Proposal for Final Capstone

* What is the problem you are attempting to solve?

Due to my background in healthcare, I’ve always been interested in keeping track of diseases, outbreaks,

and other epidemiological occurrences. Being that the flu season is upon us (in the U.S.), I’ve decided to focus my final capstone on predicting flu outbreaks. But this is no easy feat. Usually, we have to rely on gathering data in retrospect from many different types of health providers, which can take time to process before it is determined to be a full blown outbreak. But thanks to social media, we can now use real time data to analyze flu outbreaks. Enter twitter. With live tweets coming from all over the nation, we can pinpoint real time occurrence of the flu. This can bring about advances in surveillance of outbreaks of other disease, not just the flu.

* How is your solution valuable?

My solution is valuable for numerous reasons. One field where it is invaluable is public health. I can imagine local health departments using flu prediction models using twitter to track and treat flu outbreaks. Hospitals would also greatly benefit if they know they are expecting an outbreak in their area by being able to prepare staff and contain infection. By being able to predict flu outbreaks, the general public can be warned and prepare ahead of time, possibly even encouraged on getting the flu vaccine.

* What is your data source and how will you access it?

Using social media such as twitter gives us the advantage of real time data. Twitter also has geolocation attached to tweets that pinpoints us to the exact hotspots. We can gather all these data from twitter API. Access to the API requires a registration as a developer. I’ve gained access to the API and currently exploring the API.

* What techniques from the course do you anticipate using?

Using a text-based data requires the use of Natural Language Processing (NLP). Under NLP, I will explore the use of supervised and unsupervised techniques. Using multiple NLP libraries such as NLTK and Spacy is a must. I will explore using Bag of Words and Tf-Idf. Creating bigrams and trigrams would be important in understanding the text data. I’m expecting to use LSA, LDA, or NNFM for text extraction.

* What do you anticipate to be the biggest challenge you’ll face?

Text based data will have a lot of noise. Removing exogenous factors will be difficult. Cleaning the text and classifying it as flu related or not will be the biggest challenge. Another challenge will be looking at historical data from tweets. Ideally, I would like to look at past flu trends from previous years and compare it with CDC surveillance. This may or may not be doable, for the time allotted for this project.

Also, it can be hard to generalize the results since only a certain amount of the population uses twitter. For example, the elderly, who are at higher risk for catching the flu is less likely to tweet about their flu experience.

Trying to gather data from only part of the U.S. (Northeast) opens up more questions about generalization. Although Flu season generally occurs from September until April, the environment varies based on location and other factors, therefore, this could only be (hopefully) generalized in the Northeast.