Tracy Michaels

Tmichaels1 002430918

BigData Assignment 3

Setting up spark:

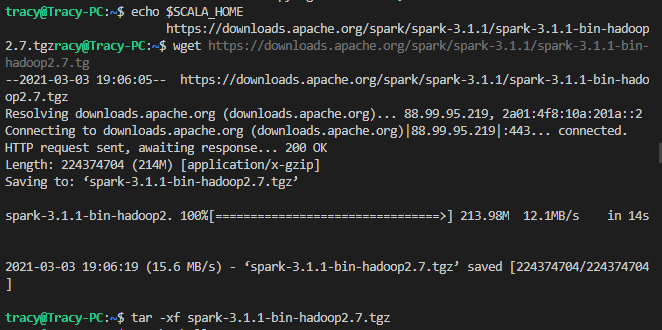
* First I made sure everything was up to date by running commands sudo apt-get update and sudo apt-get upgrade
* Then I installed scala using sudo apt install scala and confirmed the installation using scala -version



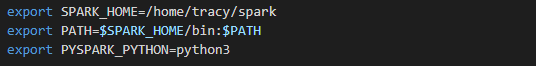
* I then updated the .bashrc to include the correct paths for scala



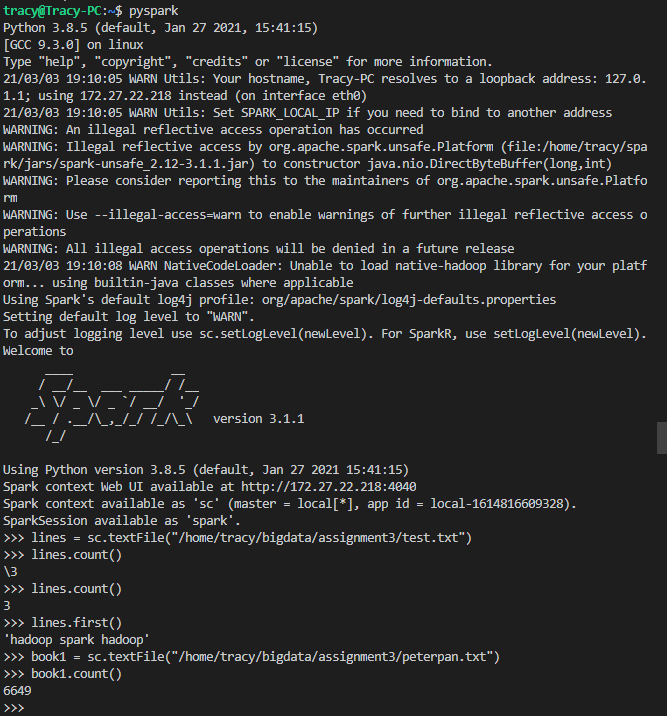
* I then downloaded spark from apache.org using a wget command and a link I found on their website
* Then I unzipped the file using tar and changed the name of the new folder to spark in the file explorer of vscode



* After downloading and unzipping these files I added their path to the .bashrc file



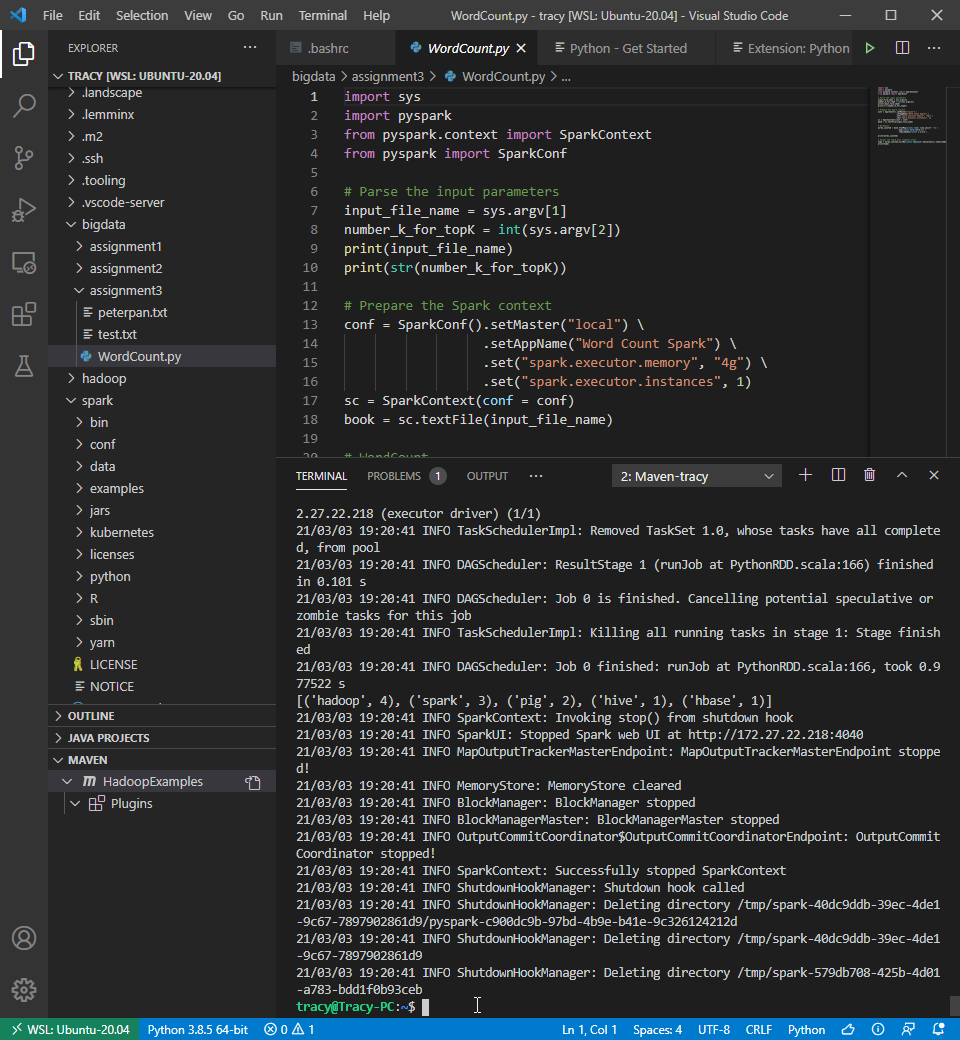
* To confirm that everything is working properly I used the command pyspark to open an interactive python spark shell and tried a few command on the files for test.txt and peterman.txt



