Homework1: Vector Space Model

B10632026 吳苡瑄

一開始沒有真的搞懂 vector space 的時候，直接把所有字拿來算個 tf-idf 就算 cosine 了，結果一直跑出像是 0.016 之類會讓人笑出來的分數，然後還一直找不到原因 (後來知道問題就是智商)，後來就一直查網路上的資料，還一直覺得沒問題，直到看到這篇 簡單的範例，才發現我前面根本理解錯了，所有字的 tf-idf 才是一個 doc 的向量，才能算 doc 和 query 的夾角。

知道算法以後就得到好多了的分數了 (記得是 0.64 之類的)，就開始調一些小地方，主要是調 tf 的部分，試過直接用詞出現的次數、除以總次數、算 log 等等，最後是 log 的結果最好，還有 idf 就直接用 log(1 + max_df / word count in the doc) 來算了，因為也想不到更好的，所以就沒什麼再改了，其他部分是先把所有檔案讀進來先存成 dict 的形式，也把每個詞在各個 doc 中出現的次數也算了，為了方便算一個字出現在幾個文件中，存成了 word - doc - num 的形式，處理完這些文字以後，就去算每個字在各個 doc 中的 tf-idf，用 dict 存，但是要說的話其實是有點像個 4000 多 doc * 50000 多 words 的一個大矩陣的感覺，然後最後再去算 cosine，大概是這樣。

感謝給我提示的 Hank 學長和讓我頓悟的那篇手算 tf-idf 範例，不然我可能真的做不出 baseline，我試到達成 baseline 以後就沒有再繼續嘗試了，希望之後的作業可以順利一點，不然我看我數學腦應該是不夠用。