Project Title: Panini Parser – Interpreting Ancient Indian Texts

Course: IST-691 – Natural Language Processing

Link:
https://drive.google.com/drive/folders/1SUzzTi-DNd5bmNXi-YgWZP_Kj1YbWHna?usp=sharing

---

Files in this Folder:

- **`tokenizer.ipynb`**
  This notebook contains the code for preprocessing Sanskrit texts and building a custom Byte Pair Encoding (BPE) tokenizer tailored to ancient vocabulary.

- **`2.tokenfinetuning.ipynb`**
  Fine-tunes a GPT-2 model using the tokenized dataset. The notebook includes loss tracking, sample outputs, and performance evaluation.

- **`text.csv`**
  Contains verses and excerpts from texts like the Bhagavad Gita and Upanishads used for model training.

- **`fine_tuned_tokenizer.json`**
  Configuration files used by the tokenizer for encoding text sequences during model fine-tuning.

Software/Tools Required to View:

- **Google Colab** *(recommended)* or Jupyter Notebook
- Python 3.x environment
- Required libraries:

```
pip install pandas torch transformers matplotlib
```