

使用 LangChain，透過 RAG 機制擴展 LLM 模型的知識庫

大型語言模型的知識只侷限在特定時間點的公開資訊，如果想要建立能夠問答個人或企業私有或是特定領域資料的 AI 應用程式，可以採用額外提供資料擴展模型知識的作法，這種方式被稱為 RAG（Retrieval-Augmented Generation），也就是以檢索文件取得的資料擴展生成能力的作法。

文／旗標科技

理論上最簡單的作法是把整份文件都送給模型參考，不過文件可能過大，超過模型可處理的 token 數量，也因為模型本身不具備記憶，所以每次問答都必須重新附上整份文件內容，耗費大量處理費用。另外，讓模型自己在一整份文件中找尋相關內容猶如大海撈針，可能會得到不夠精確的結果。

資料切片量化

RAG 通常分為資料切片量化和檢索生成兩個部分，先是將資料切片成小量的片段，針對個別片段量化後儲存下來，最常見的量化方式就是透過內嵌（embedding）轉換成向量（vector）。檢索生成回覆時，會先從量化後的片段中依照相似度找出關聯性較高的片段提供給模型參考，再生成回答。整個流程如圖 1、圖 2：

資料切片量化：

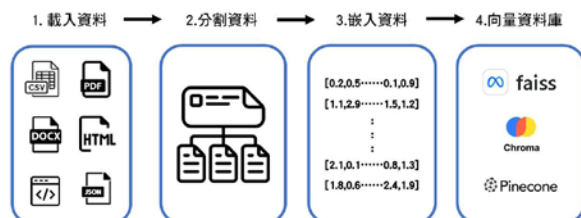


圖 1

1. 載入如 CSV、PDF、DOCX、HTML、JSON、Python 等文件
2. 將載入的資料分割成較小的內容
3. 將分割好的資料轉成向量
4. 將向量存入到向量資料庫，如：Meta 開發的 FAISS、Chroma、Pinecone 等

檢索生成：

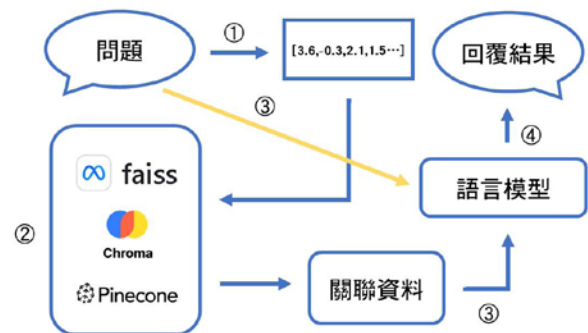


圖 2

1. 將問題轉成向量
2. 轉好的向量與資料庫中的向量比較相似度，找出最關聯的多筆資料
3. 將問題與關聯資料送給語言模型彙整
4. 得出結果並回復

文字轉向量

Embeddings（嵌入）是指將一段文字轉成向量嵌入到多維度的向量空間內，在多維度的向量空間中，可以把個別維度視為某種語意特徵，例如，一個維度可能代表快樂、悲傷等情緒特徵，而另一個維度可能代表老師、學生等角色特徵，所以轉成向量後就相當於分析了該段文字在語意上個別特徵的強度，而兩段不同的文字就可以用分析後的特徵分布判斷相關性。

LangChain 提供有搭配向量資料庫系統的類別可以直接儲存向量，並且可以使用它們的查詢功能檢索相關資料，一般推薦使用的向量資料庫系統是 **Chroma**，它是一個開源的向量資料庫系統，在 LangChain 中提供有 Chroma 類別可以搭配使用。

例如我們要準備考汽車駕照，但 AI 模型中並沒有台灣的法規，這時就可以利用 RAG 檢索汽車法規考題。只要將相關內容進行資料量化處理，並且儲存到向量資料庫，接著就來串接流程鏈把查詢到的關聯資料送給模型整理，模型就能回答出問題的答案。透過一問一答的方式，就能幫助我們學習汽車法規。

