Problem Statement

One of the magical things with Twitter is when famous users publicly converse with each other via tweets!

Find 10 such separate examples of conversations in the last 2 weeks where one example is the triple (A, B, T1, T2) such that A and B are the two users conversing and T1 and T2 are any two tweets in their conversation such that T1 is sent by A to B and T2 is sent by B to A. A should have more than a million followers on twitter, and B should have at least 125,000. Please include your code, comments on how to run it, and its output (i.e., the 10 such examples you find). For T1 and T2, include the URL, text and the time of the tweet in your output.

Hint: Use the Twitter Public API to grab tweets of a given user.

Solution

The following steps are involved in solving the above problem:-
1) Fetch user list sorted by follower count from http://twittercounter.com and dump it to a file (userList/topUsers.txt). [Optional – already prefetched]

$python getTopUsersByFollowers.py

At the backend, I essentially crawl http://twittercounter.com/pages/100/__offset__ (where 0 <=__offset__ <= 980) using YQL (Yahoo! Query Language) to extract top twitter users sorted on follower count.

2) Start fetching the conversations by executing below command:-
$python getTweets.py

a) Load the user list from "userList/topUsers.txt" into "topUserList" (all users in this list have > 1 million followers)

b) For each user (say "topUser") in topUserList, fetch 100 recent tweets (public_tweets) from user_timeline.

c)

i) For each tweet in public_ tweets i.e., T1, check if "in_reply_to_status_id" flag was set and tweetAge <= (2 weeks ago), if yes, it means that this tweet was in response to another tweet.

ii) If "yes" in previous step, fetch the original tweet i.e, T2, for which the reply was posted. In this tweet, the "author" object contains followerCount attribute, ensure that the followerCount > 125k and tweetAge <= (2 weeks ago)

iii) If both conditions c(i) and c(ii) are satisfied, then T1 and T2 is an example of conversation listed in the problem, print it to output file "result.txt"

iv) If 10 such conversations are found, break the loop, else go to Step 2b

3) Since conversations listed in the problem are sparse, we may hit Twitter API hourly rate limit (350

authenticated requests per hour), in such cases, we save the state of result and current topUser being processed in “result.txt” and “currUserIndex.txt” respectively. The processing can be resumed by following the steps listed below:-

a) To overcome the hourly rate limit, I created multiple Twitter Applications and copied the corresponding credentials in to a different file. Ideally, these credentials should be stored in some type of authenticated database rather than in plain text files.

b) Change the credentials by choosing a different credential file in “auth.py” (Line 12-19) and change the IP address (reset router or run on a different machine).

c) Run 'python getTweets.py' to resume the search and the program exits when 10 such conversations are found successfully.

4) Technologies / Libraries Used – Python 2.7, Tweepy, YQL (python-yql)

Code Explained

1) getTweets.py → main program that fetches 10 conversations as listed in problem statement
2) getTopUsersByFollowers.py → fetches userList sorted on followerCount
3) auth.py → loads credentials and gets Tweepy API handle to query Twitter API
4) helper.py → provides miscellaneous helper functions
5) result.txt → output file (contains 10 conversations as listed in problem statement)
6) consoleOutput.txt → console output
7) currUserIndex.txt → stores index of current user being processed (helps resuming search)
8) README.txt → contains on how to run the code and prerequisite installations required
9) Solution.pdf → detailed explanation of solution