ASSIGNMENT-3

BIG DATA PROGRAMING

1:

Installation of spark in kvm. As a First step check for updates using **$ sudo apt update**. Before installing spark, the spark is written in scale so scale need to be installed first. This can be done using $**sudo apt install scala** in the terminal. Once scala is installed check for its version using **$scala –version** and to conform whether it's installed successfully.

```
vmalapati1@kvm_vmalapati1:~$ scala -version
Scala code runner version 2.11.12 -- Copyright 2002-2017, LAMP/EPFL
vmalapati1@kvm_vmalapati1:~$
```

Fig: scale version

Once scale is installed update the .bachrc such that the OS will recognize the scale using below commands adding to .bashrc file in bash script.

```
export SCALA_HOME=/home/vmalapati1/SCALA_HOME

export PATH=$SCALA_HOME/bin:$PATH
```

Installation of spark can be done in two ways, Chosee the specific version of spark as per requirement from the Apache spark org and download the .tgz spark file directly into local PC or KVM. Once downloaded into local PC unzip it in local pc and copy the entire spark folder into KVM using VS code in the home of kvm. Other way is directly download into kvm and unzip it from the terminal of kvm into home of kvm.

Once spark is installed update the .bashrc such that OS will know the spark and also python3 for spark, by adding below lines of code.

```
export SPARK_HOME=/home/vmalapati1/spark
export PATH=$SPARK_HOME/bin:$PATH

export PYSPARK_PYTHON=python3
```

Once the code is added just run **$source~/.bashrc**

After these steps open spark terminal by typing pyspark command in terminal, if there exists any error related to permission use this command in terminal to give **permission chmod +x /home/rob/spark/bin/***

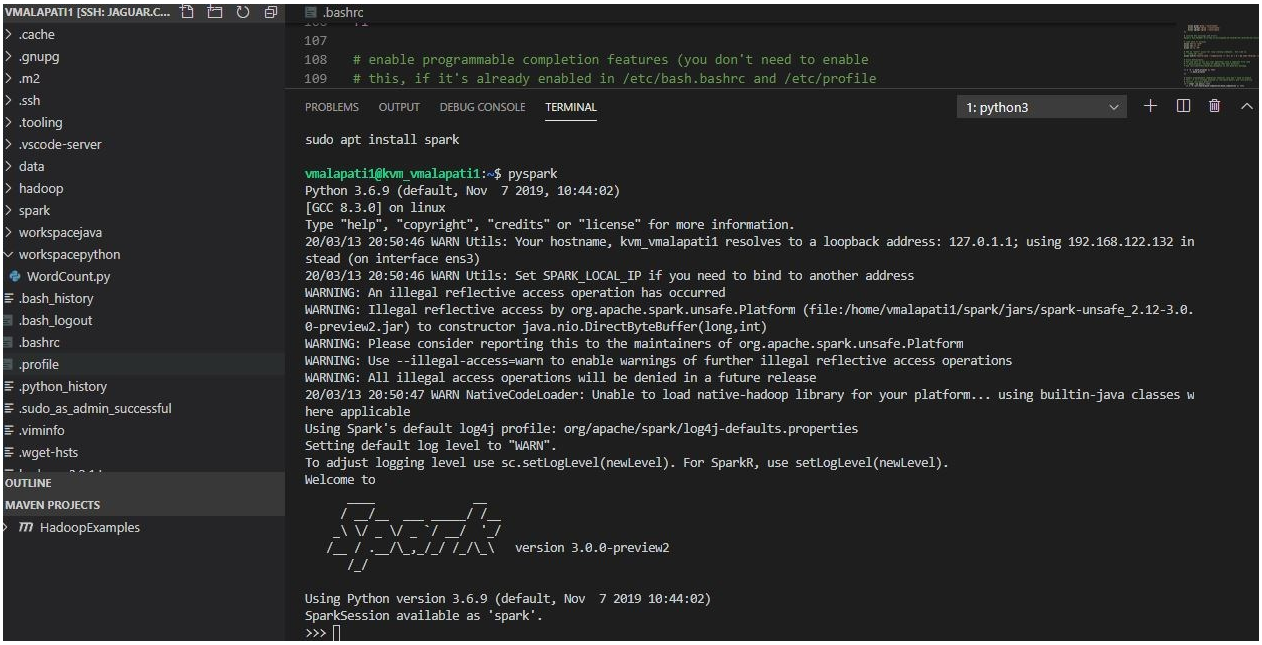

Fig: showing spark shell opened in terminal of kvm

2.

