COMP-SCI-5540
PRINCIPLES OF BIG DATA MANAGEMENT

By TEAM 14

Manasa Thipparthi(mthwf)

Sai Jyothi Gudibandi(sg583)

Sri Harsha Kumar Raja Golla(sgnt8)

Satya Harish Gumpalli(sgmt2)

Project-I

Project Goal:

Develop a system to archive a social network's data i.e. Twitter's data using Hadoop's API and HDFS where each one of the top ten tweets are stored in a separate directory.

Tasks Included:

- Collect tweets in JavaScript Object Notation (JSON) format (at least 100K record).
    - Find the list of top ten used hashtags in your collection.

- Create a directory in HDFS for each hashtag from the top ten hashtag list.
    - Create additional two directories: "Others" and "None"

- Store the tweets on files in HDFS
    - If a tweet contains a hashtag from the top ten list, store the tweet in that hashtag's directory.

- If a tweet contains one or more hashtags, but none of the hashtags are in the top ten list, store the tweet in the "Others" directory.
    - If a tweet does not contain a hashtag, store it in the "None" directory.

- Extra Requirement:
    - Implement a function that counts the number of times a keyword appears in one of two tweet JSON attributes (text and hashtags) in all of 12 directories that were created on HDFS: int countWord (String keyword, String attr)

Prerequisite Skills:

- Create, open, read, and write files using a local file system.

- Write a basic word count function.

- Read and parse a JSON file. Perform a word count on one attribute on a list of JSON objects.

Our Project Plan:

- **Module 1:** Create twitter API keys and then Collect 100K Tweets using Tweepy API using python in JSON format.

- **Module 2:** Generate the top 10 hashtags from the collected tweets using python.

- **Module 3:** Create different directories in HDFS for all these 10 hashtags along with "Others" and "None".

- **Module 4:** Now push the tweets into those directories based on the hashtags and others and none structure as defined above.

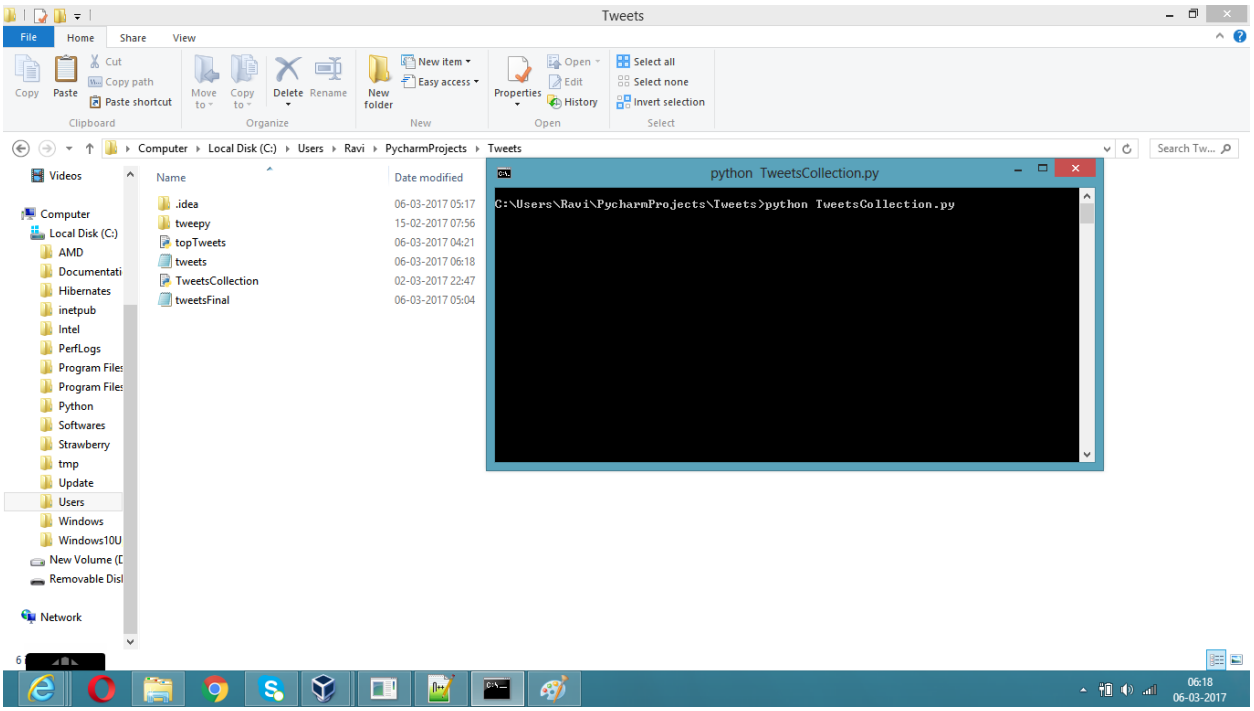
Extra Requirement:

