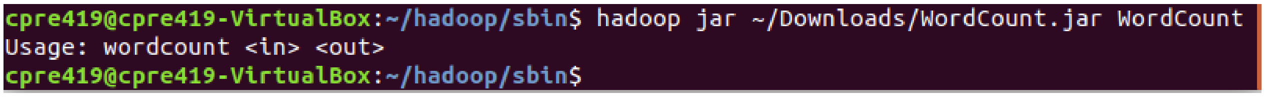
CprE419 Lab 2
Neha Maddali

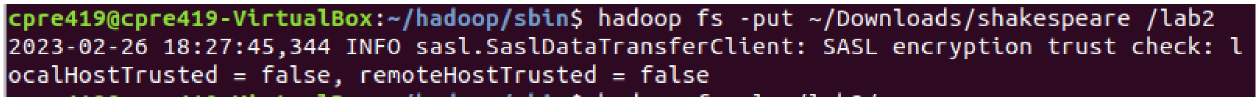
Experiment 1:

WordCount.java was downloaded and compiled with line:
hadoop jar ~/Downloads/WordCount.jar Wordcount



```
cpre419@cpre419-VirtualBox:~/hadoop/sbin$ hadoop jar ~/Downloads/WordCount.jar WordCount
Usage: wordcount <in> <out>
cpre419@cpre419-VirtualBox:~/hadoop/sbin$
```

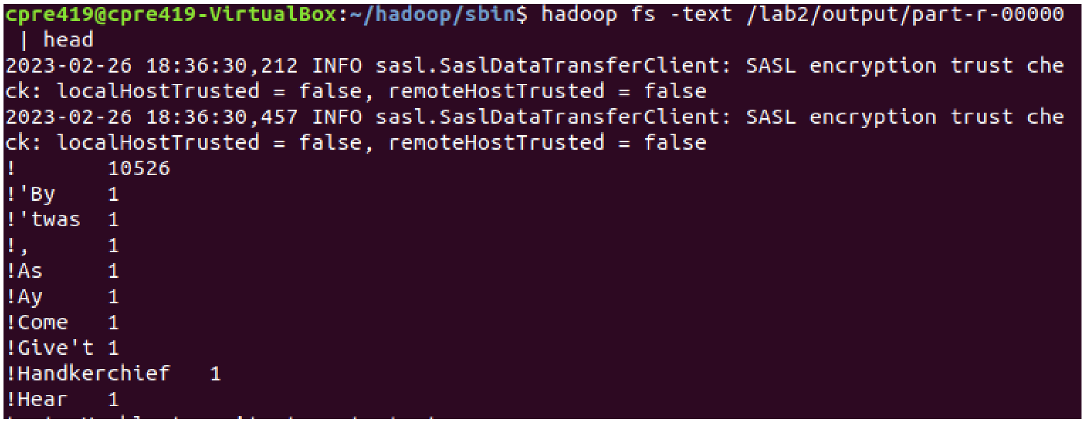
Load and move Shakespeare dataset to lab 2 folder in HDFS hadoop:
fs -put ~/Downloads/shakespeare /lab2
Check if moved: hadoop fs -ls /lab2/



```
cpre419@cpre419-VirtualBox:~/hadoop/sbin$ hadoop fs -put ~/Downloads/shakespeare /lab2
2023-02-26 18:27:45,344 INFO sasl.SaslDataTransferClient: SASL encryption trust check: l
ocalHostTrusted = false, remoteHostTrusted = false
```

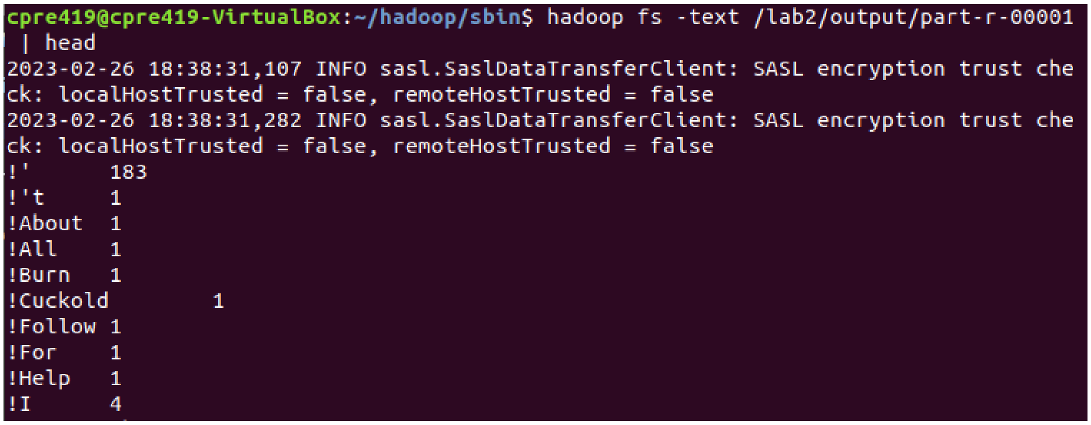
jar file was ran on shakespeare dataset with:
hadoop jar ~/Downloads/WordCount.jar WordCount /lab2/shakespeare /lab2/output

