Assignment 1

Problem 1 - ChatBot

This problem is to help you get familiar with Llama.
Please follow the guidance to get familiar with how to deploy Llama 3.2 1B or 3B to generate texts.
To create a ChatBot, please deploy a Llama chat model (e.g. `meta-llama/Llama-2-7b-chat-hf`), and follow the instructions.
Llama can be downloaded from official website. And if you are encountering problems with tokenizers, please access from Hugging Face (It takes time for HF to approve your access, so please start ASAP).
.

Problem 2 - Multilingual RAG Chatbot

This problem is to help you get familiar with larger models as well as RAG.
Please implement this problem with Google Colab.
Similar to Problem 1, the model and tokenizer can be accessed from Hugging Face. Create a vector database with documents and web pages from at least 2 different languages compatible with Llama 3.1 8B and build a chatbot using RAG.
For the vector database, embedding model and retriever, you can use existing libraries or inbuilt functions.
However, for the final QA chatbot you have to build it from scratch using only AutoModelforCausalLM from hugging face. This function should take in the retriever and question. It should then preprocess the text (this includes tokenization). Then, the prompt, question and context should be passed through the LLM. The output should then be converted back to text and shown to the user as the answer. This chat should be continued till the user types in exit.

Submission:

Please write a brief documentation to explain each step and include:

- For problem 1, please attach screenshots of a series of prompts and the answer that the ChatBot generates.
- For problem 2, please attach screenshots of one example for each language showing the prompt, context, question and answer from your chatbot.

Please submit the documentation in .pdf format and your code (.ipynb or .py) for both questions.

APPENDIX:-

1. You need approval from Hugging Face for using these models, so apply for access as soon as possible as it might take up to 48 hours.
   Once your access token is generated, you can use command
   huggingface-cli login
   To access the model.

2. For the vector database, you can use either chromaDB or FAISS.
3. For the RAG pipeline, ensure you are using a multilingual embedding model.
4. When loading the Llama model, try to use 4 bit quantisation from BitsAndBytes. This would allow you to run the models faster.