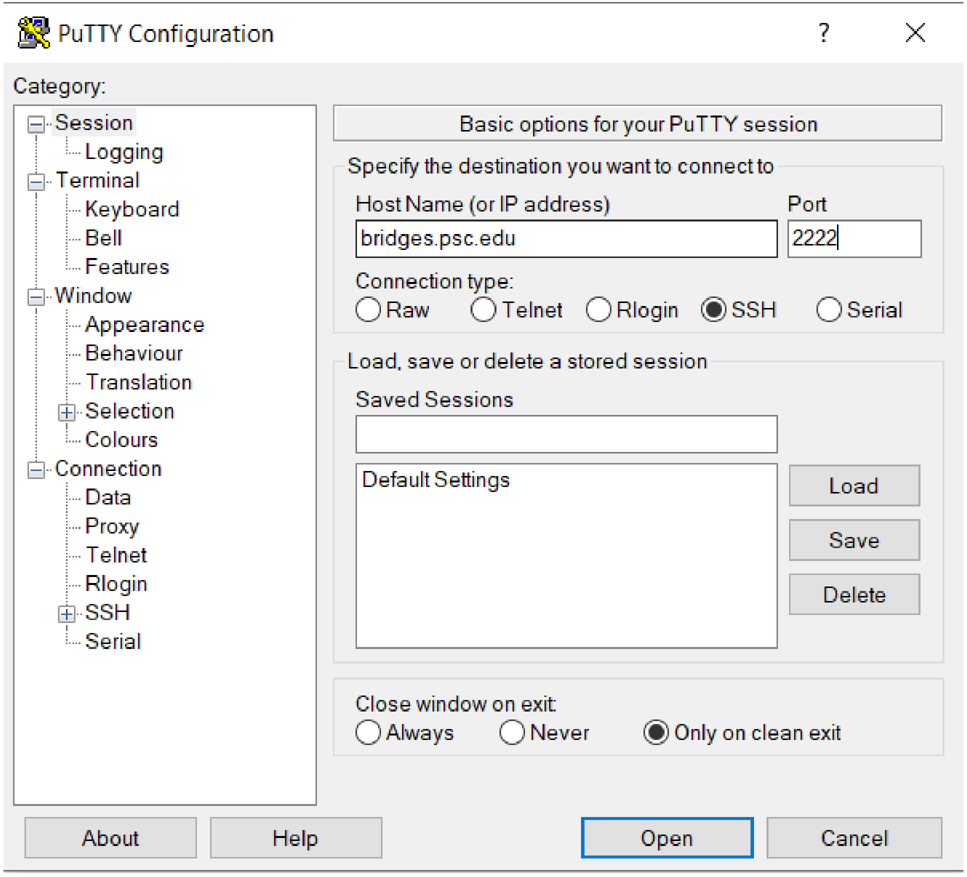
Ocean Lu
CS 4080.02
Professor Yang
11.12.2019

Project 2: MapReduce Programming on Hadoop and Spark

Task 1:

1. Log into bridges.psc.edu port 2222; used puTTY



Specify the destination you want to connect to

Host Name (or IP address): bridges.psc.edu

Port: 2222

Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

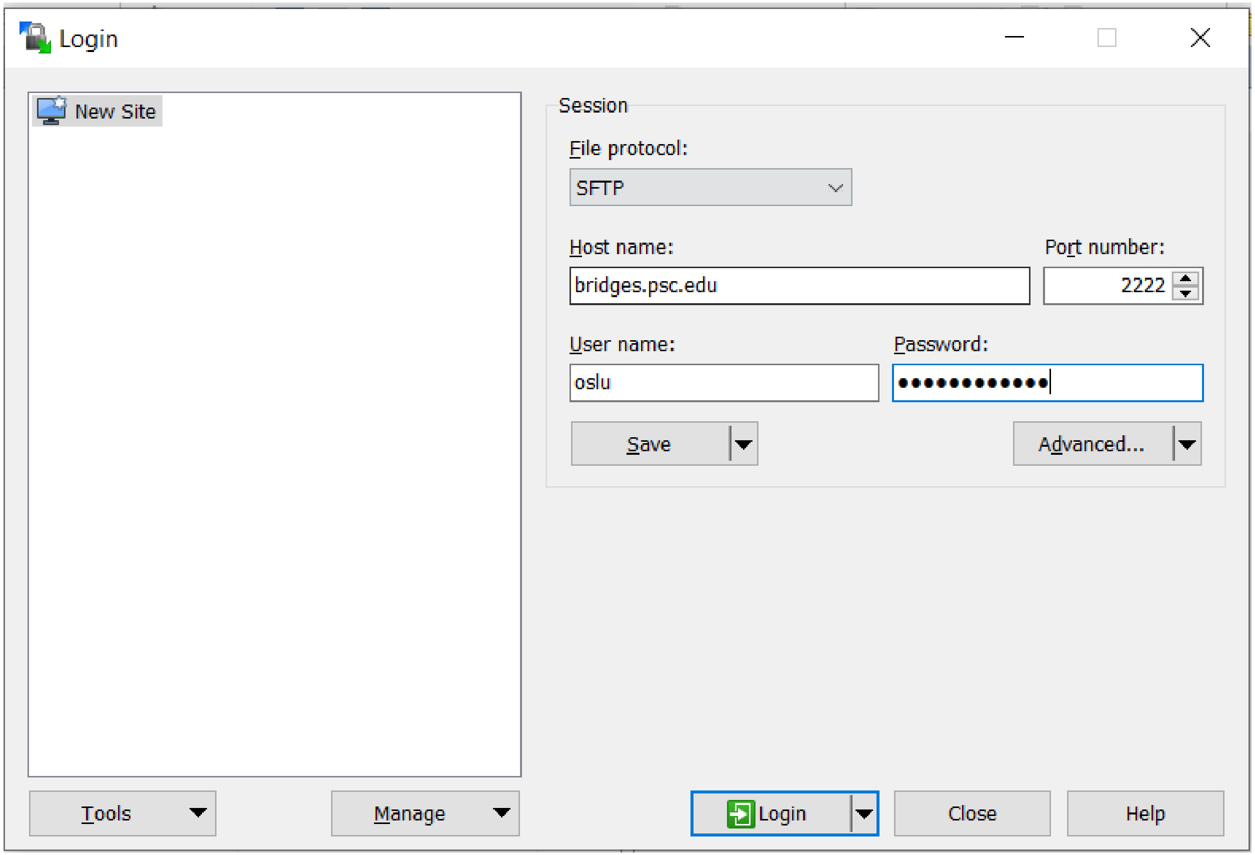
Saved Sessions:

