



UNIVERSITY OF SOUTHERN MAINE

Natural Language Processing, Spring 2023, Assignment 4

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

Due: April 17, 2023

Notes for submission:

1. Submit your file(s) with the correct naming as: NLP_Assignment4_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
-

In this assignment, we will repeat question one (similar to question two) of Assignment 3, by replacing the network with a Sentence-BERT model.

Step 1: First, without any further fine-tuning, use the Sentence-BERT model fine-tuned on Quora Duplication Question. You can find details [here](#). For all the questions, find their top 100 similar questions. Then calculate two measures: the average P@1 and the average Mean Reciprocal Rank (MRR measure explained in IR session).

Step 2: Then, using a similar split as the previous assignment (10% for testing and the remaining 90% for fine-tuning), fine-tune a Sentence-BERT model with training data on Law Stack Exchange. You may use the same positive and negative pairs as in the previous assignment. Then, on the remaining 10% report the average P@1 and MRR measures. Note that the design is your choice, you can use Bi-encoder or Cross-encoder, any pre-trained model, and any loss function.

Step 3: Apply the first step only on the 10% training set from Step 2 and calculate P@1 and MRR per each question in the test. Using the significance testing of your choice, decide whether the difference between MRR measures is statistically significant.

Extra point (25%): You will use models not trained specifically for the legal domain. However, LEGAL-BERT, is a BERT model trained specifically for the legal domain. A sample code to use this model is provided for you. Repeat the experiment above, using the LEGAL-BERT model, and report your results.