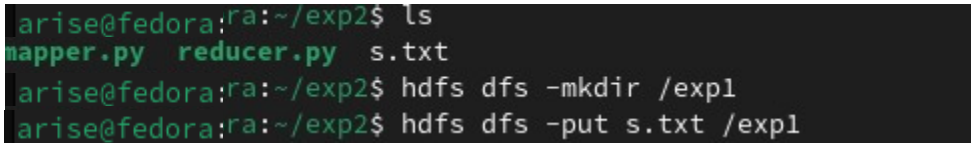


```
 GNU nano 7.2                          reducer.py
#!/usr/bin/python3
from operator import itemgetter
import sys
current_word = None
current_count = 0
word = None
for line in sys.stdin:
        line = line.strip()
        word, count = line.split('\t', 1)
        try:
                count = int(count)
        except ValueError:
                continue
        if current_word == word:
                current_count += count
        else:
                if current_word:
                        print( '%s\t%s' % (current_word, current_count))
                current_count = count
                current_word = word

if current_word == word:
        print( '%s\t%s' % (current_word, current_count))


^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify
```

4. Storing the word_count.txt in HDFS Storage.



```
arise@fedora:ra:~/exp2$ ls
mapper.py  reducer.py  s.txt
arise@fedora:ra:~/exp2$ hdfs dfs -mkdir /exp1
arise@fedora:ra:~/exp2$ hdfs dfs -put s.txt /exp1
```

5. Running the Word Count program using Hadoop Streaming.

```
arise@fedora$ hadoop jar $HADOOP_STREAMING -input /exp1/s.txt -output /exp1/output -mapper ~/exp2/mapper.py -reducer ~/exp3/reducer.py
packageJobJar: [/tmp/hadoop-unjar6064408018272369297/] [] /tmp/streamjob3446523685881352663.jar tmpDir=null
2024-10-10 20:37:55,356 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-10 20:37:56,062 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-10 20:37:58,219 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/harithaah/.staging/job_1728572703273_0001
2024-10-10 20:37:59,679 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-10 20:38:00,787 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-10 20:38:02,660 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1728572703273_0001
2024-10-10 20:38:02,660 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-10 20:38:03,651 INFO conf.Configuration: resource-types.xml not found
2024-10-10 20:38:03,655 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-10 20:38:05,195 INFO impl.YarnClientImpl: Submitted application application_1728572703273_0001
2024-10-10 20:38:05,505 INFO mapreduce.Job: The url to track the job: http://fedora:8088/proxy/application_1728572703273_0001/
2024-10-10 20:38:05,516 INFO mapreduce.Job: Running job: job_1728572703273_0001
2024-10-10 20:38:40,044 INFO mapreduce.Job: Job job_1728572703273_0001 running in uber mode : false
2024-10-10 20:38:40,104 INFO mapreduce.Job:  map 0% reduce 0%
```

```
in uber mode : false
2024-08-26 19:13:20,920 INFO mapreduce.Job:  map 0% reduce 0%
2024-08-26 19:13:35,602 INFO mapreduce.Job:  map 100% reduce 0%
2024-08-26 19:13:51,310 INFO mapreduce.Job:  map 100% reduce 100%
2024-08-26 19:13:56,305 INFO mapreduce.Job: Job job_1724678733414_0001 complete
d successfully
2024-08-26 19:13:56,572 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=97
                FILE: Number of bytes written=837208
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=414
                HDFS: Number of bytes written=71
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=23927
                Total time spent by all reduces in occupied slots (ms)=12078
                Total time spent by all map tasks (ms)=23927
                Total time spent by all reduce tasks (ms)=12078
                Total vcore-milliseconds taken by all map tasks=23927
```

```
        Total vcore-milliseconds taken by all map tasks=23927
        Total vcore-milliseconds taken by all reduce tasks=12078
        Total megabyte-milliseconds taken by all map tasks=24501248
        Total megabyte-milliseconds taken by all reduce tasks=12367872
    Map-Reduce Framework
        Map input records=7
        Map output records=10
        Map output bytes=71
        Map output materialized bytes=103
        Input split bytes=186
        Combine input records=0
        Combine output records=0
        Reduce input groups=10
        Reduce shuffle bytes=103
        Reduce input records=10
        Reduce output records=10
        Spilled Records=20
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=1759
        CPU time spent (ms)=8290
        Physical memory (bytes) snapshot=892342272
        Virtual memory (bytes) snapshot=7763681280
        Total committed heap usage (bytes)=687865856
        Peak Map Physical memory (bytes)=326397952
        Peak Map Virtual memory (bytes)=2586062848
        Peak Reduce Physical memory (bytes)=240001024
```

```
                Reduce output records=10
                Spilled Records=20
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=1759
                CPU time spent (ms)=8290
                Physical memory (bytes) snapshot=892342272
                Virtual memory (bytes) snapshot=7763681280
                Total committed heap usage (bytes)=687865856
                Peak Map Physical memory (bytes)=326397952
                Peak Map Virtual memory (bytes)=2586062848
                Peak Reduce Physical memory (bytes)=240001024
                Peak Reduce Virtual memory (bytes)=2593050624
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=228
        File Output Format Counters
                Bytes Written=71
2024-08-26 19:13:56,574 INFO streaming.StreamJob: Output directory: /exp2/outpu
t
```

Output :

```
arise@fedora:~/exp2$ hdfs dfs -cat /exp1/output/part-00000
Maria   1.0
Might   1.0
Tryna   1.0
dive    1.0
dough   1.0
in      1.0
make    1.0
my      1.0
own     1.0
the     1.0
```