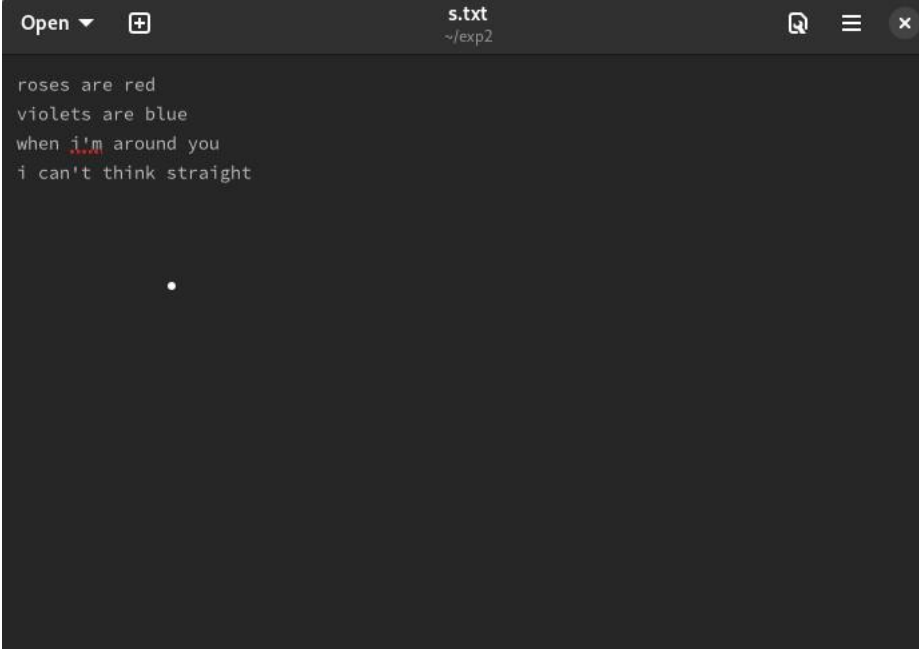


**Exp. No : 2****Word Count Map Reduce program**


1. Create s.txt file



A screenshot of a text editor window titled 's.txt' with a path of '~/.exp2'. The window contains the following text:

```
roses are red
violets are blue
when i'm around you
i can't think straight
```

2. Create mapper.py program



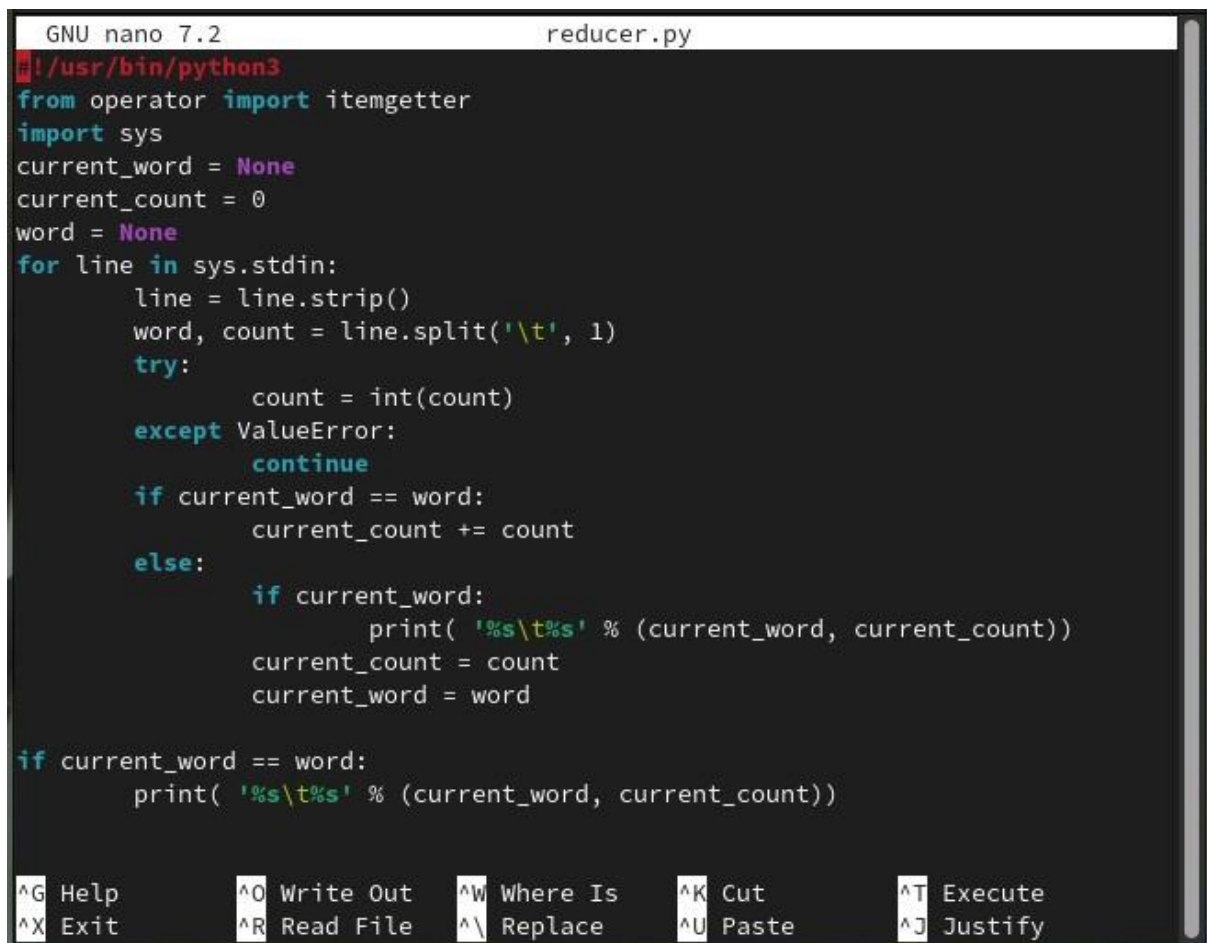
A screenshot of a nano text editor window titled 'mapper.py'. The window contains the following Python code:

```
#!/usr/bin/env python3
# import sys because we need to read and write data to STDIN and STDOUT
#!/usr/bin/python3
import sys
for line in sys.stdin:
    line = line.strip() # remove leading and trailing whitespace
    words = line.split() # split the line into words
    for word in words:
        print( '%s\t%s' % (word, 1))
```

