

**PRINCIPLES OF BIGDATA MANAGEMENT**

**PHASE 1**

**TEAM MEMBERS**

**AVINASH GANGURI 16293133**

**SRI SAI NIKHIL KANTIPUDI 16292907**

**ROSHNA TOKE 16293711**

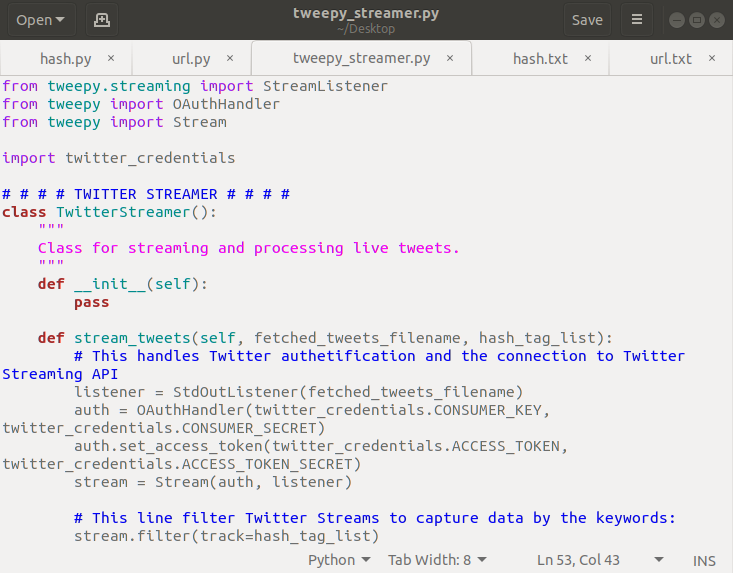
The main objective of this project is to collect twitter data by using twitter API and the extracting hashtags and URLs from the data and running wordcount on Hadoop and spark.