Below are the first 10 lines of output file part-r-00000:



```
cpre419@cpre419-VirtualBox:~/hadoop/sbin$ hadoop fs -text /lab2/output/part-r-00000
 | head
2023-02-26 18:36:30,212 INFO sasl.SaslDataTransferClient: SASL encryption trust che
ck: localHostTrusted = false, remoteHostTrusted = false
2023-02-26 18:36:30,457 INFO sasl.SaslDataTransferClient: SASL encryption trust che
ck: localHostTrusted = false, remoteHostTrusted = false
!       10526
!'By    1
!'twas  1
!,      1
!As     1
!Ay     1
!Come   1
!Give't 1
!Handkerchief   1
!Hear   1
```

Below are the first 10 lines of output file part-r-00001:



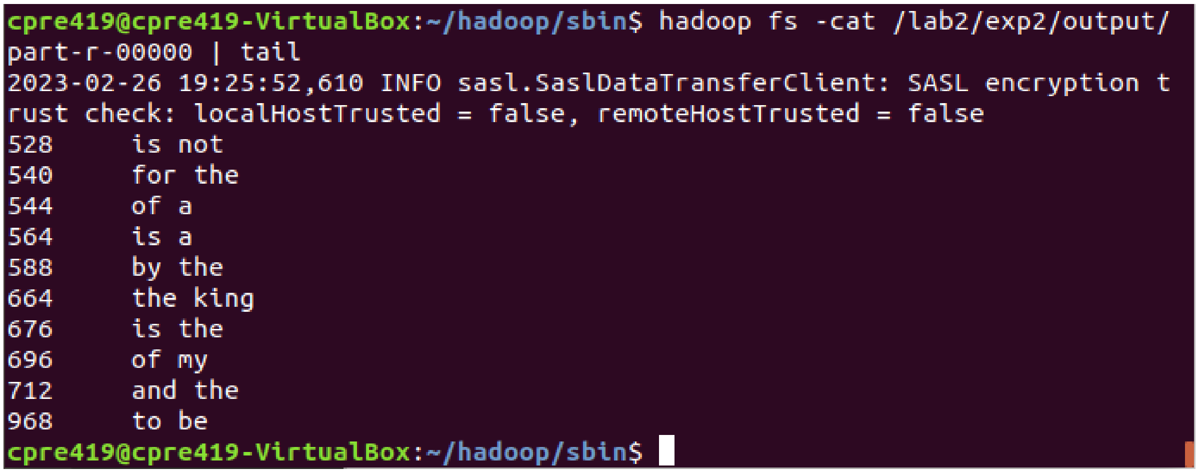
```
cpre419@cpre419-VirtualBox:~/hadoop/sbin$ hadoop fs -text /lab2/output/part-r-00001
 | head
2023-02-26 18:38:31,107 INFO sasl.SaslDataTransferClient: SASL encryption trust che
ck: localHostTrusted = false, remoteHostTrusted = false
2023-02-26 18:38:31,282 INFO sasl.SaslDataTransferClient: SASL encryption trust che
ck: localHostTrusted = false, remoteHostTrusted = false
!'      183
!'t     1
!About  1
!All    1
!Burn   1
!Cuckold        1
!Follow 1
!For    1
!Help   1
!I      4
```

Experiment 2:
Question: Think about how you might be able to get around the fact that bigrams might span lines of input. Briefly describe how you might deal with that situation?

> We can start with converting all the characters to lower case and remove all punctuation except “!”, “.”, “?”. Allow the Map method the take in a whole sentence as an input to find all bigrams in every sentence. We know that bigrams do not span from sentence to sentence. So this methodology will help find bigrams in the entire file.

