CAPSTONE THREE: PROJECT PROPOSAL

- Problem Statement Formation: How can we predict and detect as accurately as possible when a tweet concerns a current disaster apart from other usage like metaphorical? Can we build or make use of an NLP model that will help us predict with 90%+ accuracy whether the words in a tweet signify a disaster or not to enable timely disaster relief efforts?

- Context: Twitter is a social media platform that has become an important medium of communication. Oftentimes, with the ubiquity of smartphones, users can witness and notify right away of disasters that are happening around them. Various entities such as disaster relief organizations and news agencies wish to make use of the platform's tweets to respond to and report on these crises. A sentiment analysis model could be the right solution in addressing the needs of these entities in detecting which tweets are identifying a disaster and which are not.

- Criteria for Success: An accuracy in prediction rate of 90%+ that singles out which tweets are announcing a disaster and which are simply a matter of rhetoric or metaphor to take advantage of a popular social medium (Twitter) in expediting necessary relief efforts and broadcasting news alerts.

- Scope of Solution Space: The NLP sentiment analysis technique will be employed on a corpus of 10,000 collected tweets from Twitter to train the sentiment analysis model in correctly deciphering and discerning emergency tweets from non-emergency based on the words in a sentence.

- Constraints: Machines cannot as easily interpret whether a sentence is to be taken literally or figuratively unlike people. Any guarantee of a perfect prediction success rate is unrealistic and untenable.

- Stakeholders: Twitter, disaster relief organizations and news agencies need to be involved in some capacity in this endeavor. All three aforementioned groups will be presented with the results and solution for collaboration, implementation, and input of these ideas. The data is sourced from a corpus of tweets collected by a company named Figure-Eight.

- Data Sources: A corpus of 10K+ tweets collected by Figure-Eight and made available on their website ‘Data for Everyone’. Critical pieces of data needed to address the above hypothesis are the features of ‘text’ and ‘location’; for our purposes here, ‘text’ more so as this contains the text to run through the NLP model and the ‘location’ would be necessary for the disaster relief organizations and news agencies.