Default Settings

Close window on exit: ○ Always ○ Never ● Only on clean exit

2. Went to the bible site and downloaded the zip files

3. WinSCP the files of Shakespeare and apodat.txt into the bridges.psc.edu



Session

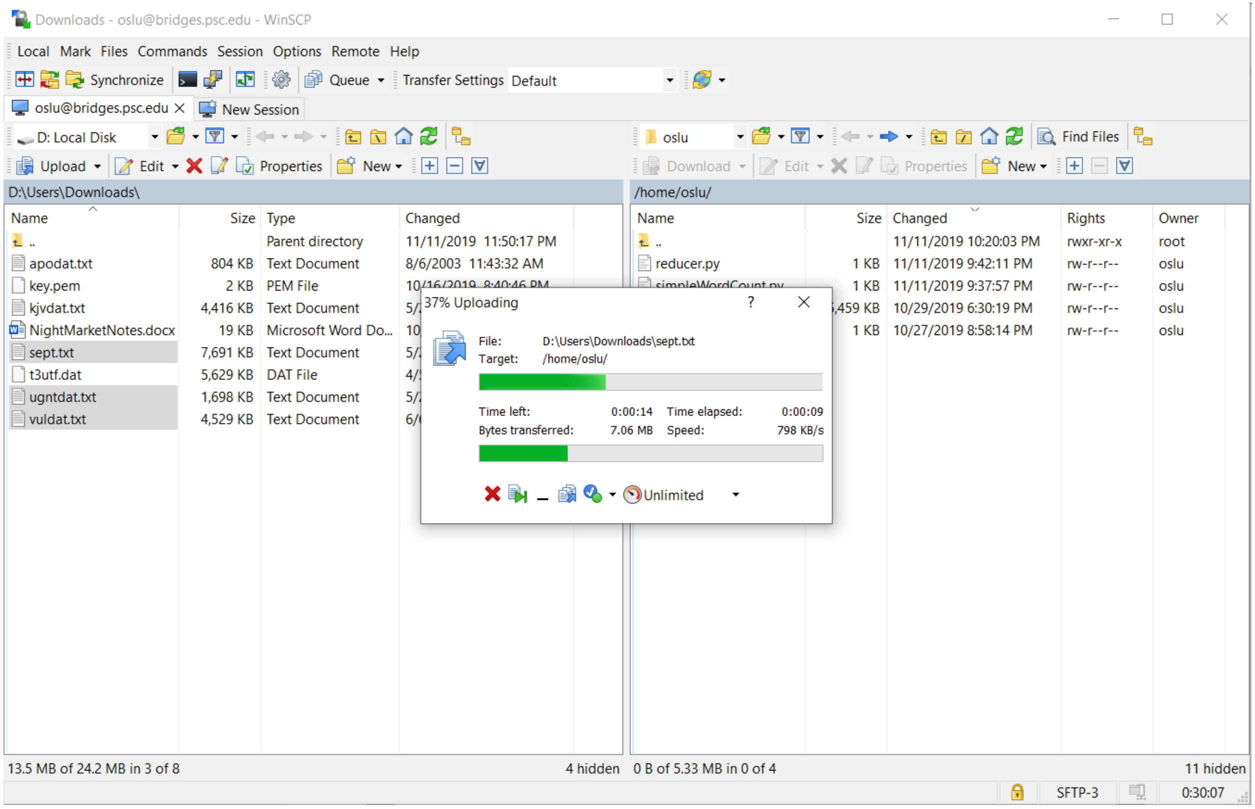
File protocol: SFTP

Host name: bridges.psc.edu

Port number: 2222

User name: oslu

Password: ●●●●●●●●●●●●



D:\Users\Downloads\

Name	Size	Type	Changed
..		Parent directory	11/11/2019 11:50:17 PM
apodat.txt	804 KB	Text Document	8/6/2003 11:43:32 AM
key.pem	2 KB	PEM File	10/16/2019 8:40:46 PM
kjvdat.txt	4,416 KB	Text Document	5/
NightMarketNotes.docx	19 KB	Microsoft Word Do...	10
sept.txt	7,691 KB	Text Document	5/
t3utf.dat	5,629 KB	DAT File	4/
ugntdat.txt	1,698 KB	Text Document	5/
vuldat.txt	4,529 KB	Text Document	6/

/home/oslu/

Name	Size	Changed	Rights	Owner
..		11/11/2019 10:20:03 PM	rwxr-xr-x	root
reducer.py	1 KB	11/11/2019 9:42:11 PM	rw-r--r--	oslu
simpleWordCount.py	1 KB	11/11/2019 9:37:57 PM	rw-r--r--	oslu
	,459 KB	10/29/2019 6:30:19 PM	rw-r--r--	oslu
	1 KB	10/27/2019 8:58:14 PM	rw-r--r--	oslu

37% Uploading

File: D:\Users\Downloads\sept.txt
Target: /home/oslu/

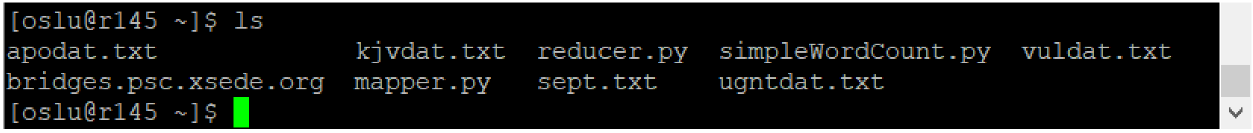
Time left: 0:00:14 Time elapsed: 0:00:09
Bytes transferred: 7.06 MB Speed: 798 KB/s

Unlimited

13.5 MB of 24.2 MB in 3 of 8 4 hidden 0 B of 5.33 MB in 0 of 4 11 hidden

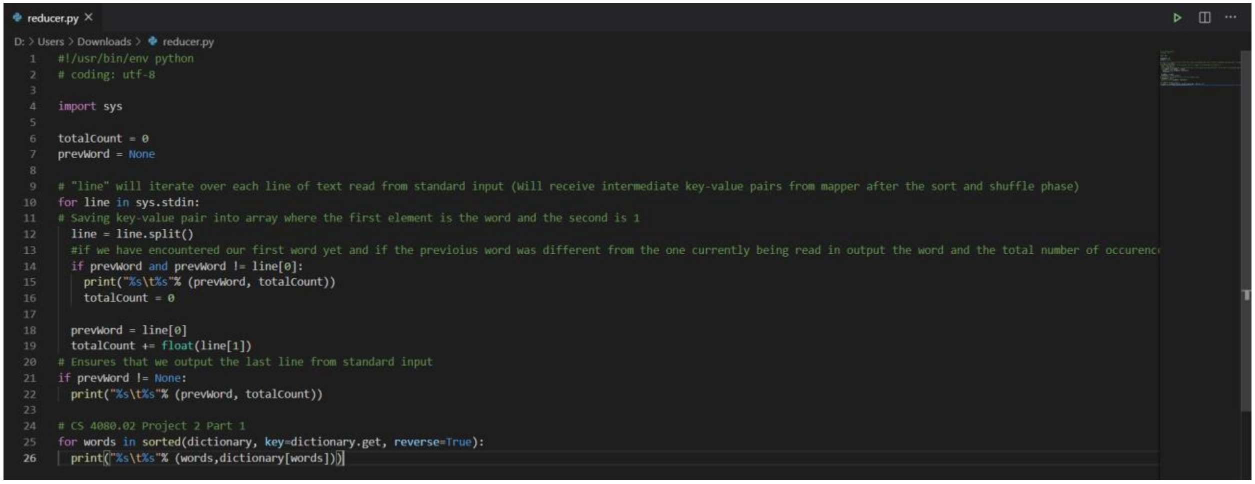
SFTP-3 0:30:07

Confirming that the file has been saved:



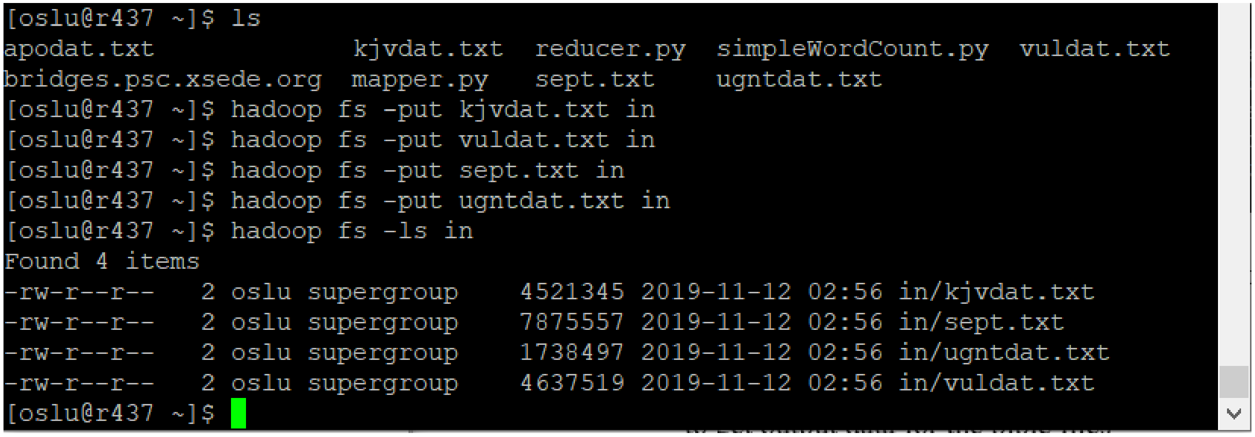
```
[oslu@r145 ~]$ ls
apodat.txt              kjvdat.txt  reducer.py  simpleWordCount.py  vuldat.txt
bridges.psc.xsede.org  mapper.py   sept.txt    ugntdat.txt
[oslu@r145 ~]$ 
```

4. I edited my reducer.py:



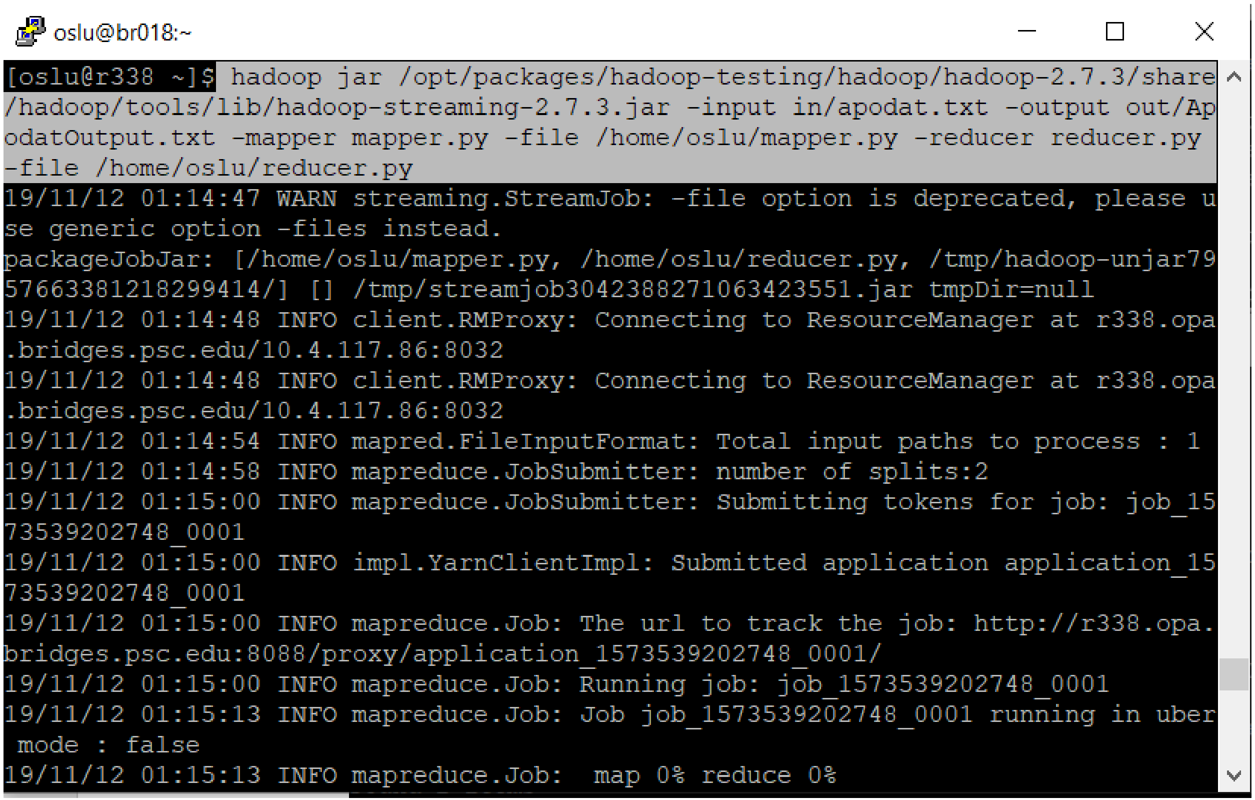
```python
#!/usr/bin/env python
# coding: utf-8

import sys

totalCount = 0
prevWord = None

# "line" will iterate over each line of text read from standard input (Will receive intermediate key-value pairs from mapper after the sort and shuffle phase)
for line in sys.stdin:
# Saving key-value pair into array where the first element is the word and the second is 1
  line = line.split()
  #if we have encountered our first word yet and if the previoius word was different from the one currently being read in output the word and the total number of occurenc
  if prevWord and prevWord != line[0]:
    print("%s\t%s"% (prevWord, totalCount))
    totalCount = 0

  prevWord = line[0]
  totalCount += float(line[1])
# Ensures that we output the last line from standard input
if prevWord != None:
  print("%s\t%s"% (prevWord, totalCount))

# CS 4080.02 Project 2 Part 1
for words in sorted(dictionary, key=dictionary.get, reverse=True):
  print("%s\t%s"% (words,dictionary[words]))
```

5. I also scped the reducer.py over with winscp
6. Startup Hadoop (interact -N 4 -t 01:00:00, module load hadoop, start-hadoop.sh)
7. Load all bible text files and reducer.py to the HDFS storing it in the 'in' directory

