**CS5540: Principles of Big Data Management**

**A system to store, analyze, and visualize Twitter’s tweets**

**Project Phase-1 Document**

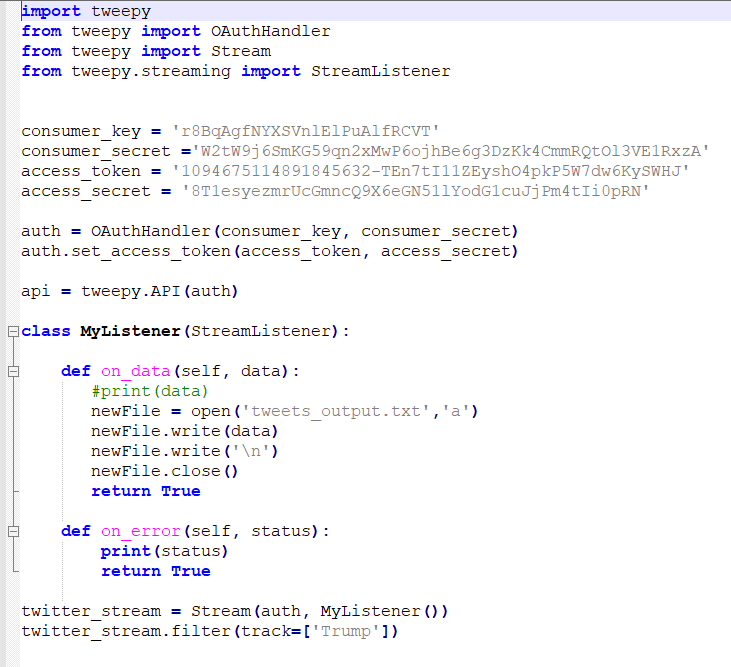
Justin Graber

Poonam Kankariya

Vidyullatha Lakshmi Kaza

**STEP 1: Collect tweets using Twitter’s Streaming API**

Python Code for Tweets Collection



The above program collects tweets into a text file by using Twitter’s streaming API Tweepy.

i. In order to connect to the Twitter’s Streaming API, we need to authenticate using credentials from our Twitter developer account.

ii. The program filters data by keyword (‘**Trump**’)

iii. The data is appended to file ‘tweets\_output.txt’.

We have collected 100k+ tweets.

