采威國際資訊股份有限公司

實習生:黃聖棋

目錄

壹、 專案概述

一、關於 RAG

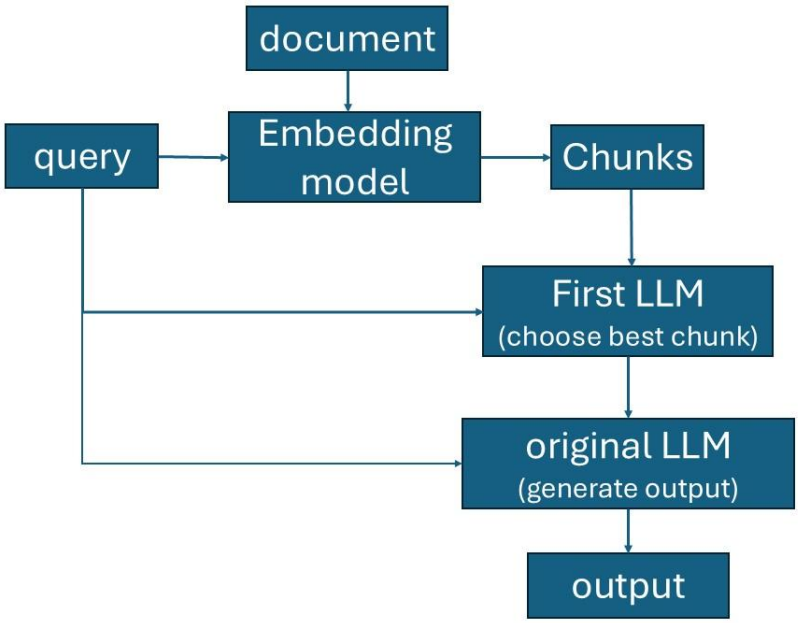
Retrieval Augmented Generation 以下簡稱 RAG，RAG 是一種技術，可讓大型語言模型(LLM) 使用從外部資訊來源擷取的支援資料來擴增使用者提示，從而產生擴充的回應。透過納入擷取的此資訊，RAG 可讓 LLM 產生更準確、更高品質的回應，而不是使用額外的內容來擴增提示。

二、RAG Data

本英文試驗文章採用諾貝爾和平獎得主馬拉拉自傳 I AM MALALA(pdf)以及 MALALA 相關新聞報導(txt)兩篇，總共約兩百頁。

中文試驗採用金庸神鵰俠侶全書。

三、模型概述



貳、 英文模型

一、Embedding model

Embedding 是一個專為文本嵌入生成的技術，基於 Sentence Transformers，它能將文本數據轉換為高維嵌入向量，這些向量是文本在語義空間中的數字表示，用於語義檢索、文本相似性比較和機器學習模型的特徵輸入。內部依賴 SentenceTransformer 模型中的 all-mpnet-base-v2，是一個基於 MPNet 的預訓練模型，能生成 768 維的嵌入向量，具有強大的語義表示能力並支持多語言。嵌入生成過程完全在本地執行，確保計算效率和數據隱私。這些嵌入可用於計算餘弦相似度，實現語義檢索或文本聚類，特別適合用於問答系統、文本分類、向量搜索等需要深度語義理解的任務場景。

二、chunk technology

在本次實驗中，我們嘗試了兩種不同的 chunk size、兩種 topk 的選擇、一種非同步的切分方法。其中，使用 chunk size=1000 且 overlap=150 的配置在本次數據集上表現最佳。這種設置不僅能有效捕捉局部資訊，還能保留整體結構。此外，對於這樣的數據量而言，計算負擔仍然在可接受的範圍內，因此非常適合此類實驗。

對於 Top-k 的 k 值設定，需要考慮的有數據的廣泛性、準確性以及運算負擔三個因素。在運算負擔可控的情況下，選擇 k=5 是最合適的，既能保證檢索結果的多樣性，又不會導致過多冗餘計算。

針對非同步的切分方法，我們主要考慮當語意剛好被切割時，如何避免目標句子被分隔到不同的 chunks。透過嘗試兩種不同的 chunk size，可以有效減少這種情況發生，提高語意連貫性。

三、before LLM

綜合以上，在嘗試不同的 chunk size 並確保 chunk 資訊正確性的基礎上，我們在切分後額外引入了一個語言模型作為驗證步驟。通過語言模型的檢查，能更精確地篩選有效的 chunks，並將這些最具語意相關性的 chunks 傳遞給最後的 LLM，進一步提高處理結果的準確性，同時有效避免資訊丟失的問題。

理想情況下，當數據量足夠時，可以選擇較大的 Top-k 值，然後通過語言模型從 Top-k 中挑選出最貼近問題的前幾個 chunk。這樣的策略不僅能保證數據的準確性，還能保留語意的廣泛性，確保關鍵信息不會因 Top-k 值過小而被忽略。同時，語言模型的篩選可以進一步提升選擇結果的相關性，為後續處理提供更加精準的上下文支持。

四、 中文模型

一、 Embedding

本次使用兩種 embedding 模型，all-mpnet-base-v2 與 text2vec-base-chinese，前者為支援多語言 embedding 之模型，後者為專門處理中文 embedding 模型。並比較簡繁中文 embedding 的向量相似度準確性，作為將 chunk 輸入 LLM 前的準確度評估依據。

(一) traditional and simplify Chinese embedding (chunk differences)

(二) differences of two embedding models

二、 chunk size

本次試驗嘗試了 chunk size 800 與 1000，

三、 before LLM