Assignment 2
Sentiment Analysis using Natural Language Processing (NLP) and Decision Tree Algorithm

The aim of this assignment is to perform a basic level of natural language processing and machine learning algorithms implementation for sentiment analysis task on the comments which are given below. The overall aim / objective of this assignment is that you should become familiar with the basic techniques/tools required to understand and pre-process data for building machine learning models before applying further methods/techniques of data science.

Sentiment analysis: Social media has opened a whole new world for people around the globe. People are just a click away from getting huge chunk of information. With information comes people's opinion and with this comes the positive and negative outlook of people regarding a topic. Sometimes this also results into bullying and passing on hate comments about someone or something.

Resources:

The following tutorials will help you to fully understand the pipeline to run the code from A to Z.
https://towardsdatascience.com/social-media-sentiment-analysis-49b395771197
https://towardsdatascience.com/social-media-sentiment-analysis-part-ii-bcacca5aaa39

Whereas the GitHub code of these links are given at following URL:

https://github.com/dD2405/Twitter_Sentiment_Analysis

Python tutorial for Natural Language Processing (NLP):
https://pub.towardsai.net/natural-language-processing-nlp-with-python-tutorial-for-beginners-1f54e610a1a0

Scikit-learn: is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, k-means etc. and designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

An introduction to machine learning with scikit-learn: https://scikit-learn.org/stable/tutorial/basic/tutorial.html

Scikit-learn documentation: https://scikit-learn.org/stable/
NLTK documentation: https://www.nltk.org/
Random Forest : https://scikit-learn.org/stable/modules/ensemble.html#forest

The assignment is to understand, and pre-process data using Python and Scikit-learn. Therefore, your task to carry out the following steps:

a. Training/testing phase:

1- Import libraries
2- Read train/test data
3- Understand data train/test data
4- Preprocess data train/test data
5- Label Encoding for train/test data
6- Feature extraction (Changing representation of data—from string to feature-vector). Apply following baseline features: bag-of-words, TF-IDF, and word count.
7- Splitting input vectors and labels
8- Train decision tree machine learning algorithms using training data
9- Evaluate decision tree machine learning algorithm using test data

b. Application phase:

1- Combine dataset (train + test)
2- Train decision tree model
3- Save the trained model in pickle file
4- Load the trained model (see last step)
5- Take input from user
6- Convert user input into feature vector (same as feature vector of trained model)
7- Apply trained model on feature vector of unseen dataset and output prediction (positive or negative) to user

Train dataset:

Index	Comment	Polarity
0	time to eat with my best buddy! #lunch	Happy
1	@user @user if they want reflection money. #ksleg	Happy
2	---Good job but I’ will expect a lot more in future. ####	Happy
3	totally dissatisfied with the service###%%@@ never used this service again	Sad
4	loved my work!!!!!	Happy
5	Worst customer care service……@@$$$angry	Sad
6	Brilliant effort guys!!!	Happy
7	@user @user you point one finger @user millions are pointed right back at you, #jewishsupremacist	Sad
8	words r free, it's how u use that can cost you! #verbal #abuse #hu #love #adult #teen @user	Happy
9	you might be a libtard if... #libtard #sjw #liberal #politics	Sad

Test dataset:

Index	Comment	Polarity
0	@user the pic says otherwise for young girls confined in that kitchen. you are void of meaning, beyond cheap publicity #topoli	Sad
1	#good night! ?? #faith ever #vaitacacommafiasdv	Happy

2	`@user when you're blocked by a troll because you promise to #blacklivesmatter &amp; let his nonsensical rants`	Sad
3	`dinner with sister!!`	Happy
4	`who else is planning on watching @user tomorrow?`	happy

Hand-in and Assessment

Upload your code on Google classroom (output in PDF) along with code of Jupyter Notebook (any other) file and data which you have used.

This assignment is worth 30% of the overall Assignments marks for CSC461. Assessment will be based on the accuracy of your answers to the questions, given the approach you have taken and on the overall quality of your code, code must be well documented. The Assignment 01 is due by **Sunday May 9th, 2021 at 23:59**. Note that if you submit the assignment after the deadline it will be considered unsubmitted and you will be **given 0 marks.** Departmental **rules concerning plagiarism and collusion will be strictly observed** – please refer to the Student Handbook for details of these.

Instructions:

Note: This assignment can be done in a group of two.

Viva date: Will be communicated latter. However, if any one fail to answer two question will get **ZERO**.