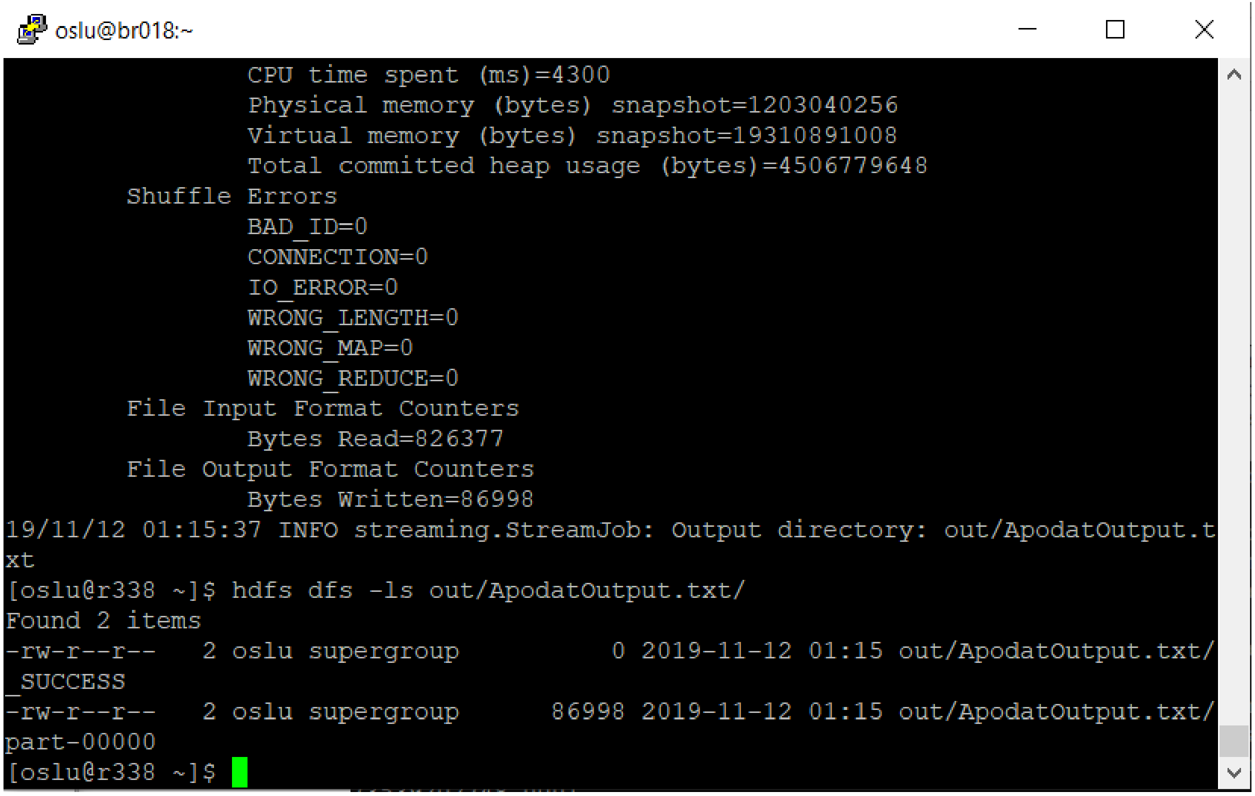


```
[oslu@r437 ~]$ ls
apodat.txt              kjvdat.txt  reducer.py  simpleWordCount.py  vuldat.txt
bridges.psc.xsede.org  mapper.py   sept.txt    ugntdat.txt
[oslu@r437 ~]$ hadoop fs -put kjvdat.txt in
[oslu@r437 ~]$ hadoop fs -put vuldat.txt in
[oslu@r437 ~]$ hadoop fs -put sept.txt in
[oslu@r437 ~]$ hadoop fs -put ugntdat.txt in
[oslu@r437 ~]$ hadoop fs -ls in
Found 4 items
-rw-r--r--   2 oslu supergroup    4521345 2019-11-12 02:56 in/kjvdat.txt
-rw-r--r--   2 oslu supergroup    7875557 2019-11-12 02:56 in/sept.txt
-rw-r--r--   2 oslu supergroup    1738497 2019-11-12 02:56 in/ugntdat.txt
-rw-r--r--   2 oslu supergroup    4637519 2019-11-12 02:56 in/vuldat.txt
[oslu@r437 ~]$ 
```

8. Run commands such as:
hadoop jar /opt/packages/hadoop-testing/hadoop/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.3.jar -input in/apodat.txt -output out/ApodatOutput.txt -mapper mapper.py -file /home/oslu/mapper.py -reducer reducer.py -file /home/oslu/reducer.py
to get output data for the bible files (same general commands)



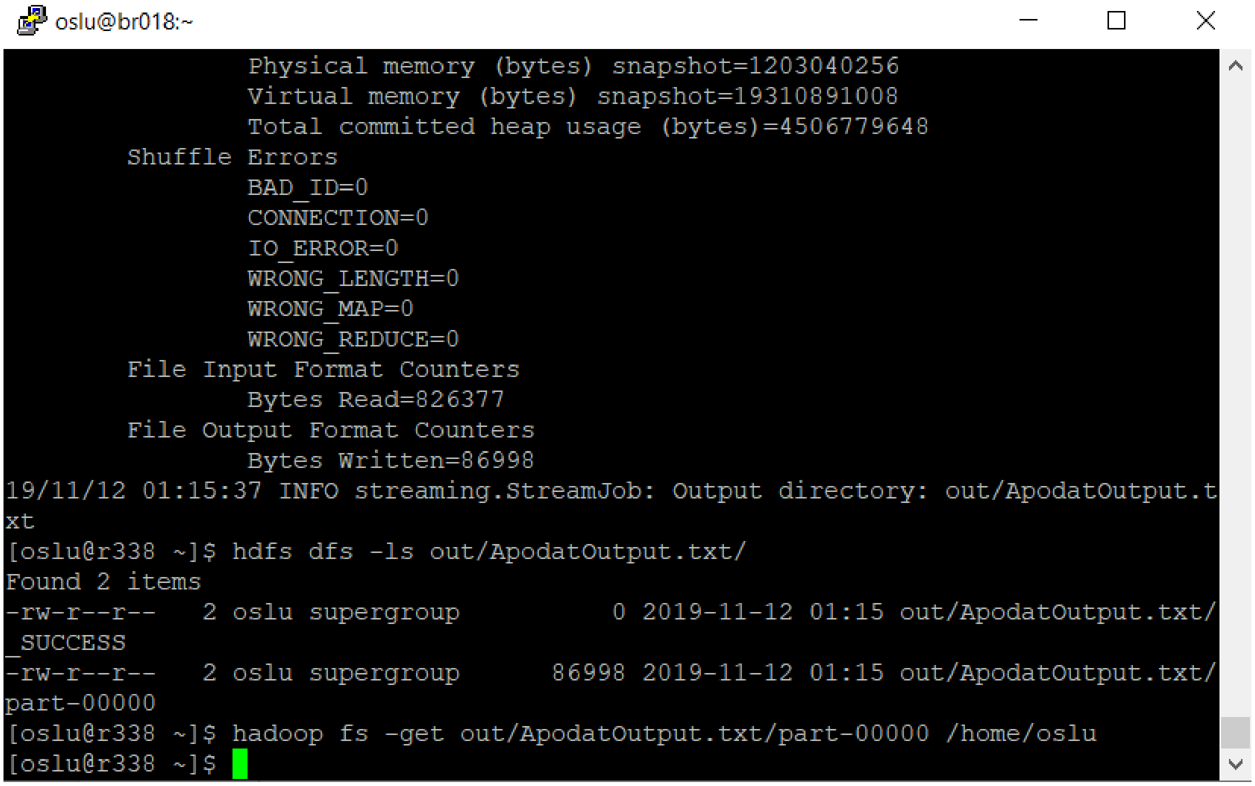
```
[oslu@r338 ~]$ hadoop jar /opt/packages/hadoop-testing/hadoop/hadoop-2.7.3/share
/hadoop/tools/lib/hadoop-streaming-2.7.3.jar -input in/apodat.txt -output out/Ap
odatOutput.txt -mapper mapper.py -file /home/oslu/mapper.py -reducer reducer.py
-file /home/oslu/reducer.py
19/11/12 01:14:47 WARN streaming.StreamJob: -file option is deprecated, please u
se generic option -files instead.
packageJobJar: [/home/oslu/mapper.py, /home/oslu/reducer.py, /tmp/hadoop-unjar79
57663381218299414/] [] /tmp/streamjob3042388271063423551.jar tmpDir=null
19/11/12 01:14:48 INFO client.RMProxy: Connecting to ResourceManager at r338.opa
.bridges.psc.edu/10.4.117.86:8032
19/11/12 01:14:48 INFO client.RMProxy: Connecting to ResourceManager at r338.opa
.bridges.psc.edu/10.4.117.86:8032
19/11/12 01:14:54 INFO mapred.FileInputFormat: Total input paths to process : 1
19/11/12 01:14:58 INFO mapreduce.JobSubmitter: number of splits:2
19/11/12 01:15:00 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_15
73539202748_0001
19/11/12 01:15:00 INFO impl.YarnClientImpl: Submitted application application_15
73539202748_0001
19/11/12 01:15:00 INFO mapreduce.Job: The url to track the job: http://r338.opa.
bridges.psc.edu:8088/proxy/application_1573539202748_0001/
19/11/12 01:15:00 INFO mapreduce.Job: Running job: job_1573539202748_0001
19/11/12 01:15:13 INFO mapreduce.Job: Job job_1573539202748_0001 running in uber
 mode : false
19/11/12 01:15:13 INFO mapreduce.Job:  map 0% reduce 0%
```



```
                CPU time spent (ms)=4300
                Physical memory (bytes) snapshot=1203040256
                Virtual memory (bytes) snapshot=19310891008
                Total committed heap usage (bytes)=4506779648
        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=826377
        File Output Format Counters
                Bytes Written=86998
19/11/12 01:15:37 INFO streaming.StreamJob: Output directory: out/ApodatOutput.t
xt
[oslu@r338 ~]$ hdfs dfs -ls out/ApodatOutput.txt/
Found 2 items
-rw-r--r--   2 oslu supergroup          0 2019-11-12 01:15 out/ApodatOutput.txt/
_SUCCESS
-rw-r--r--   2 oslu supergroup      86998 2019-11-12 01:15 out/ApodatOutput.txt/
part-00000
[oslu@r338 ~]$ 
```

