**NLP Disaster Prediction**

**Dataset Description**

What files do I need?

You'll need **train.csv**, **test.csv** and **sample\_submission.csv**.

What should I expect the data format to be?

Each sample in the train and test set has the following information:

* The text of a tweet
* A keyword from that tweet (although this may be blank!)
* The location the tweet was sent from (may also be blank)

What am I predicting?

You are predicting whether a given tweet is about a real disaster or not. If so, predict a 1. If not, predict a 0.

Files

* **train.csv** - the training set
* **test.csv** - the test set
* **sample\_submission.csv** - a sample submission file in the correct format

Columns

* id - a unique identifier for each tweet
* text - the text of the tweet
* location - the location the tweet was sent from (may be blank)
* keyword - a particular keyword from the tweet (may be blank)
* target - in **train.csv** only, this denotes whether a tweet is about a real disaster (1) or not (0)

**The final submission will consist of the following:**1. Jupyter Notebook file  
2. Test.csv file with survival predictions

3. Evaluation Metrics  
4. Clean code with comments and headings explaining all processes

**NOTE: Ensure all code is custom build no plagiarized code should be there.**

The Submissions file is evaluated using [F1](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html) between the predicted and expected answers.

F1 is calculated as follows:

*F*1=2∗*precision*∗*recall/precision*+*recall*

where:

*precision*=*TP/TP*+*FP*

*recall*=*TP/TP*+*FN*

and:

True Positive [TP] = your prediction is 1, and the ground truth is also 1 - you predicted a positive and that's true!  
False Positive [FP] = your prediction is 1, and the ground truth is 0 - you predicted a positive, and that's false.  
False Negative [FN] = your prediction is 0, and the ground truth is 1 - you predicted a negative, and that's false.

## Submission File

For each ID in the test set, you must predict 1 if the tweet is describing a real disaster, and 0 otherwise. The file should contain a header and have the following format:

id,target  
0,0  
2,0  
3,1  
9,0  
11,0