Twitter Project notes

Research Question:

* Actual Question:
  + Do the rate of Tweets about HIV match the prevalence of disease in American big cities
    - Cities with over X population?
      * How are big cities defined
      * Top 50 cities in terms of population
    - Miami, New York, LA, Chicago
  + Methods
    - ?
* Ideas
  + What factors impact how often a tweet is retweeted
    - Are funny tweets about HIV better than serious tweets?
    - Are Tweets from a health agency more effective than a tweet from a private account?
    - Problems:
      * Do we need IRB approval?
      * What ethical considerations are there?
  + Do tweets about disease match up with incidence?
  + Impact of news on awareness of diseases
  + Does the rate of Tweets about disease risk behaviors correlate with the actual rate of infectious disease in big US cities.
    - Need
      * Disease rates in big cities
      * Tweets about the disease
      * Tweets about risk factors
      * Total Tweets in a big city per (day, hour)
  + Start simple
    - Basic prevalence of tweets about disease
  + Get complex
    - Do Tweets about disease risk factors correlate to the prevalence of disease
    - Do monthly Tweets about disease risk factors correlate with future disease rates
* Questions
  + What searches do we need to run?
    - What are our search terms?
  + How long do we need to run the program for
    - Should we run it for 24hrs or 8 hrs for multi days…?

Demon: Tweets per time period

Number: Topic specific tweets

Topic of interest:

TO DO:

* Modify the code so that time is more of a factor
  + So that we can run the code for set amount of time and so that
  + We can schedule a time to run the
    - Use a while loop
      * Add vars
        + Start time
        + Run length/end time
      * If /while start LT current time then do nothing
        + Else run code for run length
* Find out if you can pull old twitter data?
  + Is there code to pull retrospective data from Twitter
  + We could validate a monthly predictive model using this
* Research Carmen
  + See healthtweets.org presentation
* See if I can pull the Florida groceries (other)
* ~~Change delimiter to something used less frequently (~)~~
* ~~Only print tweets with geocode, if geo codes re needed~~
  + ~~Get clean Tweets about HIV~~
* Review research on uses of twitter in PH research
* IF twitter does not work
  + Look for machine learning dataset
    - Find big publicly available longitudinal data sets
  + Can we do a survival analysis with machine learning
    - Framingham data?