P76084091 資工所 李宗樺

系統環境：

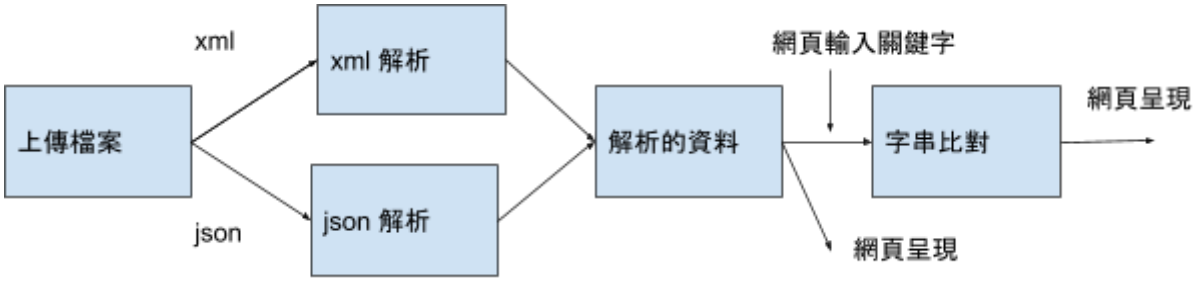
作業系統：fedora 29

主程式：Django

實做題目：

在pubmed 和 twitter 資料上實做全文搜索，並做基本字數和句數分析。

實做方法：



1.上傳檔案並執行基本分析：

自動分析上傳文件選擇對應的處理方式。

xmlParser.py：利用elementtree做xml解析，取得文章標題和大綱，利用正則表達式計算字數和句子數。

jsonParser.py：利用python json套件做解析，取得使用者和文章內容，利用正則表達式計算字數和句子數。

字數正則表達式：(w+)

句數正則表達式：(?<=[^A-Z].[.?]) +(?=[A-Z])

2.利用輸入的關鍵字進行全文比對：

full_text_match.py：進行全文比對並記下每一個匹配的位置，做為html上色用。匹配時將所有字元轉為小寫。

如果在title的地方有匹配，給一個給予權重100，在content的地方有匹配，給一個給予權重10，最後根據分數的高低排序。

問題與討論：

1. 使用element tree在解析較大檔案的時候，效能下降的非常嚴重，導致程式會卡在parsing 的地方。這部份需要再研究parsing xml 的方法。
2. 目前pubmed 和 twitter資料解析的欄位差不多，所以在同一個網頁呈現，之後會因應解析的資料決定是否要分開成現在不同網頁。
3. 還有很多基本的字串演算法和檔案處理眼算法尚未加入系統中。
4. 使用正規表達式去計算句子數量只能應付簡單狀況。