TWITTER DATA COLLECTION REPORT

Comm370 Project 1

Abstract

A project that will use Twitter Streaming API to collect tweets about the two brands over a 10-hour period (1K tweets each)

Prepared by:
Haihua Zhu 20806155
Mo Chen 52909158
Junyuan Wang 36063162

Haihuazhu1004@gmail.com
Mo.angela.chen@gmail.com
noic.junyuanwang@hotmail.com

Procedures:

For the actual code of this project, we use two major modules, “twython” and “datetime”.
Since all the required modules are already installed in the EC2 node.

A. [Code with Twitter API]

a. Our group picked two single-word brands: Microsoft and Amazon
b. Apply for Twitter developer account and get consumer key, consumer secret, access token, and access token secret to use Twitter Streaming API.
c. Applied Twython[1] module and made modification to the code to collect live tweets using Twitter Streaming API. We just simply import the “twython streamer” from “twython” to collect live tweets. The “twython streamer” would collect live tweets based on the keyword through twitter api. We have obtained our credentials beforehand and store it as a json file. This would facilitate the “twython” module to login into the twitter api.

d. Before we start, we need to modify the original code of the “twython streamer”.

e. First, we add a litter print session to show the tweets that are being collected. And every time it collects 100 tweets, we will store it into a json file. After the saving process, the twython streamer would disconnect itself from the twitter api.

f. For clarity and tidiness, we use “datetime” module to get the current year, month and date. We used the current date as the name of the saved json file so that the streamer wouldn’t overwrite the existing files.

```
g. import datetime #importing the datetime module
h. currentDT = datetime.datetime.now() #getting the current time as currentDT
i. currenttime=currentDT.strftime("%Y-%m-%d-%H") #geting the time as  YY-MM-DD-HH format
j. #storing the 100 tweets as JSON file
k. with open('{}-{}.json'.format(currenttime,keyword), 'w') as f:
       json.dump(tweets, f, indent=4)
```

B. [Work on the Cloud]

a. Use credential of amazon web service to log on to amazon server & use the same credential to set up filezilla to put relevant document under amazon server.

[1] Twython is one of many Python packages to use Twitter API. You may use other modules.

b. Remotely access the EC2 node through Terminal (Mac), and Copied the Python code into Amazon EC2 node using scp (e.g., FileZilla).
c. Made minor changes in the EC2 using text editor (e.g., nano) when we need to change relative pathnames to absolute pathnames.

C. [Data collection with cronjob]

a. Using crontab in the EC2 node, add two hourly cronjobs where each cronjob collects 100 tweets for a single keyword.
b. To create an automatic repeat schedule. We use "crontab" to set up a repetitive task that would go on very 22 minutes and 59 minutes after the hour, using two different keywords. So, the modified "twython streamer" would automatically collect tweets with the two given keywords every hour until we stop the task.

D. [What we have learned]

a. The most important thing we learned is that we could use people's code and import it or modify it, to achieve more powerful result we wish to reach. It is like we stand on the giant's shoulder to see further.

b. We didn't know that we can use an external virtual computer to help us run the code over a long time period so that we don't need to keep our computer running all day.

c. We definitely learned how to use Linux from class and data camp through some basic commands.

d. We got to know FileZilla so we could easily transfer file from the virtual computer to local, and vice versa.

e. Last but not least, we learned that the possibility of utilizing python and all the related technology is countless, we feel confident about the future of data science. And our group members expressed their interests in further pursuing data skills.