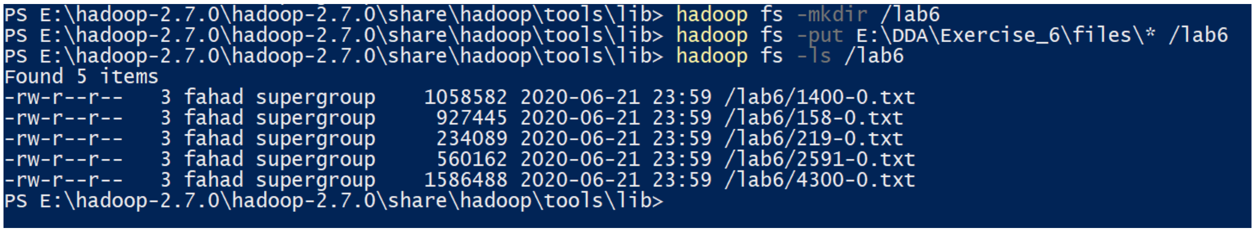
**Fahad Fiaz – (303141) – G2**

## System Info:

| | |
|---|---|
| Processor | i7-5500U , 2.40GHz |
| Cores | 4 |
| Operating system | Windows 64 Bit |
| Ram | 8GB |
| Programming Language | Python 3.7.7 |

# Q1:

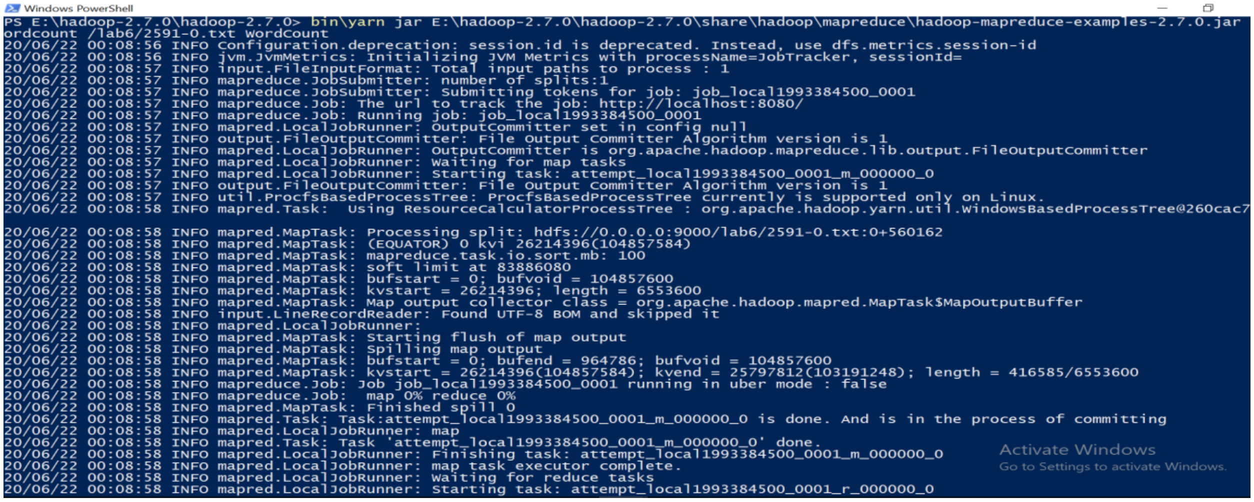
Making directory in HDFS and copied all the files I will use in exercise 1, 2, 3



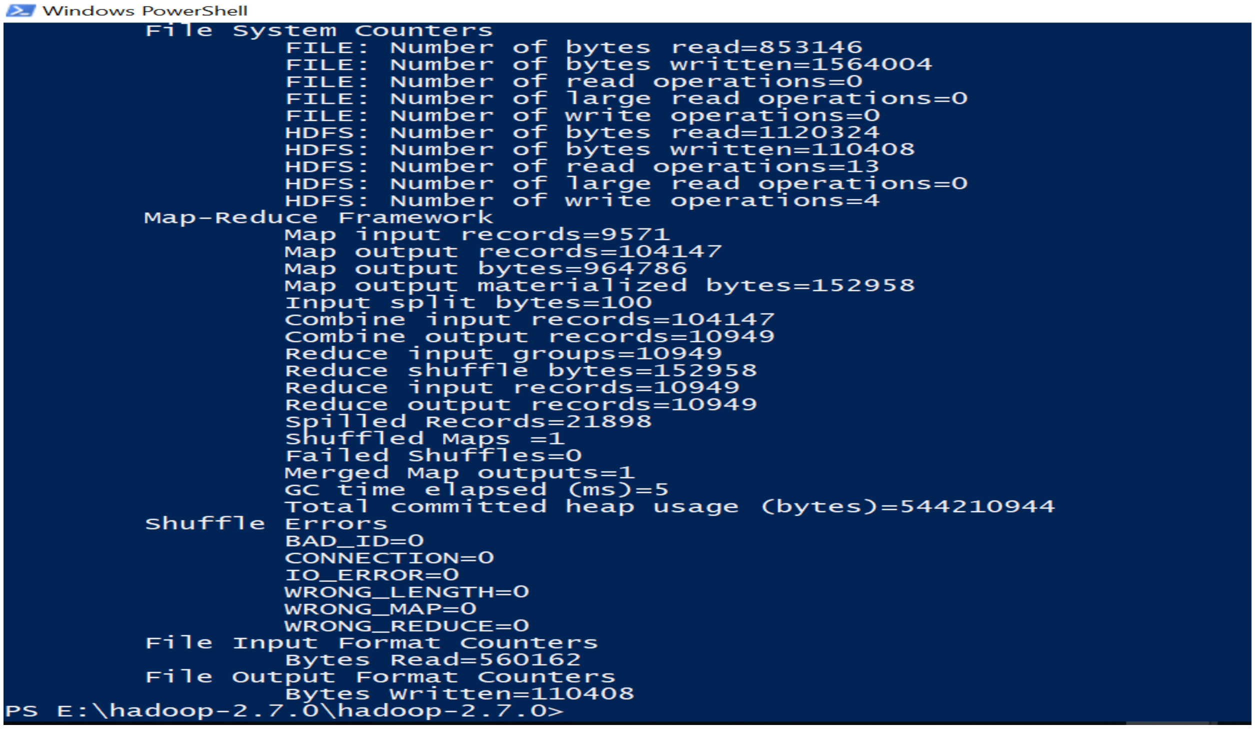
```
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop fs -mkdir /lab6
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop fs -put E:\DDA\Exercise_6\files\* /lab6
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop fs -ls /lab6
Found 5 items
-rw-r--r--   3 fahad supergroup    1058582 2020-06-21 23:59 /lab6/1400-0.txt
-rw-r--r--   3 fahad supergroup     927445 2020-06-21 23:59 /lab6/158-0.txt
-rw-r--r--   3 fahad supergroup     234089 2020-06-21 23:59 /lab6/219-0.txt
-rw-r--r--   3 fahad supergroup     560162 2020-06-21 23:59 /lab6/2591-0.txt
-rw-r--r--   3 fahad supergroup    1586488 2020-06-21 23:59 /lab6/4300-0.txt
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib>
```

Running the prebuild World count program to count occurrence of words.

I have not attached full output because output was very lengthy instead I have taken a screenshot of running the command and end of output.

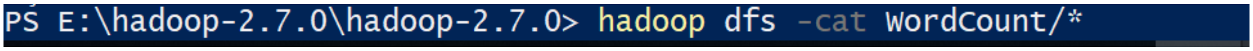


```
PS E:\hadoop-2.7.0\hadoop-2.7.0> bin\yarn jar E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\mapreduce\hadoop-mapreduce-examples-2.7.0.jar
ordcount /lab6/2591-0.txt WordCount
20/06/22 00:08:56 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
20/06/22 00:08:56 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
20/06/22 00:08:57 INFO input.FileInputFormat: Total input paths to process : 1
20/06/22 00:08:57 INFO mapreduce.JobSubmitter: number of splits:1
20/06/22 00:08:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1993384500_0001
20/06/22 00:08:57 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
20/06/22 00:08:57 INFO mapreduce.Job: Running job: job_local1993384500_0001
20/06/22 00:08:57 INFO mapred.LocalJobRunner: OutputCommitter set in config null
20/06/22 00:08:57 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/06/22 00:08:57 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
20/06/22 00:08:57 INFO mapred.LocalJobRunner: Waiting for map tasks
20/06/22 00:08:57 INFO mapred.LocalJobRunner: Starting task: attempt_local1993384500_0001_m_000000_0
20/06/22 00:08:57 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/06/22 00:08:57 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
20/06/22 00:08:58 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTree@260cac7

20/06/22 00:08:58 INFO mapred.MapTask: Processing split: hdfs://0.0.0.0:9000/lab6/2591-0.txt:0+560162
20/06/22 00:08:58 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
20/06/22 00:08:58 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
20/06/22 00:08:58 INFO mapred.MapTask: soft limit at 83886080
20/06/22 00:08:58 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
20/06/22 00:08:58 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
20/06/22 00:08:58 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
20/06/22 00:08:58 INFO input.LineRecordReader: Found UTF-8 BOM and skipped it
20/06/22 00:08:58 INFO mapred.LocalJobRunner:
20/06/22 00:08:58 INFO mapred.MapTask: Starting flush of map output
20/06/22 00:08:58 INFO mapred.MapTask: Spilling map output
20/06/22 00:08:58 INFO mapred.MapTask: bufstart = 0; bufend = 964786; bufvoid = 104857600
20/06/22 00:08:58 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 25797812(103191248); length = 416585/6553600
20/06/22 00:08:58 INFO mapreduce.Job: Job job_local1993384500_0001 running in uber mode : false
20/06/22 00:08:58 INFO mapreduce.Job:  map 0% reduce 0%
20/06/22 00:08:58 INFO mapred.MapTask: Finished spill 0
20/06/22 00:08:58 INFO mapred.Task: Task:attempt_local1993384500_0001_m_000000_0 is done. And is in the process of committing
20/06/22 00:08:58 INFO mapred.LocalJobRunner: map
20/06/22 00:08:58 INFO mapred.Task: Task 'attempt_local1993384500_0001_m_000000_0' done.
20/06/22 00:08:58 INFO mapred.LocalJobRunner: Finishing task: attempt_local1993384500_0001_m_000000_0
20/06/22 00:08:58 INFO mapred.LocalJobRunner: map task executor complete.
20/06/22 00:08:58 INFO mapred.LocalJobRunner: Waiting for reduce tasks
20/06/22 00:08:58 INFO mapred.LocalJobRunner: Starting task: attempt_local1993384500_0001_r_000000_0
```

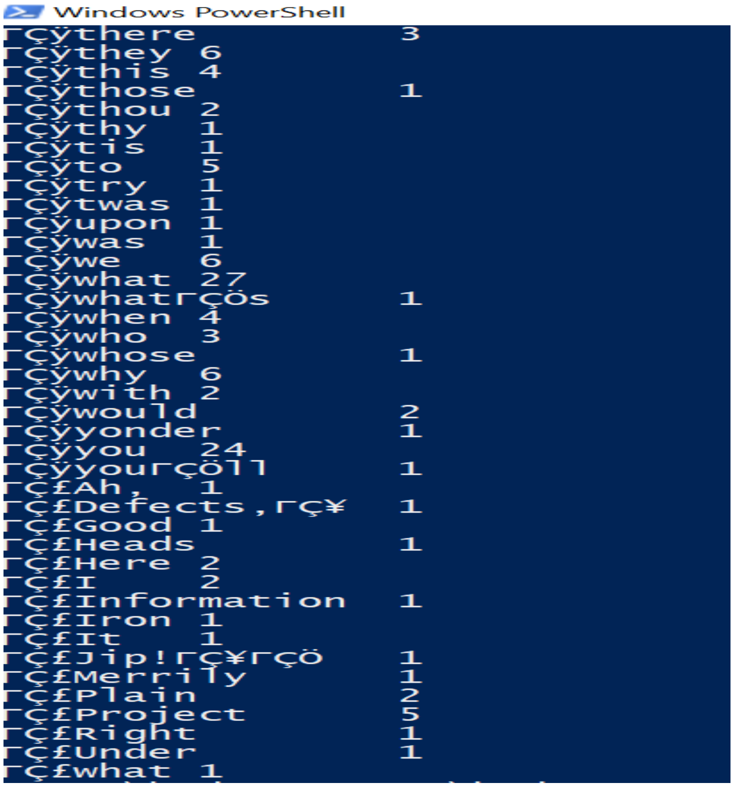


```
        File System Counters
                FILE: Number of bytes read=853146
                FILE: Number of bytes written=1564004
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1120324
                HDFS: Number of bytes written=110408
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=9571
                Map output records=104147
                Map output bytes=964786
                Map output materialized bytes=152958
                Input split bytes=100
                Combine input records=104147
                Combine output records=10949
                Reduce input groups=10949
                Reduce shuffle bytes=152958
                Reduce input records=10949
                Reduce output records=10949
                Spilled Records=21898
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=5
                Total committed heap usage (bytes)=544210944
        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=560162
        File Output Format Counters
                Bytes Written=110408
PS E:\hadoop-2.7.0\hadoop-2.7.0>
```

