Q56084048 蔡沛蓁

IIR COURSE

Homework4

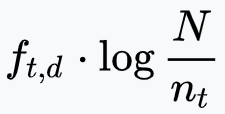
使用工具：前端(Html+Js)

後端(Python)

1. 資料蒐集：

(關鍵字**：Neoplasms/Alzheimer Disease/Influenza, Human)**

* 1. **Mesh--**使用Mesh醫學主題詞，把字詞最新版關係找出來，並從中去搜索出**pubmed**文章。
  2. **Xml—**使用**pubmed** ，隨機撈出三個**Mesh關鍵字**的資料，約100~200筆，儲存在mysql資料庫。

1. 本次目標：
2. BSBI : 先從摘要裡面的字，建立每個字的對應每篇文章id的索引表（排除停用詞和標點符號）將data拆成10個一組（最後不足10的自己計算）去找出關鍵字對應的id，最後合併起來。
3. SPIMI :關鍵字直接對獨立的每篇內容去搜尋，最後把結果合併起來。
4. 以常用的 TF-IDF : 為文件排序依據，並算出某篇文章對整體的TF-IDF，以及某句子對某篇文的TF-IDF。
5. 以cosine-similarity算出文件間的相似度

四、其他：

我先使用**Mesh**關鍵字來搜尋**pubmed**文章，發現像cancer、tumor…等字都會對應到Neoplasms，因此cancer、tumor…等字都會被視為同意字下去做計算。

在算TF-IDF前，先把每個字出現的頻率算出來外，發現像標點符號和the.and…字詞可能會影響結果。因此先把標點符號和停用詞濾掉，再做處理。

BSBI和SPIMI看起來就覺得SPIMI**處理速度**會比較快，因為他不用再去建id索引。不過索引建好了只比較**搜尋速度**的快慢，**建出id索引**的應該會是比較快的。但因早期電腦設備不發達，沒辦法一次將資料全部直接比對，故**拆成一小塊**一小塊去計算，**最後在合併**。現在由於設備進步，基本上就直接使用關鍵是直接比對文章是沒有問題的。