ADTA 5900/5770.501: Generative AI with LLMs

Thuan L Nguyen, PhD

Assignment 4

1. Overview

The rise of cloud computing has facilitated the emergence of big data. Cloud computing commodifies computing time and data storage using standardized technologies.

Big data is a term for large volumes of data that can be both structured and unstructured. These enormous volumes of data overwhelm the digital world every second. However, it is not the amount of data that matters. What we can do with the data matters: Big data analytics can provide insights that lead to better decisions and strategic moves.

The emergence of cloud computing has made it easier to provide the best technology in the most cost-effective packages. Cloud computing has reduced costs and made many applications available to companies of all sizes: small, medium, large, and giant corporations.

2. Google Cloud Platform (GCP): Service Account

Various enterprise services of Google Cloud Platform (GCP) will be used in classwork. When using cloud services to run applications, cyber-security is one of the top priority. To create and run Natural Language Processing (NLP) applications in GCP, the user must set up a service account that can be used in GCP's sophisticated authentication system.

Google Cloud Platform (GCP) Vertex AI service is used throughout the project. Each student must have a GCP account so that he/she can access and use GCP: Vertex AI services as required.

IMPORTANT NOTES:
*--) All the documents posted on the Canvas page **GOOGLE CLOUD PLATFORM: GCP for Deep Learning – TF2** can be used for HW 4, HW 5, and the final project.*

*--) All the documents posted on the Canvas page **GOOGLE CLOUD PLATFORM: GCP for Natural Language Processing (NLP)** can be used for HW 4, HW 5, and the final project.*

3. Homework 4: Assignment Format

Homework 4 is assigned as a group assignment. It means all the group's student members will collaborate while working on the assignment.

However, each student must **write** and **submit** his/her report **independently**. In other words, a student works on the assignment with the group but **writes** and **submit**s the report as if he/she had worked independently.

4. Homework 4: General Assignments

Each group is assumed to be an AI system development group in a business organization. With the explosion of popularity and widespread use of generative AI in real-world management and business activities, the corporation's leaders want the group to develop a generative AI system that the company employees can use to perform content searches, ask questions, and get answers about the contents of the organization's proprietary documents.

The group will adopt Google Cloud Platform (GCP): Vertex AI services as the primary system Integrated Development Environment (IDE) to design, build, and test the system throughout the project, including but not limited to cloud storage, vector embedding generation, vector databases management, and advanced vector search technologies. For development, the group will use Python for coding with Google Collaboratory (Colab) as the coding IDE. The group also plans to use popular generative AI techniques, including but not limited to Retrieval Augmented Generation (RAG), Sentence Transformer, and tools provided by generative AI platforms like LangChain and Hugging Face.

5. PART I: Generative AI System for Business (20 Points)

5.1 Decision on What Type of Business Organization

TO-DO

- Each group is assumed to be an AI system development group in a business organization.
- Brainstorm and discuss with group members to decide on a business organization, such as a company or a corporation, for which the group will develop a generative AI system.

IMPORTANT NOTES:
--) Each group has selected a domain expertise field to focus on while working on the semester project.
--) The group should now place the semester project in a business context of a company or corporate.

SUBMISSION REQUIREMENTS: PART I:
--) Document the group brainstorming and discussion details to decide which type of business organization will develop and deploy the generative AI system.

6. PART II: Business and Technical Requirements of the System (30 Points)

TO-DO

- Based on the business type of the business entity:
    - Document the **business values** of the generative AI system to create
        - Why does the company want to develop a generative AI system?
    - Document the **business requirements** of the generative AI system to create
        - In other words, document the business goals or objectives that the organization tries to achieve with this generative AI system.
- Document the technical requirements of the generative AI system to create.
    - For example:
        - Which AI platform will be used to create the generative AI system?
            - → GCP: Vertex AI
        - Which large language model (LLM)?
        - Which generative AI platform?
        - … and more …

