BIG DATA

SECTION D

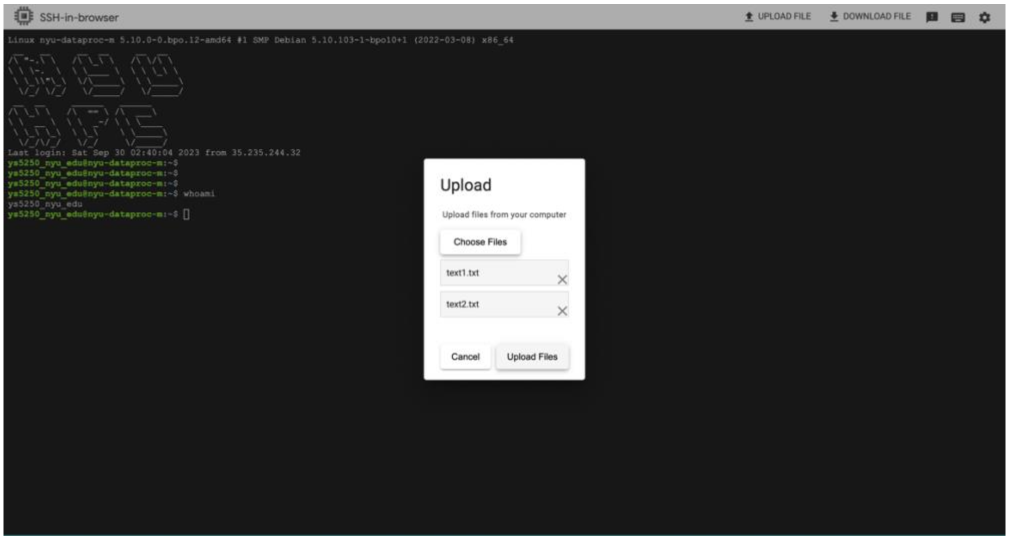
Fall’2023

Yogya Sharma

ys5250

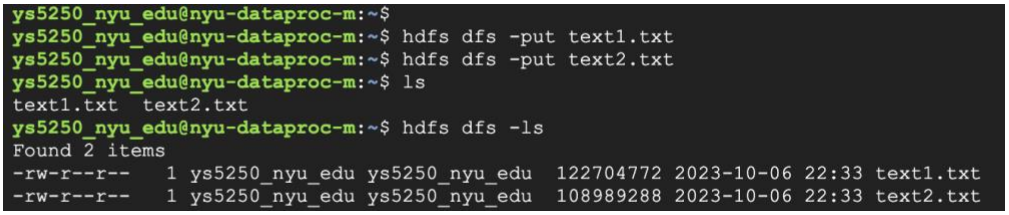
GHW#2

Uploaded the text files:



```
Linux nyu-dataproc-m 5.10.0-0.bpo.12-amd64 #1 SMP Debian 5.10.103-1-bpo10+1 (2022-03-08) x86_64
Last login: Sat Sep 30 02:40:04 2023 from 35.235.244.32
ys5250_nyu_edu@nyu-dataproc-m:~$
ys5250_nyu_edu@nyu-dataproc-m:~$
ys5250_nyu_edu@nyu-dataproc-m:~$
ys5250_nyu_edu@nyu-dataproc-m:~$ whoami
ys5250_nyu_edu
ys5250_nyu_edu@nyu-dataproc-m:~$
```

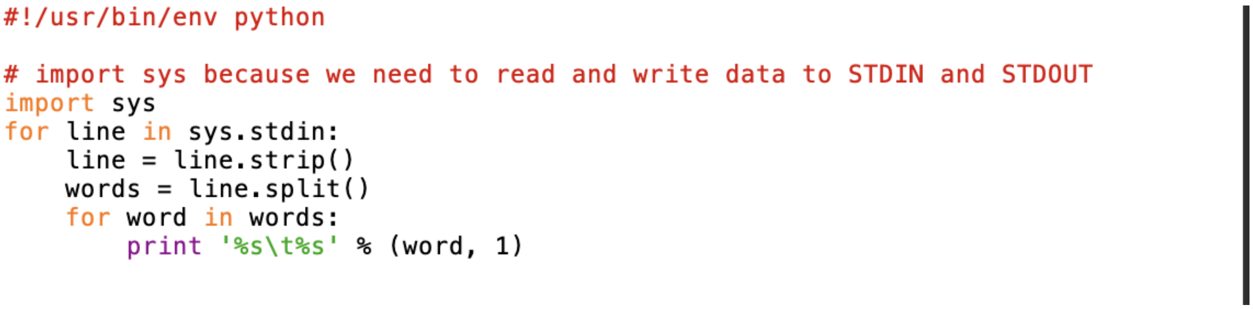
Put them on HDFS



```
ys5250_nyu_edu@nyu-dataproc-m:~$
ys5250_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put text1.txt
ys5250_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put text2.txt
ys5250_nyu_edu@nyu-dataproc-m:~$ ls
text1.txt  text2.txt
ys5250_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -ls
Found 2 items
-rw-r--r--   1 ys5250_nyu_edu ys5250_nyu_edu  122704772 2023-10-06 22:33 text1.txt
-rw-r--r--   1 ys5250_nyu_edu ys5250_nyu_edu  108989288 2023-10-06 22:33 text2.txt
```