1. Collecting twitter data using twitter API

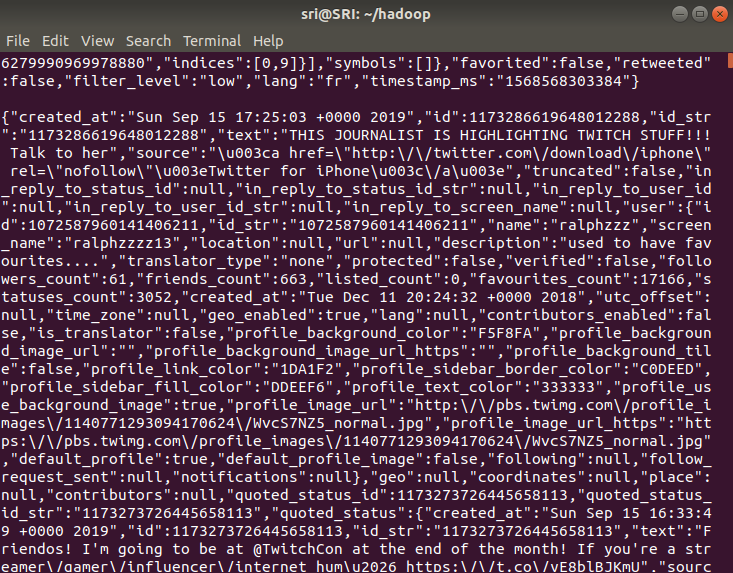
2. Extracting hashtags and URLS

3. Running wordcount program in Hadoop and spark

**Collecting twitter data using twitter API:**

****

**Twitter data extraction:**

****

Keyword used for collecting twitter data

- Cricket

- ISRO

- a

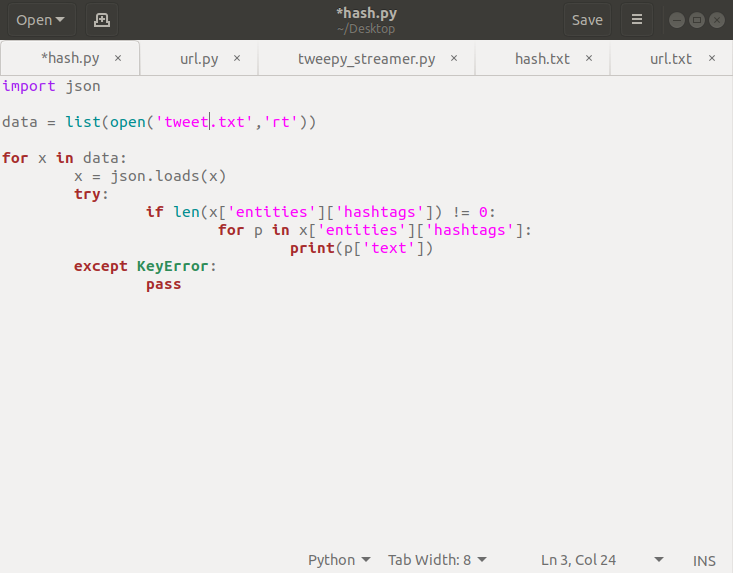
- b

**Extracting hashtags and URLS from collected twitter data:**

As the twitter data is in the text file. We need to send the text file as the input for the python code which extracts the hashtags and urls.

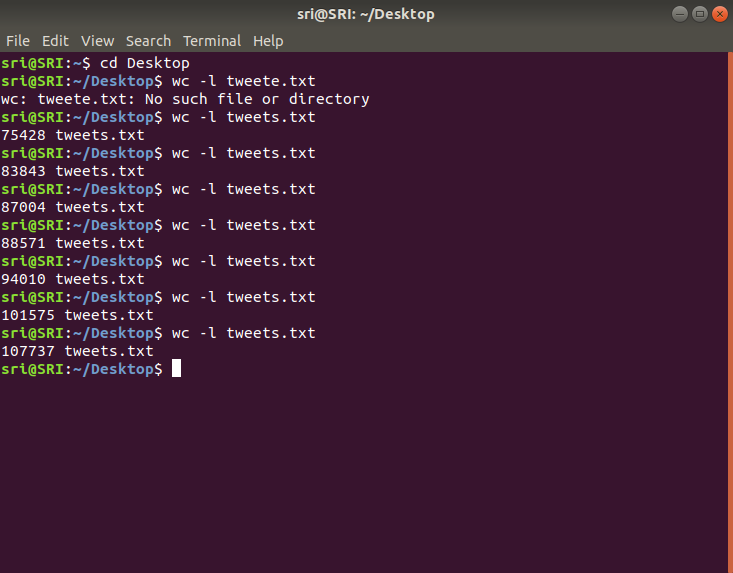
$>python3 hash.py >> hash.txt

$>python3 hash.py >> url.txt

****

hash.txt and url.txt are the text files that contains hashtags and urls. We need to run the wordcount program to count the words.

