

Exploring Streaming Twitter Data (Optional)

By the end of this activity, you will be able to:

1. View the text of Twitter data streaming in real-time containing specific words.
2. Create plots of the frequency of streaming Twitter data to see how popular a word is.

NOTE: You must complete the instructions in the Reading *Instructions for Creating a Twitter App* before you begin this activity.

Step 1. Open a terminal shell. Open a terminal shell by clicking on the square black box on the top left of the screen.



Change into the json directory:

1

```
cd Downloads/big-data-2/json
```

Run `ls` to see the files:

```
[cloudera@quickstart json]$ ls
auth  json_schema.py  LiveTweets.py  PlotTweets.py  print_json.py  twitter.json
[cloudera@quickstart json]$
```

The auth file must contain the Twitter access keys and tokens. See the Reading *Instructions for Creating a Twitter App* for instructions how to create this file.

Step 2. View real-time tweets. We can view the contents of tweets in real-time by running the *LiveTweets.py* script. Run *LiveTweets.py president* to see the tweets containing the word *president*:

```
./LiveTweets.py president
```


The output displays two columns: the first is the timestamp of the tweet, and the second is the text of the tweet.

```
May 10 02:29:24 RT @WhatSheSaid167: President & CEO of @TheRoyalMHC Dr. Merali tal
May 10 02:29:24 Vice President Business Valuation Services https://t.co/V1kPtErPwv
May 10 02:29:24 RT @PSYClaudiaB: Dapat ba parehas ang boto ng President at VP? Kung ga
May 10 02:29:24 RT @soozmyahs: literally not lehman library. Lehman hall. Barnard libr
May 10 02:29:25 RT @Raiders4Hillary: Newspaper Endorsement: The #CharlestonGazetteMail
May 10 02:29:25 RT @jomardlrs: Si BBM dinadaya nila para si Leni maging vice tapos imp
May 10 02:29:25 RT @Activities_MHS: Faith Ibones for Senior Class Vice President. http
May 10 02:29:25 Good morning Philippines! Good morning President-elect Rudy Duterte. #
May 10 02:29:25 RT @FoxNews: .@GovernorPerry: "I want...a president that doesn't belie
```

When you are done, enter *Control-c* to stop the script.

Let's run *LiveTweets.py time* to see the tweets containing the word *time*:

