Lab - 5

PRML
AY 2020-21, Trimester - III

April 4, 2021
Deadline: April 11, 2021, 11:59 PM

Q1:***Read the dataset train.csv and follow the instructions in the colab notebook: [40 marks]***

a. Data Preparation: [15 marks]
   i. Load the data
   ii. Plot the count for each target
   iii. Print the unique keywords
   iv. Plot the count of each keyword
   v. Visualize the correlation of the length of a tweet with its target
   vi. Print the null values in a column
   vii. Removing null values
   viii. Removing Double Spaces, Hyphens and arrows, Emojis, URL, another Non-English or special symbol
   ix. Replace wrong spellings with correct ones
   x. Plot a word cloud of the real and fake target
   xi. Remove all columns except text and target
   xii. Split data into train and validation
b. Compute the Term Document matrix for the whole train dataset as well as for the two classes. [5 marks]
c. Find the frequency of words in class 0 and 1. [3 marks]
d. Does the sum of the unique words in target 0 and 1 sum to the total number of unique words in the whole document? Why or why not? Explain in the report. [2 marks]
e. Calculate the probability for each word in a given class. [5 marks]
f. We have calculated the probability of occurrence of the word in a class, we can now substitute the values in the Bayes equation. If a word from the new sentence does not occur in the class within the training set, the equation becomes zero. This problem can be solved using smoothing like Laplace smoothing. Use Bayes with Laplace smoothing to predict the probability for sentences in the validation set. [6 marks]
g. Print the confusion matrix with precision, recall and f1 score. [4 marks]

The notebook is attached for your reference.
Please submit the necessary codes (Notebook) containing your output, and a PDF explaining and analyzing the steps for all the questions along with necessary plots/figures.
Notebook used in demo: Link , dataset
Note: No submission will be accepted after the final deadline.