Probattle: NLP Module

Problem Statement

February 16, 2025

Main Challenge

The primary challenge in this module is constructing a Retrieval-Augmented Generation (RAG) pipeline. The system must efficiently process and retrieve relevant information from a collection of provided links and PDF files, integrating a robust retrieval mechanism with a generative language model. The frontend will be developed using Streamlit to enable an interactive and user-friendly interface for querying and obtaining responses based on the retrieved information.

Objectives

- **Parse** the corpus of data provided.
- Generate **vector embeddings** for the parsed data.
- Store the embeddings in a **vector database** with optional metadata.
- Build a **RAG pipeline** that incorporates vector embeddings to retrieve the closest context.
- Develop a **Streamlit-based frontend** for user interaction.

Judging Criteria

- **Accuracy** of the model in generating responses.
- **Accurate retrieval of relevant context** from the vector database.
- Little to no **redundancy** in the LLM-generated responses.
- Recognition of **irrelevant context** or out-of-scope queries.
- Response time of Pipeline (This is the last thing you should worry about, more of a tiebreaker!)

Resources

- GitHub Repository: **YahyaAhmedKhan/nlp-guide**