

Project #3 for the Biomedical Information Retrieval Course

Due: Oct 27, 2026

General Guideline

This homework is basically an individual homework. Each student has to do it all by himself (or herself). The final score will be evaluated from the system performance and individual demonstration.

Homework Overview

In this project, you are asking to implement and analyze “term weighting” technology for text documents in the vector space *using the* TF-IDF and/or *modified TF-IDF methods*.

實作內容與技術要求：

任務一：文字前處理與詞項統計

任務二：實作不同的 Term Weighting

任務三：建立簡易搜尋引擎，系統必須完成：

1. 將文件轉換為詞項權重向量
2. 將使用者查詢轉換為相同詞彙空間中的向量
3. 計算查詢與文件之間的 cosine similarity
4. 依照分數由高至低排序
5. 顯示排名前 10 名的文件及其分數

任務四：設計對照實驗

任務五：搜尋結果評估，請使用至少 10 個查詢進行評估，並為每個查詢建立相關文件的判斷依據。優先使用資料集既有的相關性標籤；若沒有，可根據文件主題標籤或人工判讀建立簡單的相關性集合，並說明其限制。

System Description

1. I suggest that you can test your system in advanced using certain disease name, e.g. “GLP-1”, which contains approximately 3000 documents. Each individual student builds its own system components using basic matching scheme from the IR course.
2. In the final evaluation, each individual should present system description, running results, and system demo to verify system performance.

學習目標：

1. **理解 TF 與 IDF 的意義：** 說明詞頻、文件頻率與逆文件頻率如何反映詞項的重要性
2. **實作不同的 weighting 方法：** 至少完成原始 TF、Log TF、IDF、TF-IDF 與文件向量正規化
3. **建立向量空間檢索系統：** 將文件及查詢轉換為向量，計算相似度並產生排名
4. **分析 weighting 的影響：** 透過控制變因實驗，觀察不同權重如何改變檢索結果
5. **評估搜尋效能：** 使用 Precision@k、Recall@k、MRR 或 nDCG 等指標評估檢索結果
6. **提出有根據的結論：** 根據實驗結果，解釋不同 weighting 方法的優缺點及適用情境

核心研究問題：

本專案至少需要回答以下問題：

RQ1： 原始 TF 與 Log TF 的差異，如何影響文件的詞項權重？

RQ2： 為什麼 IDF 能降低常見詞的重要性，並提高較具辨識力詞項的權重？

RQ3： 相較於只使用 TF 或只使用 IDF，TF-IDF 是否能產生較好的搜尋排名？

RQ4： 文件向量是否需要進行 L2 normalization？正規化前後的排名有何差異？

RQ5： 不同 weighting 方法對 Precision@k、MRR 或 nDCG 等指標有何影響？

最終學習重點：

本專案最重要的不是證明某一種 weighting 方法永遠最好，而是透過可重現的實驗，理解：

- 詞頻增加，不代表詞的重要性會等比例增加。
- 出現在許多文件中的詞，通常較缺乏區分文件的能力。
- 文件向量的正規化會影響相似度計算及排名。
- 不同 weighting 方法的優劣，必須結合查詢、資料集與評估指標判斷。