- **Module 5:** Now search for a keyword in both the "text" and "hashtags" columns in a tweet in all the 12 directories and give a final count of the same word.

Module 1:

In this module, we have implemented a python program where we have used the Tweepy API and our Twitter API keys to retrieve 100K of tweets and store these in a local file.

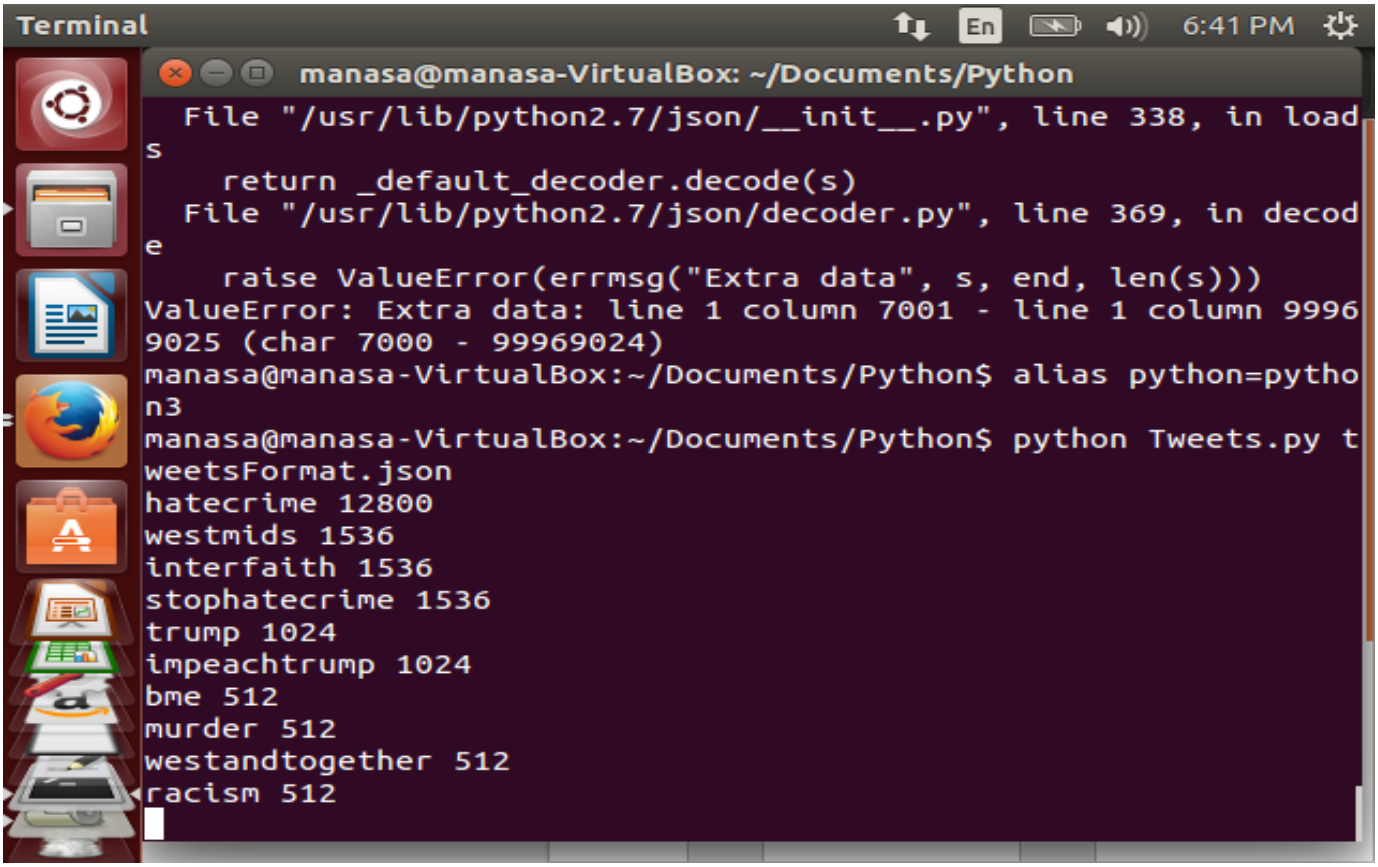
Below is the screenshot of the program:



