**Natural Language Processing with Disaster tweets**

*(Vigneshwar K – 270403)*

**AIM:**

To build a model which will predict whether a tweet is a disaster tweet or not.

**THEORY:**

We are given with two datasets – train.csv and test.csv containing the following information:

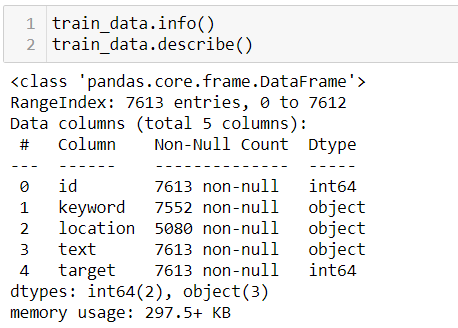
* **Id**: which is a unique identifier for each tweet
* **text**: represent the text of the tweet
* **location**: represent the location the tweet was sent from (may be blank)
* **keyword**: represent a particular keyword from the tweet (may be blank)
* **target**: represent in training set only this shows whether a tweet is about a real disaster (1) or not (0)

**IMPLEMENTATION:**

Process Followed:

* Import required libraries and read the csv files.
* Analyse the data in the datasets.
* Pre-process the data in the datasets.
  + Cleaning the data
  + Tokenization
  + Stemming
  + Vectorization
* Fit the data into suitable ML model.
* Predict the results of test dataset.

**Data Analysis:**



From this we can see that there are null values in keyword and location columns. Location columns have significant number of null values. But we cannot fill these null values intuitively and also the location information is not significant for predicting the nature of the tweets, so we choose not to utilize the location values.



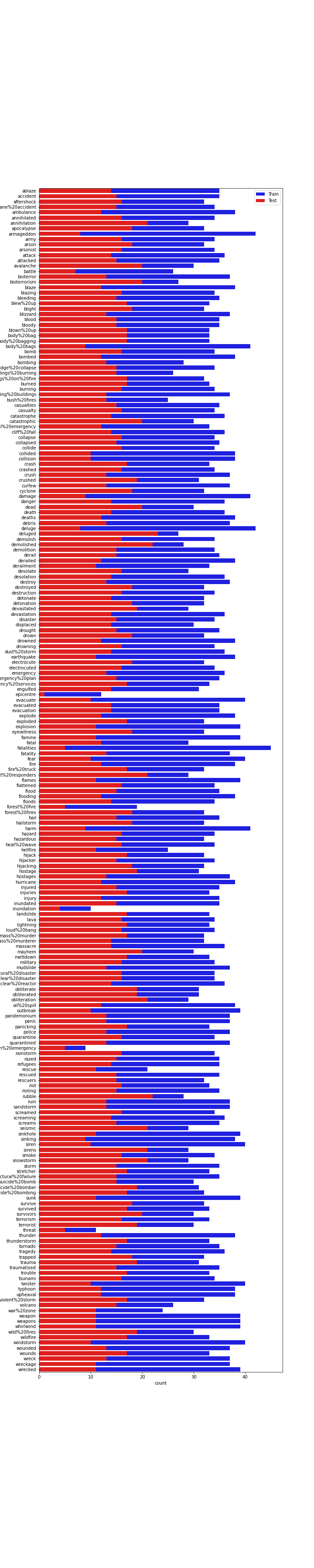
The above graph shows the count of disaster and normal tweets, with 1 being disaster tweets and 0 being normal tweets.

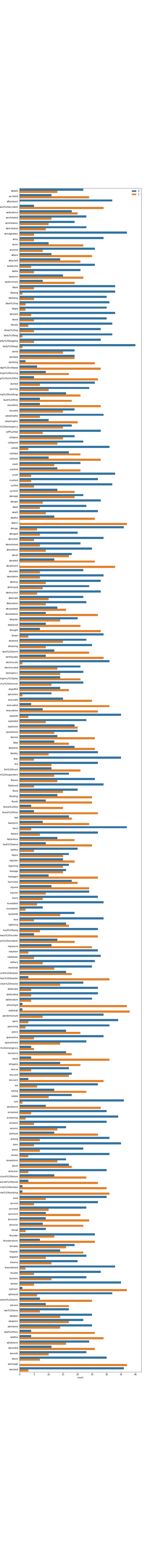
**Checking the Number of unique keywords:**

It is found that there 222 unique keywords present in the dataset, each of which corresponds to either of disaster or normal tweets. Some of the keywords are shown in the figure below.