Create a folder workspacepython to store the .py files

Get the wordcount.py file and running it on test and peterpan text files and printing the top K most frequency words on the terminal

The code takes Two input system arguments one the data file path and the other is the K value how many top k words to be printed out. The command for running python file in spark is as shown below. A python file can be run on spark by using spark-submit file location.

spark-submit /home/vmalapati1/workspacepython/WordCount.py /home/vmalapati1/data/test.txt 5

spark-submit /home/vmalapati1/workspacepython/WordCount.py /home/vmalapati1/data/peterpan.txt 30

The Three main blocks of code in wordcount.py are

A. Configure the spark APP and set this configuration to context,
B. Read the text file and apply transformation and actions on it, I mean map and reduce and splitting line by line with space
C. After this sort the top K frequency words.

The output pictures after running wordcount on above to files are shown below

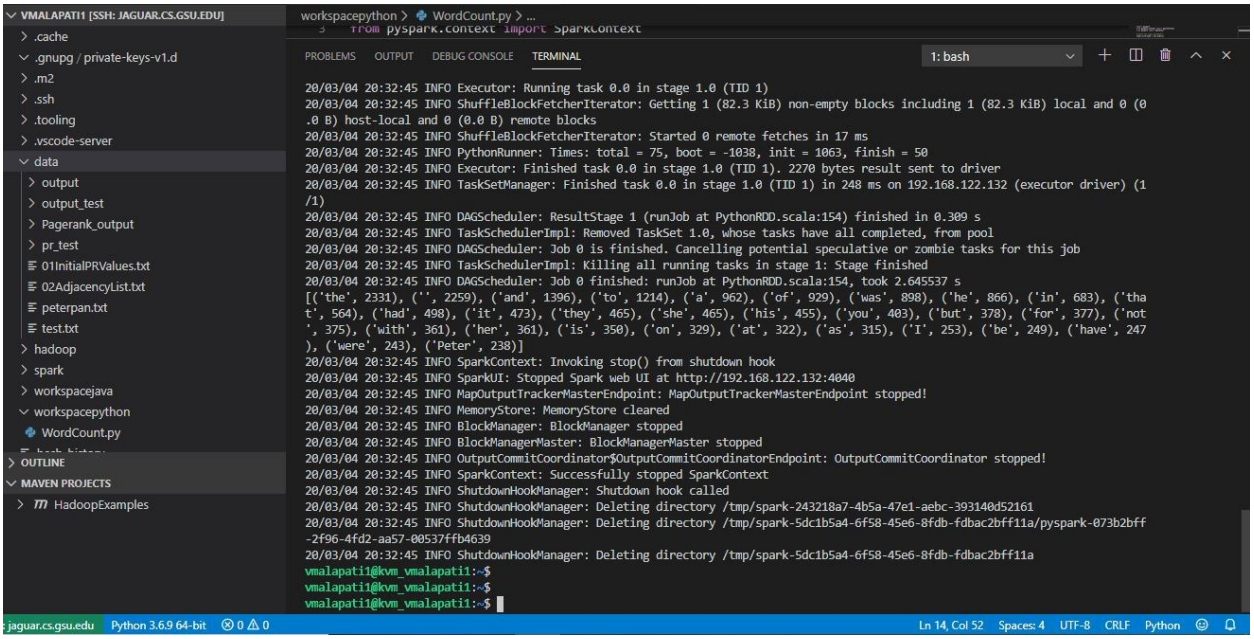

Fig: showing result on peterpan

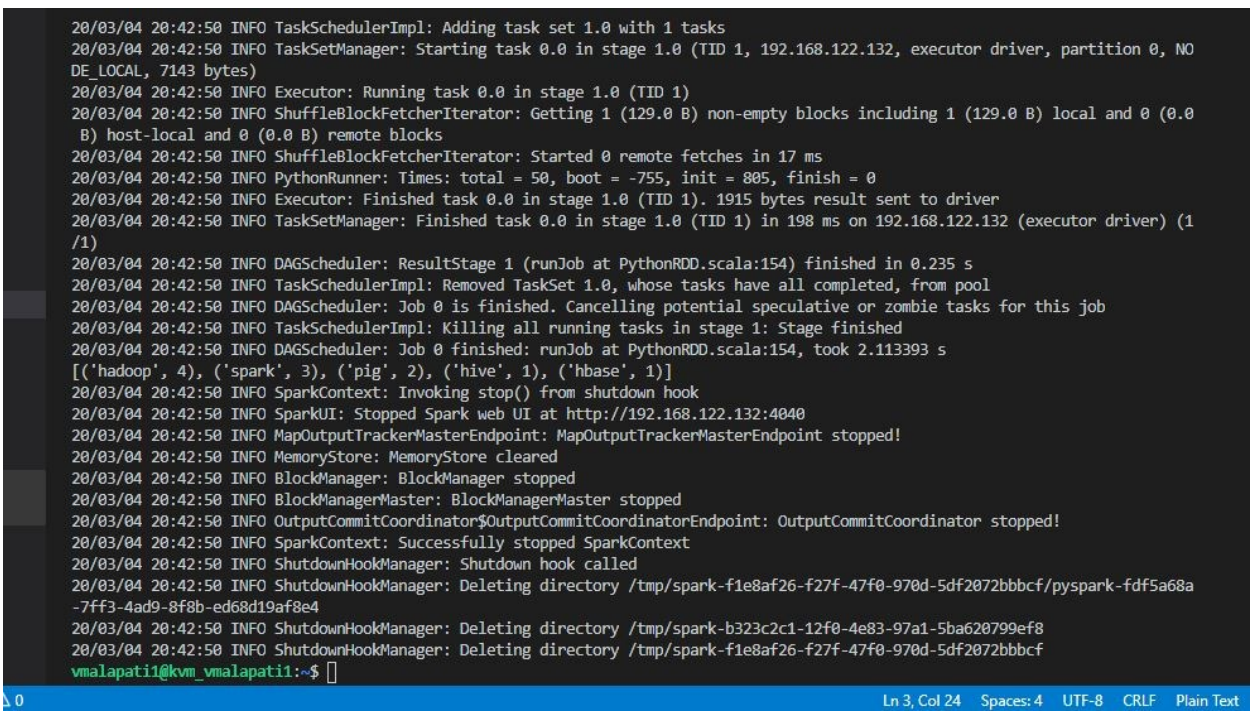

Fig: showing result on text

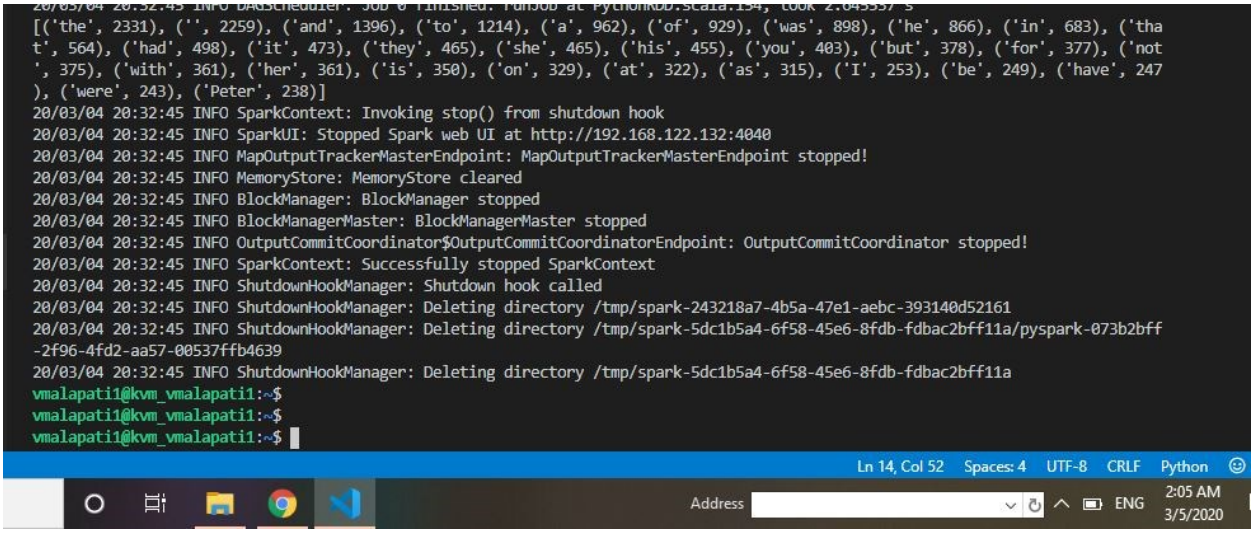

Fig: sowing result on peterpan

The output results consists of printed top K most frequency words in a list of tuples as shown in above figures.

As Further more I have implemented one more system argument that takes path for writing the output of wordcount i.e top K most frequency words in a text file in the data/sparkoutput folder these output files contain output as follows

New command for running wordcount with output file path:

spark-submit /home/vmalapati1/workspacepython/WordCount.py /home/vmalapati1/data/peterpan.txt 30 /home/vmalapati1/data/sparkoutput/peterpan_output.text