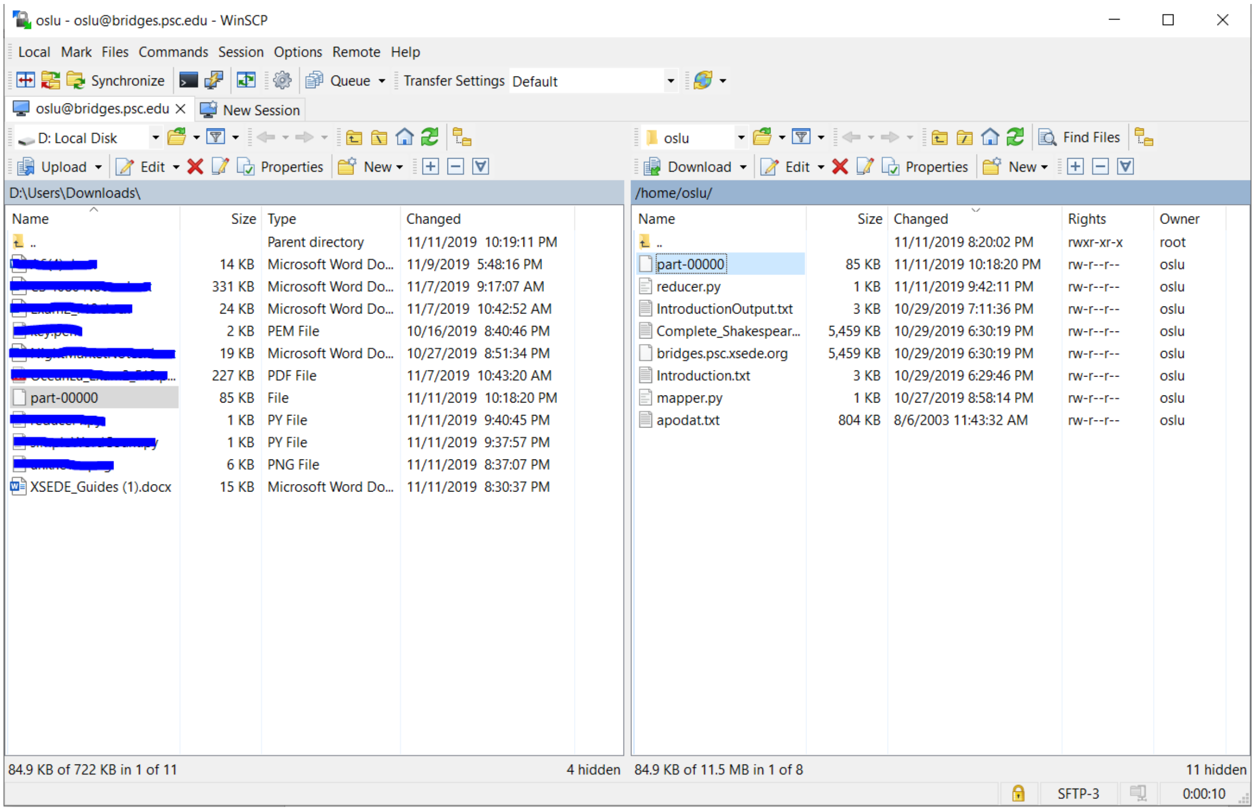
9. Retrieve the file with the command: hadoop fs -get out/ApodatOutput.txt/part-00000 /home/oslu
*Make sure part-00000 other files are deleted, this does not override current files that

already exist



```
                Physical memory (bytes) snapshot=1203040256
                Virtual memory (bytes) snapshot=19310891008
                Total committed heap usage (bytes)=4506779648
        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=826377
        File Output Format Counters
                Bytes Written=86998
19/11/12 01:15:37 INFO streaming.StreamJob: Output directory: out/ApodatOutput.t
xt
[oslu@r338 ~]$ hdfs dfs -ls out/ApodatOutput.txt/
Found 2 items
-rw-r--r--   2 oslu supergroup          0 2019-11-12 01:15 out/ApodatOutput.txt/
_SUCCESS
-rw-r--r--   2 oslu supergroup      86998 2019-11-12 01:15 out/ApodatOutput.txt/
part-00000
[oslu@r338 ~]$ hadoop fs -get out/ApodatOutput.txt/part-00000 /home/oslu
[oslu@r338 ~]$ 
```

10. WinSCP over it to local host



D:\Users\Downloads\

Name	Size	Type	Changed
..		Parent directory	11/11/2019 10:19:11 PM
[redacted]	14 KB	Microsoft Word Do...	11/9/2019 5:48:16 PM
[redacted]	331 KB	Microsoft Word Do...	11/7/2019 9:17:07 AM
[redacted]	24 KB	Microsoft Word Do...	11/7/2019 10:42:52 AM
[redacted]	2 KB	PEM File	10/16/2019 8:40:46 PM
[redacted]	19 KB	Microsoft Word Do...	10/27/2019 8:51:34 PM
[redacted]	227 KB	PDF File	11/7/2019 10:43:20 AM
part-00000	85 KB	File	11/11/2019 10:18:20 PM
[redacted]	1 KB	PY File	11/11/2019 9:40:45 PM
[redacted]	1 KB	PY File	11/11/2019 9:37:57 PM
[redacted]	6 KB	PNG File	11/11/2019 8:37:07 PM
XSEDE_Guides (1).docx	15 KB	Microsoft Word Do...	11/11/2019 8:30:37 PM

