**Unstructured Data Analytics for Policy**

**Final Project Proposal**

Team Members: Ziwen Wang, Yasuhito Morimoto, Max Weisz, Yiqi Zhou

**Policy question**

Are there ways for policy professionals or regional governments to identify the severity of instances of disaster within their region, given the tweets of a population or a group of people?

Are there serious tweets and are there not so serious tweets (either with hashtag structures or without) and is there a way to cluster them or to do topic modeling on them to come to a “severity rating” to assist in resource allocation during disasters.

**Background**

Twitter is a source for signaling that is often the source for information and crowdsourcing events, trends and other social phenomena. Twitter has been used in the past to analyze and understand certain financial trends for asset management and to understand political voting and trends. The mechanism for this is often “hashtagging”, which can be considered a form of structured data. However, often times the attached “tweet” is contained in an unstructured format and the ideas and concepts around a specific hashtag do not have the consistency seen in other structured data formats. We would like to analyze this data as it pertains to the disaster dataset on kaggle.

*Disaster Monitoring* has always put a large burden on the public sector in terms of resources used, time required and general inability to cover all things at all times. This has led municipalities and larger organizations to invest heavily in disaster recovery and in systems that will be able to monitor and advise resource allocation. Targeting regions and disaster relief can often times be the single best investment from a public finance perspective. We will attempt to develop a tool to analyze these datasets and provide organizations at every level a way to gauge the severity of a crisis/disaster.

**Data Description**

We are going to use the dataset of disasters on social media from kaggle. This dataset includes 12,000 tweets, including the text of these tweets and which label they belong to. All tweets are chosen by searching disaster related words on twitter, like “ablaze”, “quarantine”, and “pandemonium”. The label is judgment about whether the tweet referred to a disaster event (Not non-catastrophic thing like a joke or movie comment with the word).  
  
The unstructured element is the text column of all data. We would use text analysis to process the keyword of different tweets to find out which words are highly related to real disasters and which are not based on the classification.  
  
Datasets: https://www.kaggle.com/jannesklaas/disasters-on-social-media

Disasters on social m

**Proposal analysis/Evaluation of Quantitative Analyses**

1. Clustering (Gaussian Mixture Models):

* Use both of disaster relevant/non-relevant tweets
* Perform clustering all tweets into disaster relevant/non-relevant to analyze what kind of messages are used in an emergency
* Develop the filtering to find disaster relevant/non-relevant algorithm may help an emergency alerting system
* To address this, generate a document matrix from the tweet text data, and make it into a few clusters
* To evaluate the best number of clusters and the structure of relevant/non-relevant, characterize the clusters by using some test words which are likely to be used in disaster
* Test words can be picked up from frequent words from all the relevant/non-relevant tweets (expect to use word cloud)

2. Topic modeling (LDA, t-SNE, K-means):

* Use only disaster relevant tweets
* Tweet data are labeled from 221 keywords related disasters, in addition, each tweet may have some specific message such as need someone’s help or just tell the impression
* Categorizing tweet messages may let others efficiently know what is needed
* To address this, reduce dimensions from 221 keywords to a smaller number of topics (perhaps less than 10, need to check with co-occurrence) by LDA, and find what are the frequent words in each topic
* To evaluate the topic modeling is successfully done, visualize the tweets in 2-D space using by t-SNE and K-means (the number of clusters is equal to the number of topics)

3. Classification (k-NN, SVM, RF):

* Use both of disaster relevant/non-relevant tweets
* Find the best relevant/non-relevant classification model so that people can efficiently catch severe messages from a huge amount of tweets
* To address this, evaluate the train tweets data and find the best hyperparameter in each k-NN, SVM, RF method with F1 score
* To evaluate the models performance, compare their ROC curves

**Expected Finding**

1. Find out some words used most frequently in emergency.
2. Analyze the relevance of specific words and disasters.
3. Find out which types of disasters occur most frequently and in which areas.