Q56084048 蔡沛蓁

使用工具：前端(Html+Js)

後端(Python)

一、 資料蒐集：(關鍵字：influenza/cancer)

Xml —使用 pubmed ，隨機撈出influenza 與 cancer 資料各100~200 筆。

二、 本次目標：

1. 算三種 TF-IDF：(1) $f_{t,d} \cdot \log \frac{N}{n_t}$ (2) $1 + \log f_{t,d}$

(3) $(1 + \log f_{t,d}) \cdot \log \frac{N}{n_t}$

2. 以 cosine-similarity 算出文件間的相似度
3. 以兩字搜尋的方式，找出兩字的位置

三、 呈現：

1. 以第一種 TF-IDF 為文件排列順序，並算出某篇文章對整個資料集的 TF-IDF，以及某句子對某篇文的 TF-IDF
2. 點開文件呈現出，本篇文章與哪篇最相似
3. 兩字搜尋已塗色標記呈現

四、其他：

文件隨機撈：由於我覺得同一時期研究的東西可能太相似，可能會產生誤判，因此為跳著撈資料。EX 某時期可能肺癌研究特別多，若我又指撈那時期的資料，可能對肺和癌症有高相似。

在算 TF-IDF 前，除了有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。

TF-IDF (某篇文章對整個資料集)：對每篇文章算出三種 TF-IDF。

TF-IDF (某句子某文章)：對資料集的每篇文章，對其句子，和其文章做三種 TF-IDF。

相似度：發現相關的文章並不一定會互相對應。EX 第一篇文章：最相關文章為第四篇，但第四篇文章最相關的不一定是第一篇。