Module 2:

Once we are done with the Tweet collection our aim is to find out the top 10 hashtags used in the retrieved tweets. We have implemented a python code which scans the whole JSON file and increment the count value associated with each hashtag. Once the scan is completed we now sort the count values of the hashtags and find out the top 10 hashtags.

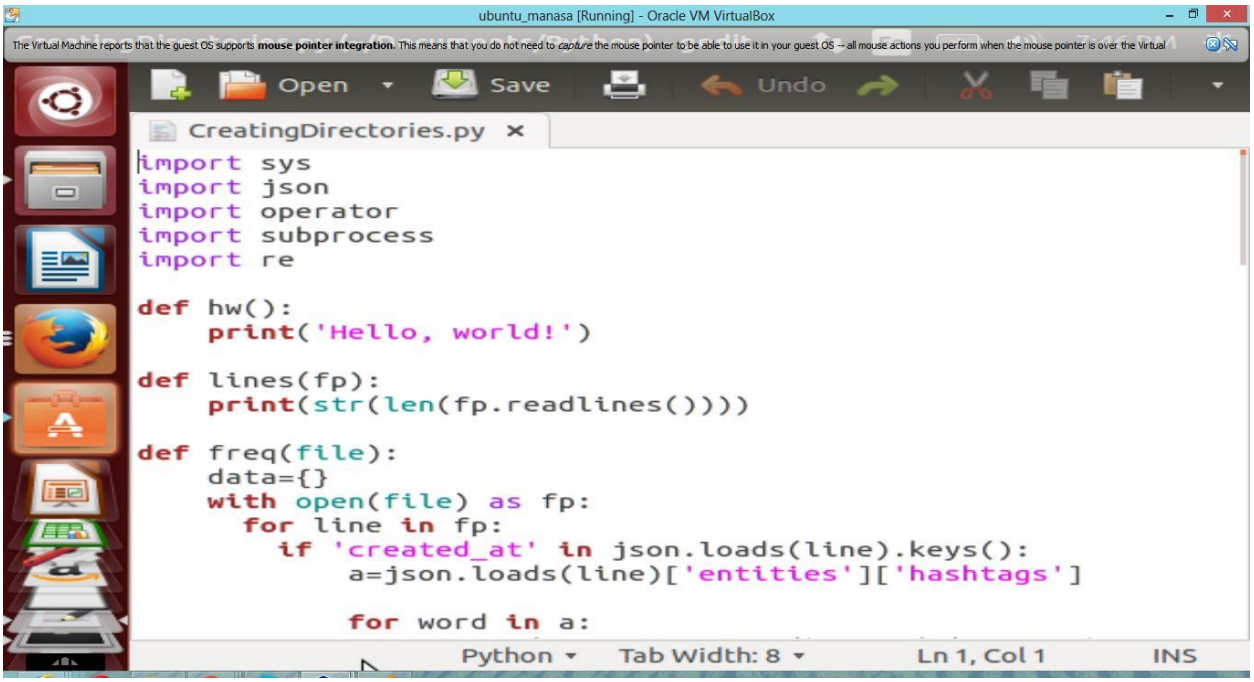
Below is the screen shot of our progra

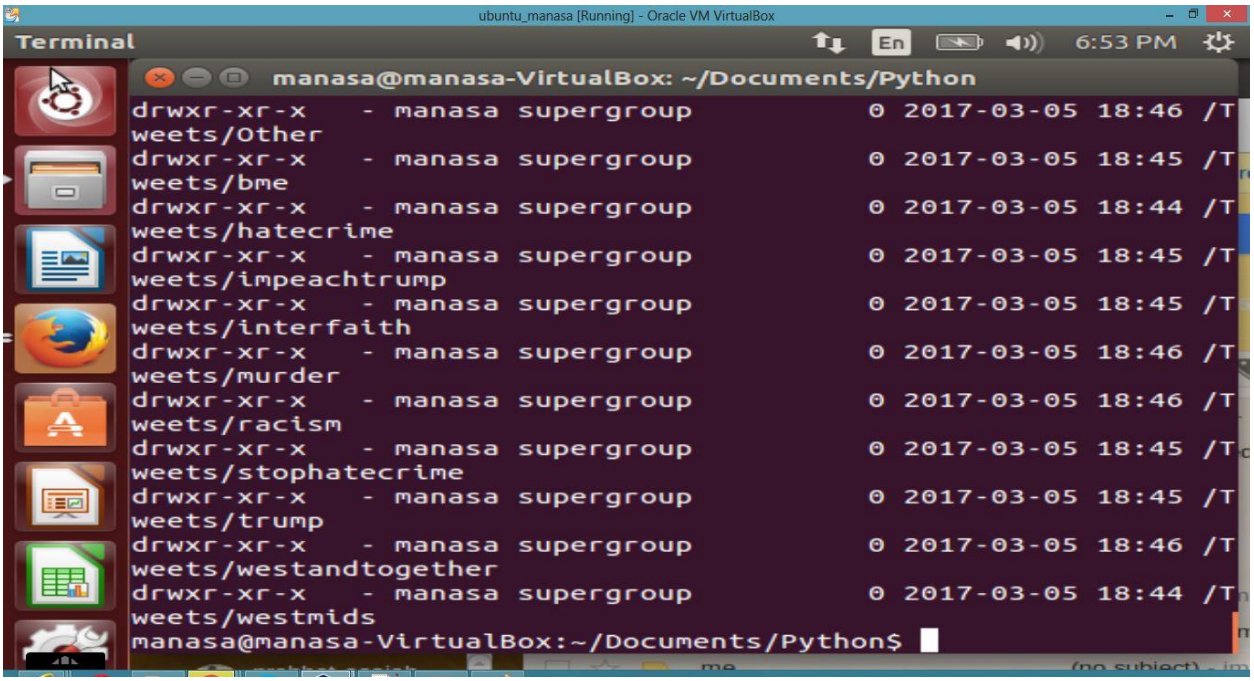


Module 3:

Based on these top 10 hashtags we now create 10 directories in HDFS with the same names and 2 other directories i.e. Others (stores all remaining tweets with hashtags) and None (stores the tweets with no hashtags).

Below is the screenshot of the program:

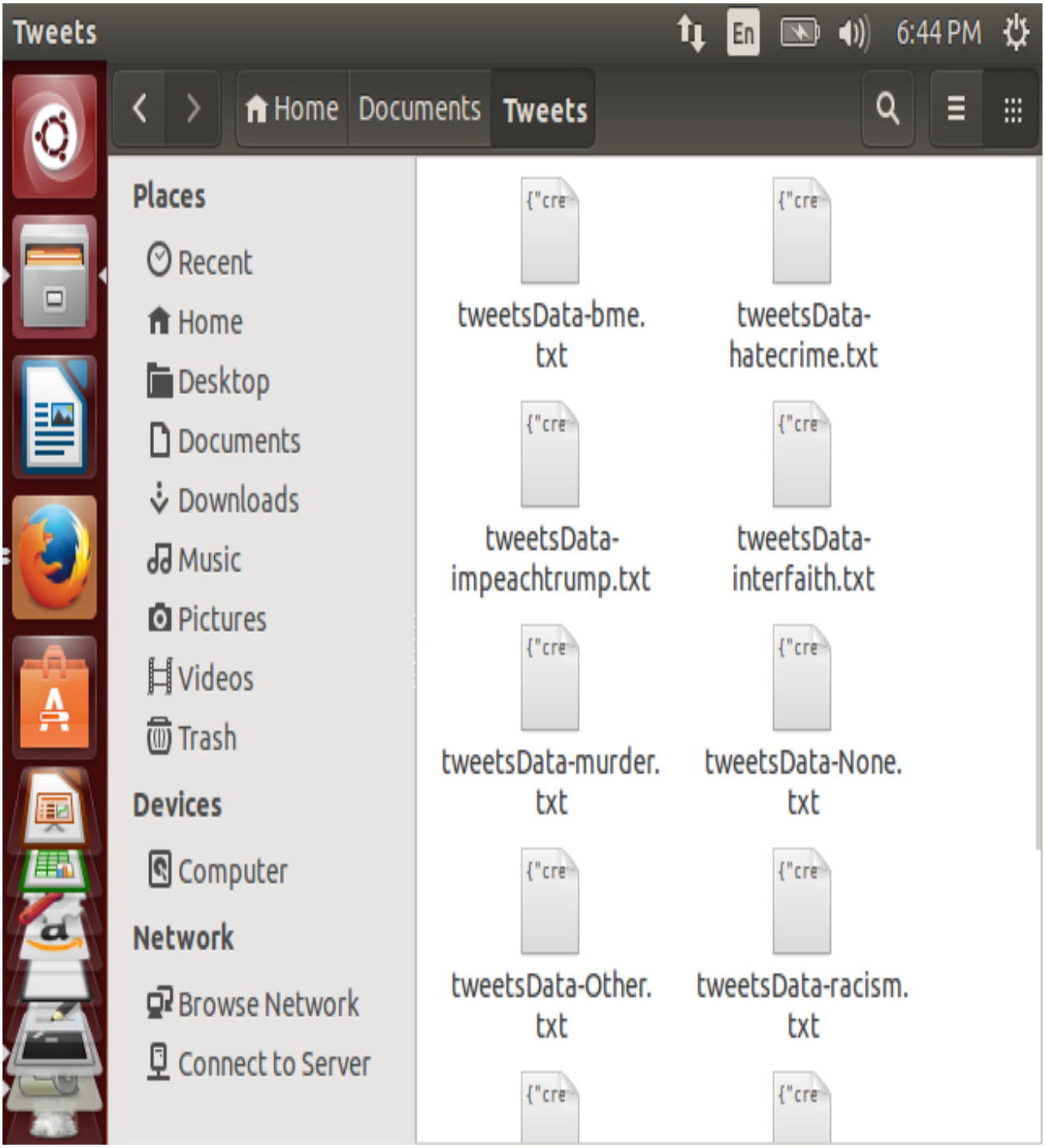


