In this project, I developed a python crawler which can efficiently crawl the tweets data in a real-time and preprocessed the raw data to extract the useful information I need including tweets text, user and time. Finally I used the linear influence model (LIM) to analysis the data and extract the hot topics.

I. Crawling raw data from twitter

“Crawler.jar” is a simple user interface to guide the user crawl data from twitter. User can follow the steps on the interface:

1. Start: Click “Crawler.jar” to start the program:

2. Step 1: user can select multiple twitter users and save it, then the crawler will crawl tweets record of these selected users.
3. Step 2: run crawler
4. Step 3: Authorization URL:

Copy the URL to browser, and login in twitter, then the PIN number will be shown:

5. Step 4: enter PIN number on the command window, then the program will begin to crawl the tweets:
6. Step 5: visualize the result: since user select the twitter users named “eminhaj” and “kevinspry”, thus the result includes these two users record.
For each user, the crawler crawl the whole tweets and create one txt file. Each line is one tweet record shown below, I use the blue color to emphasis "text" information and red color for time information.

{"favorited": false, "in_reply_to_user_id": 14779052, "contributors": null, "truncated": false, "text": "@brandonchicago Come join the fun. Plenty of room on the bandwagon!", "created_at": "Mon Oct 17 15:07:20 +0000 2011", ……}

The demo is shown on Youtube:

http://www.youtube.com/watch?v=BT4LXPUtchY

II. Preprocess the raw data

1. Extract the information about “Text”, “Created_at”, “username” attributes

There are many attributes in each tweet record, due to our need for data format, we are only interested in the “text”, “created_at”, “username” attributes. Thus, I use a python code to preprocess the data and extract these information from raw data including 1000 twitter users record (until now we have collected 1000 twitter users’ record).

2. Select the interesting topics

Then I use word count for each English word and filter out some words like “good” “Ok” which cannot be the topics. Finally we can select the interesting topics by this ranking. The 50 selected interesting topics with high frequencies are:

google twitter facebook iphone youtube ipad gutschein startup android nyc francisco mac sysoon Microsoft yahoo ios ceo php amazon wordpress samsung nokia linkedin obama gmail chrome blackberry roverpet camera tablet sony itunes miami wsj indiasmash posso skype das internetdj mit flickr har kunst reuters pas groupon Netflix sxsw starbucks kindle

3. Select the active users

Based on the selected interesting topics, I count the total times for each user who mentioned these topics, then rank the user by the frequency and select the top 200 active users out of total 1000 twitter users. These active users are people who mentioned the interesting topics most times.

4. Construct the diffusion matrix

Finally I extract the tweets information for these hot topics and organize them in the format:
In the above data format, it contains a matrix data where each row vector corresponds to a specific time index of an interesting topic. The data has N+2 columns where N is the number of nodes that we model their influence functions. The first N columns denote whether each of N nodes become infected to the concept at the time index. The N+1-th column specifies the volume of the concept at the time (count the number of users who mention that concept), and the last column (N+2-th) specifies which concept the row corresponds to. M_{u,k}(t) denotes whether node u got infected to concept k at time t.

### III. Linear Influence Model

I applied the state-of-the-art work proposed by Jaewon Yang and Jure Leskovec [1] to analyze the twitter data set. The authors focused on modeling the global influence of a node on the rate of diffusion through the (implicit) network. They model the number of newly infected nodes as a function of which other nodes got infected in the past. For each node they estimate an influence function that quantifies how many subsequent infections can be attributed to the influence of that node over time. I utilized their Linear Influence Model (LIM) to analyze our twitter data set and predicted the future hot topics. The result is shown below. In the plots, blue line represents the true volume of each topic along the timeline, red line represents the predicted volume by LIM. From the figures, we can see that LIM can approximately predict the volume of topics at time.
Also, LIM model can predict the future volume of each topic, thus we can predict the future possible hot topics. Currently, based on the twitter data set, the future 10 hot topics are:

sysoon
reuters
skype
blackberry
kindle
flickr
starbucks
ios
sxsw
groupon

There are some findings from the above interesting result:
1) Future hot topics include different types:
   Websites: sysoon, reuters, flicker, sxsw, groupon
   Electronics products: blackberry, kindle
   Computer software: skype, ios
   Cafeteria: starbucks
2) The ranking of hotness of these future predicted topics is not same as the ranking of word counts of interesting topics. In other words, although some topics have less word count from tweets, they may be the hot topics in future.
3) These hot topics can benefit the people in several ways, like finding the possible hot commercial products among twitter users.

IV. Summary

The main work of this project is collecting, understanding and analyzing social network data. In this project, I developed a python crawler which can efficiently crawl the tweets data in a real-time and organize the tweets into a structured form. Based on our well-organized structured form, I used LIM model to explore interesting patterns of social-behavior from the data and also predict the level of hotness of a certain topics from the data.

Reference