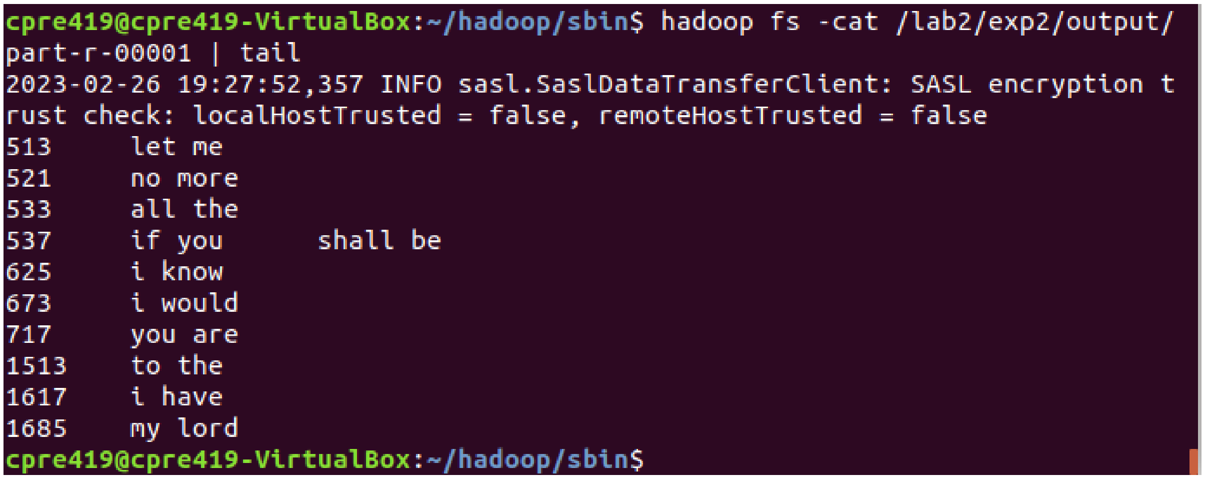
jar file was ran on shakespeare dataset with:
hadoop jar ~/Downloads/Driver.jar Driver /lab2/shakespeare /lab2/exp2

Shakespeare 10 frequent words
output file: part-r-00000



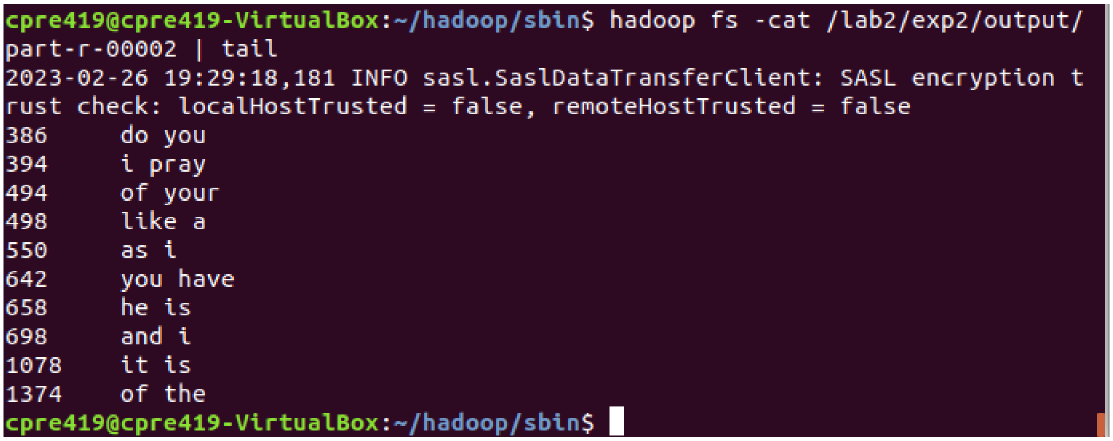
```
cpre419@cpre419-VirtualBox:~/hadoop/sbin$ hadoop fs -cat /lab2/exp2/output/
part-r-00000 | tail
2023-02-26 19:25:52,610 INFO sasl.SaslDataTransferClient: SASL encryption t
rust check: localHostTrusted = false, remoteHostTrusted = false
528     is not
540     for the
544     of a
564     is a
588     by the
664     the king
676     is the
696     of my
712     and the
968     to be
cpre419@cpre419-VirtualBox:~/hadoop/sbin$
```