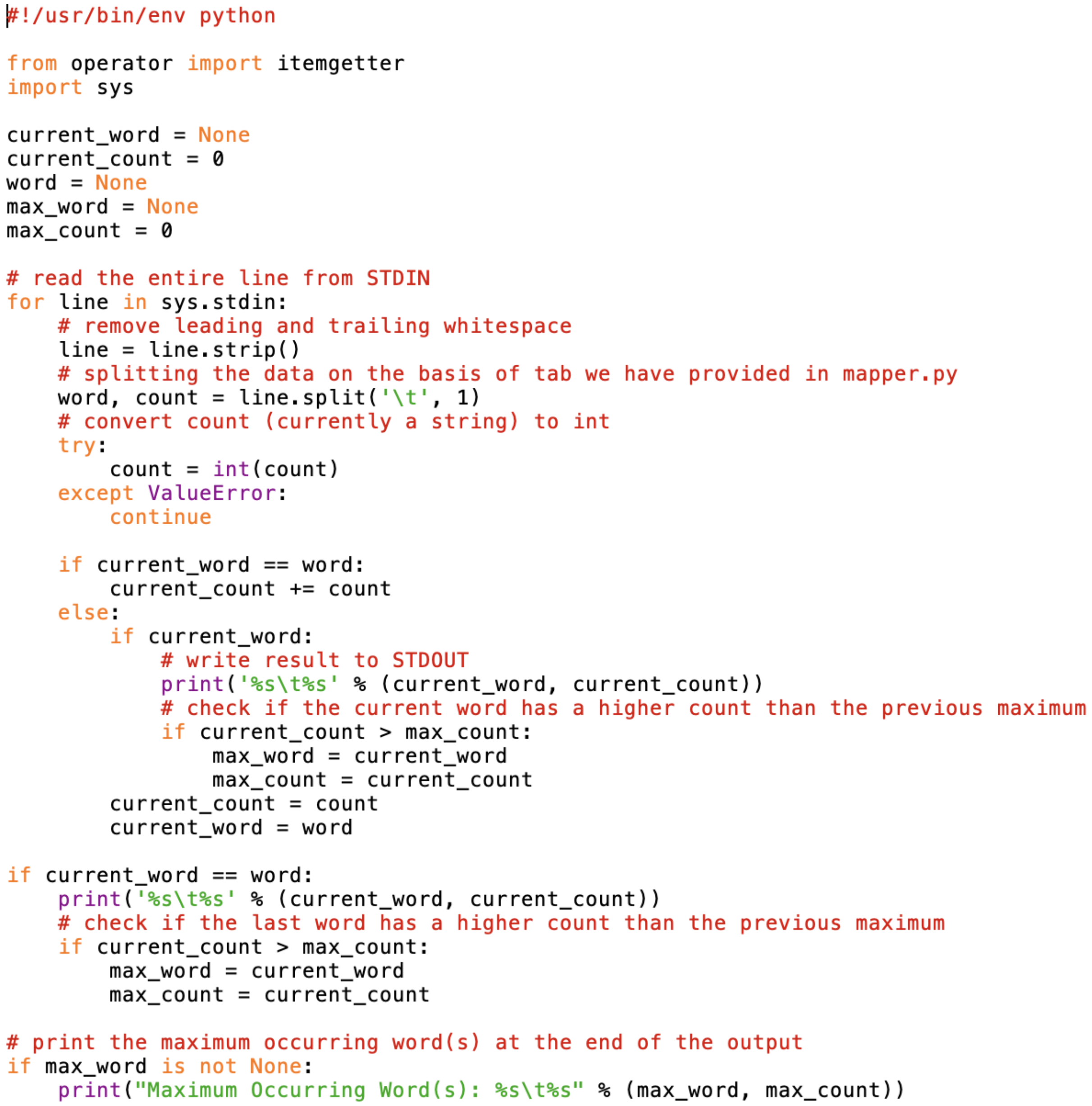
Created the mapper.py and reducer.py and uploaded them:

mapper.py:



