RAG Capstone Projects

The RAG Application Project includes four parts:

Part 1 - Source Acquisition and Preparation: You must acquire digital copies and prepare them for ingestion into the RAG system. This may involve tasks such as text cleanup, formatting, and indexing.

Part 2 - System Development: You will develop the core components of the RAG application, including the retrieval engine, language model, and the integration between the two components.

Part 3 - Iterative Refinement and Deployment: You will iteratively refine and optimize your RAG system, incorporating techniques such as retrieval quality improvement, model fine-tuning, and performance optimization. You will also need to deploy your RAG on any available platform like Streamlit Community Cloud or HuggingFace.

Part 4 - Final Evaluation and Presentation: Your RAG application will undergo a comprehensive evaluation based on the following criteria:

a. Correctness of responses: The accuracy and relevance of the answers provided by your system.

b. Proper source retrieval: The ability to correctly identify and retrieve relevant passages from the source texts.

c. Speed of retrieval: The efficiency of the retrieval process in quickly locating relevant information.

d. Speed of application: The overall responsiveness and performance of your RAG application.

You will present your final RAG system, demonstrating its capabilities and discussing the challenges encountered and techniques employed during the development process.

Learning Objectives:
Your completed RAG Application Project should demonstrate the following learning objectives:

1. Understanding of retrieval-augmented language models and their applications.
2. Proficiency in data acquisition, preprocessing, and indexing for natural language processing tasks.
3. Ability to develop and integrate retrieval engines and language models into a unified system.
4. Expertise in performance optimization and techniques for improving retrieval quality and model accuracy.
5. Effective communication skills in presenting technical projects and findings.

Option 1:

- The RAG Application Project will involve creating an AI system capable of answering questions about five classic literary works: **Moby Dick, Alice in Wonderland, Dracula, War and Peace, and Grimms' Fairy Tales.**

- Things to keep in mind:
    1. These books are very different from each other
    2. The last book is a collection of stories that are different from each other

Option 2:

- Create a RAG application that answers questions about scientific papers.
    1. Efficient Generative LLMs
    2. Feedback attention is working memory
    3. LLM Task Inference
    4. LLM Security Issues

- Things to keep in mind:
    1. The papers are around the same space
    2. The papers are all highly scientific

Option 3

- Create a RAG application that answers questions from Wikipedia about the last 10 football World Cups
- Things to keep in mind:
    1. The articles on wikipedia might be similar
    2. The articles will have an overlap for some world cups on the players and countries playing