ASSIGNMENT

Introduction to the data:
An **SSP (System Security Plan)** document is a comprehensive description of a system's security controls and procedures. It is typically required in compliance frameworks like **NIST SP 800-53** or **FedRAMP**. The document outlines how an organization implements and manages its security requirements.
Key Contents of an SSP Document:

1. **System Description**: Overview of the system, including its purpose, components, and environment.
2. **Control Summary**: Details about security controls implemented, mapped to specific standards (e.g., NIST 800-53).
3. **Roles and Responsibilities**: Identification of personnel responsible for the system's security.
4. **System Boundaries**: Defines what is included and excluded from the system scope.
5. **Implementation Details**: Step-by-step details on how each control is applied.
6. **Control Assessments**: Evidence or status of implementation for each control (e.g., "Implemented," "Partially Implemented," "Not Implemented").

Assignments:

1. Create a knowledge base using the provided data
   a. Pre-process (if necessary) and store the data in a vector store to be used for retrieval later.
2. Create a QA / Chat bot which retrieves relevant data from the knowledge base and answers the user's questions
   a. Create appropriate prompts to be used to communicate with the model
   b. Use an LLM (open source recommended) to answer the user's question
   c. Evaluate the retrieved context that will be sent to the LLM
   d. Evaluate the final answers of the LLM from the data and provide conclusions.

Instructions:

1. You may choose a vector store of your choice but we recommend using Postgres (with the PGVector extension which you may have to set up yourself)
2. Use open source LLM and embeddings model locally for your pipelines
3. Print or log the question, answers and the retrieved context from the knowledge base for evaluation
4. (Optional but recommended for better retrieval) Try extracting and storing metadata from the chunks which can be used to retrieve more relevant chunks
5. (Optional but recommended) Add more techniques for retrieval instead of just depending on Semantic Similarity Search using Embedding models. (eg: BM25, Rerankers, etc).
6. (Optional but recommended) Create an app using FastAPI which contains endpoints to perform actions mentioned in the Assignment section.
7. (Optional and only do it as the last activity if possible as it will incur personal cost) Use OpenAI LLM and Embedding models to do the same and compare the results.

Questions to ask the LLM (You may choose all or any 7 questions to answer using the LLM)

1. What is the AC-01 Control Policy?
2. What is the Parameter for AC-01 Part a or AC-01(a)?
3. What is the Implementation Status of AC-02(02)?
4. What is the Parameter for AC-02(02)-1?
5. What is the Control Origination for AC-06?
6. What is the solution for IA-08 and how is it implemented?
7. What are the Responsible Roles for AC-02(04)?
8. What are the Responsible Roles for AC-06?
9. What is the Parameter for SI-04(b)?
10. What is the Control Origination for SI-04?
11. What is the solution for SI-04(10) and how is it implemented?
12. What is the Parameter for SI-03(c)(2)-2?

13. What is the solution for SI-03 and how is it implemented?

- You may modify the questions/queries to improve them