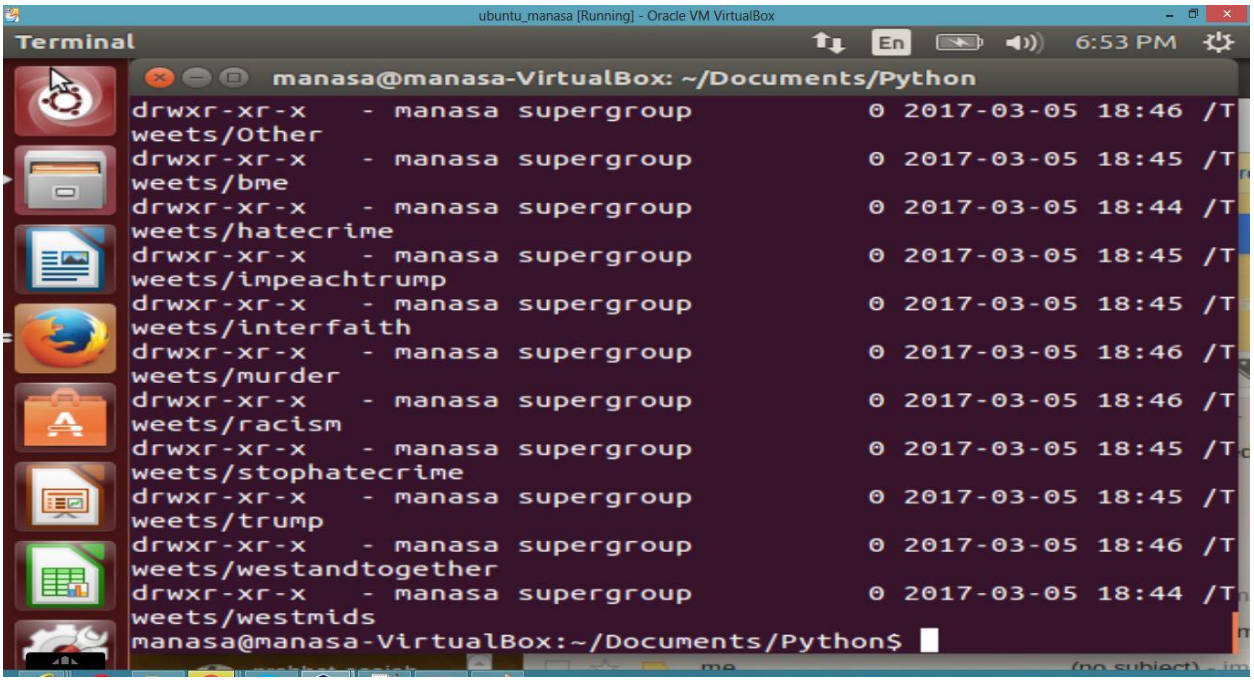