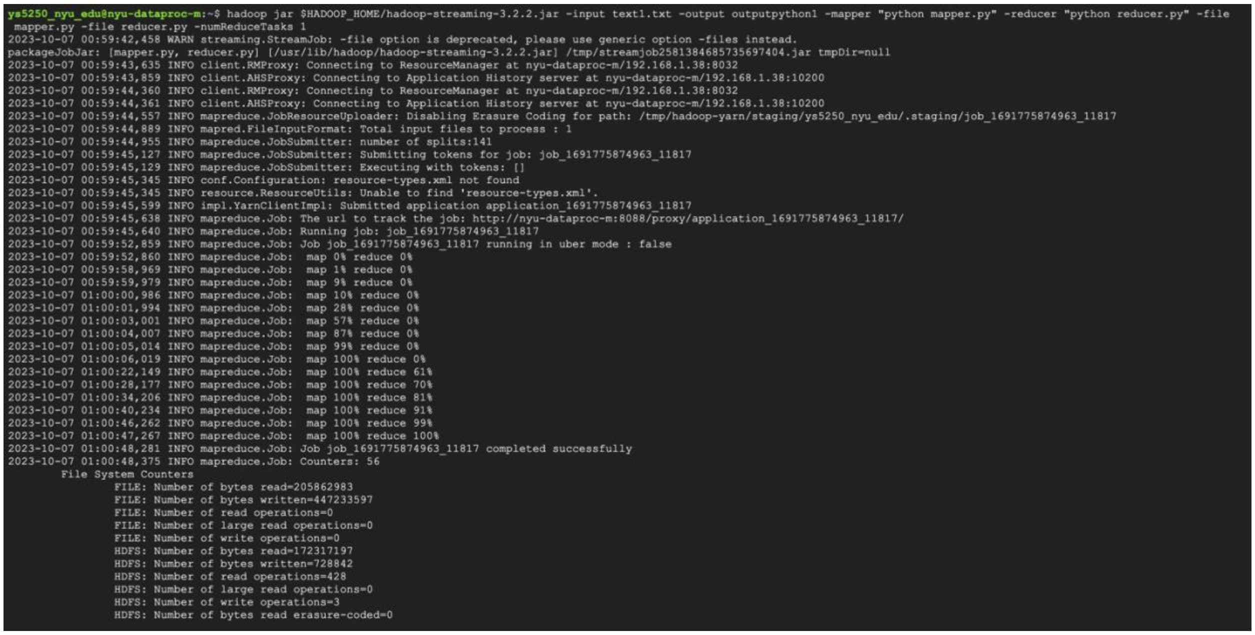
```python
#!/usr/bin/env python

# import sys because we need to read and write data to STDIN and STDOUT
import sys
for line in sys.stdin:
    line = line.strip()
    words = line.split()
    for word in words:
        print '%s\t%s' % (word, 1)
```

