**ICP2**

**BIG DATA ANALYTICS AND APPLICATIONS(CS5542)**

*Sravani Niharika Garapati(16328843)*

*Nikhila Chirumamilla(16323180)*

**QUESTION:**

Use the Twitter Data and perform sentiment Analysis on that data using one of the scikit learn classifier for text.

**What I have learned from this ICP:**

In this ICP, professor asked us to perform sentimental analysis on the twitter data by cleaning, preprocessing and building model and executing the model and making predictions. In the process of cleaning and preprocessing, I have learned any functions and procedures. Some of them are listed below.

**Nltk:** It is a library that has libraries and features which support statistical language processing.

**StopWords**: The words which are very common in the text in any language and do not add any meaning to the text.

**Stemming:** Reducing the word to their base word forms

**Lemmatization:** Same as stemming but is more powerful and gives meaningful base words by following some rules.

**Tfidf Vectorizer:** It learns the vocabulary and tokenize the words and highlights words that are more interesting i.e., words with high frequency.

**Pos\_tag ():** It is a function for marking up the words in a text to a particular part of speech (like noun, verb, adjective, etc.) based on its definition and context.

**FreqDist ():** This function is used to find the occurrences of a word within a text.

**WordCloud:** It is a library in python which gives visual representation of text data. We can adjust height, width and colors based on our interest.

**BarPlot:** A bar plot shows categorical data as rectangle bars with the height of bars proportional to values they represent.

**Sklearn:** It is most useful library in machine learning in python providing efficient tools for machine learning and statistical modelling containing regression, classification and clustering by providing interface with the python.

**Mean\_Squared\_Error ():** It shows the deviation between actual value and predicted value.

**ICP Description about the Task:**

In this task, we need to perform sentiment analysis on the given twitter data using one of the Scikit learn classifier. This is done in order of below mentioned steps:

* Imported Pandas and read the data given using pandas.
* Performed many cleaning processes such as removing twitter handles(@user), hashes (#) and numericals from the data and also removed stop words from the data
* Performed Stemming and Lemmatization so as to bring all the words to their base format.
* Then, we transformed all the data into vectors using Tfidf vectorizer so as to know how much data is there and can be used for analysis in the coming steps
* Applied Pos\_tagging on all the words to know the parts of speech (noun, verb, adjective, etc)
* Used FreqDist () to know how frequently a particular word is occurring and represents it diagrammatically using plot () function
* Then, pictorically represented all the words, positive words and negative words using WordCloud and other useful functions.
* Split the filtered data into training and test sets to fit the model.
* Trained the model using LinearSVC and calculated the predicted value
* Calculated Mean\_Squared\_error to know the differences between actual value and predicted value and also used classification\_report to know the accuracy and efficiency of the model.
* And at last checked the model by giving sample inputs.

**Challenges faced in this ICP:**

The challenges which we faced while doing this ICP are listed below:

* It became very tough for us to clean the data by removing all the punctuations, numericals, twitter handles and all the other symbols that are present in the data
* Faced trouble in writing conditions for representing positive and negative words in the diagram using WordCloud.
* It took lot of time for selecting machine learning model to do analysis and later selected LinearSVC.

**Screenshots of Execution of Code:**

Graphical user interface, text, application

Description automatically generated

*Displaying the data that is read*

*Table

Description automatically generated*

*Preprocessing the Data*

*A picture containing table

Description automatically generated*

A picture containing text

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

A picture containing text

Description automatically generated

*Removing Stop Words*

*Text

Description automatically generated*

*Stemming and Lemmatization*

*Graphical user interface, text, application

Description automatically generated with medium confidence*

*Function to remove stop words and perform stemming*

*Graphical user interface, text, application

Description automatically generated*

*Applying the function and displaying the output*

*Text

Description automatically generated with low confidence*

*Applying Tfidf vectorizer*

*Graphical user interface, text, application, email

Description automatically generated*

*POS Tagging*

*Graphical user interface, text, application, email

Description automatically generated*

*Computing Frequency of each word*

*Graphical user interface, text, application

Description automatically generated*

*Graphical Representation of FreqDist*

*Chart, line chart

Description automatically generated*

*Text, chat or text message

Description automatically generated*

*Creating WordCloud for positive words*

*A picture containing text

Description automatically generated*

*Text

Description automatically generated*

*Creating WordCloud for negative words*

*A picture containing text

Description automatically generated*

*Text, whiteboard

Description automatically generated*

*Chart, histogram

Description automatically generated*

*Splitting the Data and training the model using LinearSVC*

*Graphical user interface, text, application, email

Description automatically generated*

*Calculating predicted value and computing MSE*

*Graphical user interface, text, application, email

Description automatically generated*

*Calculating the efficiency of the model*

*Table

Description automatically generated*

*Testing the Model using Sample Test Cases*

*Graphical user interface, text, application

Description automatically generated*

**Video Link:**

*https://youtu.be/ux-3uTqRMSo*