Setting up spark on ubuntu:

1. Download Spark
    a. Apache Spark is an open-source distributed general-purpose cluster-computing framework. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance.
    b. Download spark packages using the command - **wget http://apache-mirror.8birdsvideo.com/spark/spark-3.0.0-preview2/spark-3.0.0-preview2-bin-hadoop3.2.tgz**

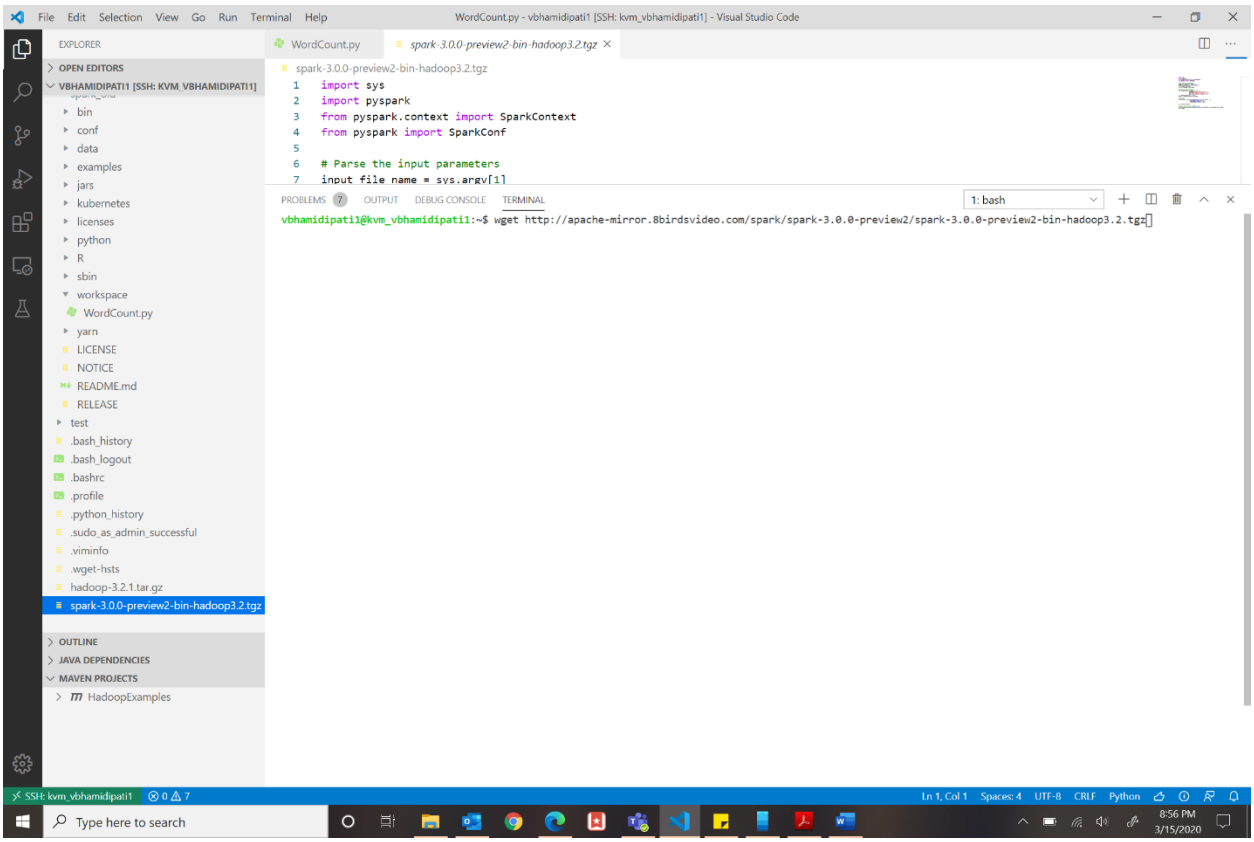

Fig 1. Downloading spark packages onto the KVM

2. Installing the downloaded package
    a. The package is in the .tgz format. After downloading we must extract the package.
    b. The spark packages are extracted using the following command - **tar -xvf spark-3.0.0-preview2-bin-hadoop3.2.tgz**