Printing output of prebuild word count program

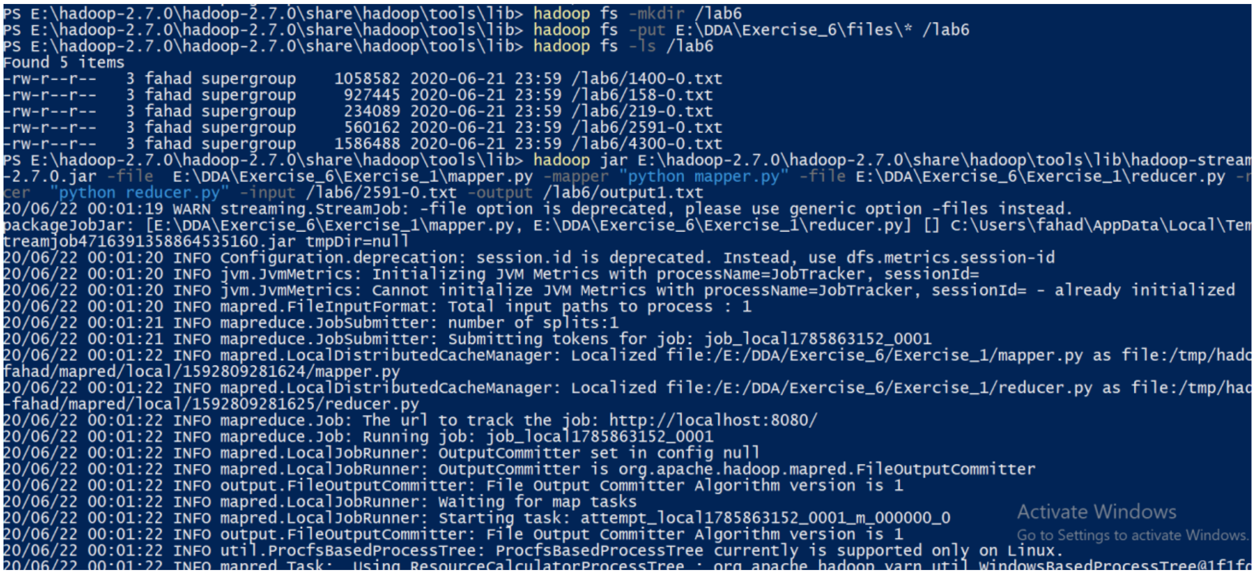


```
PS E:\hadoop-2.7.0\hadoop-2.7.0> hadoop dfs -cat WordCount/*
```

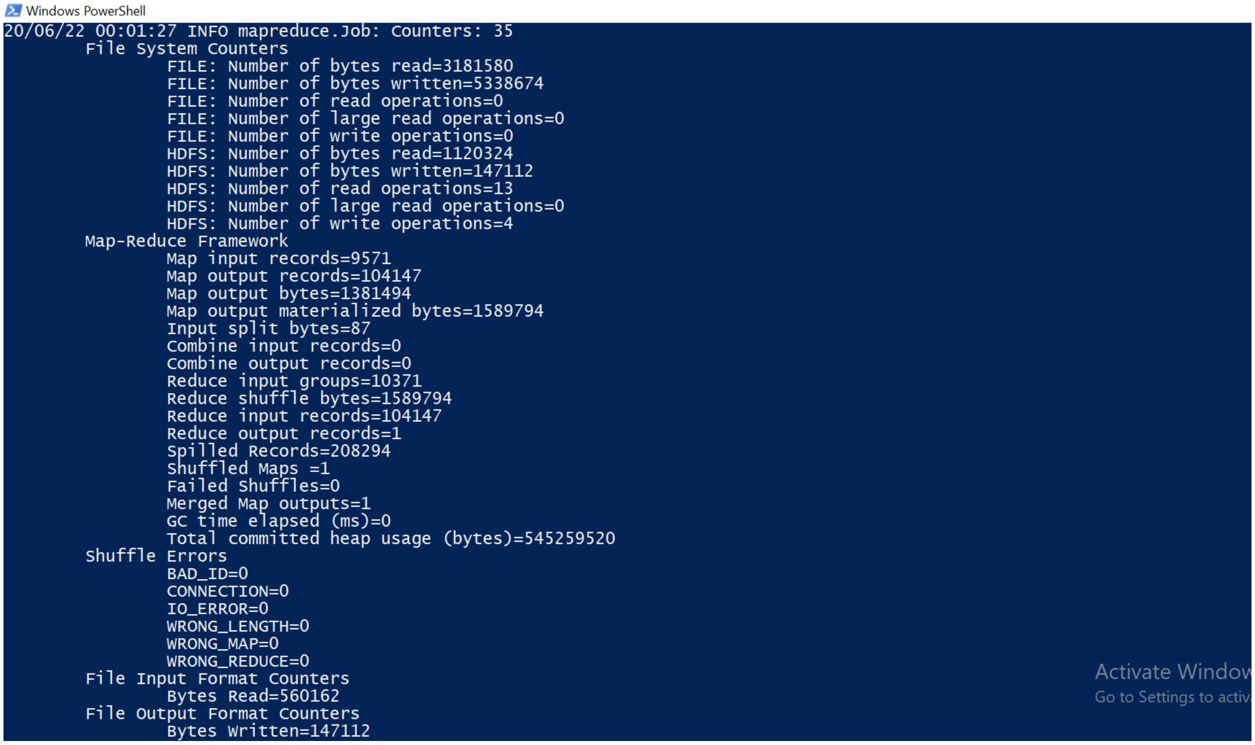


```
ΓÇÿthere         3
ΓÇÿthey  6
ΓÇÿthis  4
ΓÇÿthose         1
ΓÇÿthou  2
ΓÇÿthy   1
ΓÇÿtis   1
ΓÇÿto    5
ΓÇÿtry   1
ΓÇÿtwas  1
ΓÇÿupon  1
ΓÇÿwas   1
ΓÇÿwe    6
ΓÇÿwhat  27
ΓÇÿwhatΓÇÖs        1
ΓÇÿwhen  4
ΓÇÿwho   3
ΓÇÿwhose         1
ΓÇÿwhy   6
ΓÇÿwith  2
ΓÇÿwould         2
ΓÇÿyonder        1
ΓÇÿyou   24
ΓÇÿyouΓçöll       1
ΓÇ£Ah,   1
ΓÇ£Defects,ΓÇ¥    1
ΓÇ£Good  1
ΓÇ£Heads         1
ΓÇ£Here  2
ΓÇ£I     2
ΓÇ£Information   1
ΓÇ£Iron  1
ΓÇ£It    1
ΓÇ£Jip!ΓÇ¥Γçö      1
ΓÇ£Merrily       1
ΓÇ£Plain         2
ΓÇ£Project       5
ΓÇ£Right         1
ΓÇ£Under         1
ΓÇ£what  1
```

## Running my word count program and printing its output

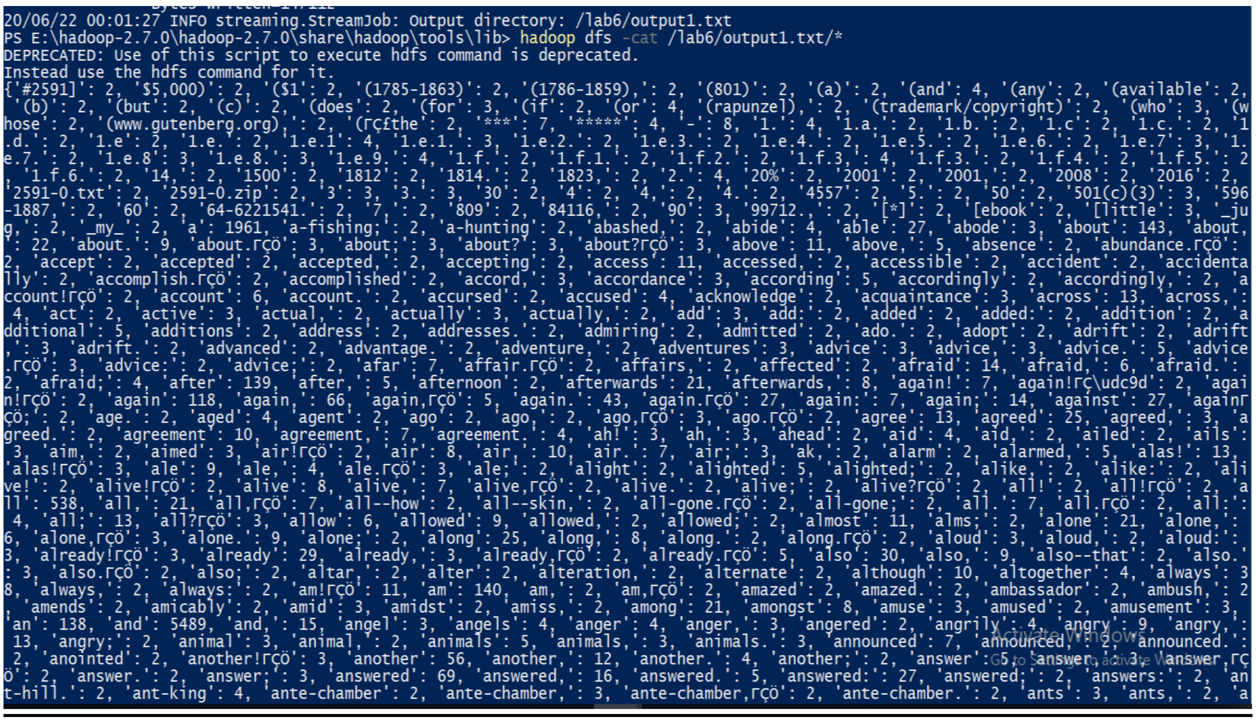


```
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop fs -mkdir /lab6
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop fs -put E:\DDA\Exercise_6\files\* /lab6
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop fs -ls /lab6
Found 5 items
-rw-r--r--   3 fahad supergroup    1058582 2020-06-21 23:59 /lab6/1400-0.txt
-rw-r--r--   3 fahad supergroup     927445 2020-06-21 23:59 /lab6/158-0.txt
-rw-r--r--   3 fahad supergroup     234089 2020-06-21 23:59 /lab6/219-0.txt
-rw-r--r--   3 fahad supergroup     560162 2020-06-21 23:59 /lab6/2591-0.txt
-rw-r--r--   3 fahad supergroup    1586488 2020-06-21 23:59 /lab6/4300-0.txt
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop jar E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib\hadoop-stream
-2.7.0.jar -file  E:\DDA\Exercise_6\Exercise_1\mapper.py -mapper "python mapper.py" -file E:\DDA\Exercise_6\Exercise_1\reducer.py -r
cer  "python reducer.py" -input /lab6/2591-0.txt -output /lab6/output1.txt
20/06/22 00:01:19 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [E:\DDA\Exercise_6\Exercise_1\mapper.py, E:\DDA\Exercise_6\Exercise_1\reducer.py] [] C:\Users\fahad\AppData\Local\Tem
treamjob4716391358864535160.jar tmpDir=null
20/06/22 00:01:20 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
20/06/22 00:01:20 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
20/06/22 00:01:20 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
20/06/22 00:01:20 INFO mapred.FileInputFormat: Total input paths to process : 1
20/06/22 00:01:21 INFO mapreduce.JobSubmitter: number of splits:1
20/06/22 00:01:21 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1785863152_0001
20/06/22 00:01:22 INFO mapred.LocalDistributedCacheManager: Localized file:/E:/DDA/Exercise_6/Exercise_1/mapper.py as file:/tmp/hadc
fahad/mapred/local/1592809281624/mapper.py
20/06/22 00:01:22 INFO mapred.LocalDistributedCacheManager: Localized file:/E:/DDA/Exercise_6/Exercise_1/reducer.py as file:/tmp/hac
-fahad/mapred/local/1592809281625/reducer.py
20/06/22 00:01:22 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
20/06/22 00:01:22 INFO mapreduce.Job: Running job: job_local1785863152_0001
20/06/22 00:01:22 INFO mapred.LocalJobRunner: OutputCommitter set in config null
20/06/22 00:01:22 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
20/06/22 00:01:22 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/06/22 00:01:22 INFO mapred.LocalJobRunner: Waiting for map tasks
20/06/22 00:01:22 INFO mapred.LocalJobRunner: Starting task: attempt_local1785863152_0001_m_000000_0
20/06/22 00:01:22 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/06/22 00:01:22 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
20/06/22 00:01:22 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTree@1f1fc
```