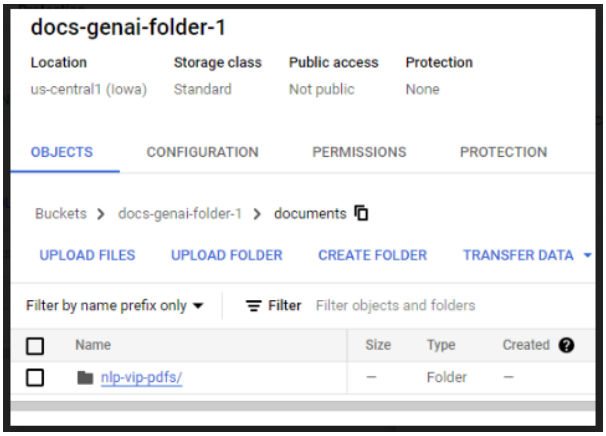
SUBMISSION REQUIREMENT: PART II:
--) Document the details of the business and technical requirements of the generative AI system to be developed.

7. PART III: Data and Cloud Data Storage Requirements (40 Points)

7.1 Create GCP Cloud Storage Bucket

TO-DO

- Access GCP Storage
- Create one bucket:
    - BUCKET Name = adta5770-docs-folder
- Inside the bucket, create a new subfolder named “documents,” in which another subfolder named “pdfs” is created.
- Capture the screenshot of the last subfolder like the following figure:



In the figure, in the bucket docs-genai-folder-1, there exists a subfolder named "documents" in which a sub-sub-folder "nlp-vip-pdfs" is created.

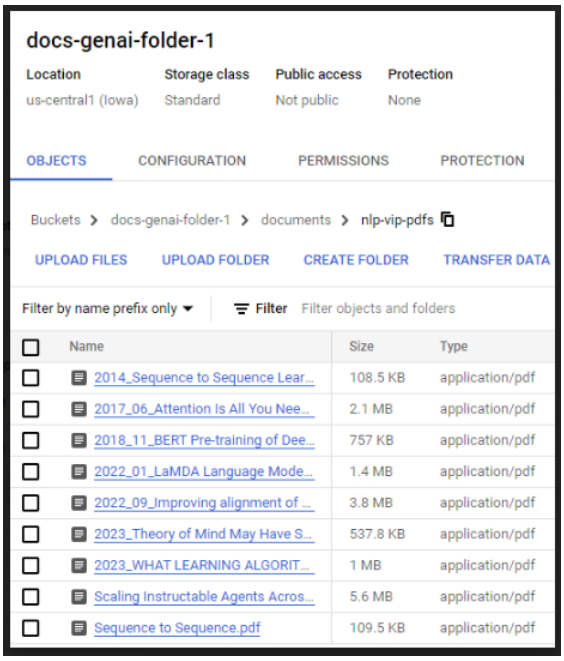
SUBMISSION REQUIREMENTS: PART III #1:
--) Document what has been done in PART III in the HW 4 report.
--) Submit the screenshot like the above, for the bucket and its subfolders

- The screenshot should show the bucket and its subfolders (**BUCKET 1/documents/pdfs)**

7.2 Upload Documents to GCP Bucket

TO-DO

- Upload the collected PDFs to the cloud folder: **BUCKET 1/documents/pdfs**
- Access the folder and take a screenshot of its contents like the following one:



SUBMISSION REQUIREMENTS: PART III #2:
--) Document what has been done in PART IV in the HW 4 report.
--) Submit the screenshot as required.

8. PART IV: Groupwork Evaluation (10 Points)

SUBMISSION REQUIREMENTS: PART IV:

Provide the information about your group activities by answering the following questions:

1. What group do you belong to? (Provide the group number)
2. Who are the members of your group?
3. Have the members organized meetings (ONLINE or IN-PERSON) to work on HW 4?
4. If **YES to #3**, which members, including the student himself/herself, showed up in the meeting?
5. If **YES to #3**, do all the members make reasonable efforts to participate actively in the group work?
6. If **NO to #5**, do you have any opinions to share about the group?

9. HOWTO Submit

The student must submit all the sections, i.e., submission requirements, in a Microsoft Word document sent to the instructor (Thuan.Nguyen@unt.edu) as an attachment to a UNT email.

The subject of the email must be:

- "ADTA 5760: Assignment 4 – Submission."

Due date & time: 11:00 PM – Wednesday 04/02/2025