汽車法規是非題

題號	答案	題目	分類編號
001	○	尊重生命是駕駛道德最重要的一點，我們開車時要處處顧及行人，尤其應該注意讓老弱婦孺身心障礙者優先通行。	10
002	X	遵守交通法規與秩序，只算是優良駕駛人，與駕駛道德無關。	10
003	○	汽油著火時，應用滅火器、泥沙或用水浸濕棉被、衣服覆蓋撲滅。	07
004	○	禮讓與寬容是駕駛道德的最好表現。	10
005	○	在狹窄道路上會車，一定要互相禮讓。	10
006	○	綠燈允許你依序通過，但駕駛人仍應注意違規闖紅燈的人和車。	01

圖3：汽車法規題庫

RAG 與圖形資料庫

圖形資料庫專門設計用來處理高度連結的資料。這種資料庫適合用於關係資訊的應用場景，其中關係資訊與資料本身同樣重要。圖形資料庫之所以在這些領域表現出色，是因為它們能夠直觀地表達和查詢複雜的關係網路，反觀傳統的關聯式資料庫在處理這類高度連結的資料時則必須建構複雜難懂的查詢方式。以電影資料庫為例，傳統資料庫要先建立一個電影表格，然後再建立與電影表格有關係的個別表格，才能透過電影編號找到相關的資料。而圖形資料庫則是在建立資料庫的時候就建立好個別欄位之間的關係，例如圖 4 是圖形資料庫作好的關係圖，可以看到 2022 年上映的臺灣動作喜劇電影《關於我和鬼變成家人的那件事》中有 4 位演員主演：

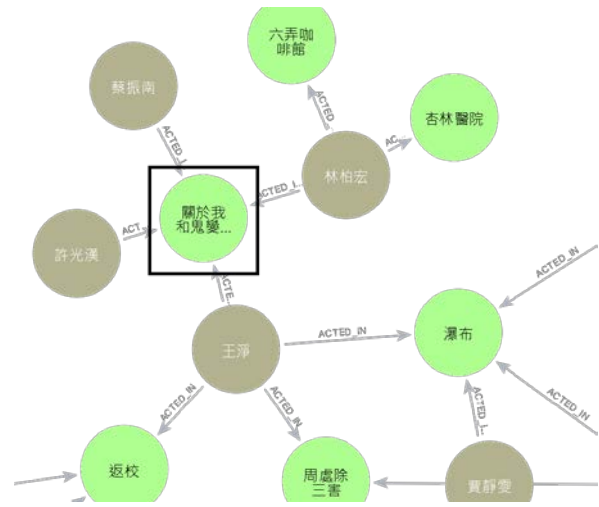


圖4：圖形資料庫會幫你建立關係圖

從圖中也可看到演員林柏宏有演出過《六弄咖啡館》和《杏林醫院》，而演員王淨有演過《返校》、《瀑布》和《周處除三害》，其他兩位演員許光漢和蔡振南在這個資料庫中沒有其他出演過的電影，從這邊就可以看出圖形資料庫幫我們釐清他們之間的關係。比起傳統資料庫，圖形資料庫可以更快速建立個別資料間的關係。

LangChain 可搭配各種圖形資料庫使用，最多人用的就是 Neo4j 圖形資料庫。Neo4j 除了圖形資料庫外也支援向量儲存，在 LangChain 中提供有 Neo4jVector 類別可以使用，所以我們可以在同一個資料庫中添加向量資料，這樣一來我們既可以使用圖形資料庫進行查詢，也可以利用向量相似度查詢關聯資料。

小結

更多關於 RAG 和圖形資料庫的使用方式，可以參考《LangChain開發手冊--OpenAI × LCEL 表達式 × Agent 自動化流程 × RAG 擴展模型知識 × 圖形資料庫 × LangSmith 除錯工具》一書，提供豐富的大型語言應用模型的開發技術與工具，可以協助企業打造專屬的問答機器人與其他生成式 AI 應用。