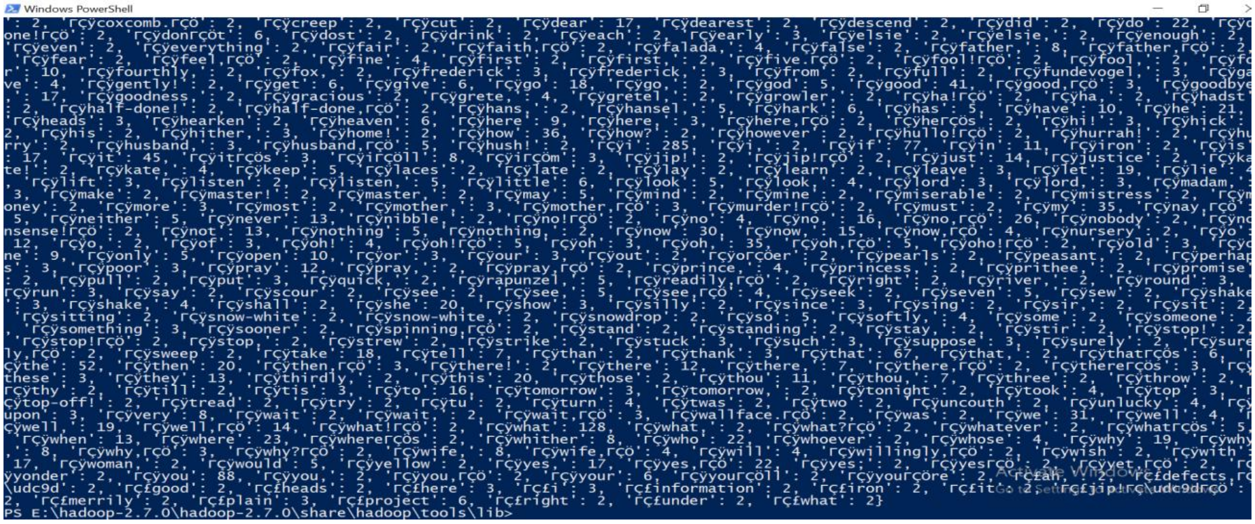


Windows PowerShell

```
20/06/22 00:01:27 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=3181580
                FILE: Number of bytes written=5338674
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1120324
                HDFS: Number of bytes written=147112
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=9571
                Map output records=104147
                Map output bytes=1381494
                Map output materialized bytes=1589794
                Input split bytes=87
                Combine input records=0
                Combine output records=0
                Reduce input groups=10371
                Reduce shuffle bytes=1589794
                Reduce input records=104147
                Reduce output records=1
                Spilled Records=208294
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=545259520
        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=560162
        File Output Format Counters
                Bytes Written=147112
```



```
20/06/22 00:01:27 INFO streaming.StreamJob: Output directory: /lab6/output1.txt
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop dfs -cat /lab6/output1.txt/*
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
{'#2591]': 2, '$5,000)': 2, '($1': 2, '(1785-1863)': 2, '(1786-1859),': 2, '(801)': 2, '(a)': 2, '(and': 4, '(any': 2, '(available': 2,
 '(b)': 2, '(but': 2, '(c)': 2, '(does': 2, '(for': 3, '(if': 2, '(or': 4, '(rapunzel),': 2, '(trademark/copyright)': 2, '(who': 3, '(w
hose': 2, '(www.gutenberg.org),': 2, '(ΓÇ£the': 2, '***': 7, '*****': 4, '-': 8, '1.': 4, '1.a.': 2, '1.b.': 2, '1.c': 2, '1.c.': 2, '1
.d.': 2, '1.e': 2, '1.e.': 2, '1.e.1': 4, '1.e.1.': 3, '1.e.2.': 2, '1.e.3.': 2, '1.e.4.': 2, '1.e.5.': 2, '1.e.6.': 2, '1.e.7': 3, '1.
e.7.': 2, '1.e.8': 3, '1.e.8.': 3, '1.e.9.': 4, '1.f.': 2, '1.f.1.': 2, '1.f.2.': 2, '1.f.3,': 4, '1.f.3.': 2, '1.f.4.': 2, '1.f.5.': 2
, '1.f.6.': 2, '14,': 2, '1500': 2, '1812': 2, '1814.': 2, '1823,': 2, '2.': 4, '20%': 2, '2001': 2, '2001,': 2, '2008': 2, '2016': 2,
'2591-0.txt': 2, '2591-0.zip': 2, '3': 3, '3.': 3, '30': 2, '4': 2, '4,': 2, '4.': 2, '4557': 2, '5.': 2, '50': 2, '501(c)(3)': 3, '596
-1887,': 2, '60': 2, '64-6221541.': 2, '7,': 2, '809': 2, '84116,': 2, '90': 3, '99712.,': 2, '[*]': 2, '[ebook': 2, '[little': 3, '_ju
g,': 2, '_my_': 2, 'a': 1961, 'a-fishing;': 2, 'a-hunting': 2, 'abashed,': 2, 'abide': 4, 'able': 27, 'abode': 3, 'about': 143, 'about,
': 22, 'about.': 9, 'about.ΓÇÖ': 3, 'about;': 3, 'about?': 3, 'about?ΓÇÖ': 3, 'above': 11, 'above,': 5, 'absence': 2, 'abundance.ΓÇÖ':
2, 'accept': 2, 'accepted': 2, 'accepted,': 2, 'accepting': 2, 'access': 11, 'accessed,': 2, 'accessible': 2, 'accident': 2, 'accidenta
lly': 2, 'accomplish.ΓÇÖ': 2, 'accomplished': 2, 'accord,': 3, 'accordance': 3, 'according': 5, 'accordingly': 2, 'accordingly,': 2, 'a
ccount!ΓÇÖ': 2, 'account': 6, 'account.': 2, 'accursed': 2, 'accused': 4, 'acknowledge': 2, 'acquaintance': 3, 'across': 13, 'across,':
 4, 'act': 2, 'active': 3, 'actual,': 2, 'actually': 3, 'actually,': 2, 'add': 3, 'add:': 2, 'added': 2, 'added:': 2, 'addition': 2, 'a
dditional': 5, 'additions': 2, 'address': 2, 'addresses.': 2, 'admiring': 2, 'admitted': 2, 'ado.': 2, 'adopt': 2, 'adrift': 2, 'adrift
,': 3, 'adrift.': 2, 'advanced': 2, 'advantage.': 2, 'adventure,': 2, 'adventures': 3, 'advice': 3, 'advice,': 3, 'advice.': 5, 'advice
.ΓÇÖ': 3, 'advice:': 2, 'advice;': 2, 'afar': 7, 'affair.ΓÇÖ': 2, 'affairs,': 2, 'affected': 2, 'afraid': 14, 'afraid,': 6, 'afraid.':
2, 'afraid;': 4, 'after': 139, 'after,': 5, 'afternoon': 2, 'afterwards': 21, 'afterwards,': 8, 'again!': 7, 'again!ΓÇ\udc9d': 2, 'agai
n!ΓÇÖ': 2, 'again': 118, 'again,': 66, 'again,ΓÇÖ': 5, 'again.': 43, 'again.ΓÇÖ': 27, 'again:': 7, 'again;': 14, 'against': 27, 'againΓ
ÇÖ;': 2, 'age.': 2, 'aged': 4, 'agent': 2, 'ago': 2, 'ago,': 2, 'ago,ΓÇÖ': 3, 'ago.ΓÇÖ': 2, 'agree': 13, 'agreed': 25, 'agreed,': 3, 'a
greed.': 2, 'agreement': 10, 'agreement,': 7, 'agreement.': 4, 'ah!': 3, 'ah,': 3, 'ahead': 2, 'aid': 4, 'aid,': 2, 'ailed': 2, 'ails':
 3, 'aim,': 2, 'aimed': 3, 'air!ΓÇÖ': 2, 'air': 8, 'air,': 10, 'air.': 7, 'air;': 3, 'ak,': 2, 'alarm': 2, 'alarmed,': 5, 'alas!': 13,
'alas!ΓÇÖ': 3, 'ale': 9, 'ale,': 4, 'ale.ΓÇÖ': 3, 'ale;': 2, 'alight': 2, 'alighted': 5, 'alighted;': 2, 'alike,': 2, 'alike:': 2, 'ali
ve!': 2, 'alive!ΓÇÖ': 2, 'alive': 8, 'alive,': 7, 'alive,ΓÇÖ': 2, 'alive.': 2, 'alive;': 2, 'alive?ΓÇÖ': 2, 'all!': 2, 'all!ΓÇÖ': 2, 'a
ll': 538, 'all,': 21, 'all,ΓÇÖ': 7, 'all--how': 2, 'all--skin,': 2, 'all-gone.ΓÇÖ': 2, 'all-gone;': 2, 'all.': 7, 'all.ΓÇÖ': 2, 'all:':
 4, 'all;': 13, 'all?ΓÇÖ': 3, 'allow': 6, 'allowed': 9, 'allowed,': 2, 'allowed;': 2, 'almost': 11, 'alms;': 2, 'alone': 21, 'alone,':
6, 'alone,ΓÇÖ': 3, 'alone.': 9, 'alone;': 2, 'along': 25, 'along,': 8, 'along.': 2, 'along.ΓÇÖ': 2, 'aloud': 3, 'aloud,': 2, 'aloud:':
3, 'already!ΓÇÖ': 3, 'already': 29, 'already,': 3, 'already,ΓÇÖ': 2, 'already.ΓÇÖ': 5, 'also': 30, 'also,': 9, 'also--that': 2, 'also.'
: 3, 'also.ΓÇÖ': 2, 'also;': 2, 'altar,': 2, 'alter': 2, 'alteration,': 2, 'alternate': 2, 'although': 10, 'altogether': 4, 'always': 3
8, 'always,': 2, 'always:': 2, 'am!ΓÇÖ': 11, 'am': 140, 'am,': 2, 'am,ΓÇÖ': 2, 'amazed': 2, 'amazed.': 2, 'ambassador': 2, 'ambush,': 2
, 'amends': 2, 'amicably': 2, 'amid': 3, 'amidst': 2, 'amiss,': 2, 'among': 21, 'amongst': 8, 'amuse': 3, 'amused': 2, 'amusement': 3,
'an': 138, 'and': 5489, 'and,': 15, 'angel': 3, 'angels': 4, 'anger': 4, 'anger,': 3, 'angered': 2, 'angrily': 4, 'angry': 9, 'angry,':
 13, 'angry;': 2, 'animal': 3, 'animal,': 2, 'animals': 5, 'animals,': 3, 'animals.': 3, 'announced': 7, 'announced,': 2, 'announced.':
 2, 'anointed': 2, 'another!ΓÇÖ': 3, 'another': 56, 'another,': 12, 'another.': 4, 'another;': 2, 'answer': 5, 'answer,': 3, 'answer,ΓÇ
Ö': 2, 'answer.': 2, 'answer;': 3, 'answered': 69, 'answered,': 16, 'answered.': 5, 'answered:': 27, 'answered;': 2, 'answers:': 2, 'an