output file: part-r-00001



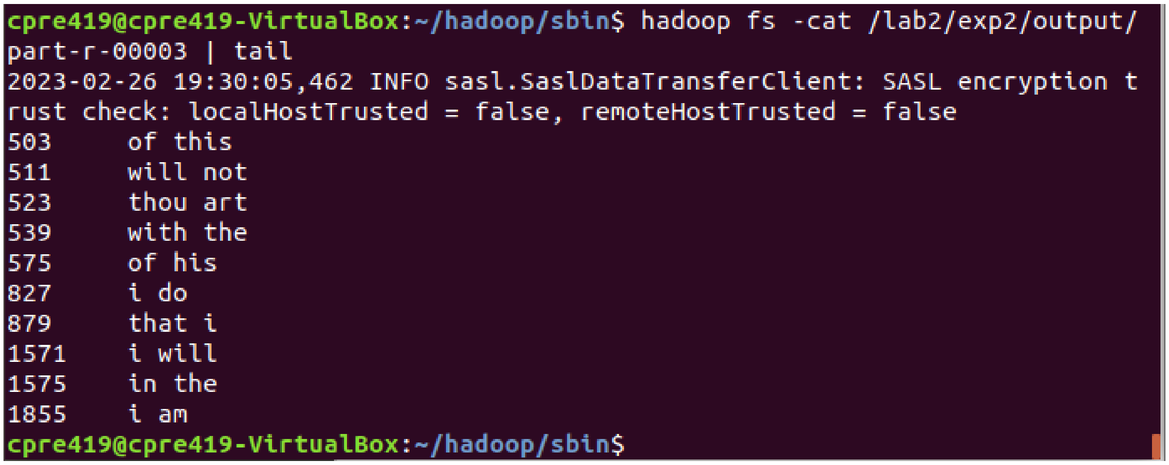
```
cpre419@cpre419-VirtualBox:~/hadoop/sbin$ hadoop fs -cat /lab2/exp2/output/
part-r-00001 | tail
2023-02-26 19:27:52,357 INFO sasl.SaslDataTransferClient: SASL encryption t
rust check: localHostTrusted = false, remoteHostTrusted = false
513     let me
521     no more
533     all the
537     if you      shall be
625     i know
673     i would
717     you are
1513    to the
1617    i have
1685    my lord
cpre419@cpre419-VirtualBox:~/hadoop/sbin$
```

output file: part-r-00002



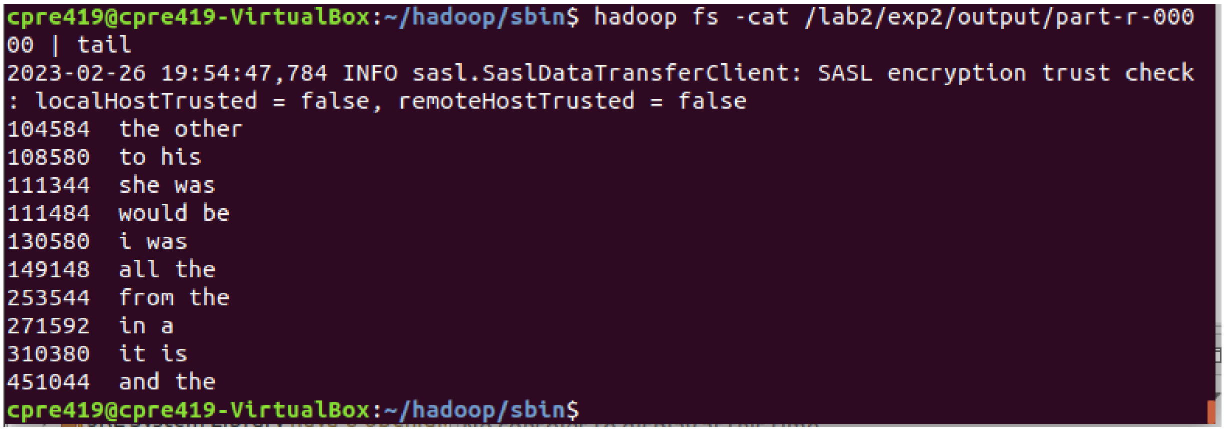
```
cpre419@cpre419-VirtualBox:~/hadoop/sbin$ hadoop fs -cat /lab2/exp2/output/
part-r-00002 | tail
2023-02-26 19:29:18,181 INFO sasl.SaslDataTransferClient: SASL encryption t
rust check: localHostTrusted = false, remoteHostTrusted = false
386     do you
394     i pray
494     of your
498     like a
550     as i
642     you have
658     he is
698     and i
1078    it is
1374    of the
cpre419@cpre419-VirtualBox:~/hadoop/sbin$ 
```

