**Big Data Analytics and Applications**

**In Class Programming- 3**

1. What you learned in the ICP.

In this ICP I have learned Pre-Processing of Twitter Data, Visualizations of Text Data, and Sentimental Analysis of the text data by using different Scikit Learn Classifiers.

1. ICP description what was the task you were performing.

In this ICP we are asked to do Data Pre-Processing and Visualizations of the Twitter Data and Perform sentiment analysis on the given twitter dataset using one of the Scikit Learn Classifiers.

c. Screen shots that shows the successful execution of each required step of your code.

First, we have imported some of the required libraries.

A screenshot of a cell phone

Description automatically generated

Then, I have imported the dataset which contains labels and tweets. I have dropped the unnecessary column id by using a drop function and created two sets for features and labels.

A screenshot of a social media post

Description automatically generated

The following screenshot is the shape of the data frame and name of the columns in the data frame. I have used shape and columns functions to display the shape and name of the columns.

A close up of a logo

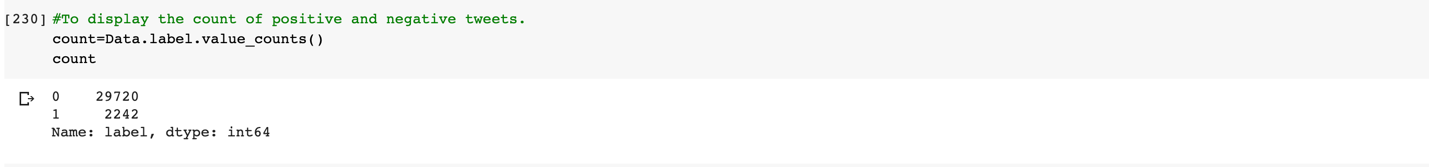
Description automatically generated

The following screenshot is the information about the data frame. For this I have used info function.

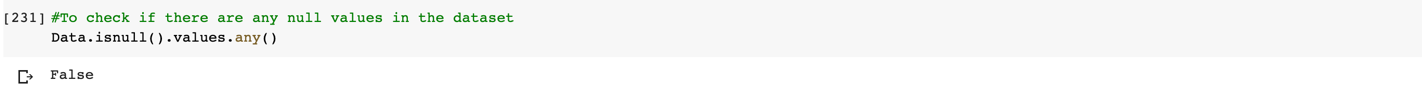
A picture containing table

Description automatically generated

I have used value\_counts function to display the count of positive and negative tweets.



In order to check if there are any null values in the data frame I have used isnull function. If the output returns false it means, there is no null value in the data frame.



Now, I have used a bar graph to visualize the counts of positive and negative tweets.

A screenshot of a social media post

Description automatically generated

I have also used a pie chart to visualize the count of positive and negative tweets.

A screenshot of a social media post

Description automatically generated

From the tweets I have removed twitter handles, operators, expressions, spaces, special characters, single quotes, single characters and multiple spaces as a part of data pre-processing.

A screenshot of a cell phone

Description automatically generated

This is the screenshot of the data after it is Pre- processed.

A screenshot of a social media post

Description automatically generated

Now in order to remove stop words from the data frame I have written a for loop so that it checks each and every word in the dataset if that is present in the English stop words set of nltk and appends it to the filtered\_sentence list if that word is not present in stop words set.

A screenshot of a social media post

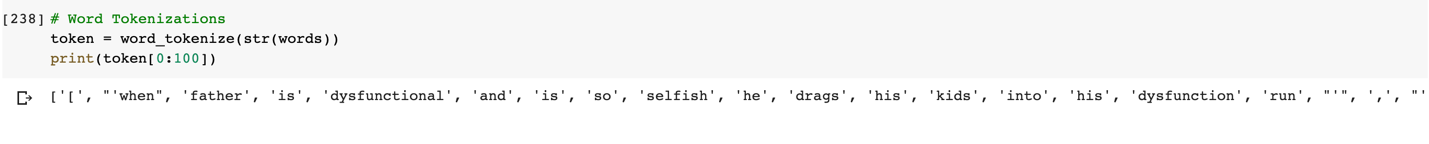
Description automatically generated

This is a screenshot of sentence tokenization.

A screenshot of a social media post

Description automatically generated

This is a screenshot of word tokenization.



This is a screenshot of word stemming.

A screenshot of a cell phone

Description automatically generated

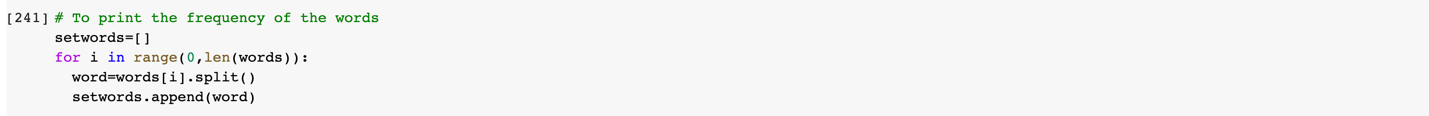
The following is the screenshot of the code to visualize the words from the tweets as a word cloud.

A screenshot of a cell phone

Description automatically generatedA close up of a logo

Description automatically generated

The following are the screenshots of the code to print the high frequency words in the tweets.



A screenshot of a social media post

Description automatically generated

The following is the screenshot where I have visualized the high frequency words using a bar graph.

A screenshot of a cell phone

Description automatically generated

Now I have applied Tfidfvectorizer on the preprocessed data and then I have splited the data into testing and training dataset.

A screenshot of a cell phone

Description automatically generated

Now I have built a Random Forest Classifier model and I have achieved 96% of accuracy.

A screenshot of a cell phone

Description automatically generatedA screenshot of a social media post

Description automatically generated

I have also built a Support Vector Machine classification model and I have achieved 95.8% of accuracy.

A screenshot of a social media post

Description automatically generated

Youtube Link: <https://youtu.be/8ODe92G1wyI>