Eli Bolotin
DS 710: Final Project
November 8th, 2018

# Final Project Proposal

1. **What is your question? You might be interested in a brand, an organization, a meme, a trend, a news story, or a place. What do you want to learn about? Do you have predictions or hypotheses about what you will find?**
   - After brainstorming a host of prompts and testing their technical feasibility (given the limits of using Twitter's API for free), I arrived at the question below.
   - My question is: are people that are vocal about their physical activity happier than people who are vocal about their media consumption?
   - Same question phrased differently: Are people that *talk* about working out happier than people who *talk* about watching media (tv, movies, internet media, etc.)?

2. **What is your audience? Who is interested in the answer to your question?**
   - My audience is diverse and includes the following groups:
     i. Business organizations that are interested in promoting health and fitness, such as gyms, health clinics, wellness centers, and insurance companies.
     ii. Media companies who want to cater an ever-increasing health-conscious consumer segment.
     iii. Academic and private institutions that study the psychology of happiness and wellbeing.

3. **What information will you gather about each tweet in order to address your research question?**
   1. First, I will collect the user_ids of current Twitter users that fall into two groups: "fitness vocalizers" and "media consumers". Each group is determined by the hashtags that they currently use in their tweets.
   2. Then, using the user_ids of the authors in both groups above, I will pull their other (**non-filtered**) tweets. These are the tweets that both groups write in general, and may or may not be fitness or media related.
   3. The aggregate dataset that I plan on using for statistical analysis will include:
      i. Sentiment analysis of tweet text
      ii. Tweet length
      iii. Hashtag text
      iv. Hashtag count

v. And other user, follower and friend parameters.

4. **How will you collect the data? Will you use the REST or Streaming APIs, or a combination of both? What specific keywords, hashtags, or other features will you search for?**

   - I will collect data using both REST and Streaming APIs.
   - For both of the groups mentioned above, I have about 25-30 key words that will be tracked in the Streaming API. My program is in active development, so this list is being tweaked.

5. **How will you analyze the data? Will you compute any proxy variables (such as number of exclamation points as an indicator of emotion, like in homework 11)? Choose at least one pair of specific hypotheses to test. What type of hypothesis test will you use?**

   - I will analyze the tweets of both groups to perform sentiment analysis to determine positivity, neutrality, negativity, and possibly the presence of other emotional states.
   - For sentiment analysis, I will (possibly) use the Natural Language Toolkit for Python, and probably another library as well.
   - **Null hypothesis** = People that are vocal about their physical fitness *are not* more or less happy than people that are vocal about their media consumption.
   - **Alternative hypothesis** = People that are vocal about their physical fitness *are* happier than people that are vocal about their media consumption.