84.9 KB of 722 KB in 1 of 11 4 hidden

/home/oslu/

Name	Size	Changed	Rights	Owner
..		11/11/2019 8:20:02 PM	rwxr-xr-x	root
part-00000	85 KB	11/11/2019 10:18:20 PM	rw-r--r--	oslu
reducer.py	1 KB	11/11/2019 9:42:11 PM	rw-r--r--	oslu
IntroductionOutput.txt	3 KB	10/29/2019 7:11:36 PM	rw-r--r--	oslu
Complete_Shakespear...	5,459 KB	10/29/2019 6:30:19 PM	rw-r--r--	oslu
bridges.psc.xsede.org	5,459 KB	10/29/2019 6:30:19 PM	rw-r--r--	oslu
Introduction.txt	3 KB	10/29/2019 6:29:46 PM	rw-r--r--	oslu
mapper.py	1 KB	10/27/2019 8:58:14 PM	rw-r--r--	oslu
apodat.txt	804 KB	8/6/2003 11:43:32 AM	rw-r--r--	oslu

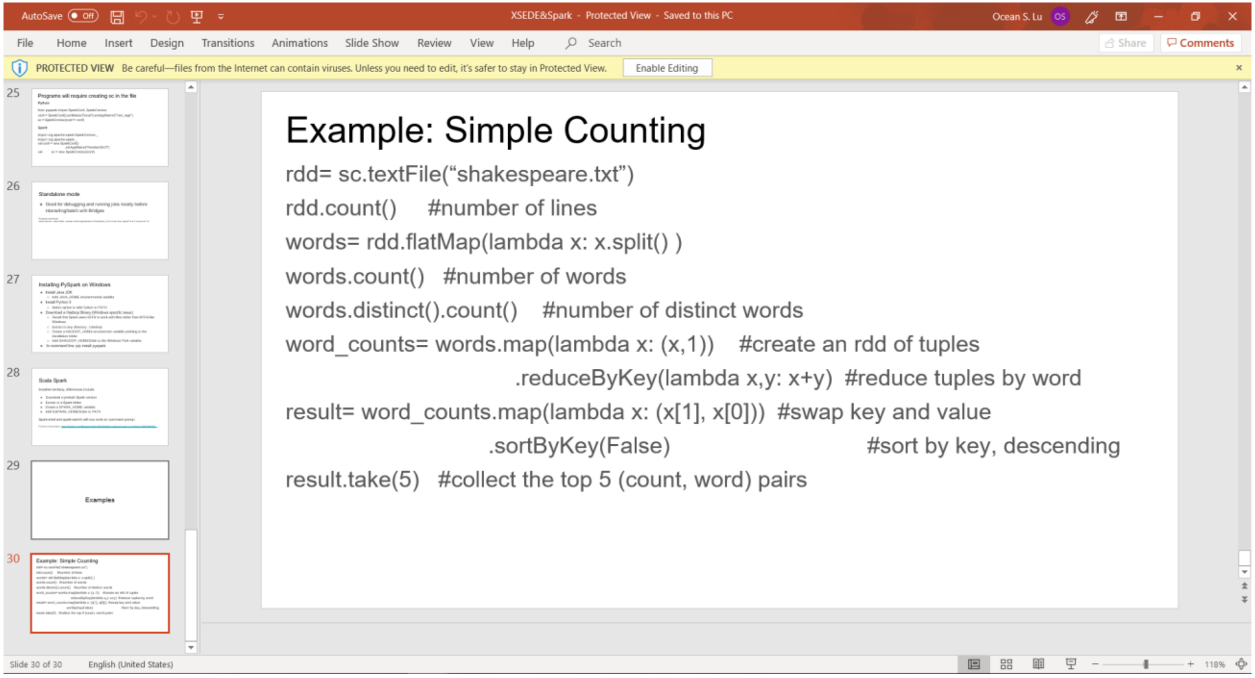
84.9 KB of 11.5 MB in 1 of 8 11 hidden

SFTP-3 0:00:10

Finished, now we have the output file of apodat.txt

Task 2:

1. Simple counting program given to us on in class slides with XSEDE & Spark notes, renamed “simpleWordCount.py”

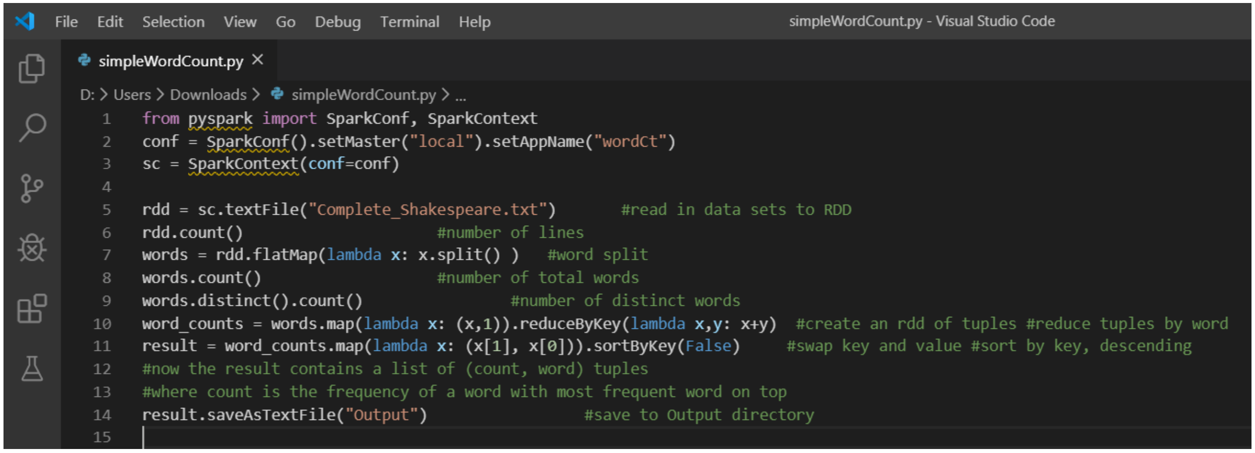


XSEDE&Spark - Protected View - Saved to this PC
File Home Insert Design Transitions Animations Slide Show Review View Help Search
PROTECTED VIEW Be careful—files from the Internet can contain viruses. Unless you need to edit, it's safer to stay in Protected View. Enable Editing
Slide 30 of 30 English (United States)

Example: Simple Counting

```
rdd= sc.textFile(“shakespeare.txt”)
rdd.count()   #number of lines
words= rdd.flatMap(lambda x: x.split() )
words.count()  #number of words
words.distinct().count()   #number of distinct words
word_counts= words.map(lambda x: (x,1))   #create an rdd of tuples
                    .reduceByKey(lambda x,y: x+y)  #reduce tuples by word
result= word_counts.map(lambda x: (x[1], x[0]))  #swap key and value
                  .sortByKey(False)                    #sort by key, descending
result.take(5)  #collect the top 5 (count, word) pairs
```

2. Take the SimpleWordCount.py and Complete_Shakespeare.txt and Bibles data files, and winscp it onto the drive
3. Edit simpleWordCount.py for Complete_Shakespeare.txt

