Github repository: <https://github.com/871234342/NYCU_AICap_FinalProject>

[**Natural Language Processing with Disaster Tweets**](https://www.kaggle.com/competitions/nlp-getting-started)

The project aims to tackle the challenge of finding tweets about disaster. The ubiquitousness of smartphones allows people to share information in real-time, like the disaster they are observing. As a result, programmatically monitoring social media, like twitter, could help gathering information.

**Dataset**

The dataset is split into training and testing set. The training set contains twitter messages and their targets, indicating a disaster tweet or not, while the testing set contains only messages. There are 7613 samples for training and 3263 for testing.

**Preprocessing**

Different from written article or spoken language, the messages may have twitter handle, http URL, hashtags, and some special characters, which need to be removed. Based on their pattern in the messages (like hashtags always start with #), we can simply remove them by looking for the pattern.

Words like ‘the’, ‘is’, and ‘me’ (or stop wards) appear commonly but carry little information, removing them could help our model focus on those informative words. To do that, we utilize the stop word list form NLTK and remove the words in the list.

In English, a word may have different spelling to under different circumstances. For example, ‘speaking’ and ‘speak’, ‘am’ and ‘are’, ‘car’ and ‘cars’. Despite the difference, they express the same semantic. Converting the words back to their original form would have positive effect on the performance.

We transform the cleaned, lemmatized sentence into a vector by words counts, i.e., how many times a word shows up in a sentence. We then use tf-idf to find some potentially important words and emphasize them. The preprocessing is applied to both training data and testing data.