output file: part-r-00003



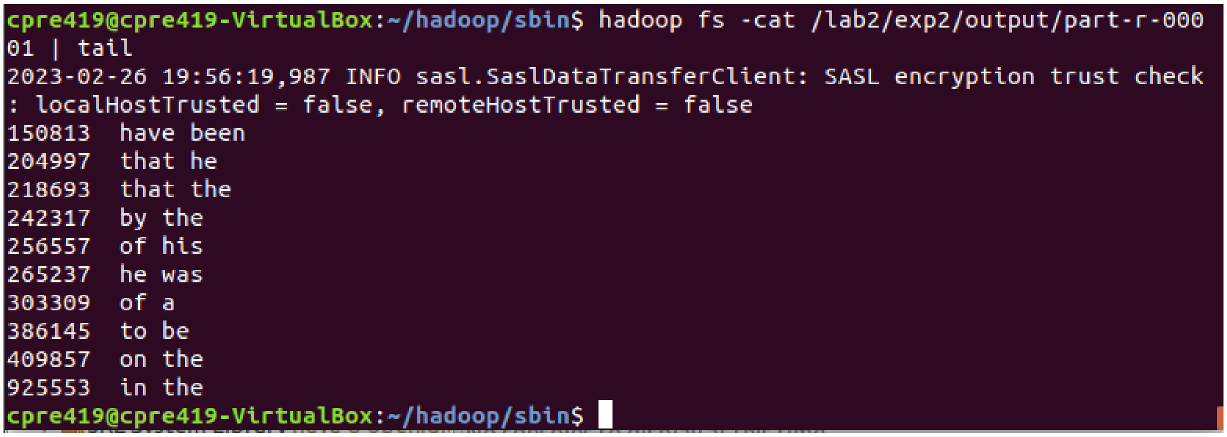
```
cpre419@cpre419-VirtualBox:~/hadoop/sbin$ hadoop fs -cat /lab2/exp2/output/
part-r-00003 | tail
2023-02-26 19:30:05,462 INFO sasl.SaslDataTransferClient: SASL encryption t
rust check: localHostTrusted = false, remoteHostTrusted = false
503     of this
511     will not
523     thou art
539     with the
575     of his
827     i do
879     that i
1571    i will
1575    in the
1855    i am
cpre419@cpre419-VirtualBox:~/hadoop/sbin$
```

Load and move gutenburg dataset to lab 2 folder in HDFS hadoop:
fs -put ~/Downloads/gutenburg /lab2
jar file was ran on gutenburg dataset with:
hadoop jar ~/Downloads/Driver.jar Driver /lab2/gutenburg /lab2/exp2
Gutenburg 10 Frequent Words
output file: part-r-00000



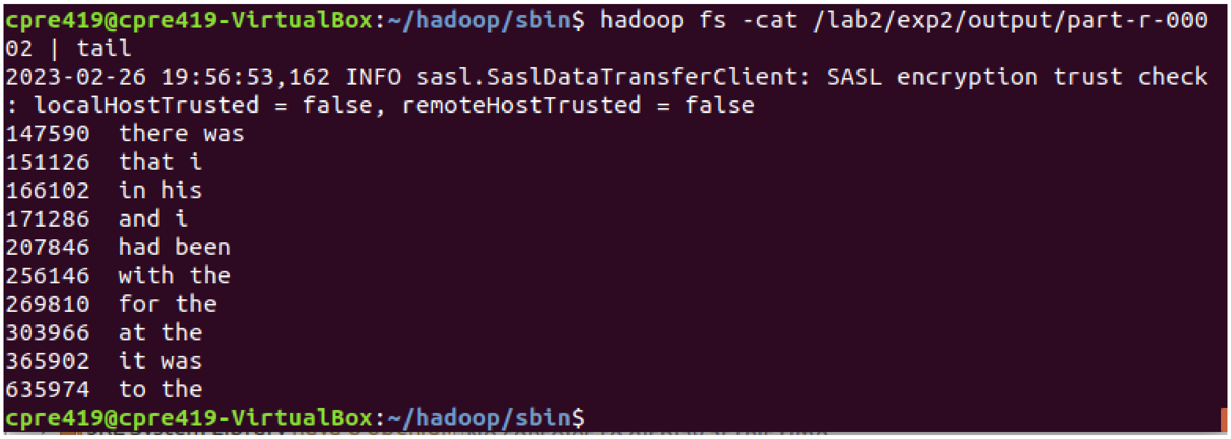
```
cpre419@cpre419-VirtualBox:~/hadoop/sbin$ hadoop fs -cat /lab2/exp2/output/part-r-000
00 | tail
2023-02-26 19:54:47,784 INFO sasl.SaslDataTransferClient: SASL encryption trust check
: localHostTrusted = false, remoteHostTrusted = false
104584  the other
108580  to his
111344  she was
111484  would be
130580  i was
149148  all the
253544  from the
271592  in a
310380  it is
451044  and the
cpre419@cpre419-VirtualBox:~/hadoop/sbin$
```

