* A text may contain non-alphanumeric characters, URLs, stop words, words with mixed cases, etc.
* They sometimes make it difficult for the algorithm to analyse text.
* To make it easy to analyze different texts and sentences, for Natural Language Processing, the text is always preprocessed.

Steps Involved in Text Pre-processing

* Removal of punctuations like . , ! ? $

Eg: (I love having pizzas in evenings!) becomes (I love having pizzas in evenings)

* Removing Stop words: Removal of the words which do not convey any special

meaning and the sentence conveys its message without them.

Eg: I love having pizzas in evenings: (love pizzas evenings) itself conveys the

message

* Converting text to lower case
* Lemmatization: to extract the root word without affecting the meaning of the word

Eg: 'enjoyed' becomes 'enjoy’

Many additional steps maybe carried out based on the type data that is preprocessed

First, we import he dataset. I’ve downloaded the dataset from Kaggle and the link to it is given here.

Pandas library of python is mainly used for data analysis. So, we’ll be using it to define the dataset and make various modifications to the data.

The dataset.columns function displays the columns in the dataset and head() function displays the first 5 rows in the dataset

So now, we start with the preprocessing.

1. The URLs are removed using the regex \S+. \S+ means “a string of non-whitespace characters” and \s+ means “a string of whitespace characters” so we replace string of non-whitespace characters starting with http with a “”.
2. All non-alphanumeric characters are replaced with spaces.
3. Convert the text to lower case (using lower() function) to maintain a uniformity in the data and to avoid the algorithm considering the same upper and lower case words different.
4. Now it’s Lemmatization to obtain the root word. NLTK is a python library which contains Natural Language prcessing tools for English. The stem in nltk package contains WordNetLemmatizer which is used for Lemmatization. averaged\_perceptron\_tagger, wordnet, punkt are downloaded from the package.
5. Stopwords are removed again using the nltk package

And have a look at the dataset now, this dataset doesn’t have any tags, URLs, any weird symbols or the stopwords but just text which the NLP algorithm finds it easy to read and analyse!