



# 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.

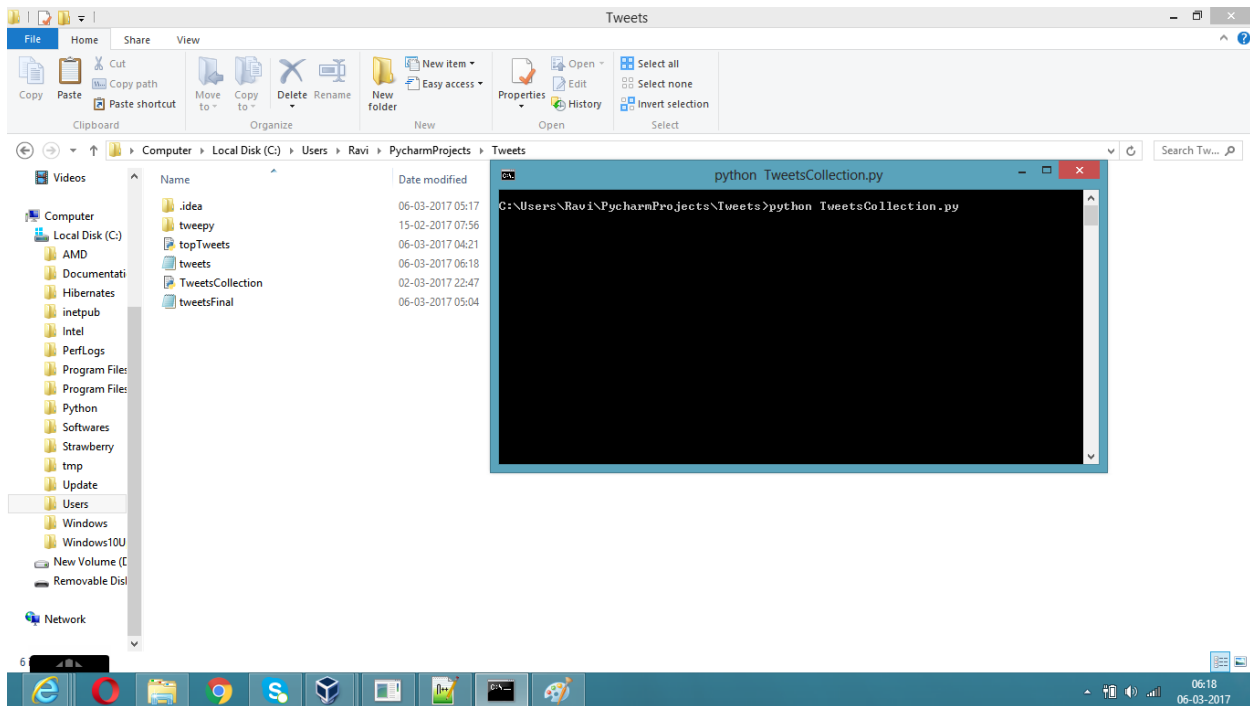
### *Extra Requirement:*

- **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.