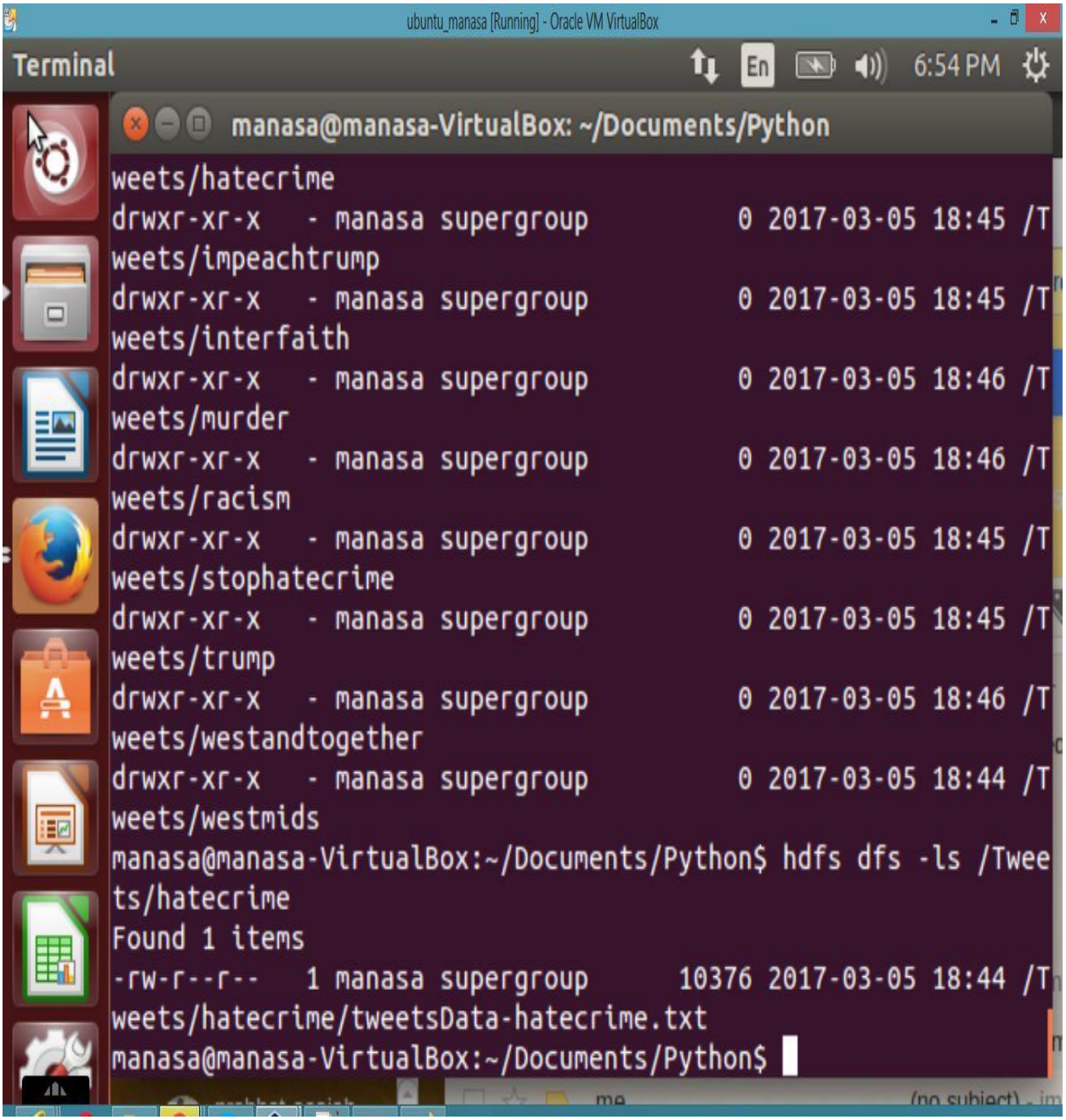
Module 4:

After the directories are created the next step in the project is to split the tweets and store them in the directories created in HDFS. We have implemented the Hadoop commands inside the python code with the help of predefined libraries.

Below is the screen shot of our program:




ubuntu_manasa [Running] - Oracle VM VirtualBox
Terminal
6:53 PM
manasa@manasa-VirtualBox: ~/Documents/Python
drwxr-xr-x - manasa supergroup 0 2017-03-05 18:46 /Tweets/Other
drwxr-xr-x - manasa supergroup 0 2017-03-05 18:45 /Tweets/bme
drwxr-xr-x - manasa supergroup 0 2017-03-05 18:44 /Tweets/hatecrime
drwxr-xr-x - manasa supergroup 0 2017-03-05 18:45 /Tweets/impeachtrump
drwxr-xr-x - manasa supergroup 0 2017-03-05 18:45 /Tweets/interfaith
drwxr-xr-x - manasa supergroup 0 2017-03-05 18:46 /Tweets/murder
drwxr-xr-x - manasa supergroup 0 2017-03-05 18:46 /Tweets/racism
drwxr-xr-x - manasa supergroup 0 2017-03-05 18:45 /Tweets/stophatecrime
drwxr-xr-x - manasa supergroup 0 2017-03-05 18:45 /Tweets/trump
drwxr-xr-x - manasa supergroup 0 2017-03-05 18:46 /Tweets/westandtogether
drwxr-xr-x - manasa supergroup 0 2017-03-05 18:44 /Tweets/westmids
manasa@manasa-VirtualBox:~/Documents/Python$