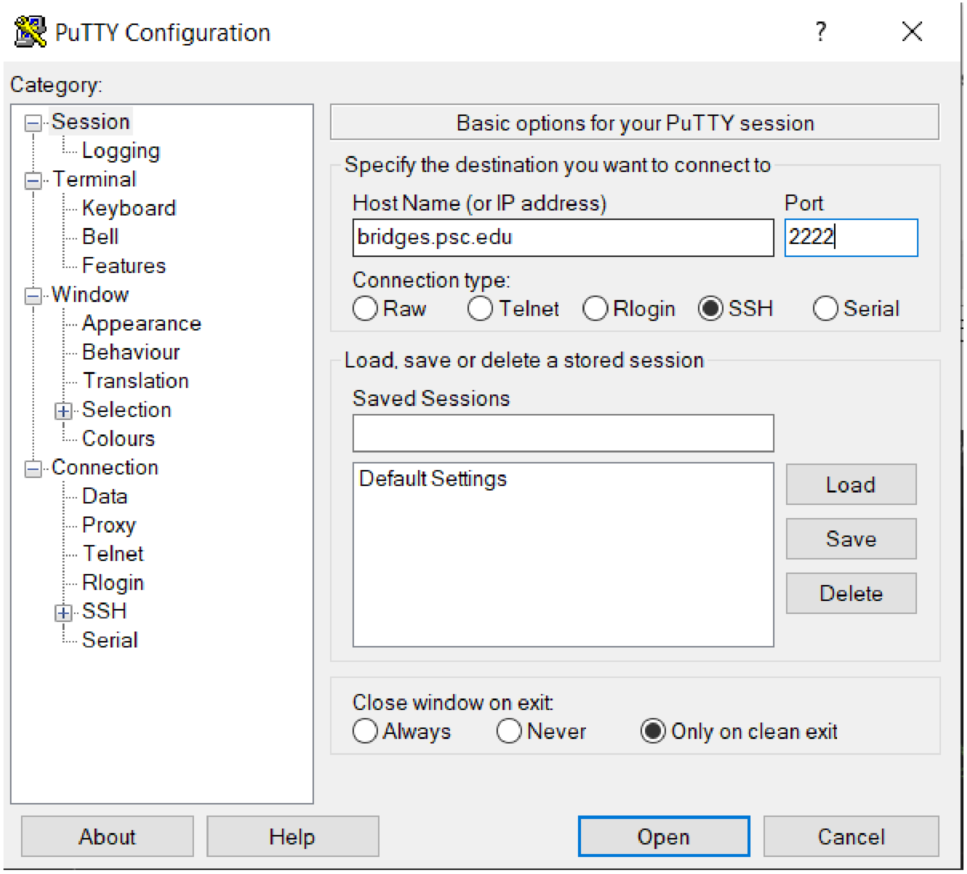


File Edit Selection View Go Debug Terminal Help
simpleWordCount.py - Visual Studio Code
simpleWordCount.py
D: > Users > Downloads > simpleWordCount.py > ...

```python
from pyspark import SparkConf, SparkContext
conf = SparkConf().setMaster("local").setAppName("wordCt")
sc = SparkContext(conf=conf)

rdd = sc.textFile("Complete_Shakespeare.txt")       #read in data sets to RDD
rdd.count()                         #number of lines
words = rdd.flatMap(lambda x: x.split() )    #word split
words.count()                       #number of total words
words.distinct().count()                    #number of distinct words
word_counts = words.map(lambda x: (x,1)).reduceByKey(lambda x,y: x+y)  #create an rdd of tuples #reduce tuples by word
result = word_counts.map(lambda x: (x[1], x[0])).sortByKey(False)     #swap key and value #sort by key, descending
#now the result contains a list of (count, word) tuples
#where count is the frequency of a word with most frequent word on top
result.saveAsTextFile("Output")                #save to Output directory
```

Edit the sc.textfile(“”) appropriately when going through the bible data files

4. Use puTTY to SSH and login appropriately



Specify the destination you want to connect to

Host Name (or IP address): bridges.psc.edu

Port: 2222

Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

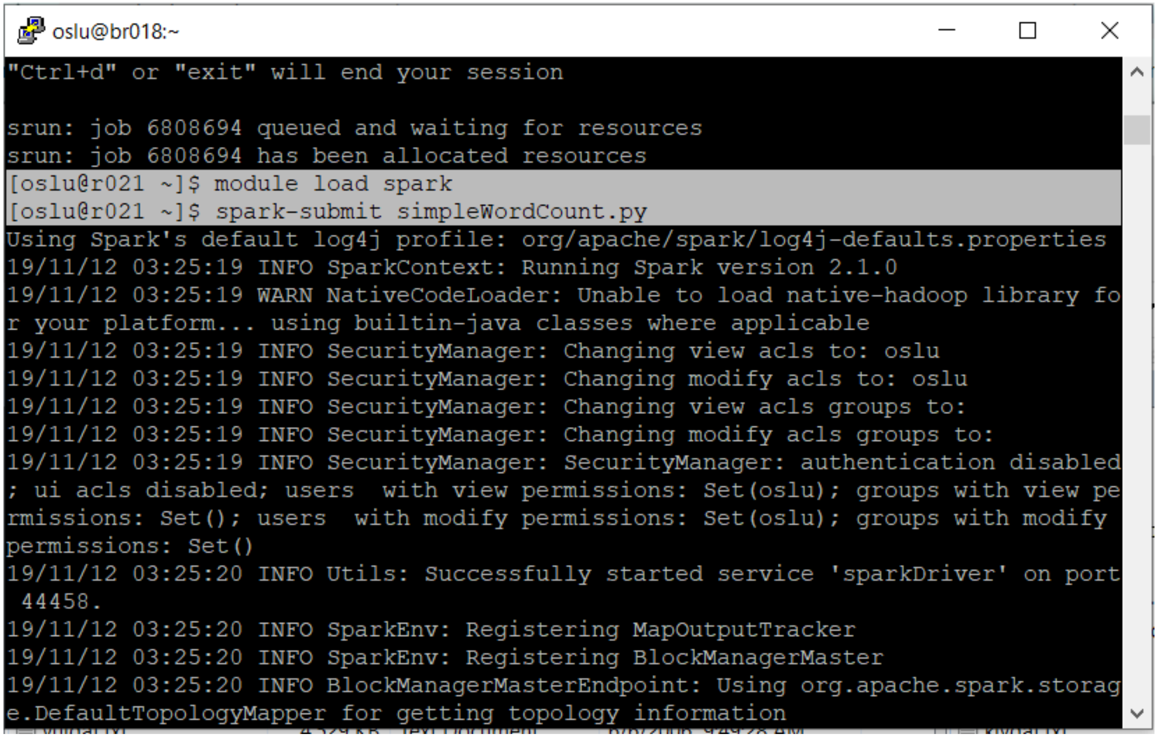
Load, save or delete a stored session

Saved Sessions

Default Settings

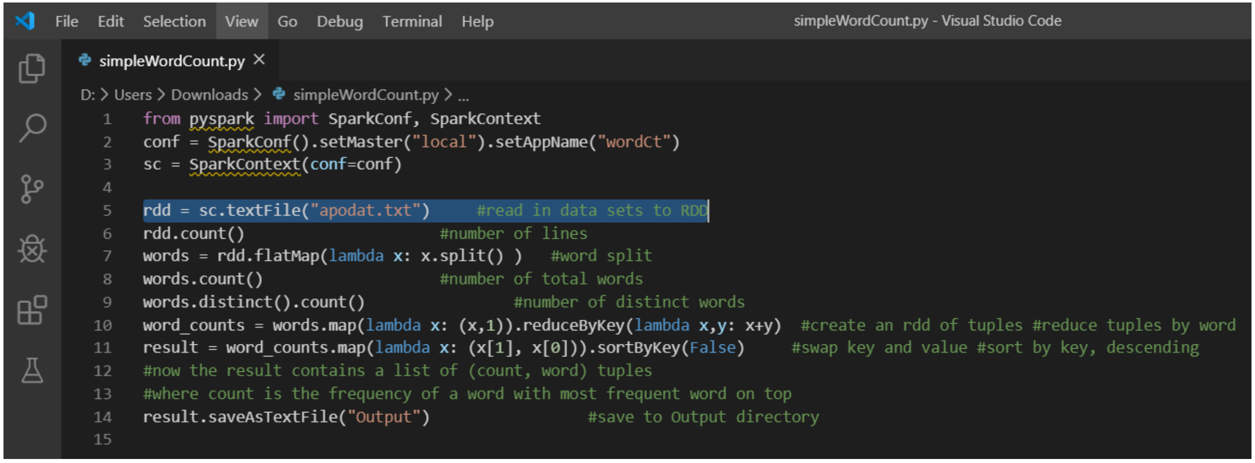
Close window on exit: ○ Always ○ Never ● Only on clean exit