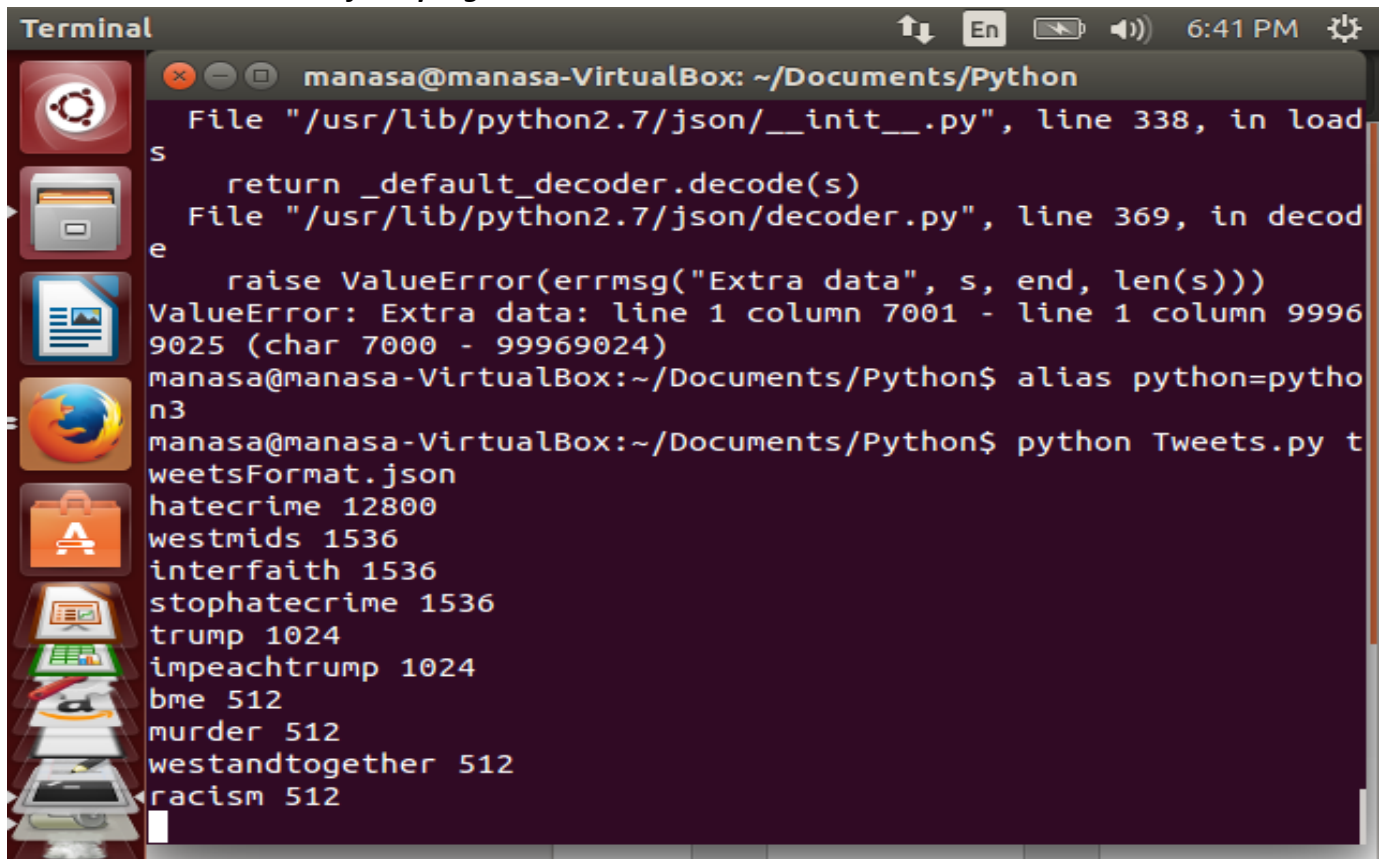
***Below is the screenshot of the program:***