t-hill.': 2, 'ant-king': 4, 'ante-chamber': 2, 'ante-chamber,': 3, 'ante-chamber,ΓÇÖ': 2, 'ante-chamber.': 2, 'ants': 3, 'ants,': 2, 'a
```



```
': 2, 'ΓÇÿcoxcomb.ΓÇÖ': 2, 'ΓÇÿcreep': 2, 'ΓÇÿcut': 2, 'ΓÇÿdear': 17, 'ΓÇÿdearest': 2, 'ΓÇÿdescend': 2, 'ΓÇÿdid': 2, 'ΓÇÿdo': 22, 'ΓÇÿ
one!ΓÇÖ': 2, 'ΓÇÿdonΓÇÖt': 6, 'ΓÇÿdost': 2, 'ΓÇÿdrink': 2, 'ΓÇÿeach': 2, 'ΓÇÿearly': 3, 'ΓÇÿelsie': 2, 'ΓÇÿelsie,': 2, 'ΓÇÿenough': 2,
'ΓÇÿeven': 2, 'ΓÇÿeverything': 2, 'ΓÇÿfair': 2, 'ΓÇÿfaith,ΓÇÖ': 2, 'ΓÇÿfalada,': 4, 'ΓÇÿfalse': 2, 'ΓÇÿfather,': 8, 'ΓÇÿfather,ΓÇÖ': 2,
 'ΓÇÿfear': 2, 'ΓÇÿfeel,ΓÇÖ': 2, 'ΓÇÿfine': 4, 'ΓÇÿfirst': 2, 'ΓÇÿfirst,': 2, 'ΓÇÿfive.ΓÇÖ': 2, 'ΓÇÿfool!ΓÇÖ': 2, 'ΓÇÿfool,': 2, 'ΓÇÿfo
r': 10, 'ΓÇÿfourthly,': 2, 'ΓÇÿfox,': 2, 'ΓÇÿfrederick': 3, 'ΓÇÿfrederick,': 3, 'ΓÇÿfrom': 2, 'ΓÇÿfull': 2, 'ΓÇÿfundevogel,': 3, 'ΓÇÿga
ve': 4, 'ΓÇÿgently!': 2, 'ΓÇÿget': 6, 'ΓÇÿgive': 6, 'ΓÇÿgo': 18, 'ΓÇÿgo,': 2, 'ΓÇÿgod': 5, 'ΓÇÿgood': 41, 'ΓÇÿgood,ΓÇÖ': 3, 'ΓÇÿgoodbye
,': 17, 'ΓÇÿgoodness,': 2, 'ΓÇÿgracious': 2, 'ΓÇÿgrete,': 4, 'ΓÇÿgretel,': 2, 'ΓÇÿgrowler,': 2, 'ΓÇÿha!ΓÇÖ': 2, 'ΓÇÿha,': 2, 'ΓÇÿhadst'
: 2, 'ΓÇÿhalf-done!': 2, 'ΓÇÿhalf-done,ΓÇÖ': 2, 'ΓÇÿhans,': 2, 'ΓÇÿhansel,': 5, 'ΓÇÿhark': 6, 'ΓÇÿhas': 5, 'ΓÇÿhave': 10, 'ΓÇÿhe': 21,
'ΓÇÿheads': 3, 'ΓÇÿhearken': 2, 'ΓÇÿheaven': 6, 'ΓÇÿhere': 9, 'ΓÇÿhere,': 3, 'ΓÇÿhere,ΓÇÖ': 2, 'ΓÇÿheΓÇÖs': 2, 'ΓÇÿhi!': 3, 'ΓÇÿhick':
2, 'ΓÇÿhis': 2, 'ΓÇÿhither,': 3, 'ΓÇÿhome!': 2, 'ΓÇÿhow': 36, 'ΓÇÿhow?': 2, 'ΓÇÿhowever': 2, 'ΓÇÿhullo!ΓÇÖ': 2, 'ΓÇÿhurrah!': 2, 'ΓÇÿhu
rry': 2, 'ΓÇÿhusband,': 3, 'ΓÇÿhusband,ΓÇÖ': 5, 'ΓÇÿhush!': 2, 'ΓÇÿi': 285, 'ΓÇÿi,': 2, 'ΓÇÿif': 77, 'ΓÇÿin': 11, 'ΓÇÿiron': 2, 'ΓÇÿis'
: 17, 'ΓÇÿit': 45, 'ΓÇÿitΓÇÖs': 3, 'ΓÇÿiΓÇÖll': 8, 'ΓÇÿiΓÇÖm': 3, 'ΓÇÿjip!': 2, 'ΓÇÿjip!ΓÇÖ': 2, 'ΓÇÿjust': 14, 'ΓÇÿjustice': 2, 'ΓÇÿka
te!': 2, 'ΓÇÿkate,': 4, 'ΓÇÿkeep': 5, 'ΓÇÿlaces': 2, 'ΓÇÿlate': 2, 'ΓÇÿlay': 2, 'ΓÇÿlearn': 2, 'ΓÇÿleave': 3, 'ΓÇÿlet': 19, 'ΓÇÿlie': 4
, 'ΓÇÿlift': 3, 'ΓÇÿlisten': 2, 'ΓÇÿlisten,': 5, 'ΓÇÿlittle': 6, 'ΓÇÿlook': 5, 'ΓÇÿlook,': 4, 'ΓÇÿlord': 3, 'ΓÇÿlord,': 3, 'ΓÇÿmadam,':
 3, 'ΓÇÿmake': 2, 'ΓÇÿmaster!': 2, 'ΓÇÿmaster,': 2, 'ΓÇÿmay': 5, 'ΓÇÿmind': 2, 'ΓÇÿmine': 2, 'ΓÇÿmiserable': 2, 'ΓÇÿmistress': 2, 'ΓÇÿm
oney': 2, 'ΓÇÿmore': 3, 'ΓÇÿmost': 2, 'ΓÇÿmother,': 3, 'ΓÇÿmother,ΓÇÖ': 3, 'ΓÇÿmurder!ΓÇÖ': 2, 'ΓÇÿmust': 2, 'ΓÇÿmy': 35, 'ΓÇÿnay,ΓÇÖ':
 5, 'ΓÇÿneither': 5, 'ΓÇÿnever': 13, 'ΓÇÿnibble,': 2, 'ΓÇÿno!ΓÇÖ': 2, 'ΓÇÿno': 4, 'ΓÇÿno,': 16, 'ΓÇÿno,ΓÇÖ': 26, 'ΓÇÿnobody': 2, 'ΓÇÿno
nsense!ΓÇÖ': 2, 'ΓÇÿnot': 13, 'ΓÇÿnothing': 5, 'ΓÇÿnothing,': 2, 'ΓÇÿnow': 30, 'ΓÇÿnow,': 15, 'ΓÇÿnow,ΓÇÖ': 4, 'ΓÇÿnursery': 2, 'ΓÇÿo':
 12, 'ΓÇÿo,': 2, 'ΓÇÿof': 3, 'ΓÇÿoh!': 4, 'ΓÇÿoh!ΓÇÖ': 5, 'ΓÇÿoh': 3, 'ΓÇÿoh,': 35, 'ΓÇÿoh,ΓÇÖ': 5, 'ΓÇÿoho!ΓÇÖ': 2, 'ΓÇÿold': 3, 'ΓÇÿo
ne': 9, 'ΓÇÿonly': 5, 'ΓÇÿopen': 10, 'ΓÇÿor': 3, 'ΓÇÿour': 3, 'ΓÇÿout': 2, 'ΓÇÿoΓÇÖer': 2, 'ΓÇÿpearls': 2, 'ΓÇÿpeasant,': 2, 'ΓÇÿperhap
s': 3, 'ΓÇÿpoor': 3, 'ΓÇÿpray': 12, 'ΓÇÿpray,': 2, 'ΓÇÿpray,ΓÇÖ': 2, 'ΓÇÿprince,': 4, 'ΓÇÿprincess,': 2, 'ΓÇÿprithee,': 2, 'ΓÇÿpromise'
: 2, 'ΓÇÿpull': 2, 'ΓÇÿput': 3, 'ΓÇÿquick,': 2, 'ΓÇÿrapunzel,': 5, 'ΓÇÿreadily,ΓÇÖ': 2, 'ΓÇÿright': 2, 'ΓÇÿriver,': 2, 'ΓÇÿround': 3, '
ΓÇÿrun': 3, 'ΓÇÿsay': 2, 'ΓÇÿscour': 2, 'ΓÇÿsee': 2, 'ΓÇÿsee,': 5, 'ΓÇÿsee,ΓÇÖ': 4, 'ΓÇÿseek': 2, 'ΓÇÿseven': 5, 'ΓÇÿsew': 2, 'ΓÇÿshake
': 3, 'ΓÇÿshake,': 4, 'ΓÇÿshall': 2, 'ΓÇÿshe': 20, 'ΓÇÿshow': 3, 'ΓÇÿsilly': 2, 'ΓÇÿsince': 3, 'ΓÇÿsing': 2, 'ΓÇÿsir,': 2, 'ΓÇÿsit': 2,
 'ΓÇÿsitting': 2, 'ΓÇÿsnow-white': 2, 'ΓÇÿsnow-white,': 2, 'ΓÇÿsnowdrop': 2, 'ΓÇÿso': 5, 'ΓÇÿsoftly,': 4, 'ΓÇÿsome': 2, 'ΓÇÿsomeone': 2
, 'ΓÇÿsomething': 3, 'ΓÇÿsooner': 2, 'ΓÇÿspinning,ΓÇÖ': 2, 'ΓÇÿstand': 2, 'ΓÇÿstanding': 2, 'ΓÇÿstay,': 2, 'ΓÇÿstir': 2, 'ΓÇÿstop!': 2,
 'ΓÇÿstop!ΓÇÖ': 2, 'ΓÇÿstop,': 2, 'ΓÇÿstrew': 2, 'ΓÇÿstrike': 2, 'ΓÇÿstuck': 3, 'ΓÇÿsuch': 3, 'ΓÇÿsuppose': 3, 'ΓÇÿsurely': 2, 'ΓÇÿsure
ly,ΓÇÖ': 2, 'ΓÇÿsweep': 2, 'ΓÇÿtake': 18, 'ΓÇÿtell': 7, 'ΓÇÿthan': 2, 'ΓÇÿthank': 3, 'ΓÇÿthat': 67, 'ΓÇÿthat,': 2, 'ΓÇÿthatΓÇÖs': 6, 'Γ
Çÿthe': 52, 'ΓÇÿthen': 20, 'ΓÇÿthen,ΓÇÖ': 3, 'ΓÇÿthere!': 2, 'ΓÇÿthere': 12, 'ΓÇÿthere,': 7, 'ΓÇÿthere,ΓÇÖ': 2, 'ΓÇÿthereΓÇÖs': 3, 'ΓÇ
these': 3, 'ΓÇÿthey': 13, 'ΓÇÿthirdly,': 2, 'ΓÇÿthis': 20, 'ΓÇÿthose': 2, 'ΓÇÿthou': 11, 'ΓÇÿthou,': 7, 'ΓÇÿthree': 2, 'ΓÇÿthrow': 2,
ΓÇÿthy': 2, 'ΓÇÿtill': 2, 'ΓÇÿtis': 3, 'ΓÇÿto': 16, 'ΓÇÿtomorrow': 3, 'ΓÇÿtomorrow,': 2, 'ΓÇÿtonight': 2, 'ΓÇÿtook': 4, 'ΓÇÿtop': 3, 'Γ
Çÿtop-off!': 2, 'ΓÇÿtread': 2, 'ΓÇÿtry': 2, 'ΓÇÿtu': 2, 'ΓÇÿturn': 4, 'ΓÇÿtwas': 2, 'ΓÇÿtwo': 2, 'ΓÇÿuncouth': 2, 'ΓÇÿunlucky': 4, 'ΓÇ
upon': 3, 'ΓÇÿvery': 8, 'ΓÇÿwait': 2, 'ΓÇÿwait,': 2, 'ΓÇÿwait,ΓÇÖ': 3, 'ΓÇÿwallface.ΓÇÖ': 2, 'ΓÇÿwas': 2, 'ΓÇÿwe': 31, 'ΓÇÿwell': 4, 'Γ
Çÿwell,': 19, 'ΓÇÿwell,ΓÇÖ': 14, 'ΓÇÿwhat!ΓÇÖ': 2, 'ΓÇÿwhat': 128, 'ΓÇÿwhat,': 2, 'ΓÇÿwhat?ΓÇÖ': 2, 'ΓÇÿwhatever': 2, 'ΓÇÿwhatΓÇÖs': 5,
 'ΓÇÿwhen': 13, 'ΓÇÿwhere': 23, 'ΓÇÿwhereΓÇÖs': 2, 'ΓÇÿwhither': 8, 'ΓÇÿwho': 22, 'ΓÇÿwhoever': 2, 'ΓÇÿwhose': 4, 'ΓÇÿwhy': 19, 'ΓÇÿwhy
,': 8, 'ΓÇÿwhy,ΓÇÖ': 3, 'ΓÇÿwhy?ΓÇÖ': 2, 'ΓÇÿwife,': 8, 'ΓÇÿwife,ΓÇÖ': 4, 'ΓÇÿwill': 4, 'ΓÇÿwillingly,ΓÇÖ': 2, 'ΓÇÿwish': 2, 'ΓÇÿwith':
 17, 'ΓÇÿwoman,': 2, 'ΓÇÿwould': 5, 'ΓÇÿyellow': 2, 'ΓÇÿyes,': 17, 'ΓÇÿyes,ΓÇÖ': 22, 'ΓÇÿyes;': 2, 'ΓÇÿyesΓÇÖ': 2, 'ΓÇÿyet,ΓÇÖ': 2, 'Γ
ÿyonder': 2, 'ΓÇÿyou': 88, 'ΓÇÿyou,': 2, 'ΓÇÿyou,ΓÇÖ': 2, 'ΓÇÿyour': 6, 'ΓÇÿyouΓÇÖll': 2, 'ΓÇÿyouΓÇÖre': 2, 'ΓÇ£ah': 2, 'ΓÇ£defects,Γ
\udc9d': 2, 'ΓÇ£good': 2, 'ΓÇ£heads': 2, 'ΓÇ£here': 3, 'ΓÇ£i': 3, 'ΓÇ£information': 2, 'ΓÇ£iron': 2, 'ΓÇ£it': 2, 'ΓÇ£jip!ΓÇ\udc9d':
2, 'ΓÇ£merrily': 2, 'ΓÇ£plain': 3, 'ΓÇ£project': 6, 'ΓÇ£right': 2, 'ΓÇ£under': 2, 'ΓÇ£what': 2}
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib>
```

## Mapper:

It will read data from STDIN, remove any leading or trailing whitespaces, split rows into words and output tuple containing word and its count to STDOUT

```
import sys

for line in sys.stdin:
    line = line.strip()
    words = line.split()  # split the line into words
    for word in words:
        word = (word, 1)
        print(word)
```

## Reducer:

It will read the results of mapper.py from STDIN (the output format of mapper.py and the expected input format of reducer.py must match).

