ADTA 5770: Generative AI with LLM
Semester Project – Coding Outlines

1. Overview

The semester project aims to develop a knowledge-based Question-Answer and Search System. The project is done using the Cloud Integrated Development Environment (IDE) System (CIDES) provided by Google Cloud Platform (GCP) Vertex Ai services.

The student should complete the following phases and steps while coding to develop the Q&A Search system using the GCP: Vertex AI platform.

2. Main Phases

The code to develop the Q&A Search system should consist of the following main phases:

1. PHASE 1: Set up the CIDES (Cloud Integrated Development Environment System (CIDES))
    a. Install all necessary system software platforms, including AI platforms, tool libraries, and utility programs

    NOTES: *MUST un-install incompatible Torch-related modules and install compatible ones*

    b. Perform the GCP authentication to connect the code in COLAB and the host GCP project
        i. Authenticate the GCP developer ID who is the author of the COLAB code

2. PHASE 2: Import all the necessary GCP cloud platforms, AI platforms, API platforms, tool libraries, and utility programs.

3. PHASE 3: Initialize GCP AI platform: aiplatform

4. PHASE 4: Create an empty vector search index and deploy it with a public end-point

5. PHASE 5: Embedding or vectorize the PDF files

6. PHASE 6: Build a vector database with embedded contents for the Q&A Search system by configuring the vector search index as an instance of the GCP: Vertex AI: Vector Store
    a. Configure the index as an instance of GCP: Vertex AI: Vector Store
    b. Upload the vector store into a GCS bucket

7. PHASE 7: Test the Q&A Search system

8. PHASE 8: Set up the query and response subsystem of the Q&A Search system
    a. Define the functions to format the prompt template
    b. Define the functions to format the responses

9. PHASE 9: Run prompts and get responses

10. Phase 10: Delete all the end-points and the indexes that have been created during the system development

3. PHASE 4: Create and Deploy Empty Index with Public End Point

3.1 Overview

PHASE 4 is considered as the most critical and significant part of developing the Q&A Search system, determining the success of the project.

The developer must ensure the following PHASE 4 steps are done correctly.

- (1) An index is created correctly with the correct index DISPLAY NAME
- (2) A public end-point is created correctly with the correct end-point DISPLAY NAME
- (3) The index is deployed correctly with the public end-point created in Step 2 above with the correct DEPLOYED ID specified by the developer (before deploying the index)

IMPORTANT NOTES:
--) After steps (1) and (2) above, the developer should print out the index ID (index.name) and the endpoint ID (end_point.name) to verify that a new index and a new endpoint have been created.

3.2 Vector Search Index: Names and IDs

A vector search index has four names or ID after having been created successfully.

3.2.1 Text display name

The developer must define the text DISPLAY name of a vector search index before creating it (the empty one).

For example: A_XYZ_DISPLAY_INDEX_NAME = "**a_string_name_here"

3.2.2 Numeric index ID

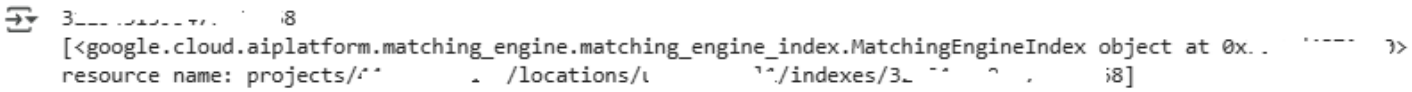
After the vector search index has been created successfully, the system assigns it a unique (GCP wide) index ID. The developer can print out the numeric index ID of an index.

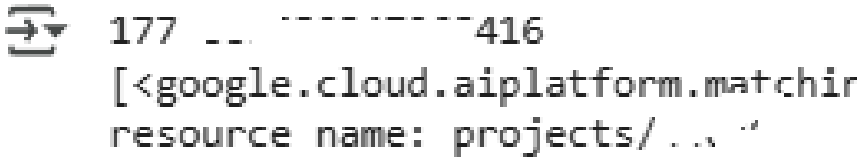
For example: 1234567891011121314151 6

3.2.3 Resource name

The system also assigns a unique (GCP-wide) resource name to the vector search index after it has been created successfully. The developer can also print out the resource name along with the numeric index ID of an index.

For example:





The complete string following the text “resource name” is the resource name of a vector search index.

3.2.4 Text DEPLOYED index ID

The developer must define the text DEPLOYED INDEX ID of a vector search index before deploying it (the empty one) with a public end point

For example: A_XYZ_DEPLOYED_INDEX_ID = "**a_string_deployed_index_id_here"

IMPORTANT NOTES:
--) Only the text DISPLAY name, the numeric index ID, and the text DEPLOYED ID are used in the semester project.

3.3 Index End Points: Names and IDs

An index end point has three names or ID after having been created successfully.

3.3.1 Text display name

The developer must define the text DISPLAY name of an index endpoint before creating it.

For example: A_XYZ_DISPLAY_END_POINT_NAME = "**a_string_name_here"

3.3.2 Numeric index endpoint ID

After the index endpoint has been created successfully, the system assigns it a unique (GCP wide) endpoint ID. The developer can print out the numeric endpoint ID of an endpoint.

For example: 1234567891011121314151 6

3.3.3 Resource name

After the index endpoint has been created successfully, the system assigns it a unique (GCP wide) endpoint resource name. The developer can also print out the endpoint resource name of an endpoint.

For example:

The complete string following the text “resource name” is the resource name of an index endpoint.

IMPORTANT NOTES:
--) Only the text DISPLAY name and the numeric end point ID are used in/ the semester project,

4. Phase 10: UN-DEPLOY Indexes and DELETE Indexes & End Points

4.1 Overview

PHASE 10 is also a very important part of developing the Q&A Search system. The developer must delete all the indexes and endpoints that have been created to avoid unnecessary charges after having completed the semester.

The developer must ensure the following PHASE 10 steps are done **IN ORDER** correctly.

- (1) Get the list of end points that have been created
- (2) Un-deploy the indexes that have been deployed with these end points
- (3) Delete all the end points that have been created
- (4): Get the list of end points that have been created
- (5) Delete all the end points that have been created