**Homework # 5**

1. We will use the BERT model to perform sentiment analysis. The sentiment is performed via a linear model applied to the output vector above the [CLS] input in BERT (that output vector is referred to as C).

Use a pretrained BERT model from:

<https://github.com/google-research/bert>

And use sentiment data from:

<https://huggingface.co/datasets/yelp_polarity>

Build a sentiment-analysis model based on BERT, using the above data. Do a detailed analysis of performance, and compare the accuracy of this model to results you achieved with the simpler baseline model from the prior homework.

Implement the model two ways:  
  
(a) Leave all BERT parameters unchanged, and just learn the linear model at the output.

(b) Fine-tune all BERT parameters, while also learning the linear layer at the output.

In your solution, provide all code and also a detailed summary of the analysis of the results.

2. This task investigates database retrieval via BERT.

Consider a database *of your choosing*, consisting of a large database of documents. Examples are (but use what you want):  
  
<https://www.kaggle.com/datasets/rowhitswami/nips-papers-1987-2019-updated>

<https://commoncrawl.org/>

<https://www.kaggle.com/datasets/crawford/20-newsgroups>

Using BERT, encode each of the documents in your corpus to a vector. Do this with “out of the box” BERT, with no changes to the base model (taken from the above GitHub)

Build code that, given a question or prompt, will pull appropriate documents from your database. This also should be done with BERT applied to the prompt, and using that write code that pulls from the database the top-10 best matches, ranked by probability of match.

In your solution, provide all code and also a detailed summary of the analysis of the results.

3. Use the GPT-2 code from:

<https://huggingface.co/openai-community/gpt2>

or

<https://github.com/openai/gpt-2>

Using results from Problem 2, put as input to GPT2 the prompt alone, and evaluate the quality of the answer. Separately, input the prompt plus the document pulled for the database. Examine and evaluate the (somewhat subjective) quality of the generated text, with and without the context provided by the pulled document.

Since the context length of GPT2 is limited, you may have to consider smaller-size documents in your context database. This can be done by considering smaller documents (e.g., abstracts from the NeurIPS) database, or breaking the larger documents into smaller pieces (and encoding each to a vector via BERT).

In your solution, provide all code and also a detailed summary of the analysis of the results. This part will require a descent level of software/coding expertise. Teaming is fine. For some this part of the assignment may be a “reach;” do your best, and focus on learning from this.