GEN AI PRINCIPLES Course Project -1:

RAG-based Interactive AI for MS in Applied Data Science Web Site

The **MS in Applied Data Science** program at the University of Chicago offers a comprehensive curriculum that prepares students for careers in data science by covering both theoretical and practical aspects of the field. Prospective students, current enrollees, and alumni often seek detailed information about the program, including course offerings, faculty expertise, admissions requirements, career outcomes, and more. However, navigating through vast amounts of information on the program's webpage can be challenging, leading to gaps in understanding and potentially missed opportunities.

To address this, a **Retrieval-Augmented Generation (RAG) system** can be implemented to facilitate better comprehension and provide instant, accurate answers to questions regarding the MS in Applied Data Science program. This system would enhance the user experience by combining the power of a retrieval-based approach with a generative language model, ensuring that users receive relevant, contextualized responses to their queries.

Knowledge Base Source: https://datascience.uchicago.edu/education/masters-programs/ms-in-applied-data-science/

Scope of the Project:

The project aims to develop a RAG-based conversational AI system that can efficiently retrieve and generate accurate responses to user inquiries about the MS in Applied Data Science program at the University of Chicago. The system will leverage both textual data from the program's webpage and a pre-trained language model to provide detailed, context-aware answers.

Key Components and Steps:

1. **Understanding and Preparing Data:**
    - **Objective:** Analyze and preprocess the textual content from the MS in Applied Data Science webpage, ensuring the data is suitable for retrieval and generation tasks.

      Web scraping, especially when dealing with a website that has multiple sublinks, involves a structured approach to ensure all relevant information is captured efficiently. Here's how you can implement web scraping to gather data from a main page and its sublinks:

      Identify the Main Page and Structure: **Web Site**

      • Start by identifying the main webpage URL, in this case, the URL for the MS in Applied Data Science program.

      • Inspect the structure of the webpage to understand how sublinks are organized. These might include links to specific sections such as curriculum details, faculty profiles, admissions information, etc.

- **Tasks:**
  - Extract and structure content from various sections of the webpage, such as program overview, curriculum details, faculty profiles, admissions criteria, and career resources.
  - Ensure data consistency and enhance content quality to optimize the performance of the RAG system.

2. **Implementing Retrieval-Augmented Generation (RAG):**
   - **Objective:** Create a RAG framework that combines efficient information retrieval with generative capabilities for dynamic question answering.
   - **Tasks:**
     - **Embedding Generation:** Use a pre-trained transformer model to generate embeddings for the structured text data, capturing the semantic meaning of different program-related sections.
     - **Embedding Storage:** Store these embeddings in a vector database (e.g., Chroma, pinecone, faiss etc.) to facilitate the fast retrieval of relevant information when queries are posed.
   - **Outcome:** A robust retrieval system that enables the language model to access accurate, contextually relevant information, thereby improving the quality of generated answers.

3. **Integrating with a Large Language Model (LLM):**
   - **Objective:** Enable the system to engage in interactive, conversational Q&A sessions using an LLM.
   - **Tasks:**
     - Integrate the retrieval mechanism with a proprietary or open-source LLM (e.g., OpenAI GPT-4, Llama, Mistral or similar models).
     - Using Lanchain or LlamaIndex like frameworks, craft effective prompt instructions and agents to generate contextually accurate responses.
     - Ensure compliance with Responsible AI practices including grounding inappropriate responses, hallucination checks, PII redaction
   - **Outcome:** A conversational interface capable of answering queries regarding the MS in Applied Data Science program with high accuracy and relevance.

4. **Developing a User Interface**
   - **Objective:** Build an intuitive user interface to allow seamless interaction with the RAG system.
   - **Tasks:**
     - Develop the interface using a suitable framework (e.g., Streamlit or Flask) that allows users to input questions and receive answers in real-time.
     - Design the interface to visually present retrieved information and generated responses in a user-friendly manner.
   - **Outcome:** An engaging platform where users can easily ask questions about the MS in Applied Data Science program and receive detailed, accurate responses.
   - Integrating the chatbot into the existing webpage is not required for this project. However, winners may have the opportunity to collaborate with program to deploy it on the actual website.

Example Interactions:

- **Text-Based Questions:**
    1. **Question:** "What are the core courses in the MS in Applied Data Science program?"
        - **Answer:** "The core courses in the MS in Applied Data Science program include Machine Learning, Data Engineering Platforms, Statistical Inference, and Applied Data Science."
    2. **Question:** "What are the admission requirements for the MS in Applied Data Science program?"
        - **Answer:** "Applicants need a bachelor's degree in a related field, with coursework in programming, statistics, and mathematics. The application also requires a personal statement, letters of recommendation, and a resume."
- **Request for Specific Program Details:**
    1. **Question:** "Can you provide information about the capstone project?"
        - **Answer:** "The capstone project is a key component of the MS in Applied Data Science program, where students work on real-world problems, applying their learned skills to develop data-driven solutions."

Data and Deliverables:

- **Data Source:** Unstructured and textual content (including diagrams, charts but not images) from the MS in Applied Data Science webpage.
- **Expected Deliverables:**
    - A fully functional RAG-based chatbot that provides accurate, context-aware answers to queries about the MS in Applied Data Science program.
    - Comprehensive documentation, including preprocessing steps, model architectures, and system design details (min 5 pages medium article like documentation).
    - A user-friendly interface that allows prospective students, current enrollees, and alumni to interact with the chatbot.
    - A power point presentation detailing the implementation process, challenges, and future improvements (10 mins long presentation).
    - Evaluation metrics such as retrieval accuracy and response relevance based on test set.