**TWITTER STREAMING API’S USING TWITTWE DATA COLLECTION**

**Sub: Principles of Big Data**

**Management**

**Phase 1- Documentation**

Manaswini Vedula (16292800)

Aarthi Nagireddy (16292448)

Yamini Saraswathi Bommineni (16292447)

Instructor

Dr. PRAVEEN RAO, Ph.D.

ABSTRACT

In Phase 1, tweets for specific Hastags are collected from twitter using Python code in the json format, the Hastags and URL’s are extracted into separate files using python code and these extracted files runs on Hadoop and Spark word count example.

**Description:**

Initially the data has been collected from twitter by creating twitter developer account. The data is separated into URL and Hashtag files and the wordcount of hashtags and URL’s is taken from extracted files using Hadoop and Spark.

**Technologies used:**

Python

Twitter API’s

Java

Apache Hadoop( Cloudera)

Apache Spark(Cloudera)

**Step Wise Procedure:**

**Step1:**Creation of twitter developer account and generate keys and tokens

**A screenshot of a computer

Description automatically generated**

**Step2:** Installation of Pycharm IDE, anaconda environment and libraries in python terminal using pip command

pip install tweepy

pip install json

pip install os-win

pip install openapi-codec

**Step3:** Download tweets using below python code .

A screenshot of a computer screen

Description automatically generated

* Screenshot of Downloaded tweets in json format.
* JSON Data google Drive Link:https://drive.google.com/file/d/18i0UiYj3nnVygQteli7jZiikHLojiL7h/view?usp=sharing

A screenshot of a computer

Description automatically generated

**Step** **5**: Extraction of Hastags and URL’S of downloaded tweets from json file using below python code.

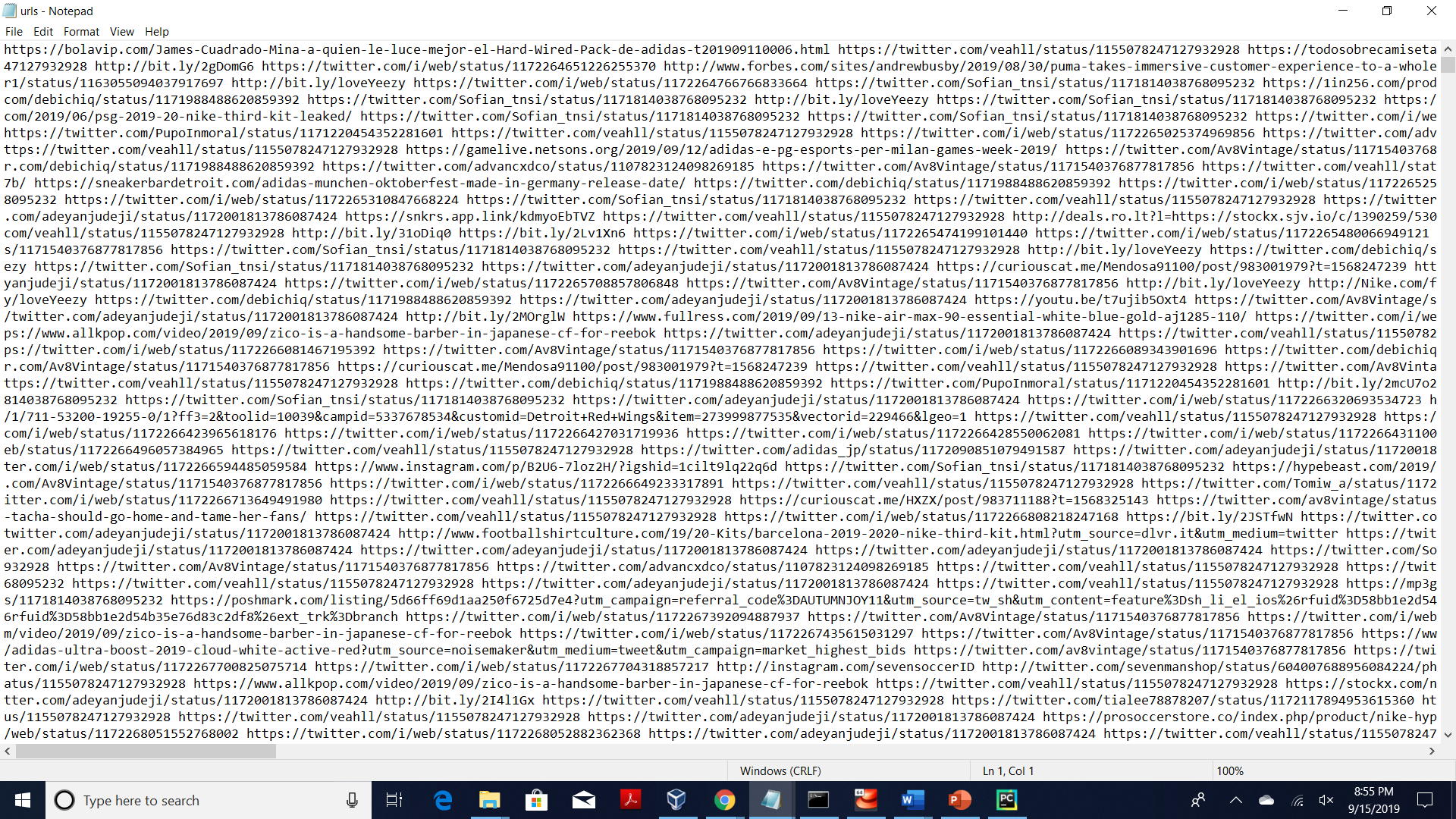
A screenshot of a computer screen

Description automatically generated

* Screenshot of extracted Hashtags and URL’s by running above python code

A screenshot of a computer

Description automatically generated



**Step 6**: Installation of Virtual Machine and Cloudera platform and copy files into virtual machine using mount command

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Step 9**: Creation of Directory named as input in Hadoop file systems using below command and copy the Hashtags and URL files into the HDFS.

Hdfs dfs -mkdir /user/cloudera/input/

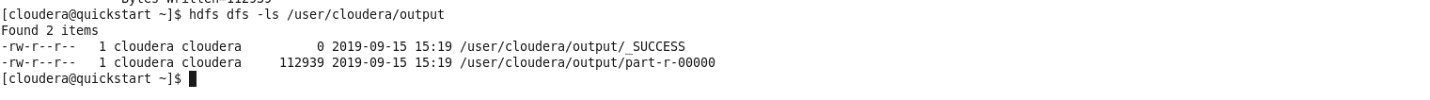
**WORDCOUNT USING HADOOP**

* Running the word count for Extracted Hashtags

A screenshot of a social media post

Description automatically generated

* Screenshot of Output for Hashtags using Hadoop wordcount



* Screenshot of output for wordcount of Hadoop in Hue for Hashtags

A screenshot of a cell phone

Description automatically generated

* Running the word count for Extracted URL’s

A screenshot of a social media post

Description automatically generated

* Screenshot of Output for URL file using Hadoop wordcount



* Screenshot of Output for wordcount of Hadoop in Hue for URL file

A screenshot of a cell phone

Description automatically generated

**WORDCOUNT USING SPARK**

* Open spark shell by using the following command

Spark-Shell

* Running the wordcount for Hashtags and URL’S files

A screenshot of a social media post

Description automatically generated

* Screenshot of Output using Spark for Hashtag file in Hue:

A screenshot of a social media post

Description automatically generated

* Screenshot of Output using Spark for URL file in Hue:A screenshot of a social media post

  Description automatically generated

The output and log files can be found on(<https://github.com/manaswinivedula/PB-twitter_data_extraction>)