自然語言處理 HW4

成功大學114資訊系 F74102022唐文蔚

|  |  |
| --- | --- |
| Environment | Colab |
| CPU | Colab |
| GPU | Colab |

Report：

Q1：Please describe the details of your implementation for the RAG system (please tell us 1. What’s in your RAG system? 2. Which retrieval model you use? 3. What’s your prompt? 4. What’s new in your code in comparison with the code from our lab course?) in this assignment and list your best score for the ten questions.

A1：

(1) 本次作業的 RAG (檢索增強生成) 系統由以下主要組件組成：

⚫ 嵌入模型 (Embedding Models): 使用 SentenceTransformer和 Hugging Face 的嵌入模型進行句子嵌入計算。

⚫ 檢索器 (Retriever): 用Chroma建立Collections查找最相似的文本。

⚫ 生成器 (Generator): 使用 Meta-Llama (Llama 3.2-1B Instruct) 模型作為語言模型，負責生成回應。

⚫ 提示語 (Prompt): 提示語結構將檢索到的上下文與使用者查詢結合，以生成相關答案。

(2) 嘗試多個模型後 ”infloat/multilingual-e5-large” 的表現最穩定。

(3) prompt包含請LLM使用數字並加上單位、簡短描述、復述檢索資料再回答等。

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述 (4) 由於修改Prompt和retriver模型就可以成功檢索資料並正確回答，我並沒有做太多修改。

best score：**10**

一張含有 文字, 螢幕擷取畫面, 字型, 文件 的圖片

自動產生的描述

Q2：Please provide analysis for the RAG performance using different prompts

A2：我嘗試過成效較好的prompt：一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

除了初始的prompt外，其餘prompt大部分都可以成功讓LLM正確回答問題。

<Bonus>

限制句數的差異：當在system prompt沒有限制回答長度時，模型會回答很長的句子且較容易回答context外較無關的內容；然而，若是限制LLM精簡答案，則可能會沒有回答到正確答案、缺失單位、或是用不同方法表示答案(2/3變成two out of three、four變成4等)，因此要讓模型回答適當長度的答案並且包含數字和單位等

Q3：Please compare the RAG performance with different retrieval models and the performance without using RAG (note that Llama 3.2 should not be fine-tuned in this assignment).

Without RAG : 4 – 8 correct

With RAG :

|  |  |
| --- | --- |
| Embedding Model | Score |
| intfloat/multilingual-e5-large | 9 - 10 |
| jinaai/jina-embeddings-v2-base-en | 8 - 10 |
| sentence-transformers/all-MiniLM-L6-v2 | 9 |
| sentence-transformers/LaBSE | 7 - 9 |
| facebook/contriever | 8 - 9 |
| princeton-nlp/sup-simcse-bert-base-uncased | 8 - 9 |

Conclusion：基本上大部分的Model產生的Vector都可以成功retrieve，因此在LLM能力足夠的情況下大部分都可以正確回答問題。