reducer.py

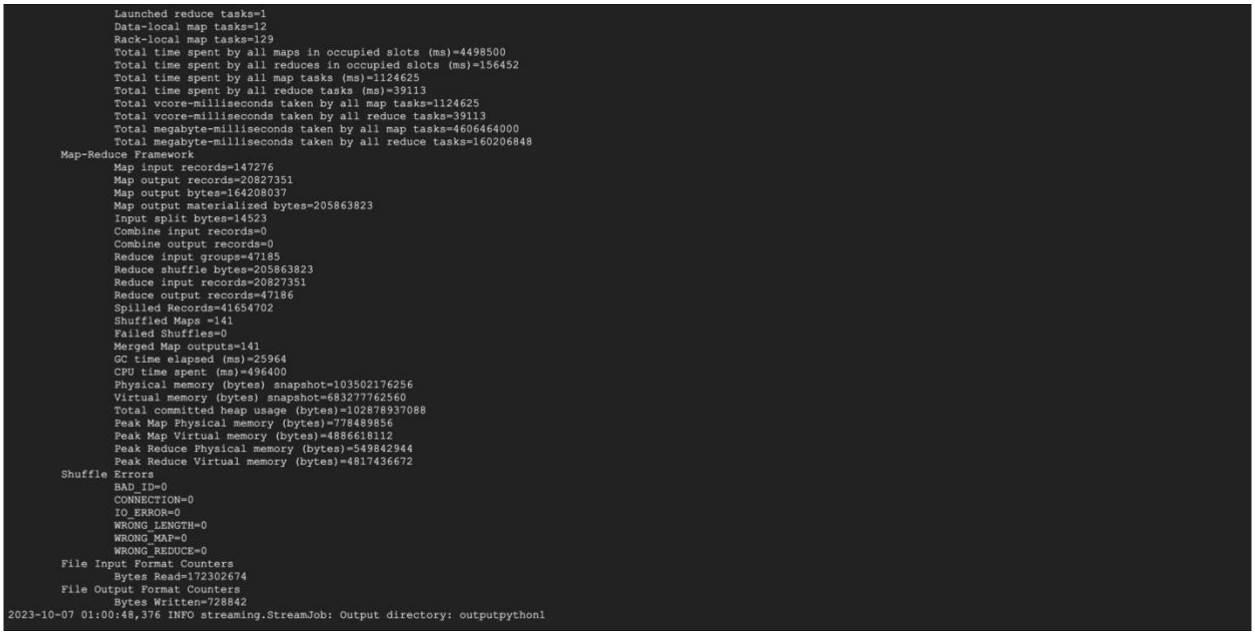


```
#!/usr/bin/env python

from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None
max_word = None
max_count = 0

# read the entire line from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # splitting the data on the basis of tab we have provided in mapper.py
    word, count = line.split('\t', 1)
    # convert count (currently a string) to int
    try:
        count = int(count)
    except ValueError:
        continue

    if current_word == word:
        current_count += count
    else:
        if current_word:
            # write result to STDOUT
            print('%s\t%s' % (current_word, current_count))
            # check if the current word has a higher count than the previous maximum
            if current_count > max_count:
                max_word = current_word
                max_count = current_count
        current_count = count
        current_word = word

if current_word == word:
    print('%s\t%s' % (current_word, current_count))
    # check if the last word has a higher count than the previous maximum
    if current_count > max_count:
        max_word = current_word
        max_count = current_count

# print the maximum occurring word(s) at the end of the output
if max_word is not None:
    print("Maximum Occurring Word(s): %s\t%s" % (max_word, max_count))
```