**Total tweets collected: 107737**

****

**Wordcount on Hadoop:**

Make sure that the data node and name node are running. Otherwise start them

$> ./start-dfs.sh

$> ./start-yarn.sh

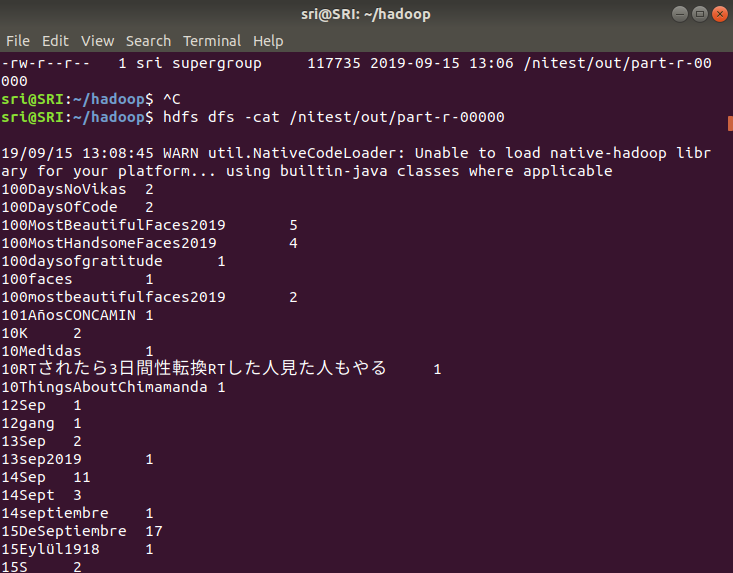
**For hashtags:**

$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar

wordcount /nitest/hash.txt /nitest/out

$ hdfs dfs -ls /nitest/out/

$ hdfs dfs -cat /nitest/out/part-r-00000



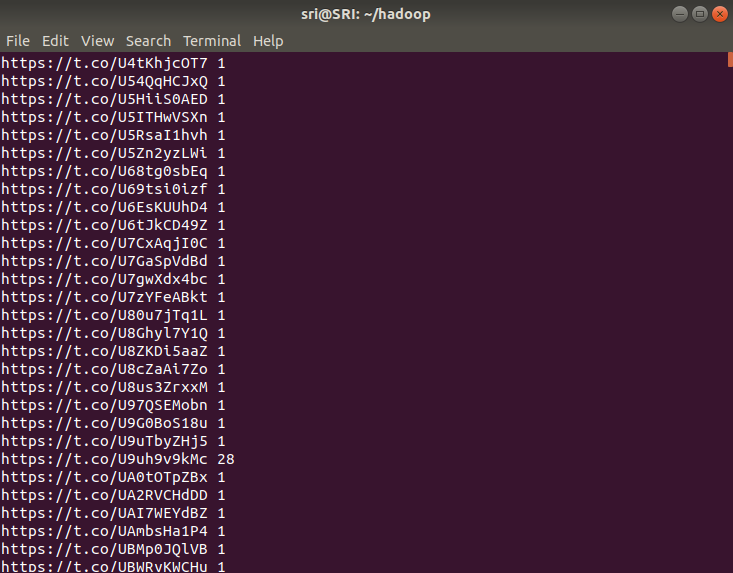
**For URLS:**

$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar

wordcount /nitest/url.txt /nitest/outurl

$ hdfs dfs -ls /nitest/outurl/

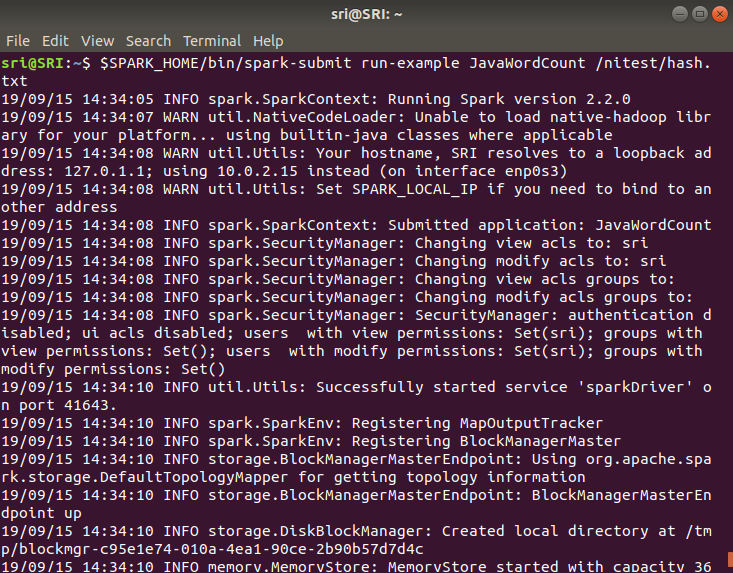
$ hdfs dfs -cat /nitest/outurl/part-r-00000