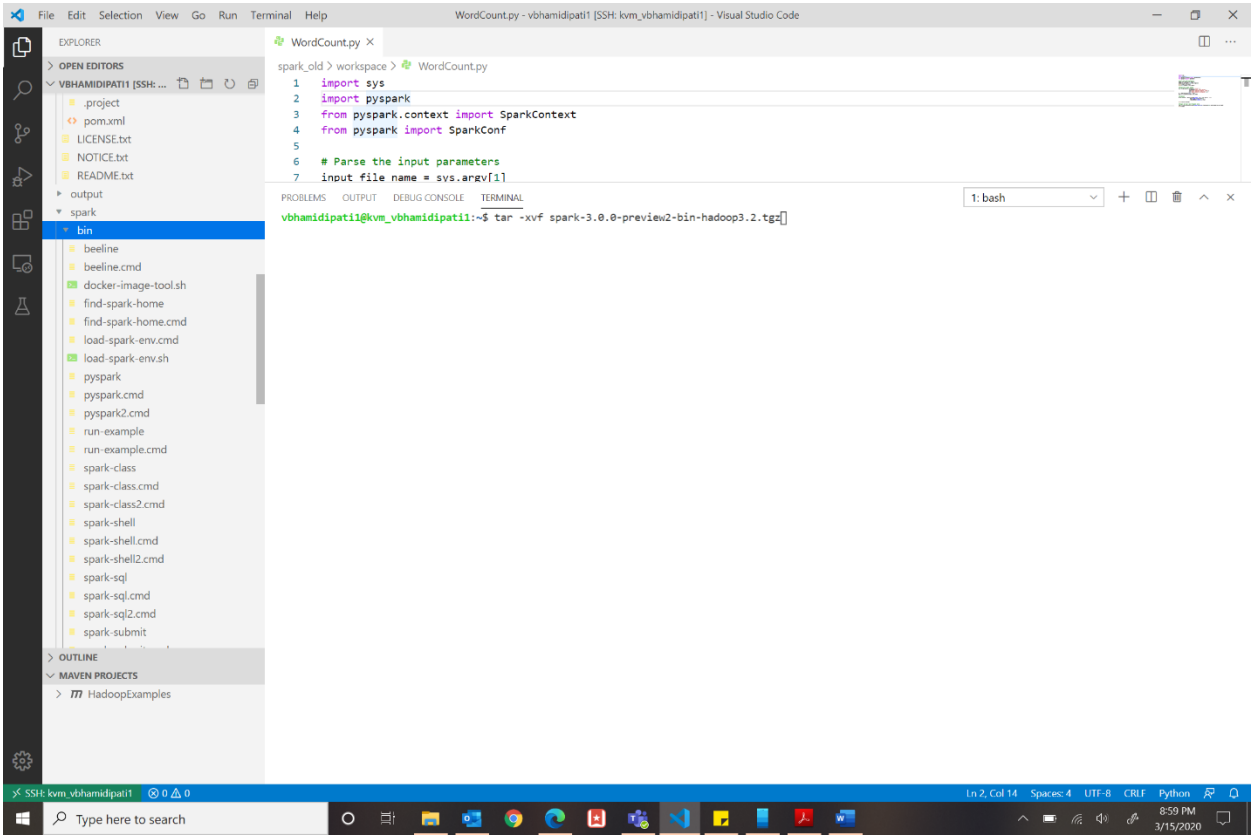

Fig 2. Extracting the Spark packages

3. Setup the path for spark and python
    a. `export SPARK_HOME=/home/vbhamidipati1/spark`
    b. `export PATH=$SPARK_HOME/bin:$PATH`
    c. `export PYSPARK_PYTHON=python3`
    d. Enter all the above statements in the bashrc and use source ~./bashrc to save the changes immediately.

