B10615023 楊傑安

NTUST_Information_Retrieval_and_Application_HW2

使用參數

- K1 = 2.75
- K2 = 25
- K3 = 0.855

使用工具

- 語言: Python 3.8.5
- package:
  - tqdm
  - math
  - multiprocessing
  - sys

設計架構

1. 先將 50 條 query 所含的 127 個字建立成 vocabulary
2. 使用 1. 得到的 vocabulary 算出 query 和 document 的 term frequency
3. 使用 1. 得到的 vocabulary 算出 document 的 document frequency
4. 計算 avg_doclen
5. 使用 2. 所得計算 idf
6. 對其中一條 query, 利用 2. 5. 所得
   a. 對於每一 doc 計算

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

   b. 將 documents 依照 a. 所得倒序排列, 即得到此 query 的 revelant documents
7. 重複 6. 即得所有 query 的 revelant documents

困難與心得

這次的作業只花了我 2 個小時寫完, 卻花了 1 個禮拜來 debug, 還上傳了將近 70 次. 中間還發現很多地方括號沒有包好, 該連乘的地方寫成連加. 最後連續四天沒有進展幾乎想要重寫的時候才有人告訴我: 這次的 tf 不除上 doc 的長度, 才在 deadline 的前一天踩到baseline.