Since the value received from STDIN is in string format so first we need to convert this string tuple to tuple.

```
dic = {}
for line in sys.stdin:
    key = literal_eval(line)[0]
```

Then I check if word exists in dictionary or not. If word already exist then take its previous count and add 1 to it and save new count. If word does not exist in dictionary then I simply add word as key and count value as 1 in dictionary.

```
    dic[key] = dic.get(key, 0) + 1
```

# Q2:

## Mapper:

It will read data from STDIN and remove any leading or trailing whitespaces. Then it read line character by character and skips any punctuations and number and concatenates other characters. Then split the line in words, check if that word exists in stop word dictionary or not. If not then output specific word with count to STDOUT

```
punctuations = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''
numbers='0123456789'
stop_words=['a','able','about','across','after','all','almost','also','am','among','a
n','and','any','are','as','at','be','because','been','but',

'by','can','cannot','could','dear','did','do','does','either','else','ever','every','
for','from','get','got','had','has','have','he',

'her','hers','him','his','how','however','i','if','in','into','is','it','its','just',
'least','let','like','likely','may','me','might',

'most','must','my','neither','no','nor','not','of','off','often','on','only','or','ot
her','our','own','rather','said','say','says',

'she','should','since','so','some','than','that','the','their','them','then','there',
'these','they','this','tis','to','too','twas',

'us','wants','was','we','were','what','when','where','which','while','who','whom','wh
y','will','with','would','yet','you','your']

no_punct_and_number = ""

for line in sys.stdin:
    line = line.strip()
    for char in line:
        if char not in punctuations and numbers:
            no_punct_and_number = no_punct_and_number + char
    words = no_punct_and_number.split()# split the line into words
    for word in words:
        if word not in stop_words:
            word = (word.lower(), 1)
            print(word)
```

Following line uses build in function of translation. maketrans first two parameters tells translate function to translate nothing to nothing and translate any punctuation or numbers to None (i.e. remove them). This function worked faster so I used this in my code. My implementation of removing punctuation and number was taking a lot of time.

```
line = line.translate(str.maketrans('','',punctuation))
line = line.translate(str.maketrans('','','1234567890'))
line = line.strip()# split the line into words
```

## Reducer:

It will read the results of mapper.py from STDIN (the output format of mapper.py and the expected input format of reducer.py must match).

Since the value received from STDIN is in string format so first we need to convert this string tuple to tuple.

```
dic = {}
for line in sys.stdin:
    key = literal_eval(line)[0]
```

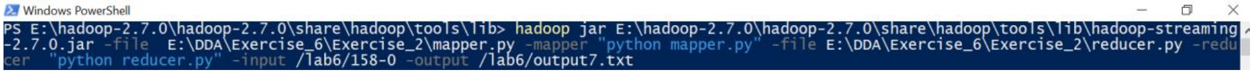
Then I check if word exists in dictionary or not. If word already exists then take its previous count and add 1 to it and save new count. If word does not exist in dictionary then I simply add word as key and count value as 1 in dictionary.

```
    dic[key] = dic.get(key, 0) + 1
```

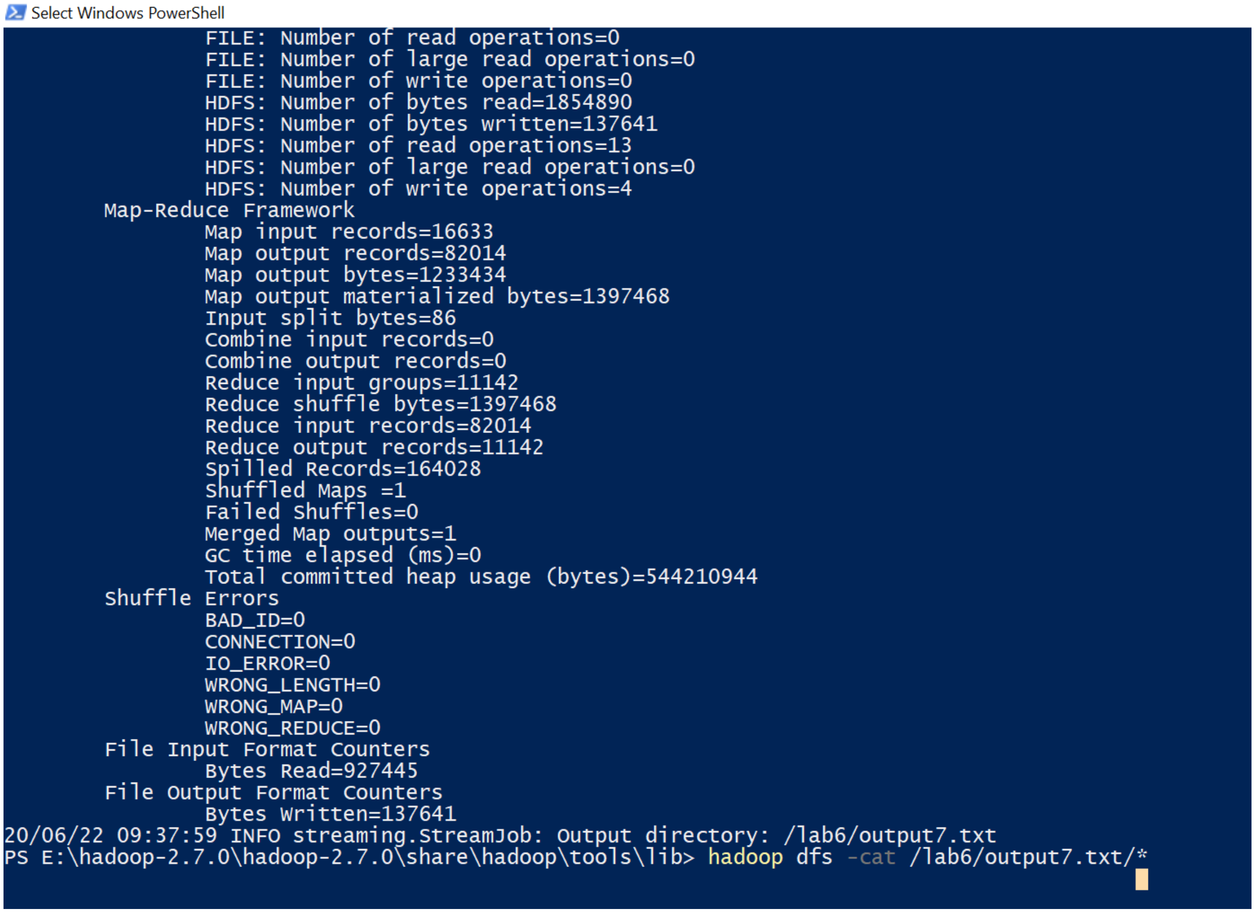
## Show the final results:

```
for k,v in dic.items():
    print('{} {}'.format(k,v))
```

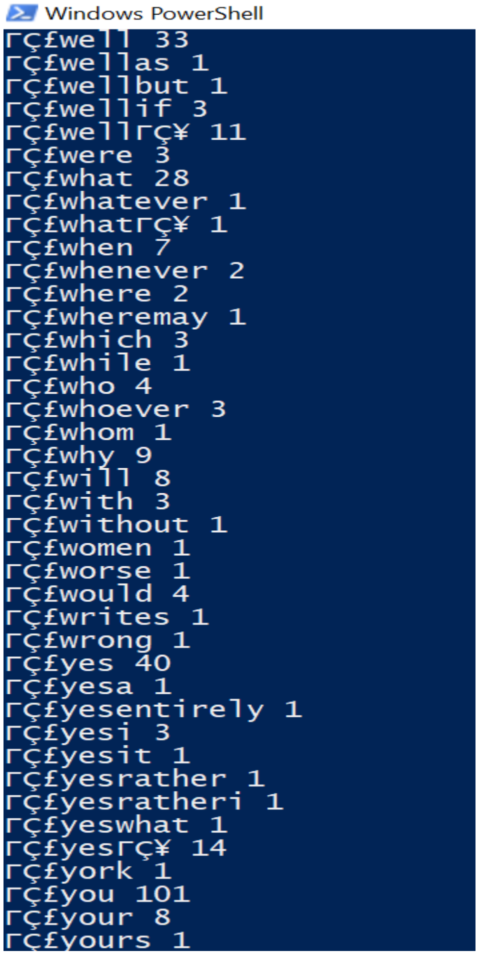
# Running commands: All 5 files



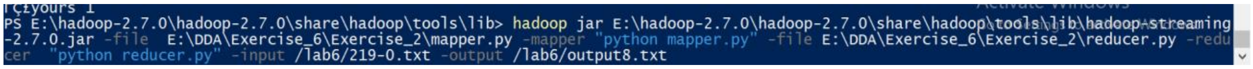
```
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop jar E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib\hadoop-streaming
-2.7.0.jar -file  E:\DDA\Exercise_6\Exercise_2\mapper.py -mapper "python mapper.py" -file E:\DDA\Exercise_6\Exercise_2\reducer.py -redu
cer  "python reducer.py" -input /lab6/158-0 -output /lab6/output7.txt
```



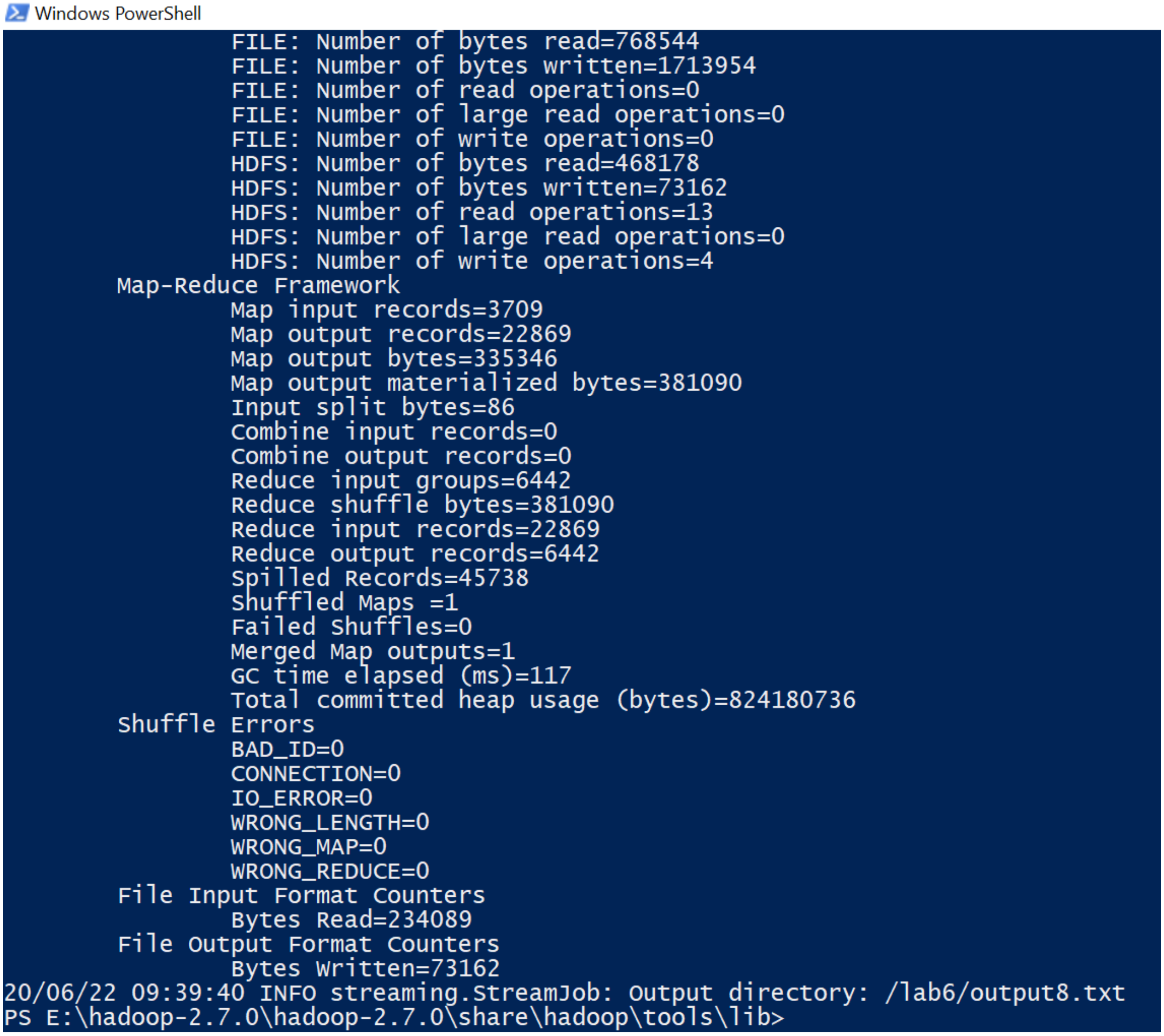
```
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1854890
                HDFS: Number of bytes written=137641
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=16633
                Map output records=82014
                Map output bytes=1233434
                Map output materialized bytes=1397468
                Input split bytes=86
                Combine input records=0
                Combine output records=0
                Reduce input groups=11142
                Reduce shuffle bytes=1397468
                Reduce input records=82014
                Reduce output records=11142
                Spilled Records=164028
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=544210944
        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=927445
        File Output Format Counters
                Bytes Written=137641
20/06/22 09:37:59 INFO streaming.StreamJob: Output directory: /lab6/output7.txt
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop dfs -cat /lab6/output7.txt/*
```