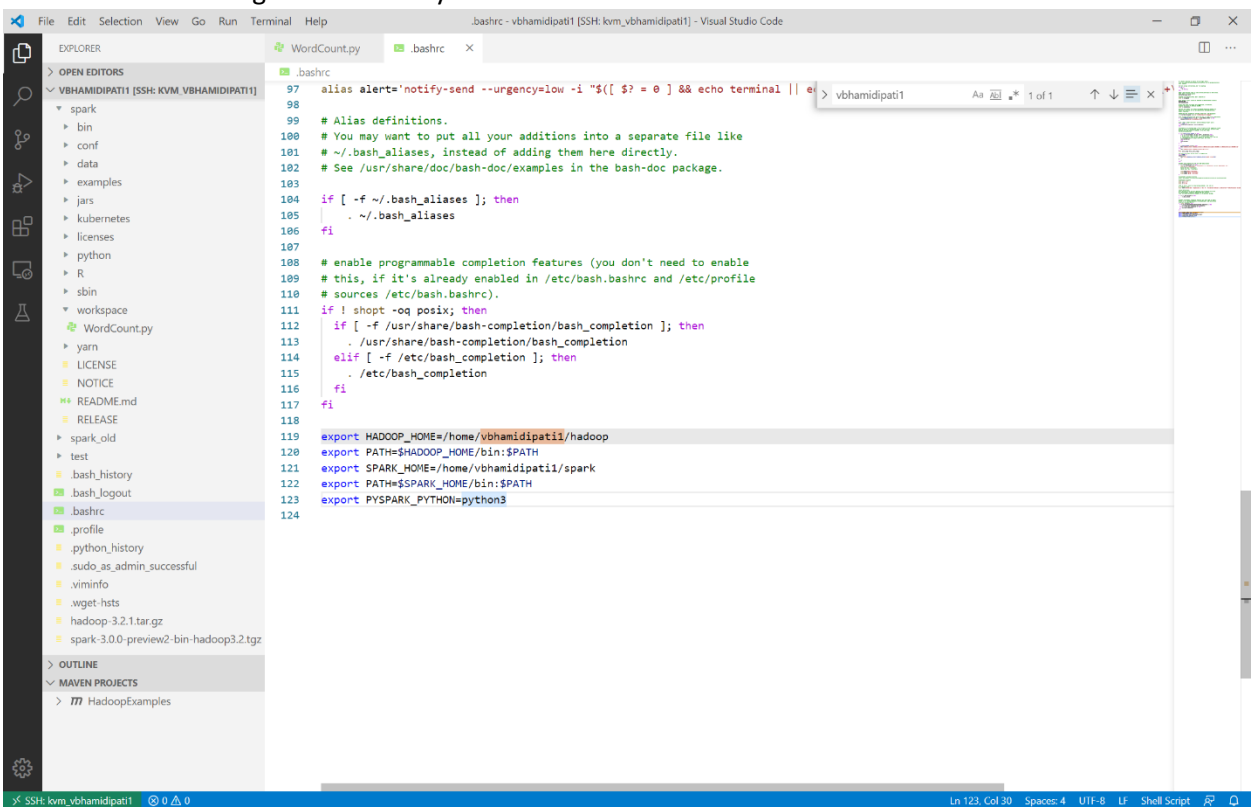

Fig 3. Setting the path of spark and python in bashrc

4. On successful installation of spark, we get the following output if we run “pyspark” command

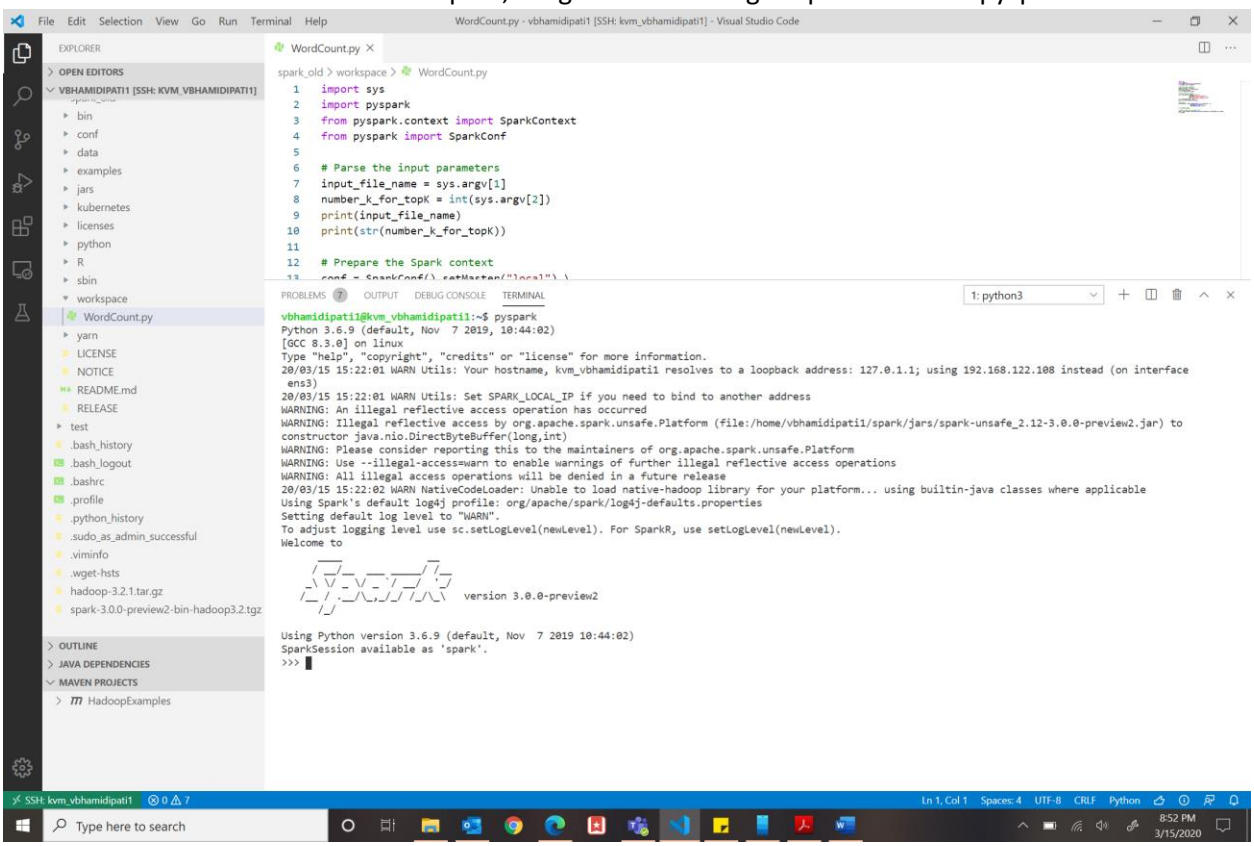

Fig 4. On Successful installation of spark, the following output is displayed

5. Once we have all the input files (peterpan.txt, test.txt) and the python program for wordcount in place, we run the following command - spark submit home/vbhamidipati1/spark/workspace/WordCount.py /home/vbhamidipati1/hadoop/data/peterpan.txt 30
    a. The last parameter gives the topmost frequent word occurrences.

