UNIVERSITY OF SOUTHERN MAINE

Natural Language Processing, Spring 2023, Assignment 2

Instructor: Behrooz Mansouri (behrooz.mansouri@maine.edu)

Due: February 24, 2023

Notes for submission:

1. Submit your file(s) with the correct naming as: NLP_Assignment2_StudentName

2. Two files should be uploaded, one .zip file having all the codes (directory is zipped, and it is named codes) and one .pdf file. There would be a penalty for uploading wrong formatted files. Any other formatting will be ignored

3. Codes should be well-structured with comments to run

4. Codes should be available on your GitHub Repo. Failure to have codes publicly available, results in a 20% reduction in your grade. Make sure to include the GitHub link in your PDF file.

5. Any assumptions made by students should be explicitly mentioned in the submitted document

6. Answers to the questions should be easy to detect. For your codes, name the .ipynb files according to the question numbers (e.g., Question1.ipynb). In your document, use the question number and just write your answer. You should not have the question itself in your document. (e.g., Question 1. Each question…)

Consider the Posts_law.xml file from the Law stack exchange (as in lab 3).

Question 1 (80%): **Naïve Bayes classification**

Each question in Law Stack Exchange can have multiple tags. Let's consider the first tag as the question's tag (class). We will split the data into training and test sets. Any question posted after 2021 is considered as a test sample, and the rest are used for training our classifier. We will consider the top-20 frequent tags for the classification task. (For this, there is a code provided on the course website)

Using the training samples, train a Naïve Bayes classifier. Then use your classifier to classify the test samples. Note that preprocessing steps should be applied (tokenization, removing HTML tags, and punctuations).

In your solution, provide micro and macro F1-scores. Discuss the class for which your model has the best and worst effectiveness. To do this, provide two examples for each, and explain why your model did or did not work properly. For the failed case, you should look at the predicted class and explain why your model had that prediction.

Note that each question has a title and body. You will repeat the experiment three times with these settings:

1. Considering only question titles
2. Considering only question bodies
3. Considering question title + body (concatenated with whitespace)

In your discussion, compare these three approaches. Does including both title and body of questions improve the effectiveness of the classification?

Grading breakdown:

Codes: 30%   Results: 20%   Results Discussion: 30%   Comparing classifiers: 20%

Question 2 (20%)**:** In future assignments, we are aiming to compare answers given to questions by users compared to machine-generated answers. Provide a solution so that one can issue a question to ChatGPT and get the answer with Python. Pick one question from Law Stack Exchange with an accepted answer and issue it to ChatGPT. Then, compare the accepted answer and the one provided by ChatGPT. What are their similarities and differences?