CSE556 NLP
Assignment 3a

Date: 11 Nov 2022
Deadline: 11:59pm 28 Nov, 2022 **Max Marks: 50**

General Instructions:

1. Allowed programming language: Python.
2. Use classroom discussion for any doubt. No query will be entertained through personal emails.
3. Each group member must do at least one of the following sections. But both should know the working of all the tasks. (Recommended: Divide the sections among yourselves.)
4. The assignment can be submitted in a group of a maximum of two members.
5. For plagiarism, institute policies will be followed.
6. You need to submit the updated A3.ipynb on google classroom with the following name: **A3a_Name1_Name2.zip. No report pdfs are needed. All coding documents should be clearly stated in the A3.ipyn**
7. Mention methodology, helper functions, preprocessing steps, any assumptions you may have, and the contribution of each member in the report.

Dataset: Semantic Text Similarity Dataset (STS). Link is provided in the sample A3.ipynb. Strictly use the dataset version specific in the A3.ipynb

Task: Given a sentence pair, predict from a range of 0-5 what is their similarity level, with 5 being the highest.

You are free to use spacy, genism, hugging face and DL frameworks like Tensorflow and Pytroch as deemed useful under specific modeling.

I. EXPERIMENTING WITH LANGUAGE MODELS [45 Marks]

Boilerplate function definitions and methods are provided in the A3.ipynb along with marks distribution for each step. You are required to complete the functions provided in the notebook, and append additional code to make the notebook functional.

There are 3 set of methods to try, each are specified in their model section in A3.py. Commonly used libraries are already enlisted in A3.ipynb

1. For Configuration 1 and Configuration 2, the training need to be a strictly sklearn

ML based regression method. Anything more advance than this will be penalised.
2. For Configuration 3 you need to fine-tune atleast 1 layer of the contextual embeddings. Not having any fine-tuning will be penalised.
3. Use the stats function of scipy to report the spearman correlation score. Reporting of any other metric to boost performance will not be marked. Documentation: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

II. IMPROVING UPON BASELINE SCORES [5 Marks]

Enlisted below are the 3 configuration specific baseline scores on the test set rounded off to 3rd decimal place. You need to get better results over the random baselines (as mentioned below).

The following improvement marks will be awarded for the respective model given that:

1. You reported improved scores should also be rounded off to 3 decimal places.
2. Model specific constraints specified in Part I and A3.ipynb should be strictly followed.
3. For configuration 3, MAX_NUM_EPOCHS = 3
4. No tampering of the test set.

Configuration	Baseline Spearman Score	Marks for Improvement
Non-contextual embeddings + ML Regression	0.216	+1.25
Contextual embeddings + ML Regression	0.208	+1.25
Fine-tuned Contextual embeddings	0.792	+2.5