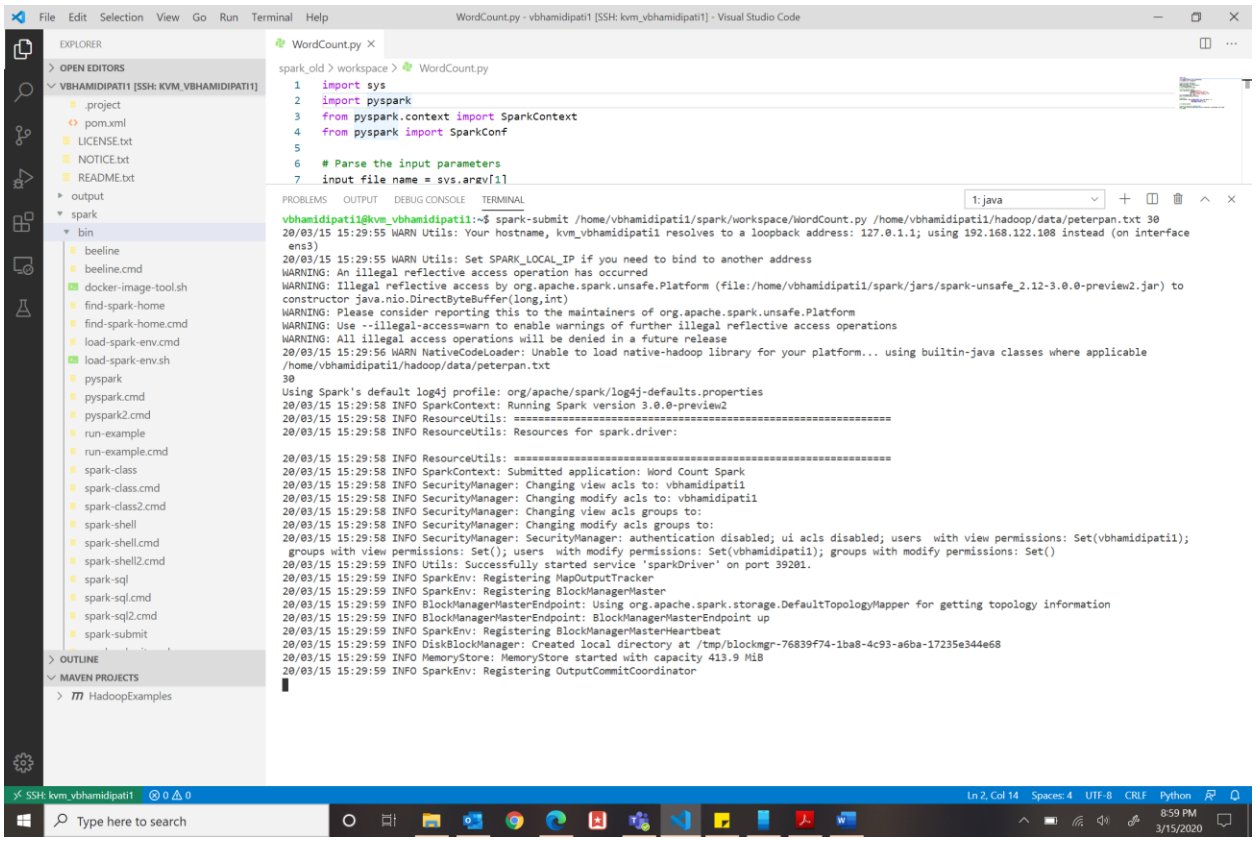

Fig 5. Command to run the word count on peterpan.txt

6. Top 30 frequent words of peterpan.txt are displayed as follows:

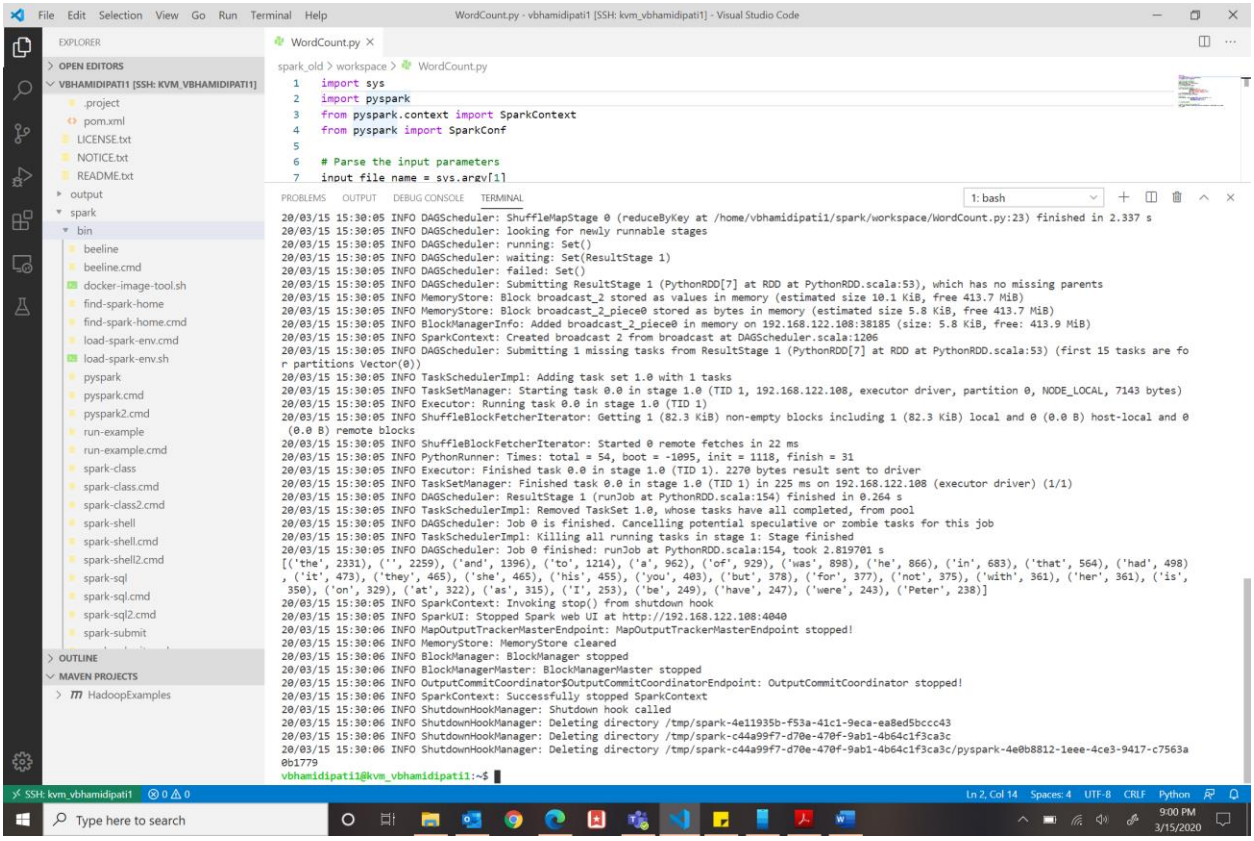