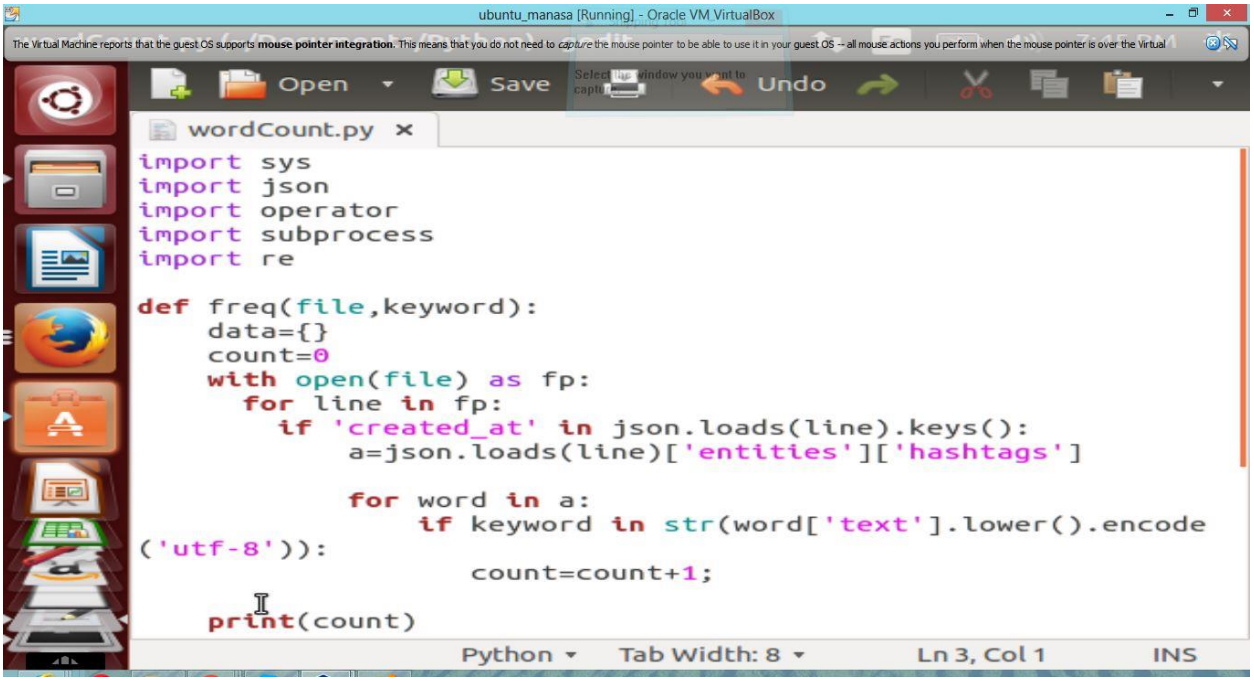


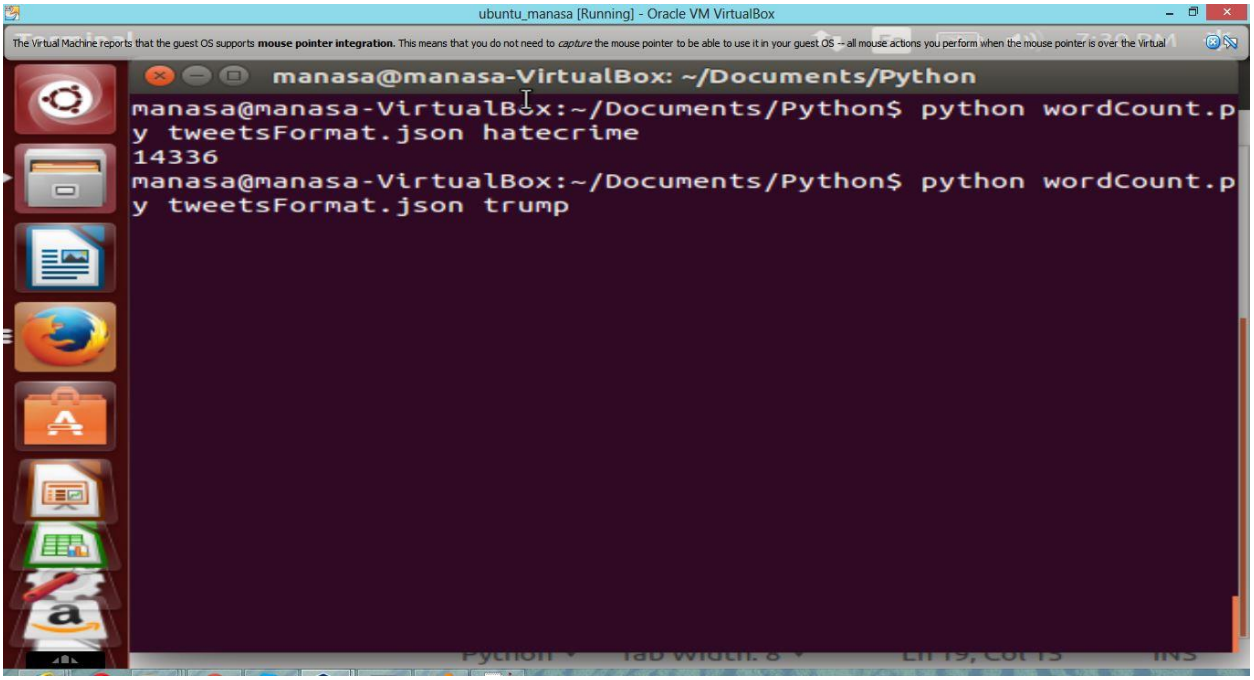
Extra Credit

Module 5:

As an extra credit, we searched for a keyword in both the “text” and “hashtags” columns in a tweet in all the 12 directories and increment the count value when it encounters an equivalent value in any of these columns i.e. “text” or “hashtag” in the tweet data.

Below is the screen shot of our program:





References:

- https://github.com/SivagamiNambi/Twitter-Sentiment-Analysis