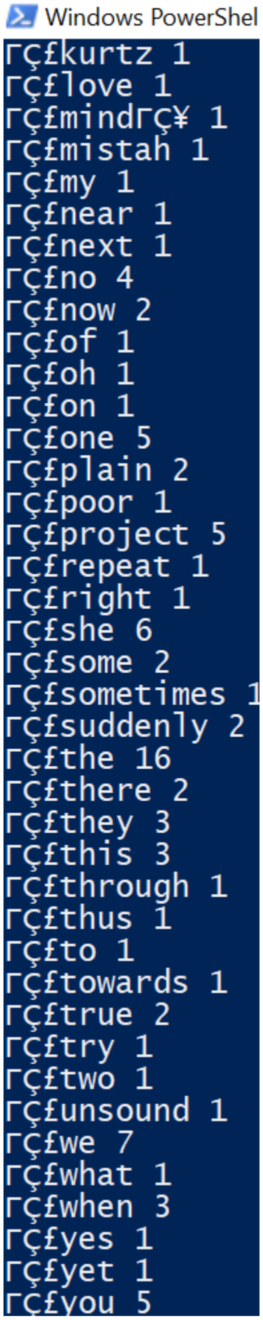
```
ΓÇ£well 33
ΓÇ£wellas 1
ΓÇ£wellbut 1
ΓÇ£wellif 3
ΓÇ£wellΓÇ¥ 11
ΓÇ£were 3
ΓÇ£what 28
ΓÇ£whatever 1
ΓÇ£whatΓÇ¥ 1
ΓÇ£when 7
ΓÇ£whenever 2
ΓÇ£where 2
ΓÇ£wheremay 1
ΓÇ£which 3
ΓÇ£while 1
ΓÇ£who 4
ΓÇ£whoever 3
ΓÇ£whom 1
ΓÇ£why 9
ΓÇ£will 8
ΓÇ£with 3
ΓÇ£without 1
ΓÇ£women 1
ΓÇ£worse 1
ΓÇ£would 4
ΓÇ£writes 1
ΓÇ£wrong 1
ΓÇ£yes 40
ΓÇ£yesa 1
ΓÇ£yesentirely 1
ΓÇ£yesi 3
ΓÇ£yesit 1
ΓÇ£yesrather 1
ΓÇ£yesratheri 1
ΓÇ£yeswhat 1
ΓÇ£yesΓÇ¥ 14
ΓÇ£york 1
ΓÇ£you 101
ΓÇ£your 8
ΓÇ£yours 1
```



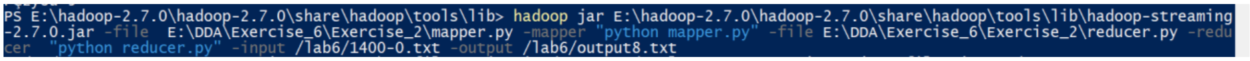
```
ΓÇ£yours 1
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop jar E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib\hadoop-streaming
-2.7.0.jar -file  E:\DDA\Exercise_6\Exercise_2\mapper.py -mapper "python mapper.py" -file E:\DDA\Exercise_6\Exercise_2\reducer.py -redu
cer  "python reducer.py" -input /lab6/219-0.txt -output /lab6/output8.txt
```



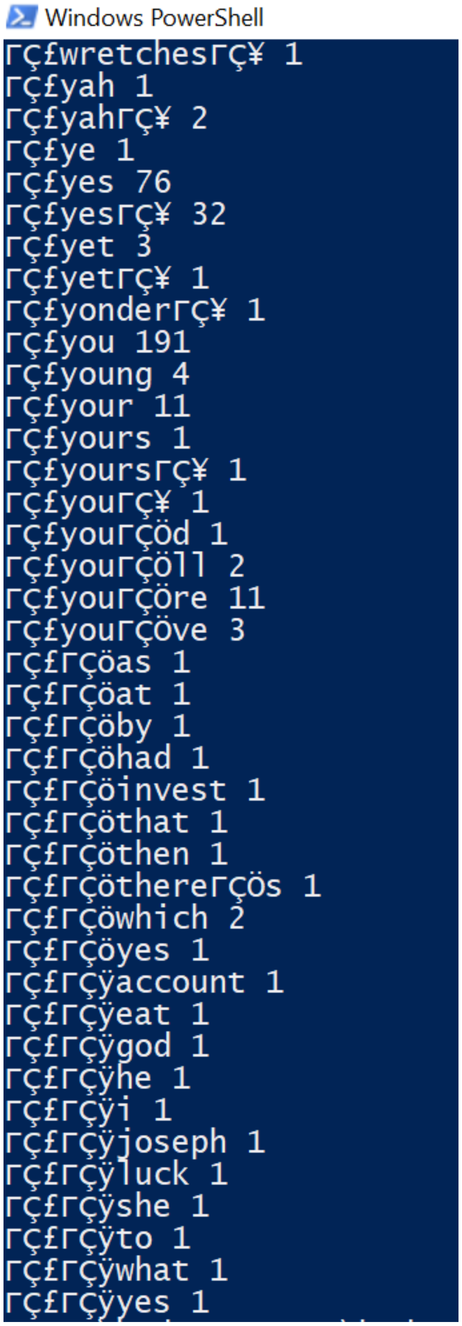
```
                FILE: Number of bytes read=768544
                FILE: Number of bytes written=1713954
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=468178
                HDFS: Number of bytes written=73162
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=3709
                Map output records=22869
                Map output bytes=335346
                Map output materialized bytes=381090
                Input split bytes=86
                Combine input records=0
                Combine output records=0
                Reduce input groups=6442
                Reduce shuffle bytes=381090
                Reduce input records=22869
                Reduce output records=6442
                Spilled Records=45738
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=117
                Total committed heap usage (bytes)=824180736
        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=234089
        File Output Format Counters
                Bytes Written=73162
20/06/22 09:39:40 INFO streaming.StreamJob: Output directory: /lab6/output8.txt
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib>
```



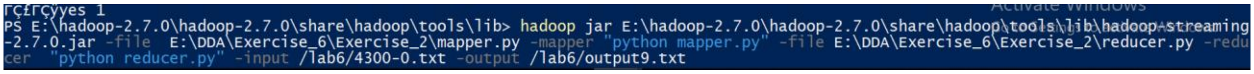
```
ΓÇ£kurtz 1
ΓÇ£love 1
ΓÇ£mindΓÇ¥ 1
ΓÇ£mistah 1
ΓÇ£my 1
ΓÇ£near 1
ΓÇ£next 1
ΓÇ£no 4
ΓÇ£now 2
ΓÇ£of 1
ΓÇ£oh 1
ΓÇ£on 1
ΓÇ£one 5
ΓÇ£plain 2
ΓÇ£poor 1
ΓÇ£project 5
ΓÇ£repeat 1
ΓÇ£right 1
ΓÇ£she 6
ΓÇ£some 2
ΓÇ£sometimes 1
ΓÇ£suddenly 2
ΓÇ£the 16
ΓÇ£there 2
ΓÇ£they 3
ΓÇ£this 3
ΓÇ£through 1
ΓÇ£thus 1
ΓÇ£to 1
ΓÇ£towards 1
ΓÇ£true 2
ΓÇ£try 1
ΓÇ£two 1
ΓÇ£unsound 1
ΓÇ£we 7
ΓÇ£what 1
ΓÇ£when 3
ΓÇ£yes 1
ΓÇ£yet 1
ΓÇ£you 5
```



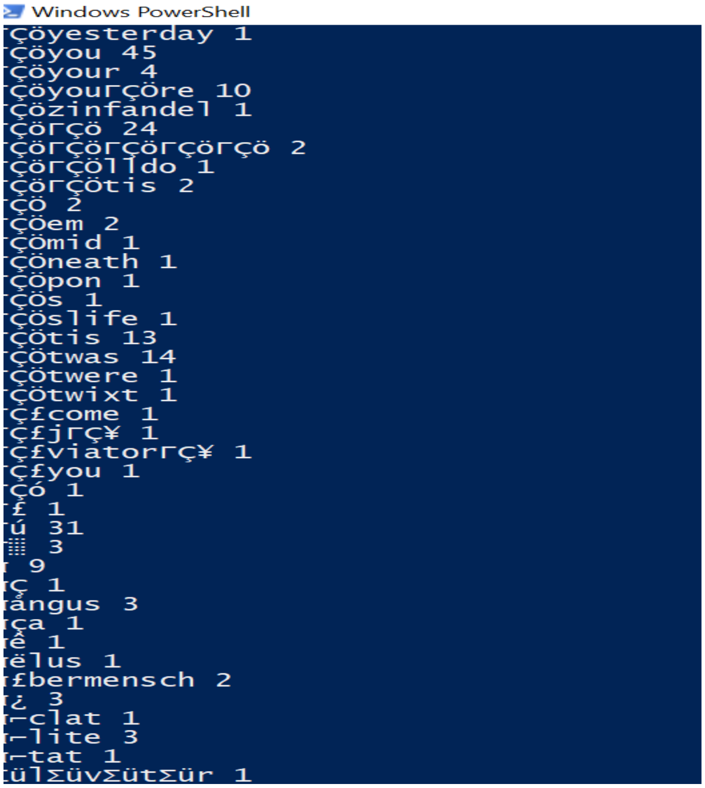
```
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop jar E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib\hadoop-streaming
-2.7.0.jar -file  E:\DDA\Exercise_6\Exercise_2\mapper.py -mapper "python mapper.py" -file E:\DDA\Exercise_6\Exercise_2\reducer.py -redu
cer  "python reducer.py" -input /lab6/1400-0.txt -output /lab6/output8.txt
```



```
ΓÇ£wretchesΓÇ¥ 1
ΓÇ£yah 1
ΓÇ£yahΓÇ¥ 2
ΓÇ£ye 1
ΓÇ£yes 76
ΓÇ£yesΓÇ¥ 32
ΓÇ£yet 3
ΓÇ£yetΓÇ¥ 1
ΓÇ£yonderΓÇ¥ 1
ΓÇ£you 191
ΓÇ£young 4
ΓÇ£your 11
ΓÇ£yours 1
ΓÇ£yoursΓÇ¥ 1
ΓÇ£youΓÇ¥ 1
ΓÇ£youΓÇÖd 1
ΓÇ£youΓÇÖll 2
ΓÇ£youΓÇÖre 11
ΓÇ£youΓÇÖve 3
ΓÇ£ΓÇöas 1
ΓÇ£ΓÇöat 1
ΓÇ£ΓÇöby 1
ΓÇ£ΓÇöhad 1
ΓÇ£ΓÇöinvest 1
ΓÇ£ΓÇöthat 1
ΓÇ£ΓÇöthen 1
ΓÇ£ΓÇöthereΓÇÖs 1
ΓÇ£ΓÇöwhich 2
ΓÇ£ΓÇöyes 1
ΓÇ£ΓÇÿaccount 1
ΓÇ£ΓÇÿeat 1
ΓÇ£ΓÇÿgod 1
ΓÇ£ΓÇÿhe 1
ΓÇ£ΓÇÿi 1
ΓÇ£ΓÇÿjoseph 1
ΓÇ£ΓÇÿluck 1
ΓÇ£ΓÇÿshe 1
ΓÇ£ΓÇÿto 1
ΓÇ£ΓÇÿwhat 1
ΓÇ£ΓÇÿyes 1
```



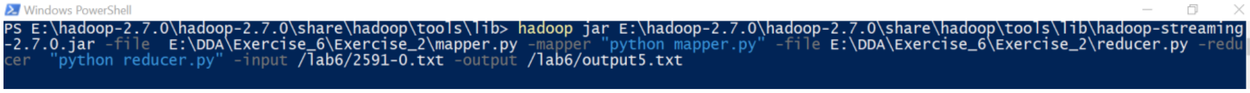
```
ΓÇ£ΓÇÿyes 1
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop jar E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib\hadoop-streaming
-2.7.0.jar -file  E:\DDA\Exercise_6\Exercise_2\mapper.py -mapper "python mapper.py" -file E:\DDA\Exercise_6\Exercise_2\reducer.py -redu
cer  "python reducer.py" -input /lab6/4300-0.txt -output /lab6/output9.txt
```



```
ÇöyesterdaY 1
```