Run the command for text1.txt

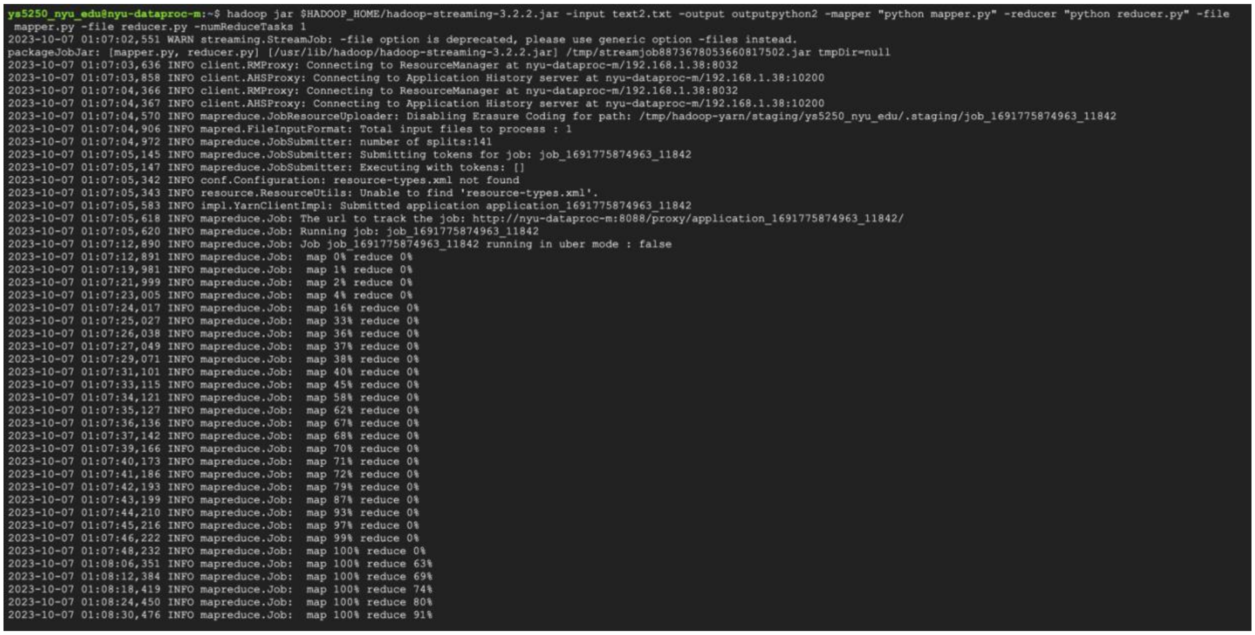


```
ys5250_nyu_edu@nyu-dataproc-m:~$ hadoop jar $HADOOP_HOME/hadoop-streaming-3.2.2.jar -input text1.txt -output outputpython1 -mapper "python mapper.py" -reducer "python reducer.py" -file
 mapper.py -file reducer.py -numReduceTasks 1
2023-10-07 00:59:42,458 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, reducer.py] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob2581384685735697404.jar tmpDir=null
2023-10-07 00:59:43,635 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.38:8032
2023-10-07 00:59:43,859 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.38:10200
2023-10-07 00:59:44,360 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.38:8032
2023-10-07 00:59:44,361 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.38:10200
2023-10-07 00:59:44,557 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ys5250_nyu_edu/.staging/job_1691775874963_11817
2023-10-07 00:59:44,889 INFO mapred.FileInputFormat: Total input files to process : 1
2023-10-07 00:59:44,955 INFO mapreduce.JobSubmitter: number of splits:141
2023-10-07 00:59:45,127 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1691775874963_11817
2023-10-07 00:59:45,129 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-10-07 00:59:45,345 INFO conf.Configuration: resource-types.xml not found
2023-10-07 00:59:45,345 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-10-07 00:59:45,599 INFO impl.YarnClientImpl: Submitted application application_1691775874963_11817
2023-10-07 00:59:45,638 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1691775874963_11817/
2023-10-07 00:59:45,640 INFO mapreduce.Job: Running job: job_1691775874963_11817
2023-10-07 00:59:52,859 INFO mapreduce.Job: Job job_1691775874963_11817 running in uber mode : false
2023-10-07 00:59:52,860 INFO mapreduce.Job:  map 0% reduce 0%
2023-10-07 00:59:58,969 INFO mapreduce.Job:  map 1% reduce 0%
2023-10-07 00:59:59,979 INFO mapreduce.Job:  map 9% reduce 0%
2023-10-07 01:00:00,986 INFO mapreduce.Job:  map 10% reduce 0%
2023-10-07 01:00:01,994 INFO mapreduce.Job:  map 28% reduce 0%
2023-10-07 01:00:03,001 INFO mapreduce.Job:  map 57% reduce 0%
2023-10-07 01:00:04,007 INFO mapreduce.Job:  map 87% reduce 0%
2023-10-07 01:00:05,014 INFO mapreduce.Job:  map 99% reduce 0%
2023-10-07 01:00:06,019 INFO mapreduce.Job:  map 100% reduce 0%
2023-10-07 01:00:22,149 INFO mapreduce.Job:  map 100% reduce 61%
2023-10-07 01:00:28,177 INFO mapreduce.Job:  map 100% reduce 70%
2023-10-07 01:00:34,206 INFO mapreduce.Job:  map 100% reduce 81%
2023-10-07 01:00:40,234 INFO mapreduce.Job:  map 100% reduce 91%
2023-10-07 01:00:46,262 INFO mapreduce.Job:  map 100% reduce 99%
2023-10-07 01:00:47,267 INFO mapreduce.Job:  map 100% reduce 100%
2023-10-07 01:00:48,281 INFO mapreduce.Job: Job job_1691775874963_11817 completed successfully
2023-10-07 01:00:48,375 INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=205862983
                FILE: Number of bytes written=447233597
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=172317197
                HDFS: Number of bytes written=728842
                HDFS: Number of read operations=428
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
```

