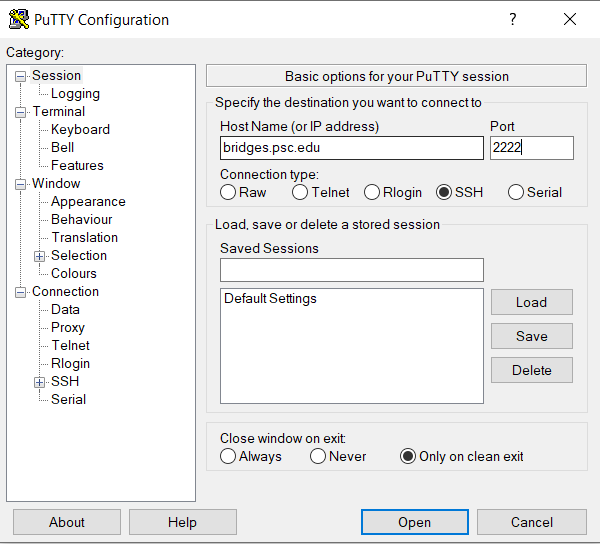
Ocean Lu  
CS 4080.02  
Professor Yang  
11.12.2019

Project 2: MapReduce Programming on Hadoop and Spark

Task 1:

1. Log into bridges.psc.edu port 2222; used puTTY  
   
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  
   A screenshot of a social media post

   Description automatically generated  
   A screenshot of a computer

   Description automatically generated  
   Confirming that the file has been saved:  
   A close up of a screen

   Description automatically generated
4. I edited my reducer.py:  
   A screenshot of a computer screen

   Description automatically generated
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  
   A picture containing text

   Description automatically generated
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)  
   A screenshot of a cell phone

   Description automatically generated  
   A screenshot of a cell phone

   Description automatically generated
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  
   A screenshot of a cell phone

   Description automatically generated
10. WinSCP over it to local host  
    A screenshot of a social media post

    Description automatically generated  
    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”  
   A screenshot of a cell phone

   Description automatically generated
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  
   A screenshot of a cell phone

   Description automatically generated  
   Edit the sc.textfile(“”) appropriately when going through the bible data files
4. Use puTTY to SSH and login appropriately  
   A screenshot of a cell phone

   Description automatically generated
5. Go into Hadoop and load spark and have the command: spark-submit simpleWordCount.py  
   A screenshot of a computer

   Description automatically generated
6. Edit simpleWordCount.py’s sc.textfile approrpiately when changing files (into bible files)  
   A screenshot of a cell phone

   Description automatically generated
7. Do it for all files, and output should be in a folder form.   
   A screenshot of a social media post

   Description automatically generated