* These commands create an RDD called lines and book1 by passing the file path of the files to the textFile() function of spark context object
* After downloading the WordCount.py from iCollege I ran the command

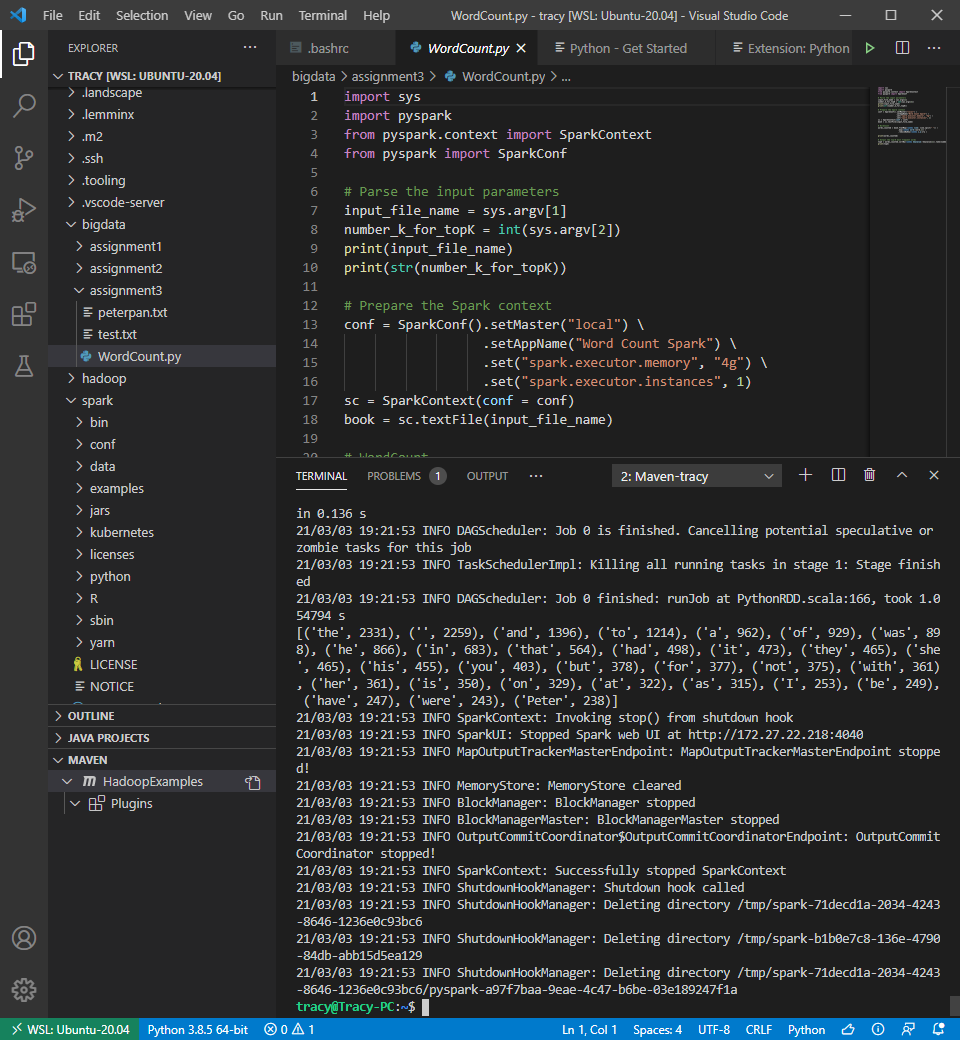
spark-submit /home/tracy/bigdata/assignment3/WordCount.py /home/tracy/bigdata/assignment3/test.txt 5

* Where the first argument is the python file that spark is going to be running, the second argument is the location of the data that the WordCount.py is going to act on and the third argument is the number of topK words that will be returned with their counts



* In this case what was returned was [('hadoop', 4), ('spark', 3), ('pig', 2), ('hive', 1), ('hbase', 1)] which are the 5 most used words in test.txt file
* I repeated this same process but for the peterpan.txt file using command

spark-submit /home/tracy/bigdata/assignment3/WordCount.py /home/tracy/bigdata/assignment3/peterpan.txt 30

* Where each argument functions the same as explained above
* This time though it will return the 30 most used words in the file
* Output: [('the', 2331), ('', 2259), ('and', 1396), ('to', 1214), ('a', 962), ('of', 929), ('was', 898), ('he', 866), ('in', 683), ('that', 564), ('had', 498), ('it', 473), ('they', 465), ('she', 465), ('his', 455), ('you', 403), ('but', 378), ('for', 377), ('not', 375), ('with', 361), ('her', 361), ('is', 350), ('on', 329), ('at', 322), ('as', 315), ('I', 253), ('be', 249), ('have', 247), ('were', 243), ('Peter', 238)]