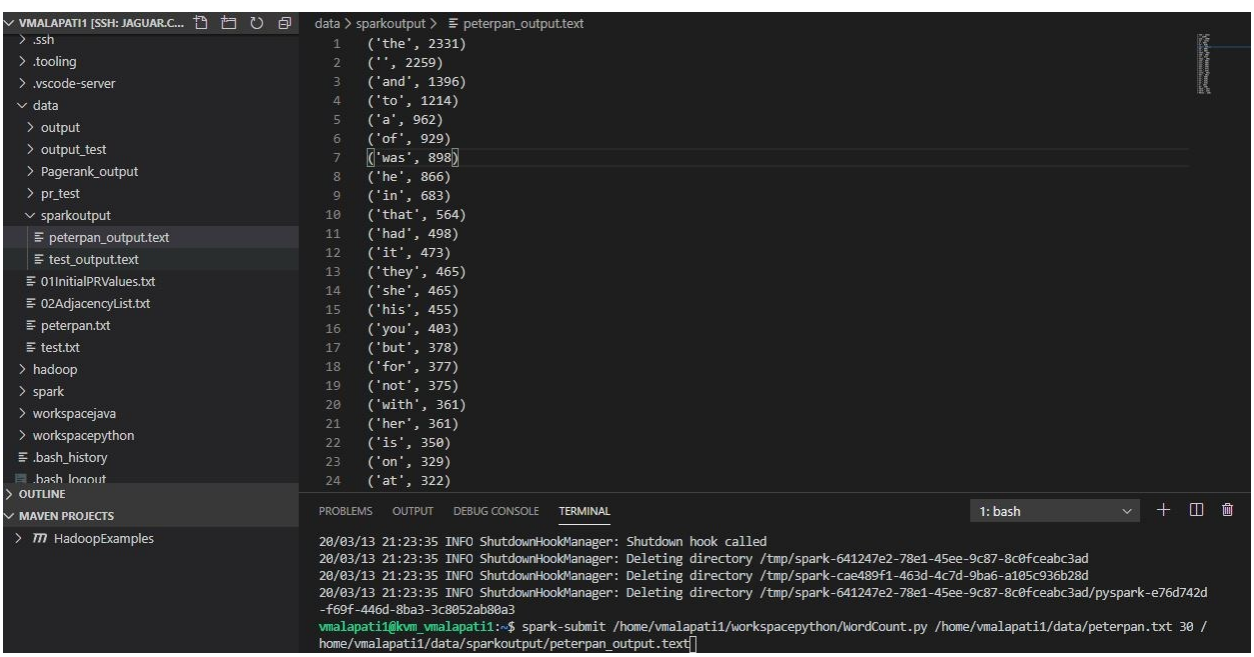

Fig: showing output file for peterpan data in KVM

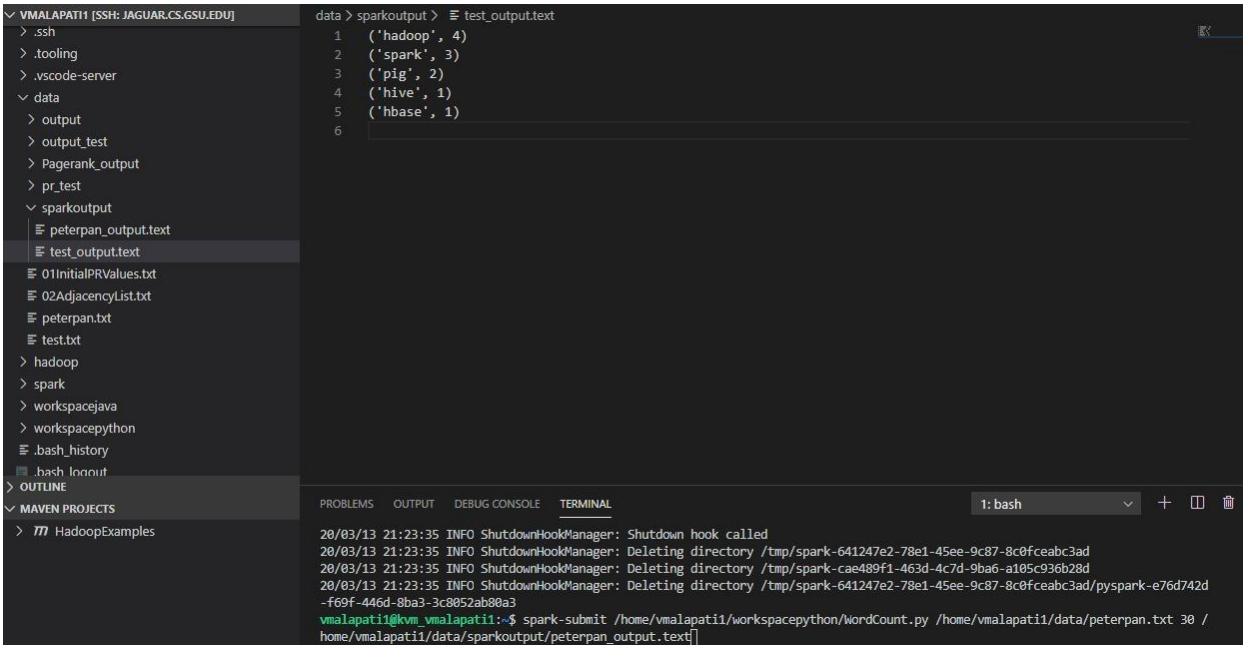

Fig: Showing output file for test