AI4ICPS IIT Kharagpur	**IIT KHARAGPUR AI4ICPS I HUB FOUNDATION** **H**ands-on **A**pproach to **AI, Cohort-2, July – October 2024** **Additional Programming Assignment 2: NLP**

Due date: Saturday 26th October 2024, EOD – IST

Important Instructions about Programming Assignments

1. Programming assignments will be evaluated automatically. **Do not** change the skeleton code provided to you.
2. Write your code **only in the designated places** in the skeleton code and process the input data provided to you in the designated variables. **Do not alter** the input-output structure in the skeleton code.
3. **Do not import** any additional libraries. **Do not use any additional files** for the processing (other than those mentioned in the skeleton code).
4. Failure to comply with these instructions may lead to you getting **zero marks** for the assignment, even if the solution is largely correct.

Question:

Objective: The objective of this assignment is to investigate the performance of two different types of sentence-level representations as well as understand the type of information encoded in these representations as produced by the bert-base-uncased model.

There are usually two ways of aggregating sentence-level representation from the bert model.

I. Using the CLS token
II. Applying some transformation on the hidden state output of the model

One such transformation is the application of max pooling.

In this assignment, you will be given a set of two sentences that will be taken as input from the system arguments with the sentences being separated " , ". You will have to use the pre-trained bert-base-uncased model to generate two different representations (as described above) for each sentence. You will then apply cosine similarity between the respective representations of each sentence and report the two values rounded off to the 2nd decimal.

The output of the assignment can be represented as,

$$cos\left(\gamma_1(f(S_1)), \gamma_1(f(S_2))\right) \qquad cos\left(\gamma_2(f(S_1)), \gamma_2(f(S_2))\right)$$

Where $f(\cdot)$ is the bert-base-uncased model, S_1 and S_2 are the two sentences, γ_1 and γ_2 are the CLS and max-pooling representations of the output of the bert-base-uncased model and lastly, $cos\ (\cdot,\cdot)$ is the cosine similarity score.

Bonus: The input cases have been provided such that the sentence similarities range from paraphrase to contradiction. Looking at the scores, try to understand which type of vector representation works best. Also, based on the last test case, do you think that the bert-base-uncase model is able to understand semantic information?

Sample Test Cases:

"input": "The brown fox jumped over the well , The brown fox did not jump over the well",

"output": "0.89 0.98"

"input": "The city of joy , better known as Kolkata",

"output": "0.51 0.76"