# Information Retrieval and Applications- HW5

M10915045 施信宏

這次實作的演算法是Rocchio，透過做3次VSM的結果測試為較高分，超過5次反而讓分數更低。同學有實作2次BM25成功，但我試的時候BM25第一次檢索分數約為0.49，第二次就會降至到0.2左右，debug了一段時間，決定使用講義上的VSM做看看。然而有上網參考使用套件就能容易的建出Rocchio，整個演算法扣除輸入及輸出的部分大約只需20行，設定在TfidfVectorizer中的max\_df = 0.95 ，min\_df = 5 在此設定下15萬個字只會剩下約4萬個較重要的字。得到tf-idf matrix 後 ，也能使用套件輕鬆的算出cosine\_similarity，在對此分數作運算，取前5篇及1篇最不相關的文章做Rocchio的query vector更新，

當中query vector能使用iteration不斷的更改。其中的超參數設定 alpha = 1 ，

Beta = 0.8 ， gemma = 0.15 ， iteration 次數則設為3 ，完成運算後上傳至kaggle在public leaderboard的分數將近0.54，也有測試只用原query向量及前面幾篇相關文章做更新，但這分數不比使用gemma減掉不相關的文章分數來的高。完成此演算法就沒有繼續實作老師後面所提的KL或RM了，想說先準備期末專案，沒想到後面又有人刷榜到0.66，這次作業分數沒有基本分的話是真的會很低…