Q56084048 蔡沛蓁

IIR COURSE

Homework1

使用工具：前端(Html+Js)

後端(Python)

1. 程式流程：

找出相關文獻標題及內文

並計算字元、字串、字串頻率、句子，與標記關鍵字

搜尋關鍵字

上傳JSON檔

上傳 XML 檔

前端部分採用JSON和XML分開方式上傳，上傳後進入同一個關鍵字搜索，判斷檔名(JSON或XML)並對他們分別做資料處理，最後相關內容回覆在網頁上。

1. 資料處理階段：

拆解：**XML--**使用**ElementTree**對來自**pubmed**資料的ID、Article、Abstract做拆解；**JSON--**使用**JSON**對**Twitter**資料的ID、Text做拆解。

文件分析：**Pubmed(XML格式)**，撈下來的資料有些有缺摘要的標籤**--只要有標題都做保留**；**Twitter(JSON格式)**，撈下來的資料有缺陷(JSON間缺少逗號、有些是刪除的資料、文件內容語言問題…等) **--把刪除的資料及非英文類剔除，最後把逗號補齊。**

1. 相關文獻顯示：

**字元、字串、字串頻率**：使用split、dict、len算出

**句子：對文件做適合他的判斷方式，Pubmed(XML格式)—**書寫**拘謹**，因此對其做**句號、問號、驚嘆號等拆解**；而**Twitter(JSON格式**)為聊天式，句子較**不拘謹**，因此用**空行方式、TAG模式做判斷。**

**標記關鍵字**：使用**replace**更改關鍵字的**CSS**

1. 問題討論：

句子判斷仍不太準確，還需加強

理想能做到不跳頁的顯示方式呈現