四資工四乙b10601002廖品捷 **資訊檢索及其應用 HW1**

使用語言：python

套件: numpy, collections, sklearn, math

這次實作分兩個部分：

第一個部分是資料預處理，將doc\_list和query\_list 及其內容存入陣列中。

接著將document的內容用label表示，好方便未來計算。第二個部分是計算TF, IDF，將結果拿來計算query和document的weight(TF-IDF)，再用sklearn計算cosine similarity。最後根據檔案與搜尋的相關程度進行排列並輸出。

結果：

一開始在運算cosine similarity時速度非常慢，要跑個五分鐘才跑得出結果。上網查詢後發現運用矩陣運算可以省去不少時間。Term frequency的部分，我是利用log normalization 可以獲得最好的成績(70以上)。IDF的部分，實作不同方法的結果差異不大。