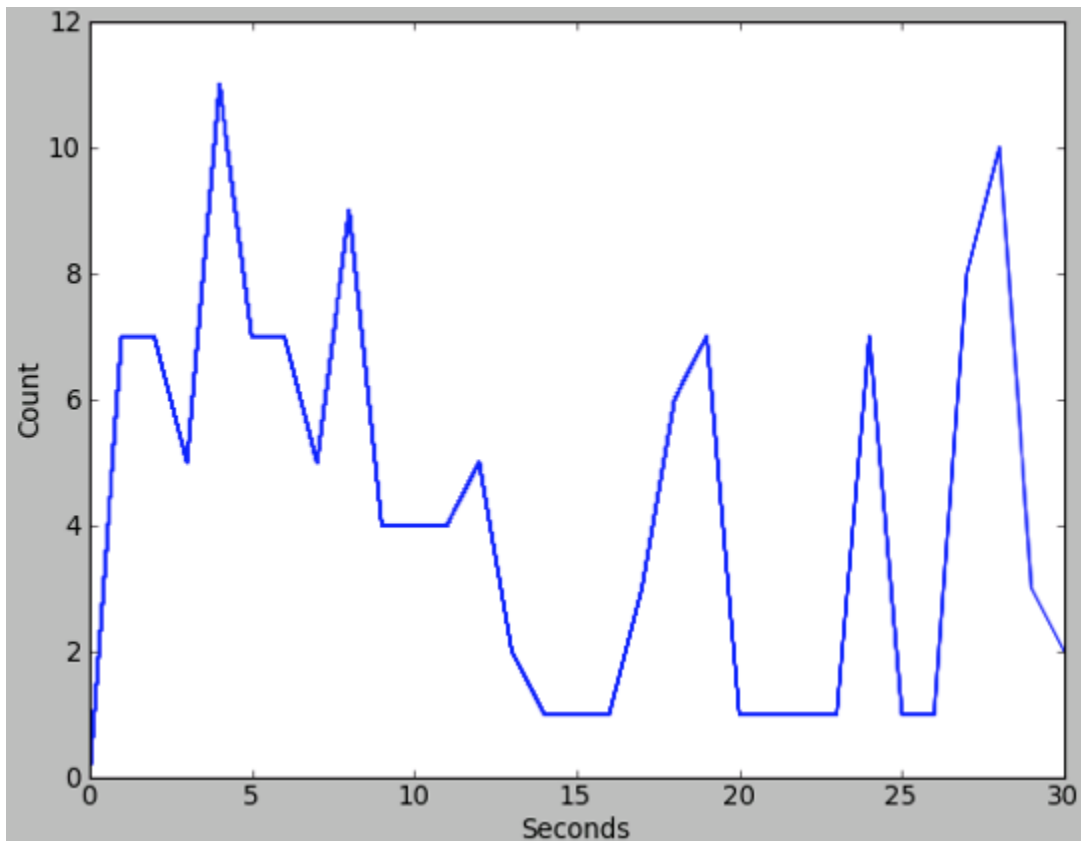
```
./LiveTweets.py time
```

May 10 02:32:22 RT @GraysonDolan: What ever it is that you're going through at the mom
May 10 02:32:22 I'll be yours 'til the end of time
May 10 02:32:22 RT @Pontifex: Jesus, ascended into heaven, is now in the lordship of G
May 10 02:32:22 Now every time I conduct 5/4 I get so confused! Lol
May 10 02:32:22 RT @Dayloveme_: I don't have time for games and please don't waste my
May 10 02:32:22 RT @FoxNews: .@GeraldoRivera: "There came a time where @HillaryClinton
May 10 02:32:22 One time I logged on and there was a gif of a man playing drums with h
May 10 02:32:22 Hi @1010XL please bring @Ballou1010xl back to the PM drive time show.
May 10 02:32:22 RT @amoghdsad: If Hilary Clinton wins, it would be the first time in
May 10 02:32:22 So glad to be home. Now time to stay for 3 more hours and study for La
May 10 02:32:22 Last time I checked I have a house 

Based on the timestamps, the word *time* appears more often than *president*.

Step 3. Plot real-time frequency of tweets. We can create a plot showing the frequency of tweets containing a specific word. Run *PlotTweets.py president* to create the plot for the word *president*:

```
./PlotTweets.py president
```

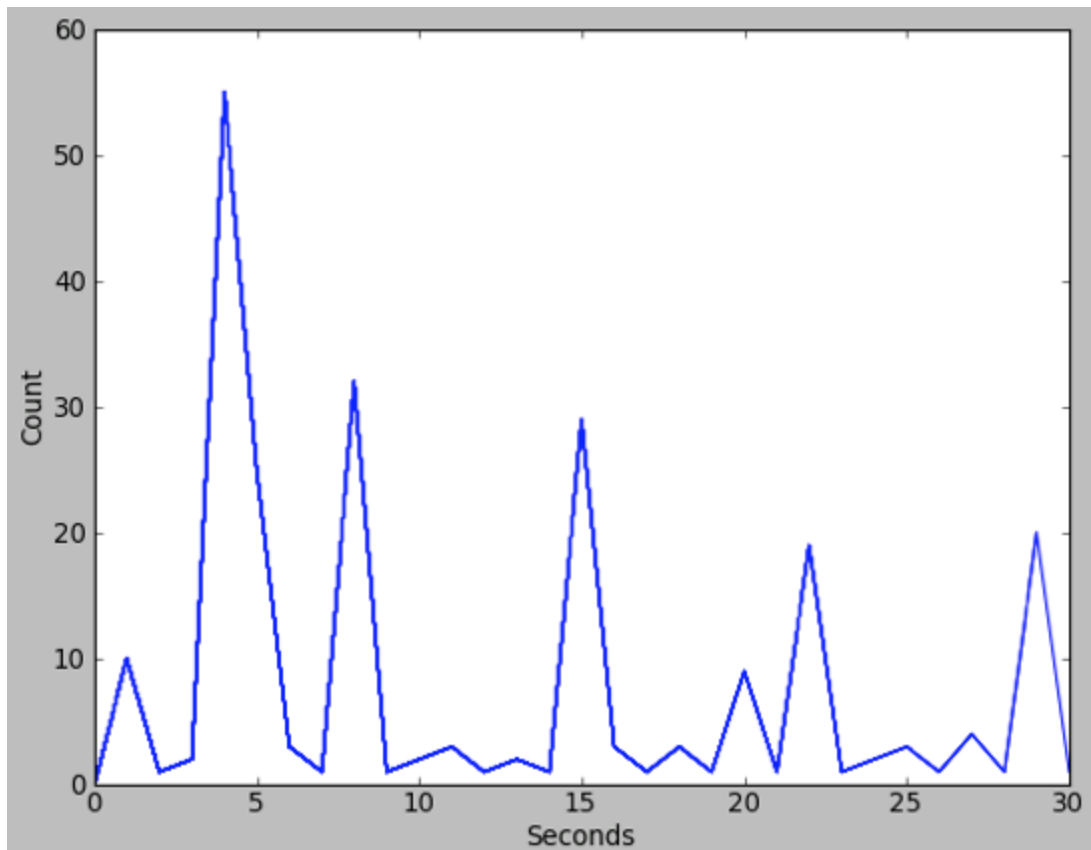


In this plot, we can see the variation over time of the number of tweets per second containing the word *president*. Over this time period, the maximum was 11 tweets per second.

When you are done looking at the plot, click in the terminal window and press *enter*.

Now let's look at the frequency plot for the word *time*:

```
./PlotTweets.py time
```



In this plot, we can see the maximum number of tweets per second containing the word *time* was about 55, which is much higher than the maximum for the word *president*.