R06944043 陳偉　網媒所碩一

(2%) Describe your VSM (e.g., parameter....)

按照Okapi BM25 給出的參考 parameter ，k1 = 1.2，k3 = 2，b = 0.75，rsv formula 選的是考慮 long query normalization 的版本，用 title 中的 unigram 找出 possible documents， 用 question + concepts (去掉'查'， '詢'和標點符號) 作為 query 的內容，對 possible documents 作rsv descending order 排序， 上傳 query 對documents rank 的前 100 篇。

(2%) Describe your Rocchio Relevance Feedback (e.g., how do you define relevant documents, parameter...)

Rocchio Relevance Feedback ， 取rank 過的前 500 個 document 作為 relevant documents， 第 500～1000 個 document 作為 irrelevant documents，alpha = 1.0, beta = 0.5, gamma = 0.5.

還實驗了BM25 的 Relevance feedback，這樣只用多紀錄relevant 和 irrelevant documents中query term 出現的次數，就可以稍微修改原來的 rsv formula 進行 feedback，而不用多加入 hyper parameter。取上一次 rank 過的前 500 個 document 作為 relevant documents， 第 500～1000 個 document 作為 irrelevant documents，parameter 和上面 Okapi BM25 一樣。

(3%) Results of Experiments

- **MAP value under different parameters of VSM**
  k1 = 1, k3 = 2.5, b = 0.75: Private Score, 0.75215 | Public Score, 0.75445
  k1 = 1, k3 = 2.5, b = 0.5: Private Score, 0.75496 | Public Score, 0.75430

- **Feedback vs. no Feedback**
  k1 = 1.2，k3 = 2，b = 0.75 的 parameters 下。
  feedback: Private Score, 0.77965 ｜ Public Score, 0.76703
  no feedback: Private Score, 0.74961 ｜ Public Score, 0.75462

- Other experiments you tried
  title + concept + narrative 作為query：Private Score, 0.76923 | Public Score, 0.75579
  不固定100個 documents，RSV 值大於 threshold 作為output(threshold的選取對結果有很大影響)：Private Score, 0.75703 ｜ Public Score, 0.74564

(1%) Discussion: what you learn in the homework.

作業過程中，對tf-idf這個 VSM 在理論和實際操作上又了更深的理解，了解了search engine背後大概的原理。同時發現了一些問題，建立字典和計算relevance的過程相當的耗費 memory 和時間，這種小型的搜索作業在搜索上的等待時間已不能容忍，對於像 google 這種 search engine 能夠同時服務這麼多人，深感其背後技術的強大，也加深了我探究其背後speed與accuracy的好奇。