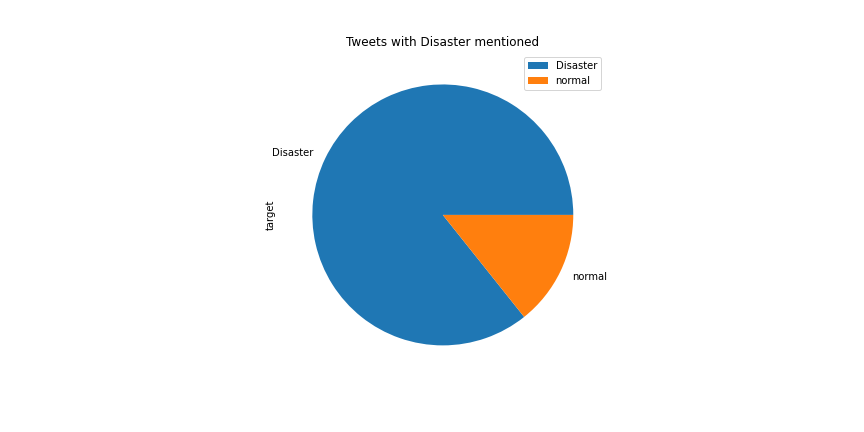
From the below figures we can infer that some of keyword indicate a greater number of disaster tweets whereas other do not. For example, keywords like “aftershock” and “blazing” have little to no disaster tweets. So, it is a good indicator for classifying tweets into disaster tweet or not.

It is also noted that a major portion of tweets with “Disaster” keyword was found out to be actually disaster tweets. A pie chart about it shown below.





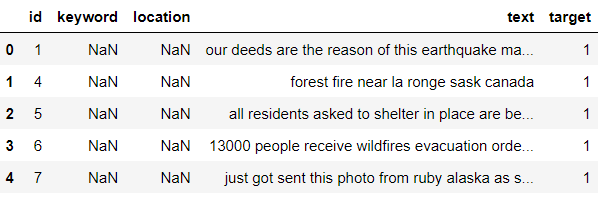
The distribution of tweets with Disaster keyword mentioned in it.



**DATA PRE-PROCESSING:**

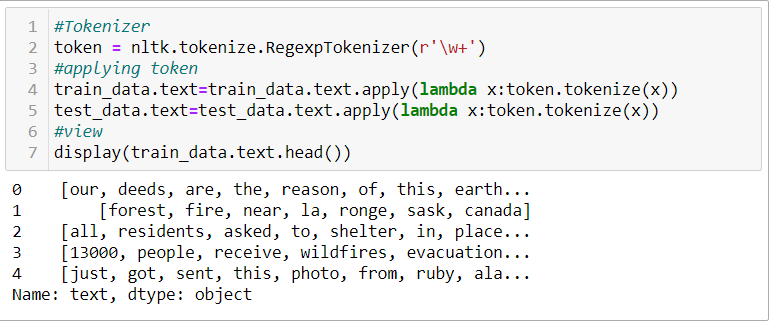
**Cleaning the data:**

Cleaning of data includes converting all text to lower case, removing hyperlinks, removing brackets and removing punctuations.



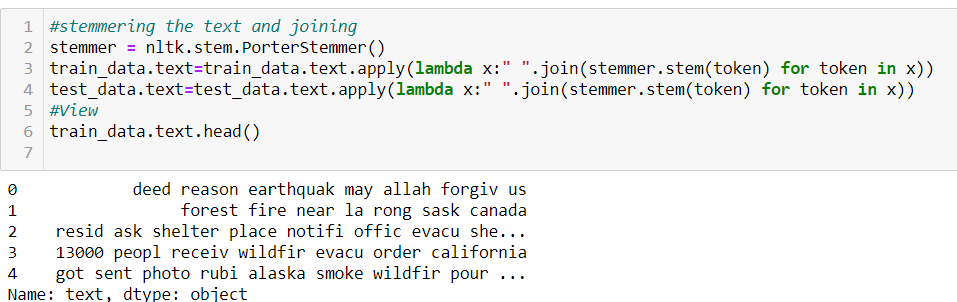
**Tokenization:**

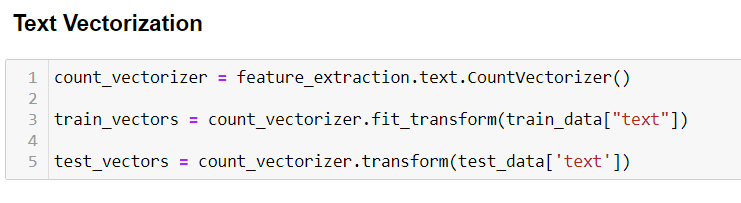
Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP.



**Stemming:**

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. For example, Run, running, runs, ran all result in the word run (as it is the stem or root word).



**Vectorization:** 

**Predicting Outcomes:**

Once the data is processed, it is time to fit the data into our model. As it is classification problem, we need to use classification algorithms.

Ridge Classifier – 78% accuracy

Logistic Regression – 79% accuracy