Chaklam Silpasuwanchai, Todsavad Tangtortan
AT82.05 Artificial Intelligence: Natural Language Understanding (NLU)

A2: Language Model

In this assignment, we will focus on building a language model using a text dataset of your choice. The objective is to train a model that can generate coherent and contextually relevant text based on a given input. Additionally, you will develop a simple web application to demonstrate the capabilities of your language model interactively.

Note: You are ENCOURAGED to work with your friends, but DISCOURAGED to blindly copy other's work. Both parties will be given 0.

Note: Comments should be provided sufficiently so we know you understand. Failure to do so can raise suspicion of possible copying/plagiarism.

Note: You will be graded upon (1) documentation, (2) experiment, (3) implementation.

Note: This is a one-weeks assignment, but start early.

Deliverables: The GitHub link containing the jupyter notebook, a README.md of the github, and the folder of your web application called 'app'.

Task 1. Dataset Acquisition - Your first task is to find a suitable text dataset. (1 points)

1) Choose your dataset and provide a brief description. Ensure to source this dataset from reputable public databases or repositories. It is imperative to give proper credit to the dataset source in your documentation.

Note: The dataset can be based on any theme such as Harry Potter, Star Wars, jokes, Isaac Asimov's works, Thai stories, etc. The key requirement is that the dataset should be text-rich and suitable for language modeling.

Task 2. Model Training - Incorporate the chosen dataset into our existing code framework. Train a language model that can understand the context and style of the text.

1) Detail the steps taken to preprocess the text data. (1 points)
2) Describe the model architecture and the training process. (1 points)

Task 3. Text Generation - Web Application Development - Develop a simple web application that demonstrates the capabilities of your language model. (2 points)

1) The application should include an input box where users can type in a text prompt.
2) Based on the input, the model should generate and display a continuation of the text. For example, if the input is "Harry Potter is", the model might generate "a wizard in the world of Hogwarts".
3) Provide documentation on how the web application interfaces with the language model.

Best of luck in developing your text-generation model!