CS 2310 Project Milestone 1

Yingze Wang
yiw32@pitt.edu

Twitter Web-Data Crawler

Due to large amounts of available social-media data nowadays, the task of collecting, understanding and exploring such data becomes increasingly important. In this project, our goal is to first develop 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, we would like to further explore interesting patterns of social-behavior from the data and also predict the level of hotness of a certain topic from the data.

Firstly, we would like to develop the crawler which collects the following information from twitter:

1. Tweet’s id
2. Tweet’s content
3. Tweet’s time
4. User’s id
5. User's screen-name
6. User’s followers count
7. User’s other information from the profile

We will try to identify a list of hot topics of public interest. We will filter out those tweets related to these hot topics and organize them in the format:

```
data=[<M_{1,1}(1)>, <M_{2,1}(1)>, ..., <M_{N,1}(1)>, <V_1(1)>, <1>;
<M_{1,1}(2)>, <M_{2,1}(2)>, ..., <M_{N,1}(2)>, <V_1(2)>, <1>;
<M_{1,1}(3)>, <M_{2,1}(3)>, ..., <M_{N,1}(3)>, <V_1(3)>, <1>;
...;
<M_{1,1}(T)>, <M_{2,1}(T)>, ..., <M_{N,1}(T)>, <V_1(T)>, <1>;
<M_{1,2}(1)>, <M_{2,2}(1)>, ..., <M_{N,2}(1)>, <V_2(1)>, <2>;
...;
<M_{1,K}(T)>, <M_{2,K}(T)>, ..., <M_{N,K}(T)>, <V_K(T)>, <K>];
```

In the above data format, it contains a matrix data where each row vector corresponds to a specific time index of a concept of interest. (If there are K concepts of interest and each concept
has T time indices, then data has KT rows.) 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. If the node u (user) mention the concept k at time t, \( M_{u,k}(t) =1 \), or if the node u doesn’t mention the concept k at time t, \( M_{u,k}(t) = 0 \). Also, the data file contains two auxiliary vectors. The (K,1) cell conceptname contains the details of K concept of interest. The (N,1) cell nodename contains the user id of each user (nodes).

Finally, we will try to carry out a few interesting data mining jobs based on the collected data. Specifically, we would like to predict the future volume for each topic. In addition, we would like to identify most active users for each topic. If possible, we would like to study the common and different characteristics among groups of active users for different hot topics.