**Pre Processing**

1. In the existence column there are Yes,No,Y,N and NaN values. Replace Yes with Y and No with N. So now there are only Y,N and NaN values.
2. Drop all rows with NaN values in existence.
3. Split the dataset 80% to train and 20% to test.
4. Use of preprocessor library <https://pypi.org/project/tweet-preprocessor/> and the function clean\_tweet to clean tweet column. Removes punctuations, converts all letters to lower case. Tweets that end up as empty after this cleaning, those rows are dropped.
5. The train and test now has cleaned data. These two are again split into x\_train, x\_test, y\_train, y\_test ; with 80% to train and 20% to test. X variables is from the tweet column and Y variables is from the existence column. Existence confidence is not used.
6. The variables x\_train, x\_test with the tweets are then tokenized using CountVectorizer of sklearn, with stop words of English not taken into account.

**Model**

1. Support Vector Classifier is used to train a model and kernel ‘sigmoid ’ showed the best accuracy score of approximately 74%.
2. KNeighbours Classifier is used to train a model and using 1 to 60 as n\_neighbours, 21 showed to be the least number with best accuracy of approximately 73%.
3. KMeans was used to make a cluster of the tweet column and elbow method showed to have a optimum of 4 clusters.