## 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***



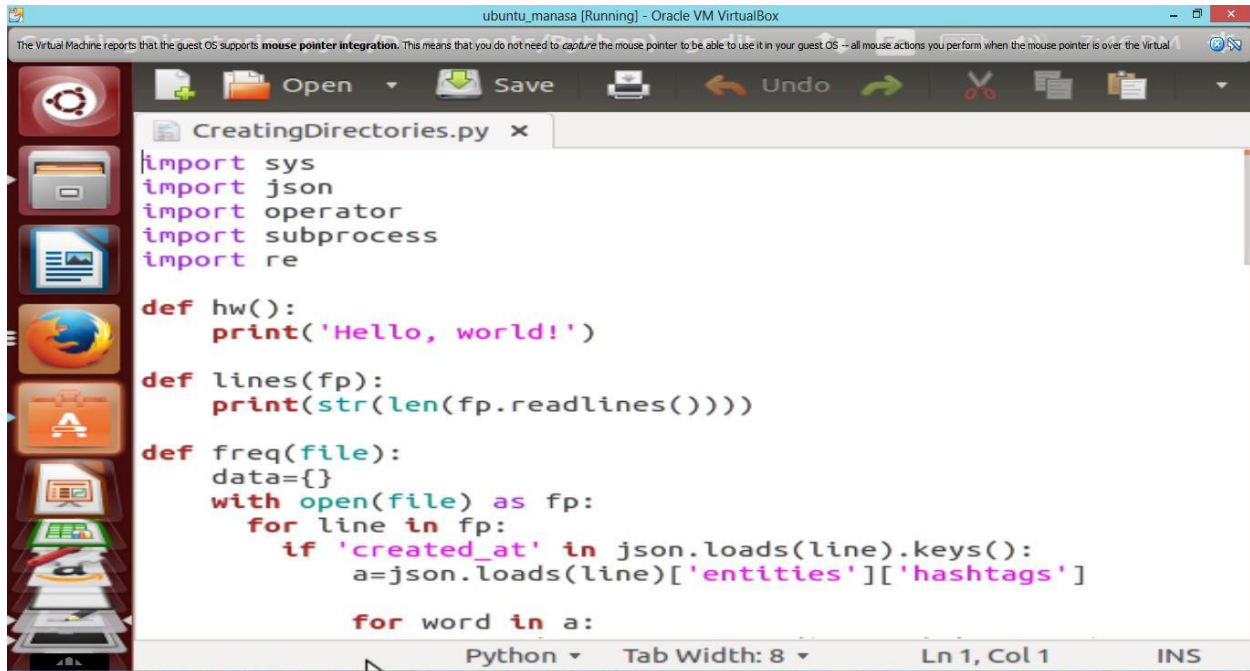
The screenshot shows a terminal window titled "Terminal" with a dark background. The prompt is "manasa@manasa-VirtualBox: ~/Documents/Python". The terminal displays the following text:

```
File "/usr/lib/python2.7/json/__init__.py", line 338, in load
s
    return _default_decoder.decode(s)
File "/usr/lib/python2.7/json/decoder.py", line 369, in decode
e
    raise ValueError(errmsg("Extra data", s, end, len(s)))
ValueError: Extra data: line 1 column 7001 - line 1 column 9996
9025 (char 7000 - 99969024)
manasa@manasa-VirtualBox:~/Documents/Python$ alias python=pytho
n3
manasa@manasa-VirtualBox:~/Documents/Python$ python Tweets.py t
weetsFormat.json
hatecrime 12800
westmids 1536
interfaith 1536
stophatecrime 1536
trump 1024
impeachtrump 1024
bme 512
murder 512
westandtogether 512
racism 512
```

### 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:***



The screenshot shows a text editor window titled 'CreatingDirectories.py' with the following Python code:

```
import sys
import json
import operator
import subprocess
import re

def hw():
    print('Hello, world!')

def lines(fp):
    print(str(len(fp.readlines())))

def freq(file):
    data={}
    with open(file) as fp:
        for line in fp:
            if 'created_at' in json.loads(line).keys():
                a=json.loads(line)['entities']['hashtags']

                for word in a:
```