Fig 6. Output of word counts of top 30 words of peterpan.txt

7. Running top 5 frequent words for test.txt

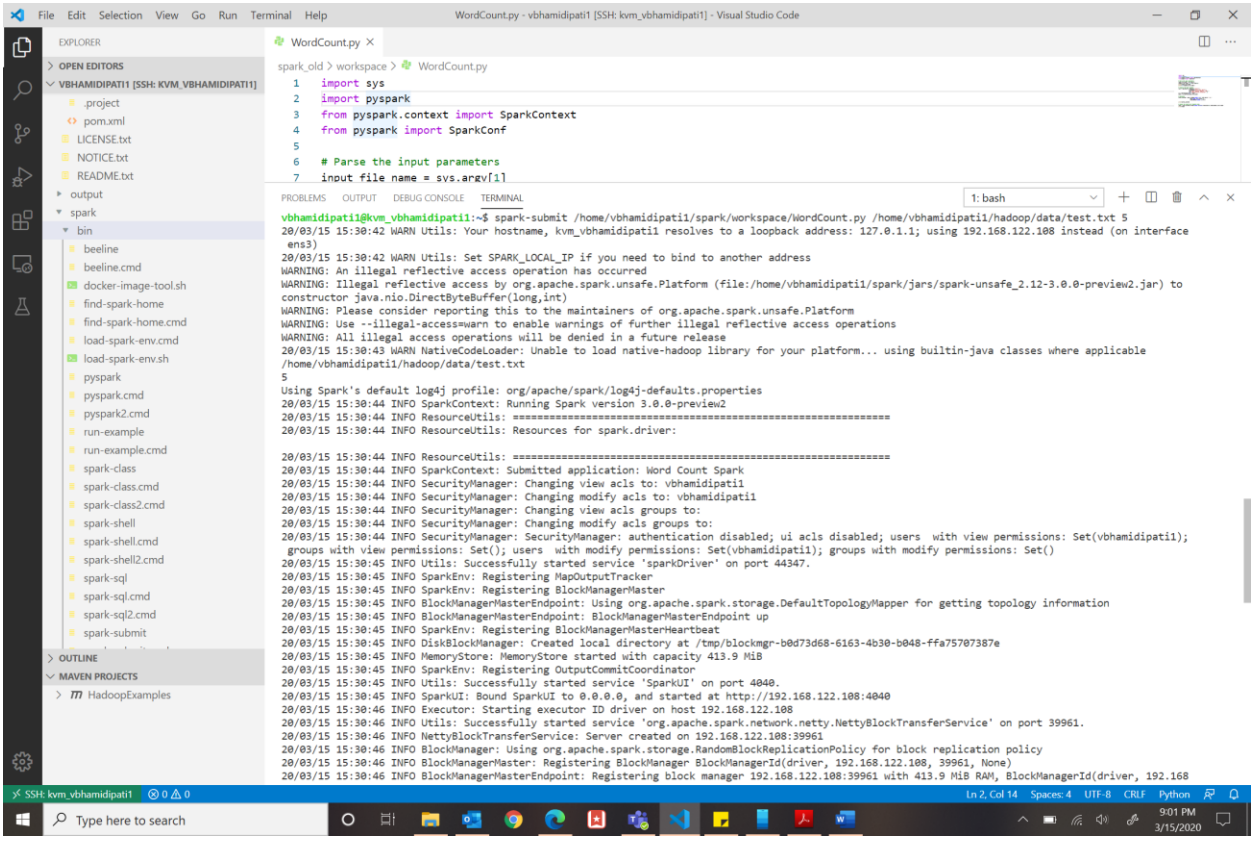