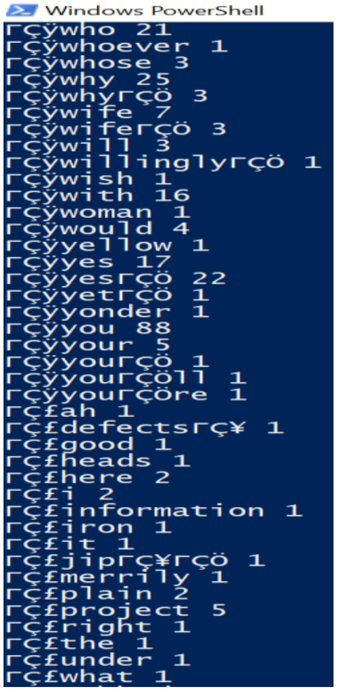
Windows PowerShell

```
Çöyesterday 1
Çöyou 45
Çöyour 4
ÇöyourΓÇÖre 10
Çözinfandel 1
ÇöΓÇö 24
ÇöΓÇöΓÇöΓÇöΓÇö 2
ÇöΓÇölldo 1
ÇöΓÇÖtis 2
ÇÖ 2
ÇÖem 2
ÇÖmid 1
ÇÖneath 1
ÇÖpon 1
ÇÖs 1
ÇÖslife 1
ÇÖtis 13
ÇÖtwas 14
ÇÖtwere 1
ÇÖtwixt 1
Ç£come 1
Ç£jΓÇ¥ 1
Ç£viatorΓÇ¥ 1
Ç£you 1
Çó 1
£ 1
ú 31
░ 3
 9
Ç 1
ángus 3
ça 1
ê 1
ëlus 1
£bermensch 2
¿ 3
⌐clat 1
⌐lite 3
⌐tat 1
ülΣüvΣütΣür 1
```



Windows PowerShell

```
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop jar E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib\hadoop-streaming
-2.7.0.jar -file  E:\DDA\Exercise_6\Exercise_2\mapper.py -mapper "python mapper.py" -file E:\DDA\Exercise_6\Exercise_2\reducer.py -redu
cer  "python reducer.py" -input /lab6/2591-0.txt -output /lab6/output5.txt
```



Windows PowerShell

```
ΓÇÿwho 21
ΓÇÿwhoever 1
ΓÇÿwhose 3
ΓÇÿwhy 25
ΓÇÿwhyΓÇö 3
ΓÇÿwife 7
ΓÇÿwifeΓÇö 3
ΓÇÿwill 3
ΓÇÿwillinglyΓÇö 1
ΓÇÿwish 1
ΓÇÿwith 16
ΓÇÿwoman 1
ΓÇÿwould 4
ΓÇÿyellow 1
ΓÇÿyes 17
ΓÇÿyesΓÇö 22
ΓÇÿyetΓÇö 1
ΓÇÿyonder 1
ΓÇÿyou 88
ΓÇÿyour 5
ΓÇÿyourΓÇö 1
ΓÇÿyourΓÇÖll 1
ΓÇÿyourΓÇÖre 1
ΓÇ£ah 1
ΓÇ£defectsΓÇ¥ 1
ΓÇ£good 1
ΓÇ£heads 1
ΓÇ£here 2
ΓÇ£i 2
ΓÇ£information 1
ΓÇ£iron 1
ΓÇ£it 1
ΓÇ£jipΓÇ¥ΓÇö 1
ΓÇ£merrily 1
ΓÇ£plain 2
ΓÇ£project 5
ΓÇ£right 1
ΓÇ£the 1
ΓÇ£under 1
ΓÇ£what 1
```

# Q3:

## Mapper:

It will read data from STDIN and get file names from environment variable "`map_input_file`". These are input files names which are passes as command line argument. Then it remove any punctuations or number from line. Then remove any leading or trailing whitespaces, split the line in words, check if that word exists in stop word dictionary or not. If not then output specific word with count to STDOUT

```
for line in sys.stdin:
    filename = os.environ["map_input_file"]
    line = line.translate(str.maketrans('','',punctuation))
    line = line.translate(str.maketrans('','','1234567890'))
    line = line.strip()# remove leading and trailing whitespace
    words = line.split() # split the line into words
    for word in words:
        if word not in stop_words:
            print((filename,word,1))
```

## Reducer:

It will read the results of mapper.py from STDIN (the output format of mapper.py and the expected input format of reducer.py must match).

Since the value received from STDIN is in string format so first we need to convert this string tuple to tuple. After converting it to tuple we only take first two arguments (filename, word) of tuple as third argument is just 1 in all cases.

```
import sys
from ast import literal_eval
dic = {}
for line in sys.stdin:
   key = literal_eval(line)[:2]
```

Then I check if word exists in dictionary or not. If word already exists then take its previous count and add 1 to it and save new count. If word does not exist in dictionary then I simply add word as key and count value as 1 in dictionary.

```
    dic[key] = dic.get(key , 0) + 1
```

Then we output the tuple as following format (filename, word, count of words in that filename).

```
for k,v in dic.items():
    print((k[0],k[1],v))
```

## Mapper2:

It will read data from STDIN (output of reducer1). Since the value received from STDIN is in string format so first we need to convert this string tuple to tuple. Then this tuple (filename, word, count of words in that filename) is passed as output to STDOUT

```
for line in sys.stdin:
    filename_word_count = literal_eval(line)
    print(filename_word_count)
```

## Reduce2:

It will read the results of mapper.py from STDIN (the output format of mapper.py and the expected input format of reducer.py must match).

Since the value received from STDIN is in string format so first we need to convert this string tuple to tuple. After converting it to tuple we calculate total words in each file. Also we save tuple that we received from STDIN which will help use later.

```
for line in sys.stdin:

    filename, word, count = literal_eval(line)
    count = int(count)
    if prev_filename == filename:
        N = N + count
    else:
        if prev_filename != None:
            file_name_with_total_wordCount[prev_filename] = N
            # print(file_name_with_total_wordCount[prev_filename])

        N = 0
        prev_filename = filename
    saved_previous_data.append(line)
file_name_with_total_wordCount[prev_filename] = N #saved last file count
```

Then we map the list of tuple that we saved previously to total words in the specific files from which the word in tuple belong.  Finally we output the tuple as following format (filename, word, count of words in that filename, total_words_count_in_specific file).

```
for line in saved_previous_data:
    filename, word, count = literal_eval(line)
    for k in file_name_with_total_wordCount.keys():
        if filename == k:
            print((word, filename, count, file_name_with_total_wordCount[k]))
```

## Mapper3:

It will read data from STDIN (output of reducer1). Since the value received from STDIN is in string format so first we need to convert this string tuple to tuple. Then this tuple (filename, word, count of words in that filename, total_words_count_in_specific file,1)  is passed as output to STDOUT

```

for line in sys.stdin:
    word_filename_count_totalcount = literal_eval(line)

print((word_filename_count_totalcount[0],word_filename_count_totalcount[1],word_filename_count_totalcount[2],word_filename_count_totalcount[3],1))

```

## Reduce3:

It will read the results of mapper.py from STDIN (the output format of mapper.py and the expected input format of reducer.py must match).

Since the value received from STDIN is in string format so first we need to convert this string tuple to tuple. After converting it to tuple we calculate number of documents that contains token t. Also we save tuple (word, filename) that which will help use later.

```
for line in sys.stdin:
    word, filename, wordCount, total_WordCount, count = literal_eval(line)
    if prev_word == word:
        total_count = total_count + int(count)
    else:
        if prev_word != None:
            df[prev_word] = (wordCount, total_WordCount, total_count)
            word_filename = (prev_word, filename)
            saved_previous_data.append(word_filename)
        total_count = 1
        prev_word = word

df[prev_word] = (wordCount, total_WordCount, total_count)
```

```
word_filename = (prev_word, filename)
saved_previous_data.append(word_filename)
```

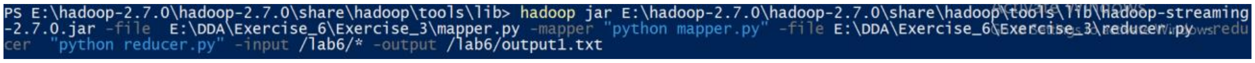
Now we have a list of tuple which contain word and filename. Then we have dictionary that has word as key and tuple (wordCount, total_WordCount, total_count_of_tokenK_i n document) as value.

Then first I take token (word) and get its data from dictionary.Then I used this data to find tfifd of token using formula given in exercise

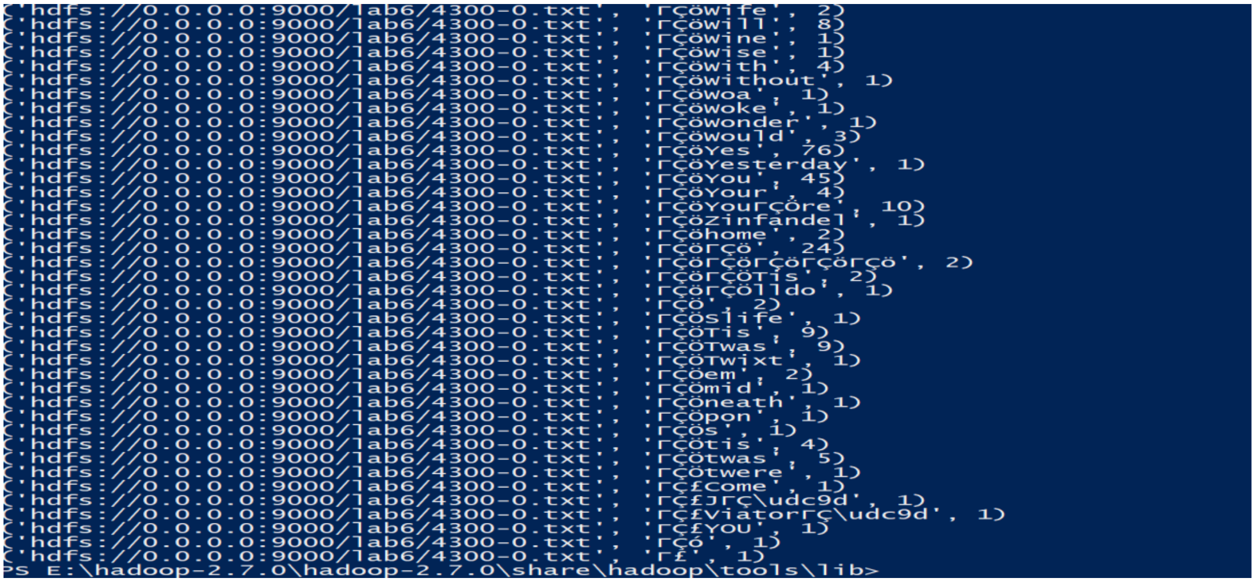
```
for line in saved_previous_data:
    word, filename = line
    for k in df.keys():
        if word == k:
            wordCount, total_WordCount, total_count_of_word_in_different_document = 
df[k]
            tfidf = (wordCount / total_WordCount) * log10(5 / 
total_count_of_word_in_different_document)
            print('{}: {}'.format(word, tfidf))
```

## Running commands:

## First mapper, reducer:

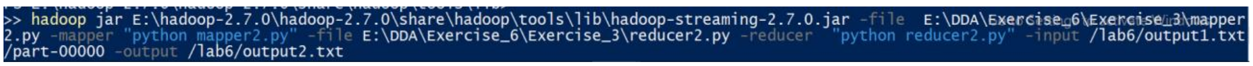


```
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop jar E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib\hadoop-streaming
-2.7.0.jar -file  E:\DDA\Exercise_6\Exercise_3\mapper.py -mapper "python mapper.py" -file E:\DDA\Exercise_6\Exercise_3\reducer.py -redu
cer  "python reducer.py" -input /lab6/* -output /lab6/output1.txt
```