5. Go into Hadoop and load spark and have the command: spark-submit simpleWordCount.py



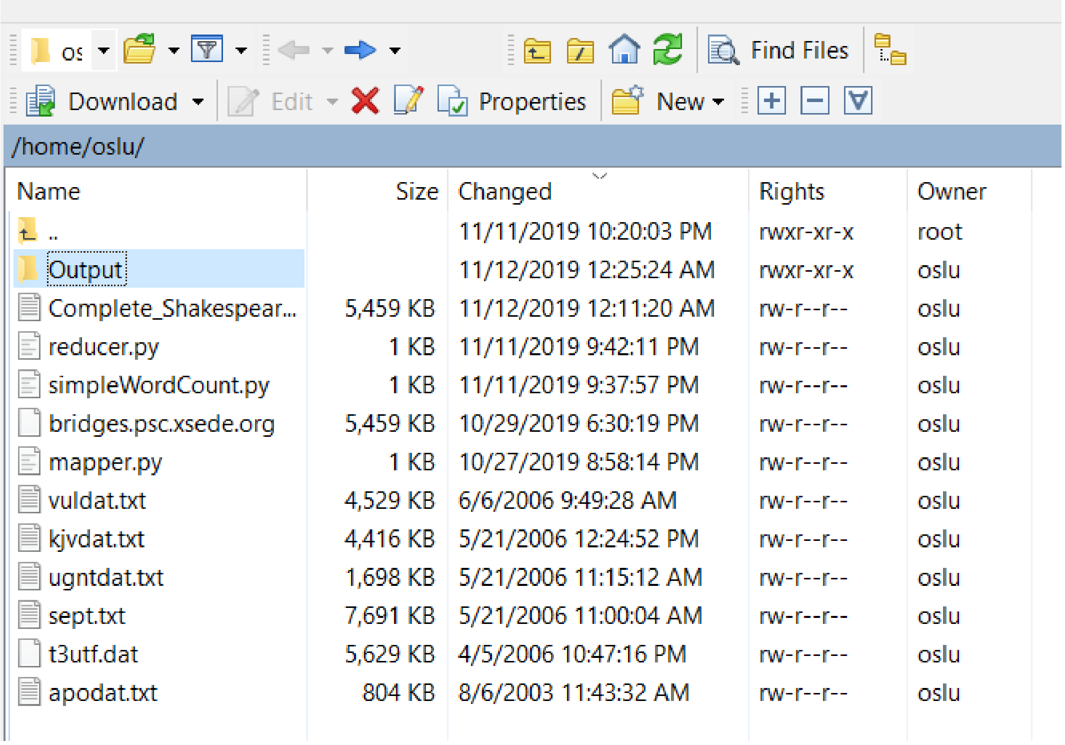
```
"Ctrl+d" or "exit" will end your session

srun: job 6808694 queued and waiting for resources
srun: job 6808694 has been allocated resources
[oslu@r021 ~]$ module load spark
[oslu@r021 ~]$ spark-submit simpleWordCount.py
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
19/11/12 03:25:19 INFO SparkContext: Running Spark version 2.1.0
19/11/12 03:25:19 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
19/11/12 03:25:19 INFO SecurityManager: Changing view acls to: oslu
19/11/12 03:25:19 INFO SecurityManager: Changing modify acls to: oslu
19/11/12 03:25:19 INFO SecurityManager: Changing view acls groups to:
19/11/12 03:25:19 INFO SecurityManager: Changing modify acls groups to:
19/11/12 03:25:19 INFO SecurityManager: SecurityManager: authentication disabled
; ui acls disabled; users  with view permissions: Set(oslu); groups with view pe
rmissions: Set(); users  with modify permissions: Set(oslu); groups with modify
permissions: Set()
19/11/12 03:25:20 INFO Utils: Successfully started service 'sparkDriver' on port
 44458.
19/11/12 03:25:20 INFO SparkEnv: Registering MapOutputTracker
19/11/12 03:25:20 INFO SparkEnv: Registering BlockManagerMaster
19/11/12 03:25:20 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storag
e.DefaultTopologyMapper for getting topology information
```

6. Edit simpleWordCount.py’s sc.textfile approrpiately when changing files (into bible files)


File
Edit
Selection
View
Go
Debug
Terminal
Help
simpleWordCount.py - Visual Studio Code
simpleWordCount.py
D: > Users > Downloads > simpleWordCount.py > ...

```python
from pyspark import SparkConf, SparkContext
conf = SparkConf().setMaster("local").setAppName("wordCt")
sc = SparkContext(conf=conf)

rdd = sc.textFile("apodat.txt")     #read in data sets to RDD
rdd.count()                     #number of lines
words = rdd.flatMap(lambda x: x.split() )   #word split
words.count()                   #number of total words
words.distinct().count()                #number of distinct words
word_counts = words.map(lambda x: (x,1)).reduceByKey(lambda x,y: x+y)  #create an rdd of tuples #reduce tuples by word
result = word_counts.map(lambda x: (x[1], x[0])).sortByKey(False)     #swap key and value #sort by key, descending
#now the result contains a list of (count, word) tuples
#where count is the frequency of a word with most frequent word on top
result.saveAsTextFile("Output")                 #save to Output directory

```

7. Do it for all files, and output should be in a folder form.



Name	Size	Changed	Rights	Owner
..		11/11/2019 10:20:03 PM	rwxr-xr-x	root
Output		11/12/2019 12:25:24 AM	rwxr-xr-x	oslu
Complete_Shakespear...	5,459 KB	11/12/2019 12:11:20 AM	rw-r--r--	oslu
reducer.py	1 KB	11/11/2019 9:42:11 PM	rw-r--r--	oslu
simpleWordCount.py	1 KB	11/11/2019 9:37:57 PM	rw-r--r--	oslu
bridges.psc.xsede.org	5,459 KB	10/29/2019 6:30:19 PM	rw-r--r--	oslu
mapper.py	1 KB	10/27/2019 8:58:14 PM	rw-r--r--	oslu
vuldat.txt	4,529 KB	6/6/2006 9:49:28 AM	rw-r--r--	oslu
kjvdat.txt	4,416 KB	5/21/2006 12:24:52 PM	rw-r--r--	oslu
ugntdat.txt	1,698 KB	5/21/2006 11:15:12 AM	rw-r--r--	oslu
sept.txt	7,691 KB	5/21/2006 11:00:04 AM	rw-r--r--	oslu
t3utf.dat	5,629 KB	4/5/2006 10:47:16 PM	rw-r--r--	oslu
apodat.txt	804 KB	8/6/2003 11:43:32 AM	rw-r--r--	oslu