Fig 7. Command to run the word count on test.txt

8. Top 5 frequent words of test.txt are displayed as follows:

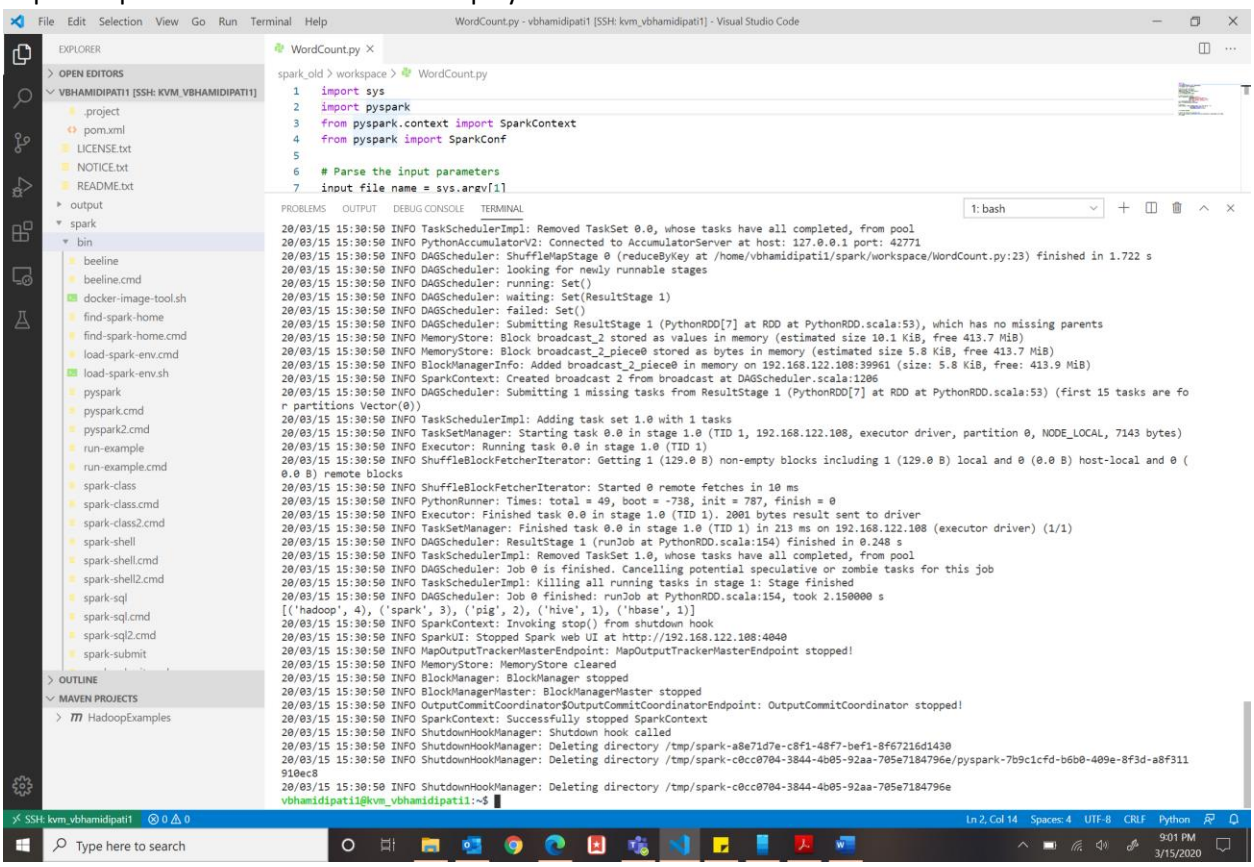

Fig 8. Output of word counts of top 5 words of test.txt