Case study 5

**Assignment Title:** Development of a Structured Database and Text Extraction System for Finance Professional Development Resources

**Course:** Big Data and Intelligent Analytics

**Assignment Due Date:** April 12th, 2024 (5.00pm)

**Overview:** You are an enterprise experimenting with the use of Models as a service APIs to build intelligent applications for knowledge retrieval and Q/A tasks.

In this assignment, we will leverage Pinecone and OpenAI api's for:

1. Creating knowledge summaries using OpenAI's GPT
2. Generating a knowledge base (Q/A) providing context
3. Using a vector database to find and answer questions.
4. Use the knowledge summaries from 1 to answer questions.

**HYPTOTHESIS:**

In enterprises, there are often use cases where information needs to be curated from multiple sources.While no single solution can meet the requirement, experiments need to be done to check which approach bodes well for the use case. We will do 4 experiments to see if the results can actually be operationalized in the enterprise.

**Assignment Description:**

Your task is to use the assigned documents from the CFA Institute's website for your assignment.Each team will work on 3/4 documents. You should have already scraped

the Introduction, Summary, LOS for each of these topics. You may use that as a starting point.

**Requirements:**

**Build a multipage streamlit for each of the 4 usecases.**

1. **Creating knowledge summaries using OpenAI's GPT**

**Goal: To build a knowledge base given topic summaries and objectives (LOS)**

**Use context:** The summary, introduction and LOS.

**Target user:** A financial analyst with an MBA interested in learning more about the LOS

- Given a LOS, create a technical note that summarizes the key Learning outcome statement (LOS). Note: Be sure to include tables, figures and equations as you see fit.
- Repeat with all the LOSs.
- Consolidate the entire note into a document in markdown format.
- Chunk each LOS and the generated note and store it into Pinecone.

2. **Generating a knowledge base (Q/A) providing context**

**Goal: To build a question/answer set to reinforce learning from topic summaries.**

**Use context:** The summary, introduction and LOS.

**Target user:** A financial analyst with an MBA interested in learning more about the LOS

Generate a question bank (50 questions - Set A) each with 4 options with one correct answer from the “Summary” section of each of the ¾ topics assigned to you.

Ensure the questions are similar to the ones in complexity and type of : https://www.cfainstitute.org/-/media/documents/support/programs/cfa/sample-level-II-itemset-questions.pdf

https://www.cfainstitute.org/-/media/documents/support/programs/cfa/sample-level-I-questions.pdf

https://www.cfainstitute.org/-/media/documents/support/programs/cfa/sample-level-III-itemset-questions.pdf

This means, you will have to parse all these three documents (including any equations) and provide them as context to generate questions.

Rerun this and generate another 50 questions and answers and keep it aside. (We will call it set B)

Store the first 50 in pinecone with separate namespaces for questions and answers (See 3)

3. Using a vector database to find and answer questions.

We will use RAG to to search for similar questions and see if we can answer questions in set B accurately just by using the question and answers we created and stored in pinecone(Set A)

We won't provide the answers. But we will take a question from Set B and find 3 questions in Set A that are most similar. We will pass the answers of these 3 questions to GPT-4 along with the question under consideration from Set B along with the 4 answer choices and ask GPT to only use the information provided to get back an answer and justify why it provided that answer. Compare the answer to the answer it correctly determined in step 3.

How many of the 50 questions were answered correctly? Explain and report

4. Use the knowledge summaries from 1 to answer questions.

For each of the questions in Set A and Set B (along with the answer choices), search for similar embeddings and LOS that have the answer using RAG in the pinecone vector database created in Step 1 and tabulate how many of the 100 questions/topics were correctly answered.

Discuss whether this or the approach discussed in 3 is a better option. Comment on the design and discuss any other approaches you may think may lead to a better design to answer questions from Set A and Set B

**Submission Guidelines:**

- Prepare a GitHub repository to host all the materials related to this assignment.
- Ensure your submission is professional, with clear documentation and comments within your code to explain your process and logic.
- The assessment of your assignment will be based on the correctness of your implementation, adherence to the project requirements, the efficiency of your code, and the organization and professionalism of your GitHub repository.

**Evaluation Criteria:**

- Accuracy and completeness of the datasets created.
- Efficiency and organization of the code.
- Adherence to the project requirements and deadlines.
- Clarity and professionalism of the GitHub repository and documentation.

**Additional Notes:**

- Ensure the privacy and security settings of your S3 bucket are appropriately configured to prevent unauthorized access. Same goes with API keys (OpenAI, Pinecone etc.)
- Consider the scalability of your solution, as the techniques and methodologies you develop for this assignment may be applied to larger datasets or similar projects in the future.

**Good luck, and we look forward to reviewing your innovative solutions!**

## References:

**Review**


1. https://docs.pinecone.io/guides/getting-started/quickstart
2. https://github.com/openai/openai-cookbook/blob/main/examples/vector_databases/pinecone/Using_Pinecone_for_embeddings_search.ipynb
3. https://github.com/openai/openai-cookbook/blob/main/examples/vector_databases/pinecone/Gen_QA.ipynb
4. https://github.com/openai/openai-cookbook/blob/main/examples/vector_databases/pinecone/GPT4_Retrieval_Augmentation.ipynb
5. https://github.com/pinecone-io/examples/blob/master/learn/search/hybrid-search/ecommerce-search/ecommerce-search.ipynb
6. https://github.com/openai/openai-cookbook/blob/main/examples/vector_databases/pinecone/Semantic_Search.ipynb
7. https://github.com/pinecone-io/examples/tree/master/learn/search/question-answering