Hackathon Problem Set Sponsor: Data Science Club

Problem:
Data summarization and extracting answers from documents are crucial in today's world due to the sheer volume of information generated daily. With vast amounts of data, from business reports to research papers, it's challenging to sift through information to find what's relevant. Summarization condenses data, making it easier and faster to grasp the core insights without reading everything in detail. Similarly, extracting specific answers from documents allows individuals and organizations to quickly retrieve essential information, supporting timely decision-making and efficient knowledge management. This process is invaluable for enhancing productivity, reducing information overload, and making informed, data-driven choices in a fast-paced environment.

Challenge:
Design and develop a prototype for a Smart Document Summarization and Q&A System that enables users to upload documents, receive concise summaries, and get accurate answers to their questions about the content. The solution should be user-friendly, efficient, and capable of handling various document types, such as PDFs, text files, and Word documents.

Key Features and Objectives:

1. **Document Summarization:**
    - Develop a feature to create summaries of uploaded documents, highlighting key insights, essential points, and notable data.
2. **Question and Answer Extraction:**
    - Build a Q&A tool where users can ask specific questions about the document content and receive precise, context-aware answers.
3. **Natural Language Processing (NLP) Capabilities:**
    - Leverage NLP techniques to parse and understand natural language in both summarization and Q&A.
4. **Security and Privacy:**
    - Implement basic security protocols to ensure data confidentiality, especially if sensitive documents are uploaded.

Brownie Points:

1. **User Interface (UI) and Experience (UX):**
    - Create a user-friendly interface that allows easy document upload and interaction with the summarization and Q&A features.
2. **Incorporate sentiment analysis or tone detection as an optional feature for reports or customer feedback documents.**
3. **Teams are expected to demonstrate a working model that showcases an example of user questions and answers along with a document summary.**

- Participants are encouraged to use any resources at their disposal and are advised to explore open-source LLMs for this project. You may also use Ollama to access various LLM models if you wish to incorporate language models into your solution.
- You can use NLP Techniques like transformers, Hugging Face for accessing and selecting different models, PDF readers to read the PDF files for processing.

Checkpoints:

- Ability to read and display content from a PDF or Excel file.
- Implement data preprocessing to prepare content for model input.
- Develop logic to retrieve accurate answers based on user questions.
- Demonstrate the system by answering at least two questions.

Note: You may use an Excel sheet with predefined questions and answers to showcase responses if necessary. However, preference will be given to solutions that dynamically answer various questions from a PDF file.