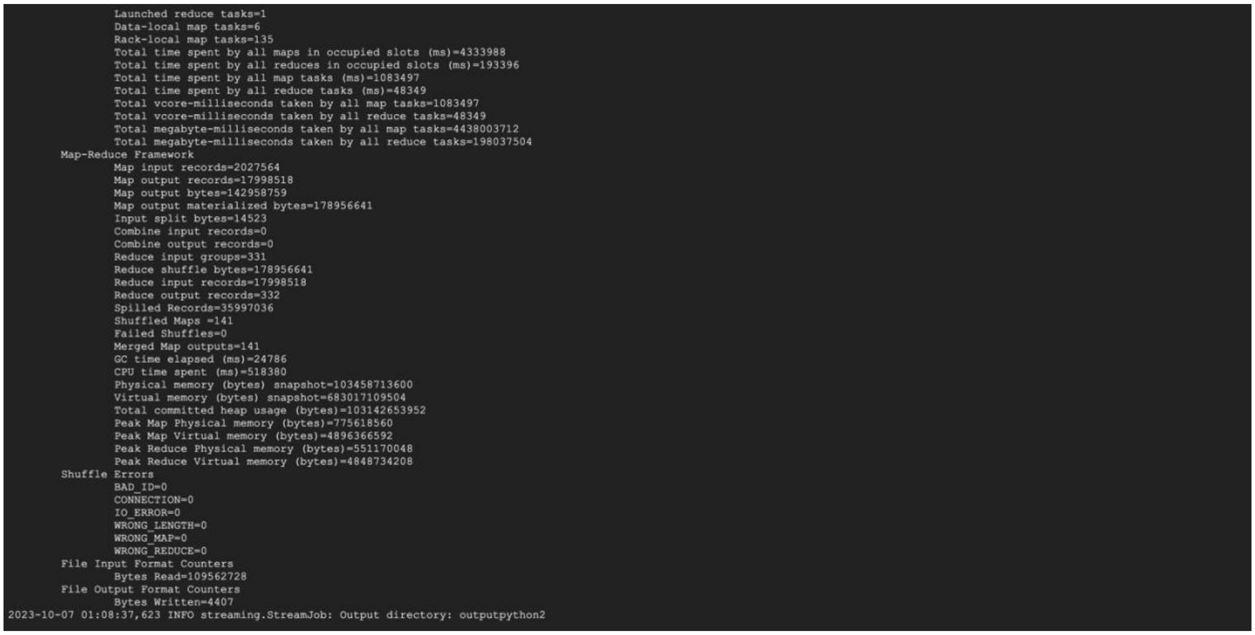


```
                Launched reduce tasks=1
                Data-local map tasks=12
                Rack-local map tasks=129
                Total time spent by all maps in occupied slots (ms)=4498500
                Total time spent by all reduces in occupied slots (ms)=156452
                Total time spent by all map tasks (ms)=1124625
                Total time spent by all reduce tasks (ms)=39113
                Total vcore-milliseconds taken by all map tasks=1124625
                Total vcore-milliseconds taken by all reduce tasks=39113
                Total megabyte-milliseconds taken by all map tasks=4606464000
                Total megabyte-milliseconds taken by all reduce tasks=160206848
        Map-Reduce Framework
                Map input records=147276
                Map output records=20827351
                Map output bytes=164208037
                Map output materialized bytes=205863823
                Input split bytes=14523
                Combine input records=0
                Combine output records=0
                Reduce input groups=47185
                Reduce shuffle bytes=205863823
                Reduce input records=20827351
                Reduce output records=47186
                Spilled Records=41654702
                Shuffled Maps =141
                Failed Shuffles=0
                Merged Map outputs=141
                GC time elapsed (ms)=25964
                CPU time spent (ms)=496400
                Physical memory (bytes) snapshot=103502176256
                Virtual memory (bytes) snapshot=683277762560
                Total committed heap usage (bytes)=102878937088
                Peak Map Physical memory (bytes)=778489856
                Peak Map Virtual memory (bytes)=4886618112
                Peak Reduce Physical memory (bytes)=549842944
                Peak Reduce Virtual memory (bytes)=4817436672
        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=172302674
        File Output Format Counters
                Bytes Written=728842
2023-10-07 01:00:48,376 INFO streaming.StreamJob: Output directory: outputpython1
```

