Project
NLP Using Transformers

A. Nagar

Due Date Specified on eLearning

Instructions

- This assignment requires you to use the transformer model to do NLP tasks, like text summarization, question answering, and machine translation.
- You should store your dataset on a public location, such as Github, AWS S3, etc. You can also use the link directly from Project Gutenberg. Do not submit the dataset (which could be quite large) on eLearning.
- You are allowed to work in teams of maximum two students. Please write the names and NetIDs of each group member on the cover page. Only 1 final submission per team.
- **You have a total of 4 free late days for the entire semester. You can use at most 2 days for any one assignment. After four days have been used up, there will be a penalty of 10% for each late day. The submission for this assignment will be closed 2 days after the due date.**
- Please ask all questions on Piazza, not via email.

Dataset

For this project, you will select an English language book from the Project Gutenberg: `https://www.gutenberg.org`

You have to select only one such book and use its text format for use in this assignment.

NLP Tasks Using Transformers

You will use the transformer model discussed in class to perform one of the NLP tasks mentioned below. You are free to use a pre-built/pre-trained model from a source, such as HuggingFace or from any other source such as Keras, PyTorch.

You have to perform *one* of the following NLP tasks using the transformer model:

- Text Summarization
- Machine Translation to any language of your choice
- Developing a question-answering system (chatbot) on the text
- Text generation starting from a small seed text from the book

The following are important requirements for this part:

1. You have to do only one of the above mentioned NLP tasks on one of the books selected from the Gutenberg project.
2. You can use Google Colab or your local computer for coding.
3. Please do not use any local paths, use only global URLs while accessing your data.
4. You will need to evaluate your model using appropriate evaluation metrics. For example, if you are doing text summarization or machine translation, you can use the ROUGE score. For machine translation tasks, you can also consider the BLEU score. For text generation tasks, you may want to consider the perplexity metric.
5. If you are unsure, please ask the instructor through Piazza.

Submission

You need to output the at least the following along with your code:

1. A report explaining how transformers are used in the specific task that you chose, details of your transformer architecture, model hyper-parameters tuning, results obtained, and analysis of results. You should consider this as a professional data science project and develop a professional and polished report.
2. Link to your code on Google Colab or include the code as part of your submission.

If you have made any assumptions, please state them completely. Also include instructions on how to compile and run your code in a README file.