Assignment 2 : Dataset Preparation and Fine-Tuning LLMs

February 06th

Due Date: February 13th

Part A: Data Preparation for LLMs

Exercise 1:

Q1: Name three sources where you can find datasets for training or fine-tuning LLMs. Briefly describe the advantages and limitations of each.

Q2: Explain why data preprocessing is crucial when preparing a dataset for an LLM. Mention at least three preprocessing techniques and describe their impact on model performance.

Q3: What is tokenization in NLP? And Why is subword tokenization widely used in modern LLMs?

Q4: Explain what dataset bias is and how it can impact an LLM's predictions. What steps can be taken to mitigate bias when preparing training data?

Exercise 2:

Prepare a dataset for fine-tuning an LLM using Hugging Face's "**datasets**" library.

Task:

- Load a dataset or a subset of it from Hugging Face (e.g.: **IMDB Movie Reviews** or any other text dataset).
- Apply basic text preprocessing, including:
  - Lowercasing
  - Data Cleaning (Remove Punctuation, special characters, and tags if any to reduce noise).
  - Tokenization
    - Use AutoTokenizer from Hugging Face to tokenize the text.
    - Ensure padding and truncation for uniform sequence lengths.
- Print few processed examples to verify correctness

- Save the dataset for later use (If needed).

Part B : Fine-Tuning LLMs

Exercise 1:

Q1: How does fine-tuning differ from pre-training? Provide an example of a real-world use case where fine-tuning an LLM is preferable to using a pre-trained model as-is.

Q2: What are three common evaluation metrics for fine-tuned LLMs?

Q3: Why is parameter-efficient fine-tuning useful when working with large-scale models?

Exercise 2:

Use **Hugging Face's Trainer** to fine-tune a **distilbert-base-uncased** model.

Task:

- Use the dataset you prepared in Exercice1 or any other dataset.
- Tokenize the dataset of your choice.
- Load and Fine-tune distilbert-base-uncased on the dataset using **Hugging Face’s Trainer**.
- Set up Training Arguments (Use the notebooks we explained during the online session)
- Evaluate the fine-tuned model on the test dataset.

Hint:

```
# Load model directly from HuggingFace

from transformers import AutoTokenizer, AutoModelForSequenceClassification

tokenizer =
AutoTokenizer.from_pretrained("distilbert/distilbert-base-uncased")
```

```
model =
AutoModelForSequenceClassification.from_pretrained("distilbert-base-uncased
", num_labels=2)
```

Submission Guidelines

- Submit your answers to **Section 1 (Conceptual Questions) as a PDF or DOCX**.
- Submit your **code from Section 2** as a Python notebook (.ipynb) or a Python script (.py).