Run the command for text2.txt

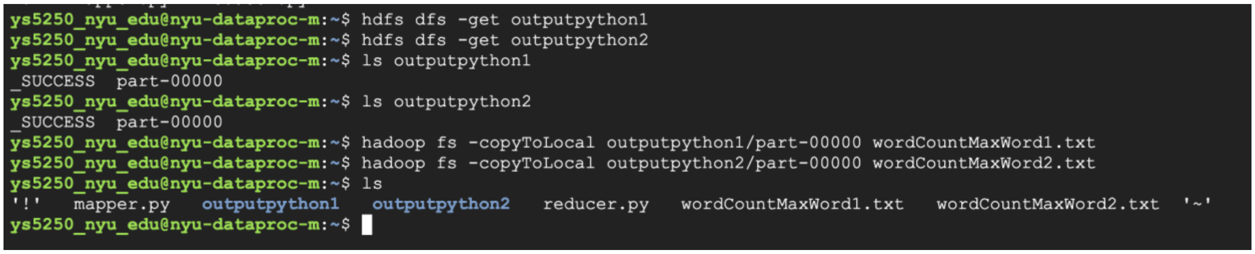


```
ys5250_nyu_edu@nyu-dataproc-m:~$ hadoop jar $HADOOP_HOME/hadoop-streaming-3.2.2.jar -input text2.txt -output outputpython2 -mapper "python mapper.py" -reducer "python reducer.py" -file
 mapper.py -file reducer.py -numReduceTasks 1
2023-10-07 01:07:02,551 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, reducer.py] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob8873678053660817502.jar tmpDir=null
2023-10-07 01:07:03,636 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.38:8032
2023-10-07 01:07:03,858 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.38:10200
2023-10-07 01:07:04,366 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.38:8032
2023-10-07 01:07:04,367 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.38:10200
2023-10-07 01:07:04,570 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ys5250_nyu_edu/.staging/job_1691775874963_11842
2023-10-07 01:07:04,906 INFO mapred.FileInputFormat: Total input files to process : 1
2023-10-07 01:07:04,972 INFO mapreduce.JobSubmitter: number of splits:141
2023-10-07 01:07:05,145 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1691775874963_11842
2023-10-07 01:07:05,147 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-10-07 01:07:05,342 INFO conf.Configuration: resource-types.xml not found
2023-10-07 01:07:05,343 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-10-07 01:07:05,583 INFO impl.YarnClientImpl: Submitted application application_1691775874963_11842
2023-10-07 01:07:05,618 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1691775874963_11842/
2023-10-07 01:07:05,620 INFO mapreduce.Job: Running job: job_1691775874963_11842
2023-10-07 01:07:12,890 INFO mapreduce.Job: Job job_1691775874963_11842 running in uber mode : false
2023-10-07 01:07:12,891 INFO mapreduce.Job:  map 0% reduce 0%
2023-10-07 01:07:19,981 INFO mapreduce.Job:  map 1% reduce 0%
2023-10-07 01:07:21,999 INFO mapreduce.Job:  map 2% reduce 0%
2023-10-07 01:07:23,005 INFO mapreduce.Job:  map 4% reduce 0%
2023-10-07 01:07:24,017 INFO mapreduce.Job:  map 16% reduce 0%
2023-10-07 01:07:25,027 INFO mapreduce.Job:  map 33% reduce 0%
2023-10-07 01:07:26,038 INFO mapreduce.Job:  map 36% reduce 0%
2023-10-07 01:07:27,049 INFO mapreduce.Job:  map 37% reduce 0%
2023-10-07 01:07:29,071 INFO mapreduce.Job:  map 38% reduce 0%
2023-10-07 01:07:31,101 INFO mapreduce.Job:  map 40% reduce 0%
2023-10-07 01:07:33,115 INFO mapreduce.Job:  map 45% reduce 0%
2023-10-07 01:07:34,121 INFO mapreduce.Job:  map 58% reduce 0%
2023-10-07 01:07:35,127 INFO mapreduce.Job:  map 62% reduce 0%
2023-10-07 01:07:36,136 INFO mapreduce.Job:  map 67% reduce 0%
2023-10-07 01:07:37,142 INFO mapreduce.Job:  map 68% reduce 0%
2023-10-07 01:07:39,166 INFO mapreduce.Job:  map 70% reduce 0%
2023-10-07 01:07:40,173 INFO mapreduce.Job:  map 71% reduce 0%
2023-10-07 01:07:41,186 INFO mapreduce.Job:  map 72% reduce 0%
2023-10-07 01:07:42,193 INFO mapreduce.Job:  map 79% reduce 0%
2023-10-07 01:07:43,199 INFO mapreduce.Job:  map 87% reduce 0%
2023-10-07 01:07:44,210 INFO mapreduce.Job:  map 93% reduce 0%
2023-10-07 01:07:45,216 INFO mapreduce.Job:  map 97% reduce 0%
2023-10-07 01:07:46,222 INFO mapreduce.Job:  map 99% reduce 0%
2023-10-07 01:07:48,232 INFO mapreduce.Job:  map 100% reduce 0%
2023-10-07 01:08:06,351 INFO mapreduce.Job:  map 100% reduce 63%
2023-10-07 01:08:12,384 INFO mapreduce.Job:  map 100% reduce 69%
2023-10-07 01:08:18,419 INFO mapreduce.Job:  map 100% reduce 74%
2023-10-07 01:08:24,450 INFO mapreduce.Job:  map 100% reduce 80%
2023-10-07 01:08:30,476 INFO mapreduce.Job:  map 100% reduce 91%
```



```
                Launched reduce tasks=1
                Data-local map tasks=6
                Rack-local map tasks=135
                Total time spent by all maps in occupied slots (ms)=4333988
                Total time spent by all reduces in occupied slots (ms)=193396
                Total time spent by all map tasks (ms)=1083497
                Total time spent by all reduce tasks (ms)=48349
                Total vcore-milliseconds taken by all map tasks=1083497
                Total vcore-milliseconds taken by all reduce tasks=48349
                Total megabyte-milliseconds taken by all map tasks=4438003712
                Total megabyte-milliseconds taken by all reduce tasks=198037504
        Map-Reduce Framework
                Map input records=2027564
                Map output records=17998518
                Map output bytes=142958759
                Map output materialized bytes=178956641
                Input split bytes=14523
                Combine input records=0
                Combine output records=0
                Reduce input groups=331
                Reduce shuffle bytes=178956641
                Reduce input records=17998518
                Reduce output records=332
                Spilled Records=35997036
                Shuffled Maps =141
                Failed Shuffles=0
                Merged Map outputs=141
                GC time elapsed (ms)=24786
                CPU time spent (ms)=518380
                Physical memory (bytes) snapshot=103458713600
                Virtual memory (bytes) snapshot=683017109504
                Total committed heap usage (bytes)=103142653952
                Peak Map Physical memory (bytes)=775618560
                Peak Map Virtual memory (bytes)=4896366592
                Peak Reduce Physical memory (bytes)=551170048
                Peak Reduce Virtual memory (bytes)=4848734208
        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=109562728
        File Output Format Counters
                Bytes Written=4407
2023-10-07 01:08:37,623 INFO streaming.StreamJob: Output directory: outputpython2
```

