**Pandemic Tweet Challenge**

**Group Names:** Yudong Lu, Fengyi Zhao, Javan Reuto, and Bradley R. [Mascarenhas](https://umd.instructure.com/courses/1314706/users/4280535)

**1) Description of Research Question or Problem**

The project is to identify the nature of the Covid Tweets pulled from Twitter during the coronavirus period based on Natural Language Processing. We will need to clean the tweets and classify each one using Text Classification into 1 of the 5 category class labels which are neutral, positive, negative, extremely positive, and extremely negative.

**2) Data Sources and Description**

We are not collecting data on our own but directly using the data that Kaggle provides us with. The data consisted of the training and testing dataset. The training dataset consists of 41157 tweets with corresponding sentiments. The test dataset has 3798 tweets with corresponding sentiments. The training dataset’s columns include Username, ScreenName, Location, Tweet At (Date), Original Tweet, Sentiment. There are five categories for the sentiment of tweets, Neutral, Positive, Negative, Extremely Positive and Extremely Negative. We are going to train our model using the training dataset and then use the test dataset to predict the Sentiment for each tweet. Kaggle also provides a sample submission file including original tweets for us to insert the predicted sentiment class label.

**3) Possible Methods**

In the project, we will use LSTM or Fully connected feedforward network for model training.

TFIDF will allow us to examine the relevance of words across tweets. Count vectorizer incorporates the tokenization process and returns the frequency of words. The data set covers the pandemic topic; using relevance in TFIDF would be useful to differentiate which words in the same topic are relevant or useless.

**4) Expected Outcome**

By applying the testing data to the model trained, we should expect to have a high accuracy rate in classifying the text into the five sentiment categories, Neutral, Positive, Negative, Extremely Positive and Extremely Negative.

Kaggle link:

[Pandemic Tweet Challenge | Kaggle](https://www.kaggle.com/c/pandemic-tweet-challenge/overview)