steve_project_proposal

April 16, 2024

1 Milestone 1

2 Project Proposal

2.1 Description

My intended audience is lobbyists and any other individual firm who may seek to influence the passage of legislation by identifying and communicating with specific Members of Congress. My intended audience is also anyone who is interested in how individual congressmen use social media and seek to study the relationship between social media use and the movement of legislation through congress. I seek to study different measures of what it means to be an influential twitter user: summary type statistics (retweeets, likes follows, etc.) as well as social network analysis if it can be done.

2.2 Questions

1.) What are the different ways one can measure influence in a community of twitter users?

2.) What kind of derived relationships can I create among the twitter data we get?

3.) Are there going to be any pitfalls (say, pitfalls associated to computational complexity) with respect to determining metrics of influence? Do we have the computational resources necessary to do this?

4.) How many retweets does each user have to his or her credit?

5.) How many follows does each user have to his or her credit?

6.) Does tweet volume correlated to these and other measures of influence?

7.) Can we derive usable social networks from user accounts associated to specific hashtags in tweet body text?

8.) Can use derive usable social networks from user accounts associated to specific handles in tweet body text?

9.) What other measures of influence are known among people who analyze twitter data?

2.3 Hypotheses

1.) The more tweets one produces, the more collateral twitter traffic (retweets, follows, etc.) one will have.

2.) There will not be enough hashtags and twitter handles in the twitter traffic to generate a usable social network, but I'll try something anyway.

3.) The presence of a lot of null values in the data will make it difficutlt to generate metrics that feel reliable.

4.) There might be different screennames belonging to the same account to account for in the twitter data.

5.) I expect that the in_reply_to_screen_name field, when quantized, should be bounded above by a user's retweet count.

6.) I'm interested in the relationship between statuses_count and other measures of popularity.

2.4 Approach

1.) I plan to use joins on the userid columns (which have different names) to correlate userids with specific metrics associated to screennames.

2.) I will extract hashtags and twitter handles from the text field, and I will attempt to build a social network with that data.

3.) I don't know how exactly the in_response_to_status_id field can be used, but I will see if I can correlate anything to it.

4.) I intend to explore the 'favorite_count', 'id', 'in_reply_to_screen_name', 'in_reply_to_status_id', 'in_reply_to_user_id', 'quoted_status_id','retweet_count','screen_name','text', and 'user_id' from the tweet data. In the user data, I'll explore the 'description', 'following', 'followers_count', 'friends_count', 'id','listed_count','screen_name','statuses_count', and 'verified' fields from the user data.

5.) I intend to do a lot of summary statistics and simple aggregation to count retweets, likes, follows, etc. for each twitter user.