Get and download the files:



```
ys5250_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -get outputpython1
ys5250_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -get outputpython2
ys5250_nyu_edu@nyu-dataproc-m:~$ ls outputpython1
_SUCCESS  part-00000
ys5250_nyu_edu@nyu-dataproc-m:~$ ls outputpython2
_SUCCESS  part-00000
ys5250_nyu_edu@nyu-dataproc-m:~$ hadoop fs -copyToLocal outputpython1/part-00000 wordCountMaxWord1.txt
ys5250_nyu_edu@nyu-dataproc-m:~$ hadoop fs -copyToLocal outputpython2/part-00000 wordCountMaxWord2.txt
ys5250_nyu_edu@nyu-dataproc-m:~$ ls
'!'   mapper.py   outputpython1   outputpython2   reducer.py   wordCountMaxWord1.txt   wordCountMaxWord2.txt  '~'
ys5250_nyu_edu@nyu-dataproc-m:~$ 
```

1. Show the output for both the files for the above modifications by attaching the output text file.

Answer) The output for the files are in the same folder as this document

For text1.txt : wordCountMaxWord1.txt
For text2.txt : wordCountMaxWord2.txt

2. The Mapper and Reducer code of language of your choice

Answer) The mapper and reducer are in the same folder as mapper.py and reducer.py

3. What difference do you notice in the output directory after running the map reduce job?

Answer) The difference in the output directories (outputpython1 and outputpython2) have only one part unlike multiple parts that were concatenated in the previous exercise.

4. Why do you think there is a difference in the output despite the input files being of similar size?

Answer) The reason for the variance in the output between the two files primarily lies in the differences within their content. Although the files may have similar sizes, disparities in word usage and frequencies can lead to varying word count results. Several factors contribute to these discrepancies:

Firstly, the presence of unique words in one file, not found in the other, significantly influences the word count. These distinctive terms contribute to the divergence in the final counts.

Secondly, the distribution of word frequencies can differ substantially. One file may contain a few words that appear very frequently, whereas the other might exhibit a more even distribution of word frequencies. This variation directly impacts the overall word count figures.

Lastly, the order of words within the files can also play a pivotal role. If the sequencing of words varies significantly between the two files, it can lead to disparities in word counts. As a graduate student, understanding these factors provides valuable insights into the intricacies of text analysis and the potential sources of variation in word count results.