The editor interface includes a menu bar with 'Open', 'Save', 'Undo', and other options. The status bar at the bottom indicates 'Python', 'Tab Width: 8', 'Ln 1, Col 1', and 'INS'.

```

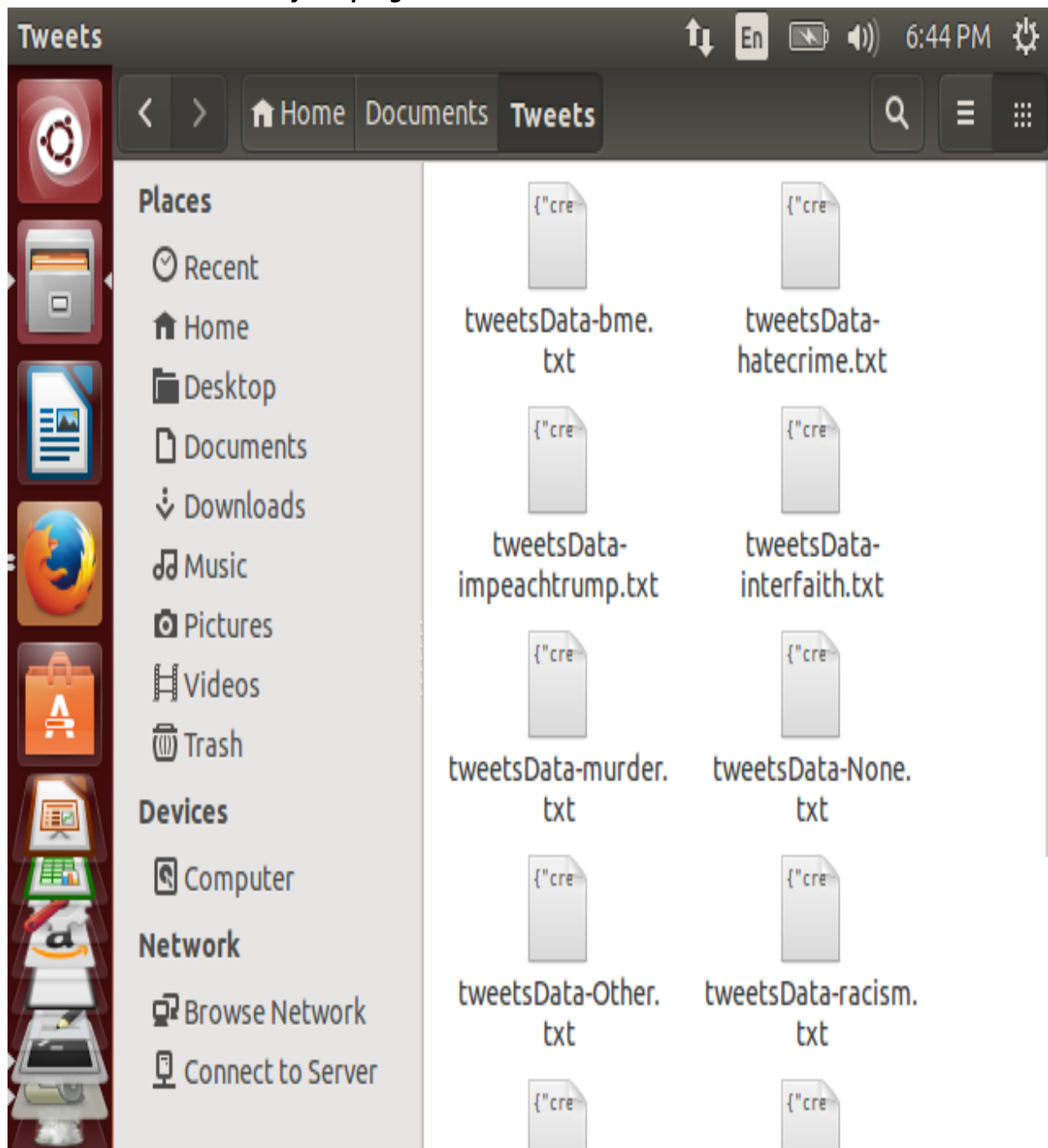
Terminal
ubuntu_manasa [Running] - Oracle VM VirtualBox
manasa@manasa-VirtualBox: ~/Documents/Python
drwxr-xr-x  - manasa supergroup 0 2017-03-05 18:46 /T
weets/Other
drwxr-xr-x  - manasa supergroup 0 2017-03-05 18:45 /T
weets/bme
drwxr-xr-x  - manasa supergroup 0 2017-03-05 18:44 /T
weets/hatecrime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

