PDF Reading Assistant

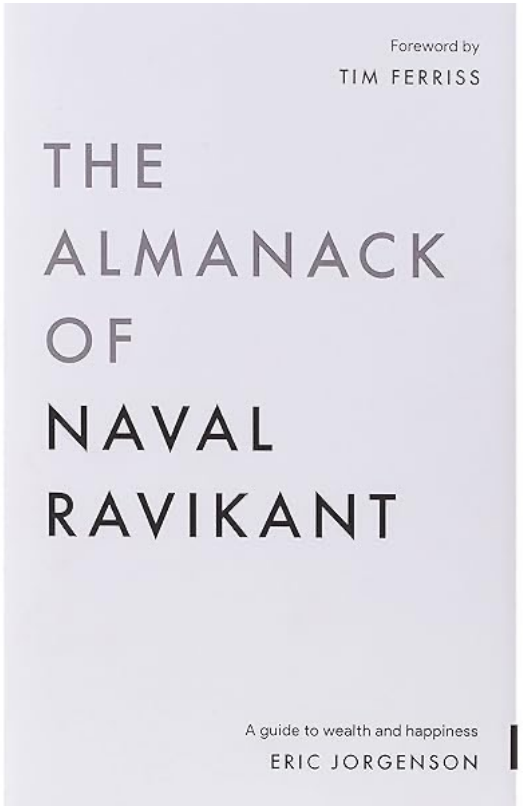
報告人：李盛廉

Background

- 專案目標：建立一個基於PDF的 AI 諮詢機器人，能夠以PDF內容價值相關的語氣和原則回答問題。
- PDF: 《The Almanack of Naval Ravikant》
- Tools:
  - LangChain (框架)
  - Pinecone (向量資料庫)
  - OpenAI (LLM & Embedding)
  - Cohere (Rerank 模型)。

Data

- The Almanack of Naval Ravikant.PDF
  - 由其推特金句和訪談精華而成
  - 涵蓋多個主題：
    - 創造財富
    - 快樂生活
    - 清晰思考
    - 建立長期思維...



Text Processing

- Read PDF
- 文本切分策略
  - RecursiveCharacterTextSplitter。
  - chunk_size=1000 and chunk_overlap=100。
  - 因為這本書很多章節是由推特金句和訪談精華組成且橫跨多個主題，因此使用較小的Chunk與ovelap。

Embedding and Storage

- Model: text-embedding-3-large
- 維度: 3072
- Uploaded vectors to Pinecone

檢索

- gpt-4o 進行Query Expansion
- Pinecone 查詢回傳各query TOP 15 相關結果
- 以Cohere Cross-Encoder重新依與query的相關性排序Pinecone回傳結果，並取前5個。

生成

- Model: gpt-4o
- 自訂prompt
  - 根據PDF回答
  - 風格控制
  - 引用要求
- 輸入：自訂prompt + 檢索回傳的相關資料