Information Retrieval and Applications

HW02-Best Model 25

四資工四甲B10515047陳牧凡

此次作業運用到的MODEL是best model 25(BM25)，其是融合BM11以及BM15並透過可調整參數進行調整。BM系列跟VSM最大的差別主要就在於BM會去考慮long document penalty，在考慮docuent TF時會除以long document normalization以抵銷長的文章所帶來的一些問題，像是一篇100字文章和一篇1000字文章，他們符合query的字數同樣都是20個字，理論上100字的那篇相關性應該要比另一篇來的高，但是依照一般TF的算法無法去區分兩者的優先順序。

下圖是BM25的公式:

- To sum up, the BM25 model can be written as:

$$sim_{BM25}(d_j, q) \equiv \sum_{w_i \in \{d_j \cap q\}} \frac{(K_1 + 1) \times tf_{i,j}}{K_1 \left[(1 - b) + b \times \frac{len(d_j)}{avg_{doclen}}\right] + tf_{i,j}} \times \frac{(K_3 + 1) \times tf_{i,q}}{K_3 + tf_{i,q}} \times log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

Tunable Parameters
Document Length Normalization
Term Frequency
Inverse Document Frequency

簡單來說就是: $TF(doc) * TF(query) * IDF$ ， 而 $TF(doc)$ 和 $TF(query)$ 透過K1、K3、b三個參數進行調整， 且$TF(doc)$ 會去考慮long document penalty。

另外在我自己嘗試過程中，發現在計算IDF時，若 $\frac{N-ni+0.5}{ni+0.5}$ 太小的話會造成結果會出現負數(log(X)若X<10則log(X)<0)，這樣在最後計算BM25 similarity時會不準確，因此我把IDF公式改成: $log(1 + \frac{N-ni+0.5}{ni+0.5})$

最後最好的正確率是:58.2%
參數調整: k1 = 1.5, k3 = 5, b = 1