IR 新聞立場檢索技術獎金賽

R07922003 劉濬慶

此次所使用到的參數共有 3 個

1. tf
2. idf
3. query_tf

以下為跑參數的結果:

1. Score: 0.1595371

   document_scores[doc] = query_tf * idf * doc_tf

2. Score: **0.1311584**

   document_scores[doc] = query_tf * np.log(idf) *np.log(doc_tf)

   有取 log 比較差一點

3. Score: **0.1761016**

   TF(t,d) = (k+1) * count(t,d) / (count(t,d)+k)

   k=0.8

   此次改 doc_tf 公式

4. Scroe: **0.1820000**

   TF = (BM25_k + 1)*CountInDoc / (CountInDoc + BM25_k*(1 - BM25_b +

   BM25_b* DocLen / Avg_Doc_Len))

k=0.8

b=0.2

此次改 doc_tf 公式以及計算平均文章的長度，利用 inverted file 計算

5. Score: **0.2083505**

   TF = np.log(1 + np.log(1+CountInDoc)) / (1 - BM25_b + BM25_b* DocLen /Avg_Doc_Len)

   k=0.7

   b=0.3

   此次改 doc_tf 公式以及計算平均文章的長度

6. Score: **0.2107674**

   TF = np.log(1 + np.log(1+CountInDoc)) / (1 - BM25_b + BM25_b* DocLen /Avg_Doc_Len)

   k=0.7

   b=0.3

   新增一個條件:

   若此 word 在所有文章中出現數量 >= 100 才計算進入 Score 中

Conclusion:

這次的參數試了很多次，也試了很多 TF 的變形公式，覺得很有趣，尤其加入了限制 word 出現數量才計算效果變好有點驚訝!