**自然語言處理**

**Natural Language Processing**

**Assignment 4: Retrieval-Augmented Generation with LangChain**

|  |  |
| --- | --- |
| 開設學校： | 國立清華大學 |
| 就讀校名： | 國立臺北科技大學 |
| 就讀班級： | 資工碩一 |
| 修課學號： | 113598043 |
| 修課學生： | 張育丞 |
| 授課老師： | 高宏宇 教授 |

Dec 22, 2024

# **前言**

在現今資訊爆炸的時代，如何從大量的非結構化數據中提取有用資訊成為一項重要的技術挑戰。檢索增強生成 (Retrieval-Augmented Generation, RAG) 系統作為結合檢索與生成模型的新型技術，為解決此問題提供了一種高效的方法。本次作業的目標是運用 RAG 技術，結合 Llama 3.2 語言模型和 Chroma 向量資料庫，建構一個能準確回答問題的系統。透過對提示詞設計、檢索模型選擇及嵌入技術的實驗與優化，我們不僅能深入理解 RAG 系統的運作原理，還能進一步探索其在不同設計情境下的效能差異。此報告將詳細說明系統的架構、實作過程與效能分析，並提出改進建議，期望為未來相關應用提供參考依據。

# **程式實作**

作業四可將程式實作分為 8 的區段，分別為基礎環境設置、 Hugging Face 登入、 Ollama 平台設置、資料集載入與預處理、嵌入模型與向量儲存、提示詞設置、建構 RAG 、問答測試、系統優化與改進。

1. 基礎環境設置

* 安裝與載入必要套件，包括 langchain、huggingface\_hub、rank-bm25 等工具，確保程式運行環境完整。
* 使用 nvidia-smi 檢查 GPU 狀態，以確認可用資源是否支持模型運行。

1. Hugging Face登入

* 使用 Hugging Face 提供的 API token 進行授權，允許存取如 Llama 3.2 等需要認證的模型。
* 運行 huggingface-cli whoami 確認登入狀態，保證後續模型載入正常執行。

1. Ollama 平台設置

* 安裝 Ollama並且啟用，在Colab上運行 Llama 3.2模型。
* 取得 Llama 3.2 模型的權重並且啟動相關服務，使得能夠進行問題的查詢與回答的生成。

1. 資料集載入與預處理

* 使用 cat-facts.txt 作為基礎的資料集，將每一行的文字進行獨立處理，存放於 refs 的List當中。
* 將處理後的資料轉為 LangChain 所需的 Document 格式，方便進行後續的向量儲存。

1. 嵌入模型與向量儲存

* 本實驗選用Jina提供的嵌入模型 (jinani/jina-embeddings-v2-base-en) 進而生成文件的嵌入向量。
* 使用 Chroma 作為向量儲存工具，建立檢索數據庫，並設置為支援mmr檢索模式，以利提升結果的多樣性。

1. 提示詞設置

* 設計一個清晰且明確的提示詞模板，包含以下：
  + 使用上下文來回答問題。
  + 若無法回答，直接回應「不知道」。
  + 回答限制在三句內，使其簡潔。
* 使用 ChatPromptTemplate 作為 RAG 的模板。

1. 建構 RAG

* **回答鏈 (QA Chain)：**基於 LangChain 所提供的 create\_stuff\_documents\_chain 之函數，組合語言模型與提示詞生成答案。
* **檢索鏈 (Retrieval Chain)：**結合檢索器與問答鏈，確保系統能基於檢索到的文件上下文生成精準的回應。

1. 問答測試

* 定義一組具有 10 個問題的查詢列表，涵蓋存貓的行為到歷史事件等多種問題。
* 使用 RAG 進行查詢，並依據回應與正確答案匹配度計算正確率。
* 設計程式邏輯以允許答案比對，確保測試的公平性。

# **環境建置資訊**

本次實驗由於環境 Ollama 搭建不易，因此全部採用 Colab 進行運行(參考表1)。

表1 開發平台與工具

|  |  |
| --- | --- |
| **Component** | **Specification** |
| Running environment | The program is run on Colab.  GPU: T4 |
| Python version | The program is run on Colab.  Python: 3.10 |

# **問答解析**

1. What’s in your RAG System?

* **Language Model:** Utilized Llama 3.2 provided by Ollama.
* **Embedding Model:** Jina Embeddings (v2-base-en) was employed for sentence embeddings.
* **Retrieval Model:** Chroma was used as the vector store, supporting multiple retrieval types (with "mmr" used in this assignment).
* **Dataset**: Processed "cat-facts.txt," converting it into document format and building the Chroma vector store.

1. Which retrieval model do you use?

* Chroma vector store with the retrieval type set to "mmr" (Maximal Marginal Relevance).

1. What’s your prompt?

**The prompt is structured as follows:**

Use the given context to answer the question.

If you don't know the answer, say you don't know.

Use three sentence maximum and keep the answer: concise.