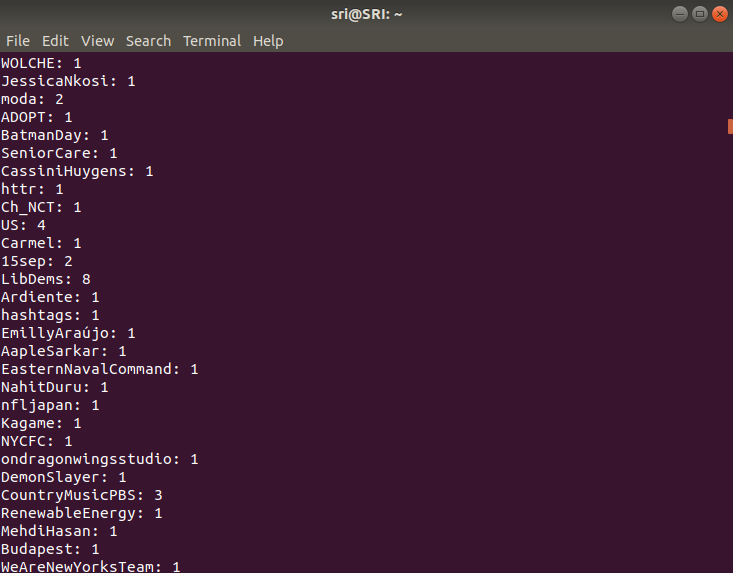
****

**Wordcount using spark:**

**For hashtags:**

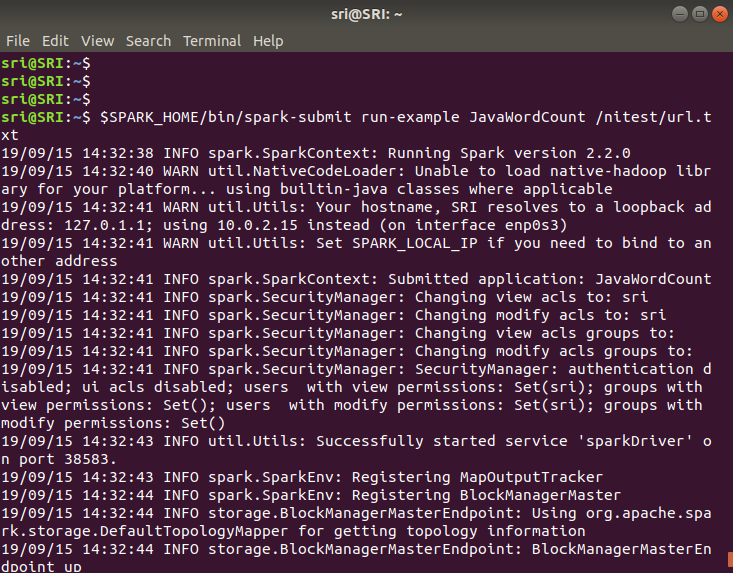
$SPARK\_HOME/bin/spark-submit run-example JavaWordCount /nitest/hash.txt

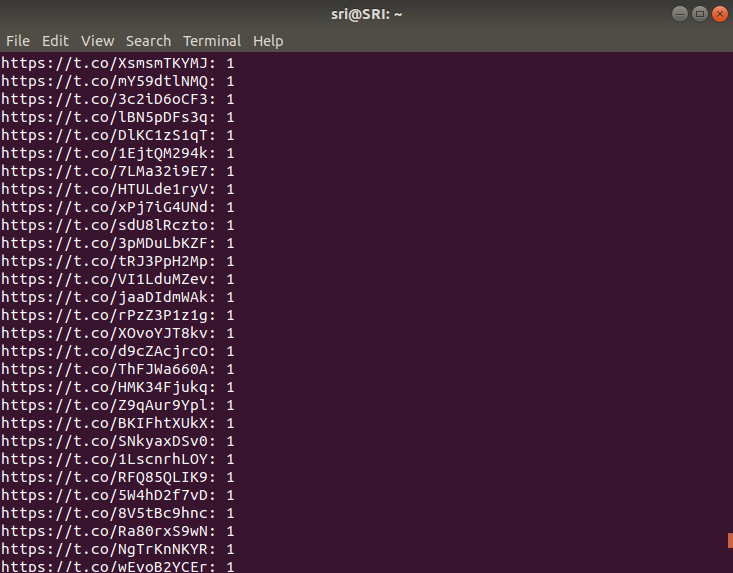




**For URLS:**

$SPARK\_HOME/bin/spark-submit run-example JavaWordCount /nitest/url.txt





Google drive link

https://drive.google.com/open?id=13dJ2lr0RD6n63UQkkeFzBLgjzznbnfUv