At the bottom of the window, there is a status bar showing '[ Read 9 lines ]' and a table of keyboard shortcuts:

^G Help	^O Write Out	^W Where Is	^K Cut	^T Execute	^C Location
^X Exit	^R Read File	^_ Replace	^U Paste	^J Justify	^_ Go To Line

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. Running the Word Count program using Hadoop Streaming

```

kaarokki@fedora:~$ hadoop jar $HADOOP_STREAMING -input /exp1/s.txt -output /exp1/output1 -mapper ~/exp2/mapper.py -reducer
~/exp2/reducer.py
packageJobJar: [/tmp/hadoop-unjar4345487188727211757/] [] /tmp/streamjob2795676177018606624.jar tmpDir=null
2024-10-20 10:09:27,083 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-20 10:09:27,397 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-10-20 10:09:28,201 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/kaa
rokki/.staging/job_1729433151498_0001
2024-10-20 10:09:29,400 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-20 10:09:29,995 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-20 10:09:30,314 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729433151498_0001
2024-10-20 10:09:30,314 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-20 10:09:30,606 INFO conf.Configuration: resource-types.xml not found
2024-10-20 10:09:30,607 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-20 10:09:31,218 INFO impl.YarnClientImpl: Submitted application application_1729433151498_0001
2024-10-20 10:09:31,300 INFO mapreduce.Job: The url to track the job: http://fedora:8088/proxy/application_1729433151498_00
01/
2024-10-20 10:09:31,302 INFO mapreduce.Job: Running job: job_1729433151498_0001
2024-10-20 10:09:42,667 INFO mapreduce.Job: Job job_1729433151498_0001 running in uber mode : false
2024-10-20 10:09:42,670 INFO mapreduce.Job: map 0% reduce 0%
2024-10-20 10:09:50,910 INFO mapreduce.Job: map 100% reduce 0%
2024-10-20 10:09:57,012 INFO mapreduce.Job: map 100% reduce 100%
2024-10-20 10:09:58,067 INFO mapreduce.Job: Job job_1729433151498_0001 completed successfully
2024-10-20 10:09:58,237 INFO mapreduce.Job: Counters: 54
    File System Counters
        FILE: Number of bytes read=278

```

```

2024-10-20 10:09:58,237 INFO mapreduce.Job: Counters: 54
    File System Counters
        FILE: Number of bytes read=278
        FILE: Number of bytes written=835554
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=396
        HDFS: Number of bytes written=175
        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)=11182
        Total time spent by all reduces in occupied slots (ms)=4117
        Total time spent by all map tasks (ms)=11182
        Total time spent by all reduce tasks (ms)=4117
        Total vcore-milliseconds taken by all map tasks=11182
        Total vcore-milliseconds taken by all reduce tasks=4117
        Total megabyte-milliseconds taken by all map tasks=11450368
        Total megabyte-milliseconds taken by all reduce tasks=4215808
    Map-Reduce Framework
        Map input records=7

```

## Map-Reduce Framework

```
Map input records=7
Map output records=30
Map output bytes=212
Map output materialized bytes=284
Input split bytes=168
Combine input records=0
Combine output records=0
Reduce input groups=24
Reduce shuffle bytes=284
Reduce input records=30
Reduce output records=24
Spilled Records=60
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=361
CPU time spent (ms)=3480
Physical memory (bytes) snapshot=885071872
Virtual memory (bytes) snapshot=7768154112
Total committed heap usage (bytes)=684195840
Peak Map Physical memory (bytes)=325566464
Peak Map Virtual memory (bytes)=2589315072
Peak Reduce Physical memory (bytes)=234557440
Peak Reduce Virtual memory (bytes)=2593460224
```

## Shuffle Errors

```
BAD_ID=0
```

```
Peak Map Virtual memory (bytes)=2589315072
Peak Reduce Physical memory (bytes)=234557440
Peak Reduce Virtual memory (bytes)=2593460224
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=175
```

```
2024-10-20 10:09:58,237 INFO streaming.StreamJob: Output directory: /expl/output1
```

**Output :**

```
kaarokki@fedora:~$ hdfs dfs -cat /expl/output/part-00000
Callin 1
Finally 1
LA 2
Lookin 1
Lost 1
Made 1
Maria 2
Might 1
Trynnna 1
dive 1
dough 1
for 2
in 2
it 1
make 1
marina 1
my 1
own 1
the 2
though 1
to 1
weed 1
without 1
yeah 2
```