**Assignment 1 – Problem Statement 1**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Instructions**

1. **Follow the instructions in each question carefully.**
2. **Only two files should be uploaded in canvas without zipping them. One is ipynb file and other one html output of the ipynb file. No other files should be uploaded**
3. **Any assignment submitted using other python IDEs are not considered for grading.**
4. **Incorrect Assignment Set submitted will not be considered.**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

1. .

**Dataset Link:**

<https://naco.gov.in/sites/default/files/Standard%20Operating%20Procedures%20for%20SCM%20under%20NACP.pdf>

**Problem Statement:**

Develop and implement and Advanced RAG system using Vector store/DB of choice for retrieval and pretrained LLM. You need to implement 2 services:

1. **Ingestion Service:** Ingest the data from files into a vector store**.[Only text data].** Images and graphics are out of scope.
2. **Retrieval Service:** Takes user query as input and retrieve relevant data from knowledge base and generates answer to the query.
3. **Q&A System:** Use LLM and generate Q&A answers on user query.
4. You can use LLM and embedding models of your choice.
5. You have to use Lang Chain or LLamaindex for implementation of the above services.

**Tasks:**

**Ingestion Service** [5 Marks]

1. Design the approach and extract text data from input pdf file. Explain the step by step process of data parsing mechanism for text extraction [1 Mark]
2. Leverage any open source OCR extraction library to extract the content from unstructured document. [1 Mark]
3. Convert text into chunks and explain the logic of chunking strategy. [1 Mark]
4. Implement vectorization of chunks into embedding vectors. Explain which embedding model you have used [1 Mark]
5. Ingest data to the vector store. Use vector DB like croma DB or any other open source vector DB as a knowledge base [1 Mark]

**Retrieval Service** [10 Marks]

1. Design the retrieval pipeline of the framework that will be leveraged for the end to end pipeline [2 Marks]
2. Leverage open source LLM to generate user query for information retrieval [2 Marks]
3. Retrieve relevant chunks from vector store using Langchain or LLAMAindex frameworks [3 Marks]
4. Implement semantic similarity approach to proceed with ranking to select most relevant chunks. [1 Mark]
5. Response Generation: [2 Marks]

* Generate responses as part of the output.
* Add required guardrails to improve the retrieval mechanism.