**Finding Similar Strangers on Twitter**

A person-to-follow recommender based on content

1. **Motivation**

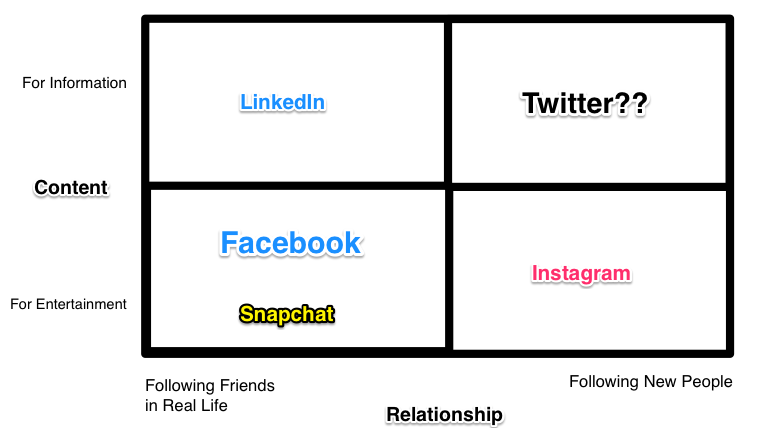
Do you know of someone who created a Twitter account but rarely logs on?

Twitter’s biggest differentiators are:

1. Connections can be based on interests, not personal friendships.
2. The ability to share substantive content.

However, Twitter’s current “Who-To-Follow” algorithms are based on Triad Closure and a similarity-to algorithm, both of which are based on existing connections. This creates a discovery and “cold-start” problem for finding people.

**2 x 2 Analysis of Market Positioning of Popular Social Networks**



Currently, finding “interesting strangers” is a time consuming process. There is nothing that facilitates communities around similar interests. With this project, I aim to create a recommender for people-to-follow based on content, with the goal of creating new conversations and conversations.

1. **The Project Overview**
2. **The input**
   1. *Option 1*: User authorize the app their Twitter. The app will scrape their tweets and build a “user model” and find other people similar to this model.
   2. *Option 2*: User will select their favorite tweets from “most popular tweets” and the app will build a “user model” and find other people similar to this model.
   3. *Option 3* (probably most feasible): User inputs a twitter handle and clicks on “find similar”.
3. **The Features**
   1. Featurization of User Attributes. Each of the following represents a column.
      1. Reading level of the tweets
      2. # of followers
      3. # following
      4. # tweets / day
      5. # of pictures
      6. # of news articles
      7. “tone” -- ?? would be nice to have sentiment analysis
   2. Featurization of Tweets by topic. Each of the following represents a column and in each column, there will be count (normalized) of # of tweets in the following topics.
      1. Current events
      2. World News
      3. US News
      4. Economics & business
      5. Comedy
      6. Entertainment
      7. Science
      8. Arts
      9. Strange and aberrant, Etc…
4. **The Models**
   1. TF-IDF and NLP to featurize tweets by topic
   2. K-means, Non-Negative Matrix Factorization
5. **The Output**:
   1. List of similar people.
   2. If I have time, I’d like to create a “twitter page” with tweets from these people.
6. **The Data**

All the necessary data is available by API.

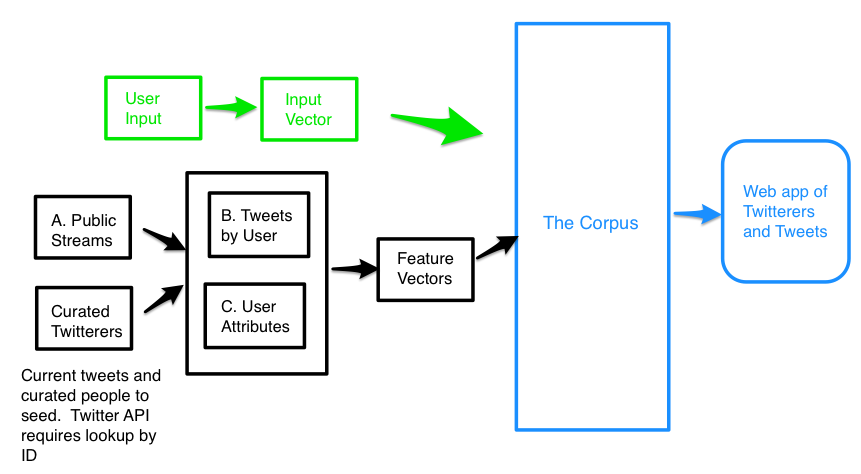
A. Public Streams**:** <https://dev.twitter.com/streaming/public>

B. Tweets per User:<https://dev.twitter.com/rest/reference/get/statuses/user_timeline>

C. User Attributes: https://dev.twitter.com/rest/reference/get/users/show

D. Information on Followers: <https://dev.twitter.com/rest/reference/get/followers/ids>

**The Data Pipeline**

****

Currently missing: rows of data, but I plan to have this information by end of day 12/22.

1. **Possible Hurdles**
   1. ***Twitter API difficulties***
      1. Start gathering data immediately
      2. Enlisting help from friends who are programmers to help
   2. ***Weird suggestions***. Given that it is a complex and contextual corpus, it may be hard to get helpful suggestions. The mitigation plan is:
      1. Limit the scope by only focusing on a narrow domain (for example: news recommendations, or comedy) so that I really tune the parameters to suit the topic.
      2. Seed the corpus with popular / suitable “interesting strangers”.
      3. Get data for another (easier) project if this fails?

**References**

Ashish Goel, Aneesh Sharma, Dong Wang, Zhijun Yin: *Discovering Similar Users on Twitter*.

<http://snap.stanford.edu/mlg2013/submissions/mlg2013_submission_20.pdf>

Scott A Golder, Sarita Yardi: *Structural Predictors of Tie Formation in Twitter: Transitivity and Mutuality.*

http://www.redlog.net/papers/golder\_yardi\_2010.pdf