Context: {context}

1. What’s new in your code compared to the code from our lab course?

* **Embedding Model:** Switched to Jina Embeddings and adjusted embedding parameters (trust\_remote\_code and normalize\_embeddings).
* **Retrieval Model:** Opted for "mmr" retrieval to improve result diversity.
* **Prompt Optimization:** Limited response length and added explicit instructions to state "I don't know" if the answer is not available.

1. Shown my best score for the ten questions?

The RAG system achieved a correct answer count of

|  |  |  |
| --- | --- | --- |
| No. | Type | Text |
| QA1 | Query | How much of a day do cats spend sleeping on average? |
| Response | Cats spend approximately 16 to 18 hours per day sleeping on average. This translates to around 2/3 of the entire day, as mentioned in the context. However, it's also worth noting that this sleep time includes periods when they're still alert and responsive to stimuli while asleep. |
| Status | Correct! |
| QA2 | Query | What is the technical term for a cat's hairball? |
| Response | The technical term for a cat's hairball is a "bezoar." A bezoar is a small mass of material that forms in a cat's stomach when they swallow their own fur. It can become lodged and cause a blockage if it passes through the digestive tract. |
| Status | Correct! |
| QA3 | Query | What do scientists believe caused cats to lose their sweet tooth? |
| Response | Scientists believe that a mutation in a key taste receptor is the reason why cats lack a sweet tooth. |
| Status | Correct! |

|  |  |  |
| --- | --- | --- |
| QA4 | Query | What is the top speed a cat can travel over short distances? |
| Response | A cat can travel at a top speed of approximately 31 mph (49 km) over a short distance. This speed allows cats to quickly escape, hunt, and navigate their surroundings. This impressive speed is typical for many domesticated breeds. |
| Status | Correct! |
| QA5 | Query | What is the name of the organ in a cat's mouth that helps it smell? |
| Response | The Jacobson’s organ is the additional organ that allows cats to smell besides using their nose. It is located in the upper surface of a cat's mouth. This organ aids the sense of smell, particularly when a cat smells itself. |
| Status | Correct! |
| QA6 | Query | Which wildcat is considered the ancestor of all domestic cats? |
| Response | The African Wild Cat is considered the ancestor of all domestic cats. This species still exists today, implying a direct line of evolution between it and domesticated cats. It lived around 30 million years ago. |
| Status | Correct! |
| QA7 | Query | What is the group term for cats? |
| Response | The group of words associated with cat stem from the Latin catus, and are not a collective noun. A "clowder" is actually a collective noun for cats, referring to a large group of them. |
| Status | Correct! |
| QA8 | Query | How many different sounds can cats make? |
| Response | Cats make about 100 different sounds. This is significantly more than dogs, which only make about 10 sounds. The exact range of possible cat sounds is not definitively known, but it is clear that they are much more vocal than dogs. |
| Status | Correct! |
| QA9 | Query | What is the name of the first cat in space? |
| Response | The first cat in space was a French cat named Felicette (a.k.a. "Astrocat"). She was launched into outer space by France in 1963. Electrodes implanted in her brains sent neurological signals back to Earth, and she survived the trip. |
| Status | Correct! |

|  |  |  |
| --- | --- | --- |
| QA10 | Query | How many toes does a cat have on its back paws? |
| Response | I don't know. Cats are born with four toes on each back paw, but the question isn't asking about cats being born with extra toes. |
| Status | Correct! |

1. Please provide an analysis of the RAG performance using different prompts.

* **Performance with the Default Prompt:** The prompt is clear and concise, but due to its brevity, responses may sometimes lack sufficient detail.
* **Improvement Suggestions:** Test more detailed prompts, such as adding, "Try to provide a comprehensive answer using the context without unnecessary assumptions."

1. Please compare the RAG performance with the different retrieval models and the performance without RAG (note that Llama 3.2 should not be fine-tuned in this assignment).

* **Comparison of Different Retrieval Models:**
  + Using Chroma with "mmr" retrieval improved response diversity and accuracy.
  + Testing other retrieval models like BM25 or Faiss would provide further insights, especially for large datasets.
* **Performance Without RAG:** The LLM's accuracy dropped significantly, as it lacked contextual information provided by the retrieval process.

1. Anything that can strengthen your report.

* Add more challenging test questions (e.g., questions with ambiguous language).
* Explore multi-turn interactive QA to test the system's dynamic response capabilities.
* Experiment with additional retrieval models (e.g., Faiss or Dense Passage Retrieval) for performance comparisons.

# **實驗心得**

這個實驗相較於先前三實驗更加簡單且容易上手，先後我詳閱了實驗說明影片，並複習課程資料之RAG\_Tutorial\_1.pdf及RAG\_Tutorial\_2.pdf，依兩簡報針對RAG訓練及Ollama的部署，將課程所需之環境成功搭建，這將使得我的實驗更加快速完成。整體內容以貓的知識資料集為主軸，套用RAG技術，對於LLM接觸如蜻蜓點水的我來說，還是挺有魅力且喜歡的，眼見這是該課程的最後一次作業還是有點不捨的感受，也感謝課程激起我對於LLM的初探與熱情。