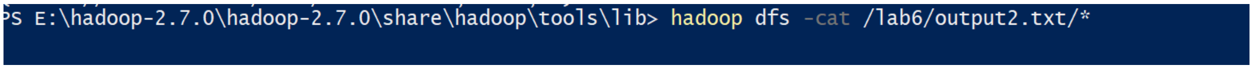


```
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöwife', 2)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöwill', 8)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöwine', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöwise', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöwith', 4)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöwithout', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöwoa', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöwoke', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöwonder', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöwould', 3)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöYes', 76)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöYesterday', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöYou', 45)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöYour', 4)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöYouΓÇÖre', 10)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöZinfandel', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöhome', 2)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöΓÇö', 24)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöΓÇöΓÇöΓÇöΓÇö', 2)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöΓÇÖTis', 2)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöΓÇÖlldo', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇö', 2)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇöslife', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇÖTis', 9)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇÖTwas', 9)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇÖTwixt', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇÖem', 2)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇÖmid', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇÖneath', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇÖpon', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇÖs', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇÖtis', 4)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇÖtwas', 5)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇÖtwere', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇ£Come', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇ£JΓÇ\udc9d', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇ£ViatorΓÇ\udc9d', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇ£YOU', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'ΓÇó', 1)
('hdfs://0.0.0.0:9000/lab6/4300-0.txt', 'Γ£', 1)
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib>
```

## Second mapper reducer:

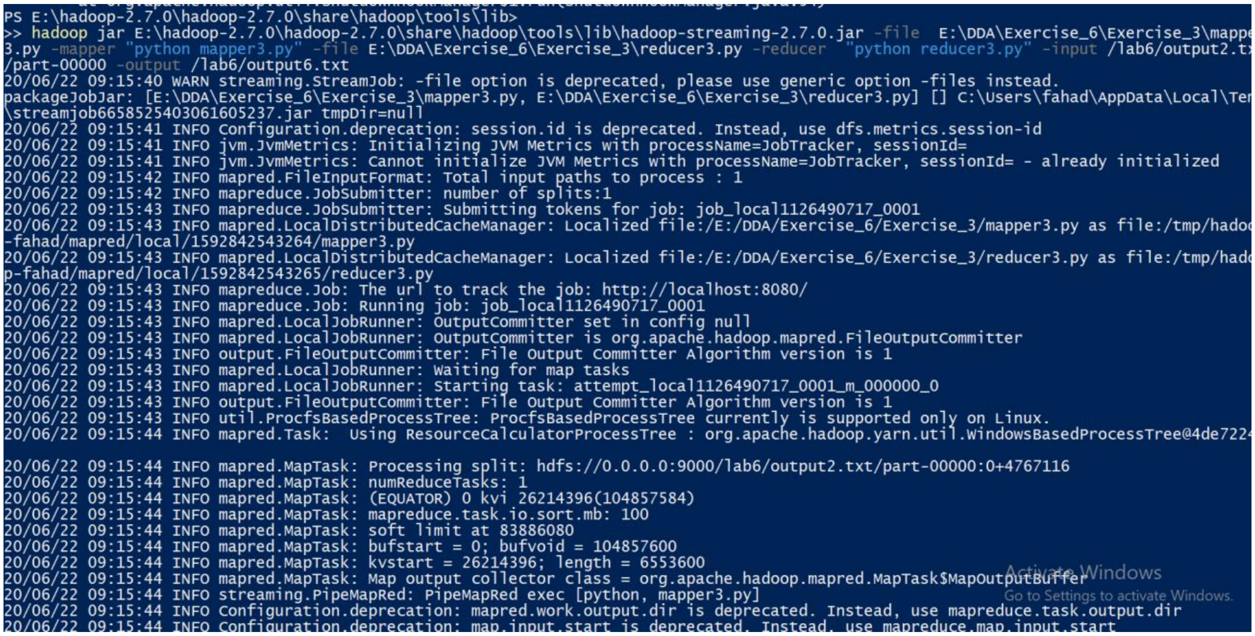


```
>> hadoop jar E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib\hadoop-streaming-2.7.0.jar -file  E:\DDA\Exercise_6\Exercise_3\mapper
2.py -mapper "python mapper2.py" -file E:\DDA\Exercise_6\Exercise_3\reducer2.py -reducer  "python reducer2.py" -input /lab6/output1.txt
/part-00000 -output /lab6/output2.txt
```

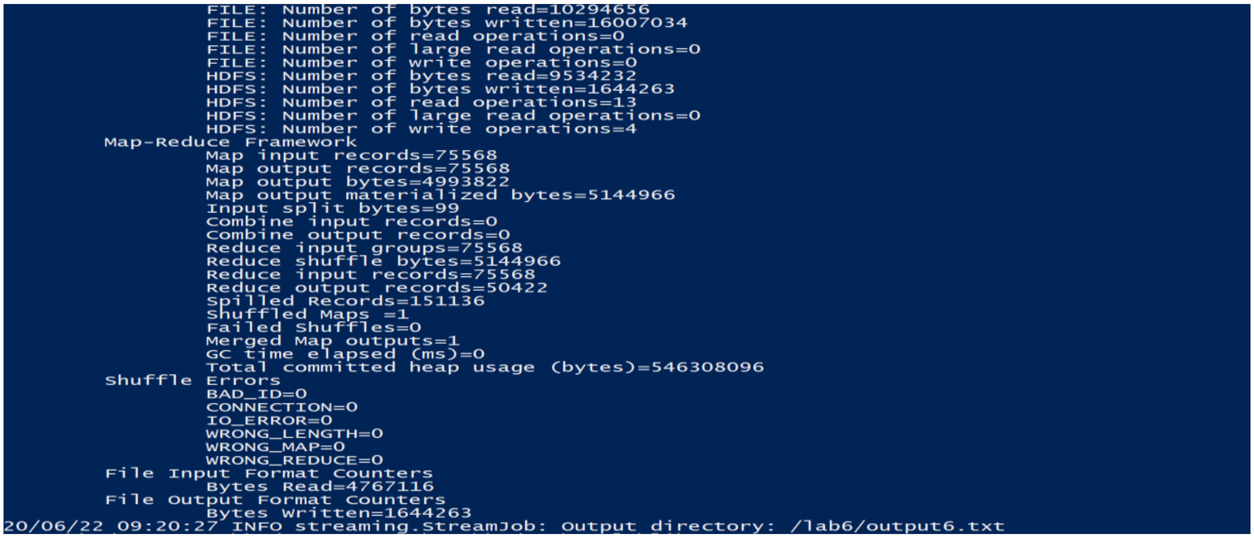


```
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop dfs -cat /lab6/output2.txt/*
```

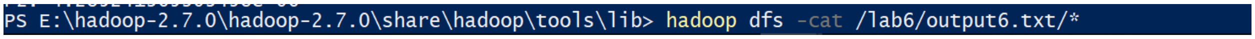
## Third mapper reducer:



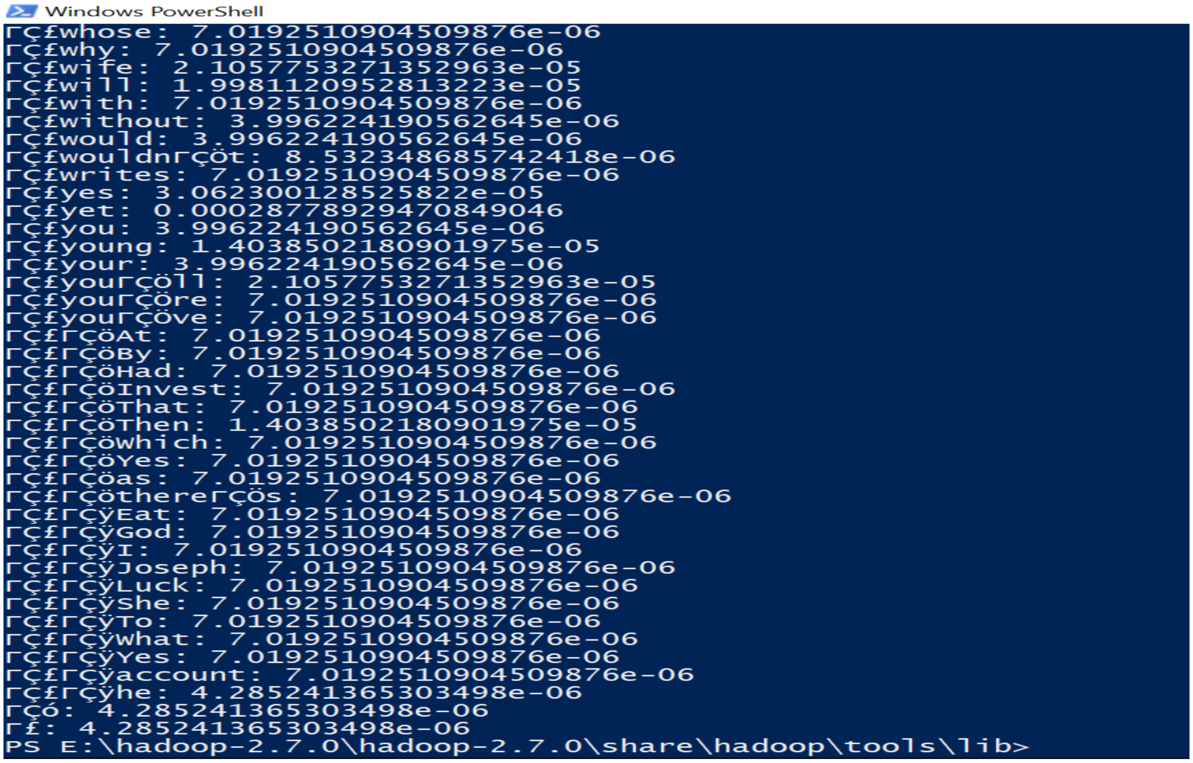
```
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib>
>> hadoop jar E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib\hadoop-streaming-2.7.0.jar -file  E:\DDA\Exercise_6\Exercise_3\mappe
3.py -mapper "python mapper3.py" -file E:\DDA\Exercise_6\Exercise_3\reducer3.py -reducer  "python reducer3.py" -input /lab6/output2.t
/part-00000 -output /lab6/output6.txt
20/06/22 09:15:40 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [E:\DDA\Exercise_6\Exercise_3\mapper3.py, E:\DDA\Exercise_6\Exercise_3\reducer3.py] [] C:\Users\fahad\AppData\Local\Te
\streamjob6658525403061605237.jar tmpDir=null
20/06/22 09:15:41 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
20/06/22 09:15:41 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
20/06/22 09:15:41 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
20/06/22 09:15:42 INFO mapred.FileInputFormat: Total input paths to process : 1
20/06/22 09:15:42 INFO mapreduce.JobSubmitter: number of splits:1
20/06/22 09:15:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1126490717_0001
20/06/22 09:15:43 INFO mapred.LocalDistributedCacheManager: Localized file:/E:/DDA/Exercise_6/Exercise_3/mapper3.py as file:/tmp/hado
-fahad/mapred/local/1592842543264/mapper3.py
20/06/22 09:15:43 INFO mapred.LocalDistributedCacheManager: Localized file:/E:/DDA/Exercise_6/Exercise_3/reducer3.py as file:/tmp/had
p-fahad/mapred/local/1592842543265/reducer3.py
20/06/22 09:15:43 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
20/06/22 09:15:43 INFO mapreduce.Job: Running job: job_local1126490717_0001
20/06/22 09:15:43 INFO mapred.LocalJobRunner: OutputCommitter set in config null
20/06/22 09:15:43 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
20/06/22 09:15:43 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/06/22 09:15:43 INFO mapred.LocalJobRunner: Waiting for map tasks
20/06/22 09:15:43 INFO mapred.LocalJobRunner: Starting task: attempt_local1126490717_0001_m_000000_0
20/06/22 09:15:43 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/06/22 09:15:43 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
20/06/22 09:15:44 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTree@4de722

20/06/22 09:15:44 INFO mapred.MapTask: Processing split: hdfs://0.0.0.0:9000/lab6/output2.txt/part-00000:0+4767116
20/06/22 09:15:44 INFO mapred.MapTask: numReduceTasks: 1
20/06/22 09:15:44 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
20/06/22 09:15:44 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
20/06/22 09:15:44 INFO mapred.MapTask: soft limit at 83886080
20/06/22 09:15:44 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
20/06/22 09:15:44 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
20/06/22 09:15:44 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
20/06/22 09:15:44 INFO streaming.PipeMapRed: PipeMapRed exec [python, mapper3.py]
20/06/22 09:15:44 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
20/06/22 09:15:44 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
```



```
                FILE: Number of bytes read=10294656
                FILE: Number of bytes written=16007034
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=9534232
                HDFS: Number of bytes written=1644263
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=75568
                Map output records=75568
                Map output bytes=4993822
                Map output materialized bytes=5144966
                Input split bytes=99
                Combine input records=0
                Combine output records=0
                Reduce input groups=75568
                Reduce shuffle bytes=5144966
                Reduce input records=75568
                Reduce output records=50422
                Spilled Records=151136
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=546308096
        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=4767116
        File Output Format Counters
                Bytes Written=1644263
20/06/22 09:20:27 INFO streaming.StreamJob: Output directory: /lab6/output6.txt
```



```
PS E:\hadoop-2.7.0\hadoop-2.7.0\share\hadoop\tools\lib> hadoop dfs -cat /lab6/output6.txt/*
```



```
ΓÇ£whose: 7.01925109045098 76e-06
```