```

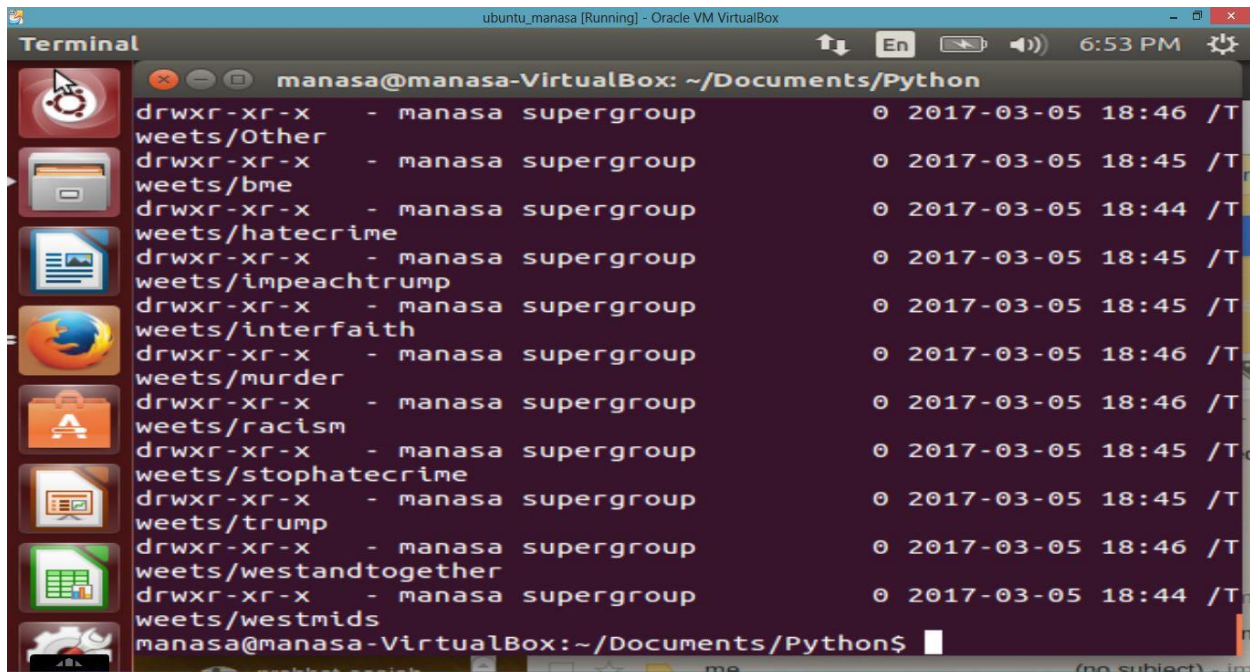
## Module 4:

