Project Abstract: Information Retrieval using Language Models through Question Answering for Machine Learning and Artificial Intelligence

(CS 410: Text Information Systems , Fall 2023)

(Theme 5: Free Topics)

(Team: Lexical Wizard)

Team : Lexical Wizards			
#	Name	ID	Email
1	Amrit Kumar (Captain)	amritk2	amritk2@illinois.edu
2.	Gattu, Sudha Mrudula	sudhamg2	sudhamg2@illinois.edu
3.	Ishika Awachat, Ishika	awachat2	awachat2@illinois.edu
4.	Madhavan, Siddharth	sm120	sm120@illinois.edu

Table of Contents

Abstract:

In the rapidly evolving fields of Machine Learning and Artificial Intelligence, where efficient access to knowledge and information is critical, this project proposal aims to develop an information retrieval system that leverages open-source language models for question answering. Our primary objective is to create a valuable resource for students, researchers, and professionals in the Machine Learning and Artificial Intelligence domains. With research papers as the primary retrieval source, our emphasis on open-source models ensures that the project aligns with the evolving landscape of these fields while maintaining accessibility to knowledge seekers.

Project Objective:

The primary objective of this project is to develop an information retrieval system with the capability to extract knowledge from research papers and provide precise answers to user queries. Our goal is to achieve the following key capabilities:

Web Scraping: We will create a web scraping component to systematically extract research papers from renowned online repositories, including sources such as ArXiv, conference proceedings, and other relevant platforms. This data collection process will build a comprehensive and up-to-date database of research papers, encompassing URLs like "https://arxiv.org/abs/1706.03762" and "https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf," among others.

Text Chunk Embedding: We will leverage cutting-edge natural language processing models like BERT, RoBERTa, or equivalent open-source models to convert the textual content within research papers into dense vector representations. These embeddings will serve as the foundation for our indexing and retrieval system, ensuring efficient and accurate access to information.

Database Management: The development of a robust indexing system will be a critical component, enabling the persistent storage of these embeddings within a database. This approach ensures that the information is easily searchable and retrievable, supporting a seamless user experience.

Interactive Chat Interface: To facilitate user interaction, we will implement an interactive chat interface that empowers knowledge seekers to ask questions in natural language. This user-friendly interface will bear resemblance to platforms like ChatGPT, making the process of querying and accessing information straightforward and intuitive.

Question Answering with Open-Source Language Models: The project's central feature is the use of open-source language models, such as DistilBERT or similar models, to generate responses to user queries. Unlike proprietary large language models, this approach ensures accessibility and aligns with the project's open-source focus. The language model will draw upon the indexed information from research papers to provide contextually relevant answers, thereby enriching the information retrieval experience.

Through these capabilities, the project aims to build an efficient and comprehensive information retrieval system that meets the critical need for extracting knowledge from research papers in the fields of Machine Learning and Artificial Intelligence.

Technologies:

The project would rely primarily on Python. It will be used for web scraping, data processing and system development.

The project will consist of several interconnected components:

- **Web Scraper:** Written in Python, this component will use libraries like Beautiful Soup and Scrapy to extract research papers from online sources.
- **Text Embedding:** Modern NLP models, such as BERT or RoBERTa or LangChain embeddings will be used to convert research paper chunks into dense embeddings.
- **Database Management:** We will use database technologies like Elasticsearch /OpenSearch or similar solutions to efficiently store and index the embeddings.
- **Interactive Chat Interface:** For the user interface, we will create a web-based chat platform using python UI libraries.
- **Question Answering with Large Language Models:** The core of the project will revolve around the integration of open-source language models, such as DistilBERT or equivalent models, to process user queries and retrieve contextually relevant information from the indexed research papers. This approach ensures accessibility and aligns with the project's open-source focus, enriching the information retrieval experience.

Expected Outcome:

Our expected outcome is to finish hopefully most of the goals we've set for ourselves in the project proposal, specifically a comprehensive information retrieval system which meets the need for extracting knowledge from Machine Learning and Artificial Intelligence related research papers.

Evaluation:

To evaluate the effectiveness of our information retrieval system, we will primarily rely on key information retrieval metrics such as precision, recall and F1 Score. Precision will assess the proportion of the retrieved documents that are relevant. Recall will measure the proportion of relevant documents that are successfully retrieved. F1 Score will offer a comprehensive view of the system's performance.

These metrics will aid in accurately determining the system's ability to provide accurate and relevant answers to user's queries.

WorkLoad:

The workload for the main tasks and their estimated workloads are as following:

- Data collection and Preprocessing **(12 hours)**
    - Developing web scraping components
    - Performing data cleaning and preprocessing
- Model Integration and Development **(30 hours)**
    - Integrating open source model
    - Develop basic text embedding for research papers
    - Implement elastic search and indexing

- User Interface Development **(15 hours)**
    - Simple chat feature for question and answer
- Question and answering component **(15 hours)**
    - Integrating open source language models for question answering
- Optimization **(12 hours)**
    - Tweaking parameters to improve answering capabilities
- Documentation **(6 hours)**

Approximately, just the main tasks will take us about **90 hours** to complete.