output file: part-r-00001



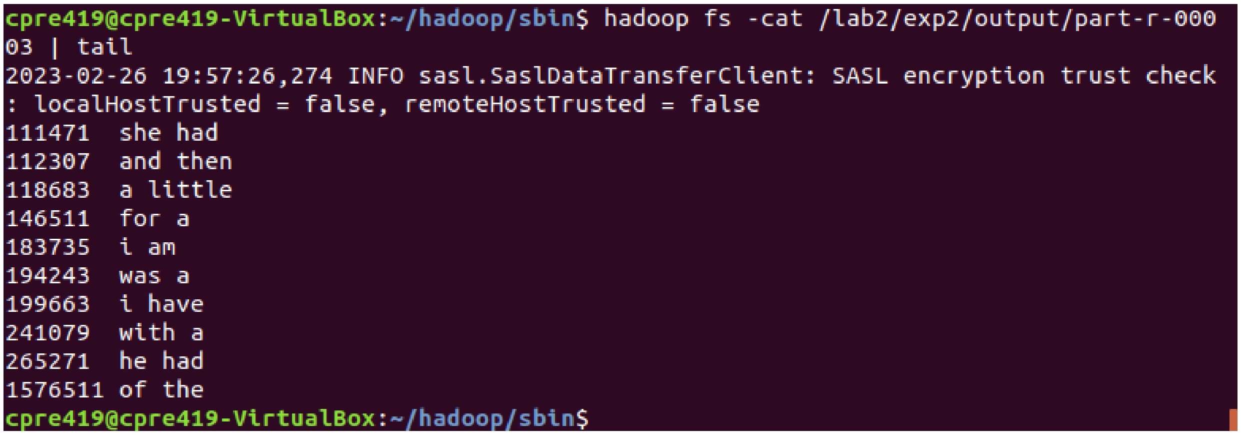
```
cpre419@cpre419-VirtualBox:~/hadoop/sbin$ hadoop fs -cat /lab2/exp2/output/part-r-000
01 | tail
2023-02-26 19:56:19,987 INFO sasl.SaslDataTransferClient: SASL encryption trust check
: localHostTrusted = false, remoteHostTrusted = false
150813  have been
204997  that he
218693  that the
242317  by the
256557  of his
265237  he was
303309  of a
386145  to be
409857  on the
925553  in the
cpre419@cpre419-VirtualBox:~/hadoop/sbin$
```

output file: part-r-00002



```
cpre419@cpre419-VirtualBox:~/hadoop/sbin$ hadoop fs -cat /lab2/exp2/output/part-r-000
02 | tail
2023-02-26 19:56:53,162 INFO sasl.SaslDataTransferClient: SASL encryption trust check
: localHostTrusted = false, remoteHostTrusted = false
147590  there was
151126  that i
166102  in his
171286  and i
207846  had been
256146  with the
269810  for the
303966  at the
365902  it was
635974  to the
cpre419@cpre419-VirtualBox:~/hadoop/sbin$
```

output file: part-r-00003



```
cpre419@cpre419-VirtualBox:~/hadoop/sbin$ hadoop fs -cat /lab2/exp2/output/part-r-000
03 | tail
2023-02-26 19:57:26,274 INFO sasl.SaslDataTransferClient: SASL encryption trust check
: localHostTrusted = false, remoteHostTrusted = false
111471  she had
112307  and then
118683  a little
146511  for a
183735  i am
194243  was a
199663  i have
241079  with a
265271  he had
1576511 of the
cpre419@cpre419-VirtualBox:~/hadoop/sbin$
```