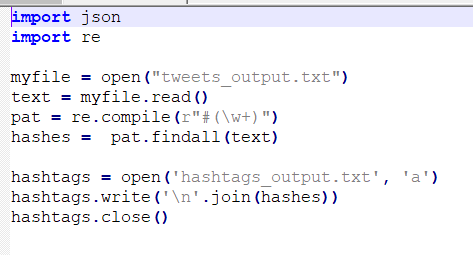
**Output File:**

**tweets\_output.txt**



**STEP 2: Extracting Hashtags**

**Hashtags.py**

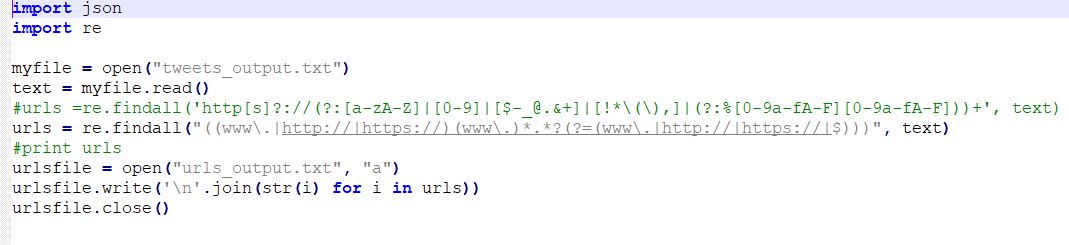


**OUTPUT File:**

**Hashtags\_output.txt**



**STEP 3: Extracting URL’s**

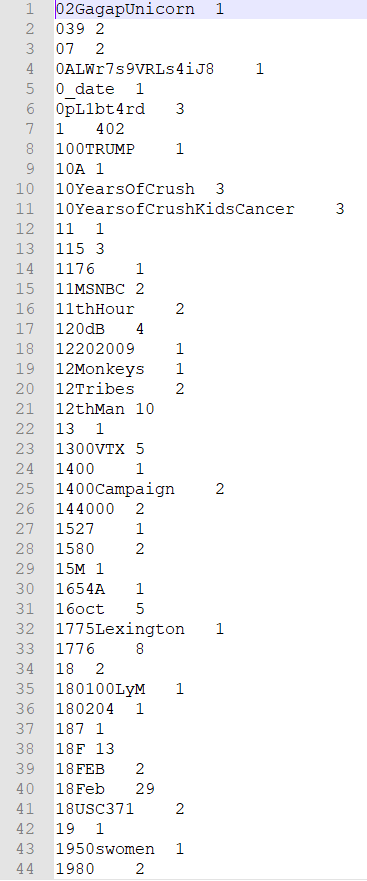


**OUTPUT File:**

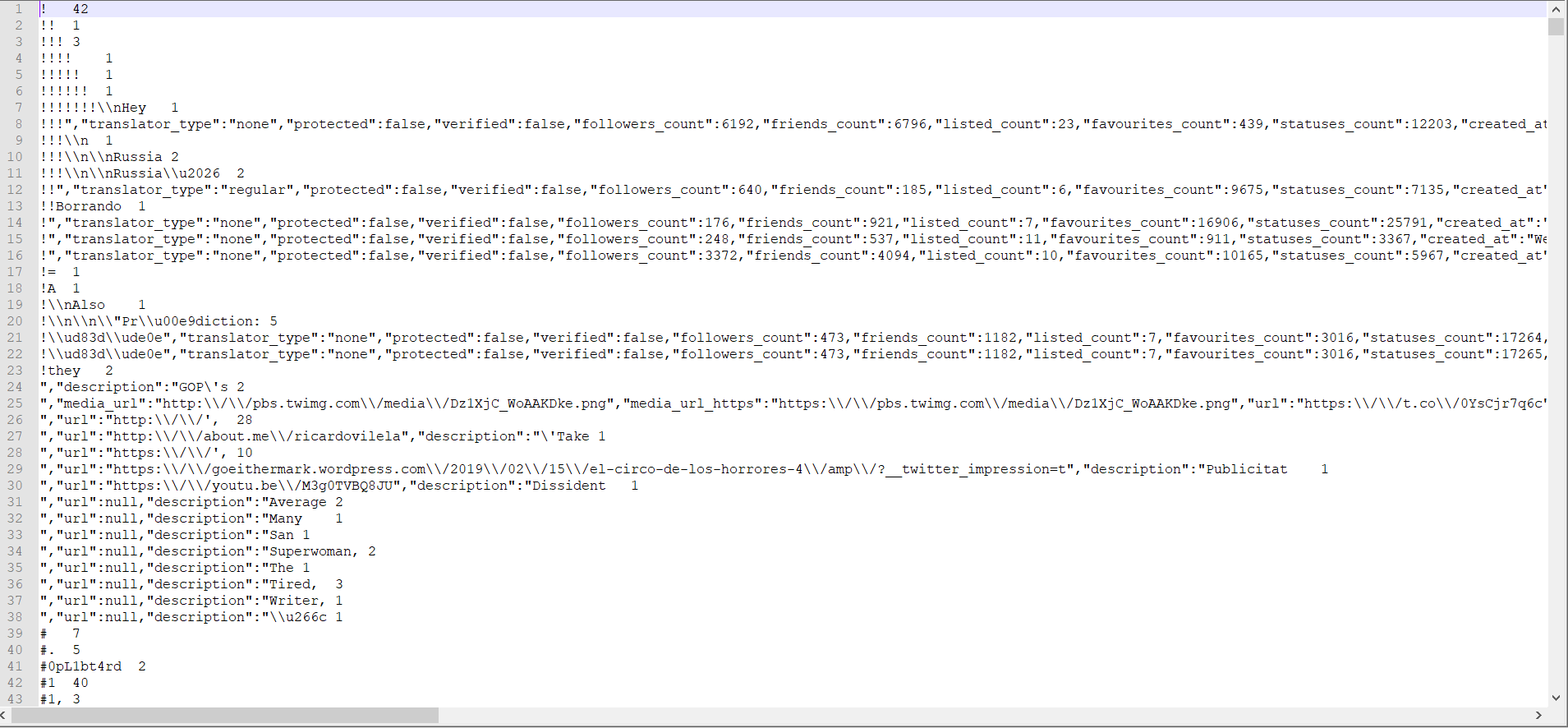
**urls\_output.txt**



**Hashtag’s Word Count using Hadoop:**



**URL’s Word Count using Hadoop:**



**Hashtag’s Word Count using Spark:**



**URL’s Word Count using Spark:**