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:*







The screenshot shows a terminal window titled "Terminal" with the subtitle "ubuntu\_manasa [Running] - Oracle VM VirtualBox". The terminal displays the output of a command, likely `ls -l`, showing file permissions, owner, group, size, date, and path for various directories. The directories listed are: weets/Other, weets/bme, weets/hatecrime, weets/impeachtrump, weets/interfaith, weets/murder, weets/racism, weets/stophatecrime, weets/trump, weets/westandtogether, and weets/westmids. All directories have permissions of drwxr-xr-x, are owned by manasa, and belong to the supergroup. The size is 0 bytes, and the date is 2017-03-05. The path is /T. The terminal prompt is manasa@manasa-VirtualBox: ~/Documents/Python\$.

```
manasa@manasa-VirtualBox: ~/Documents/Python
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:46 /T
heets/Other
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:45 /T
heets/bme
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:44 /T
heets/hatecrime
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:45 /T
heets/impeachtrump
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:45 /T
heets/interfaith
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:46 /T
heets/murder
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:46 /T
heets/racism
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:45 /T
heets/stophatecrime
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:45 /T
heets/trump
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:46 /T
heets/westandtogether
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:44 /T
heets/westmids
manasa@manasa-VirtualBox: ~/Documents/Python$
```

```

manasa@manasa-VirtualBox: ~/Documents/Python
ls -l /Tweets/hatecrime
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:45 /Tweets/hatecrime
ls -l /Tweets/impeachtrump
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:45 /Tweets/impeachtrump
ls -l /Tweets/interfaith
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:46 /Tweets/interfaith
ls -l /Tweets/murder
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:46 /Tweets/murder
ls -l /Tweets/racism
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:45 /Tweets/racism
ls -l /Tweets/stophatecrime
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:45 /Tweets/stophatecrime
ls -l /Tweets/trump
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:46 /Tweets/trump
ls -l /Tweets/standtogether
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:44 /Tweets/standtogether
ls -l /Tweets/westmids
drwxr-xr-x  - manasa supergroup      0 2017-03-05 18:44 /Tweets/westmids
manasa@manasa-VirtualBox:~/Documents/Python$ hdfs dfs -ls /Tweets/hatecrime
Found 1 items
-rw-r--r--   1 manasa supergroup    10376 2017-03-05 18:44 /Tweets/hatecrime/tweetsData-hatecrime.txt
manasa@manasa-VirtualBox:~/Documents/Python$

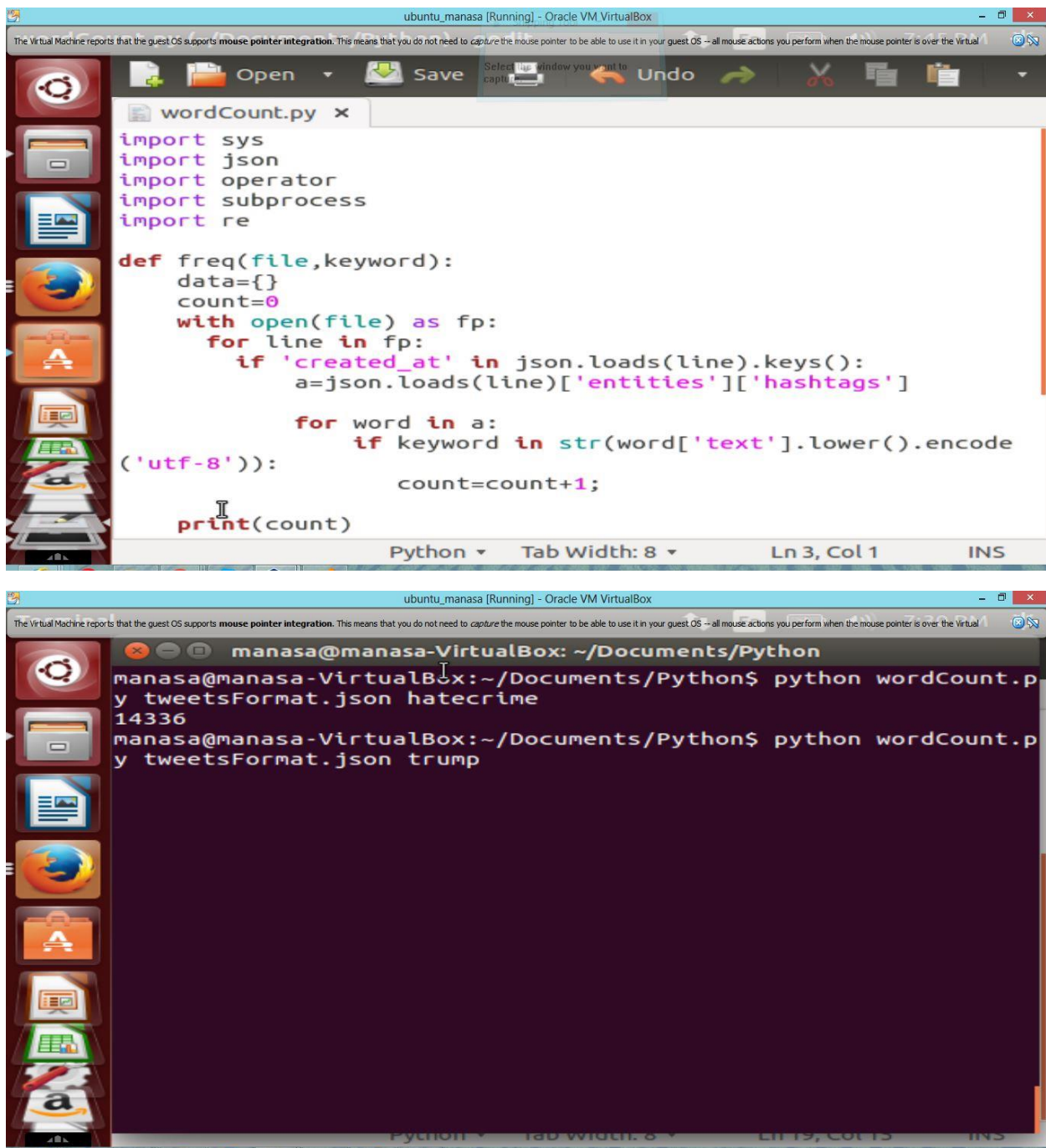
```

## 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:***



The first screenshot shows a text editor window titled 'wordCount.py' with the following Python code:

```
import sys
import json
import operator
import subprocess
import re

def freq(file,keyword):
    data={}
    count=0
    with open(file) as fp:
        for line in fp:
            if 'created_at' in json.loads(line).keys():
                a=json.loads(line)['entities']['hashtags']

                for word in a:
                    if keyword in str(word['text']).lower().encode
('utf-8')):
                        count=count+1;

    print(count)
```

The second screenshot shows a terminal window with the command prompt 'manasa@manasa-VirtualBox: ~/Documents/Python'. The user has executed the command 'python wordCount.py tweetsFormat.json hatecrime' and the output is '14336'. The user has then executed the command 'python wordCount.py tweetsFormat.json trump'.

## References:

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