Project 1 for the Biomedical Information Retrieval Course

資工所 P78101514 黃仁鴻

https://github.com/toonnyy8/ncku/tree/bir2021/hw1

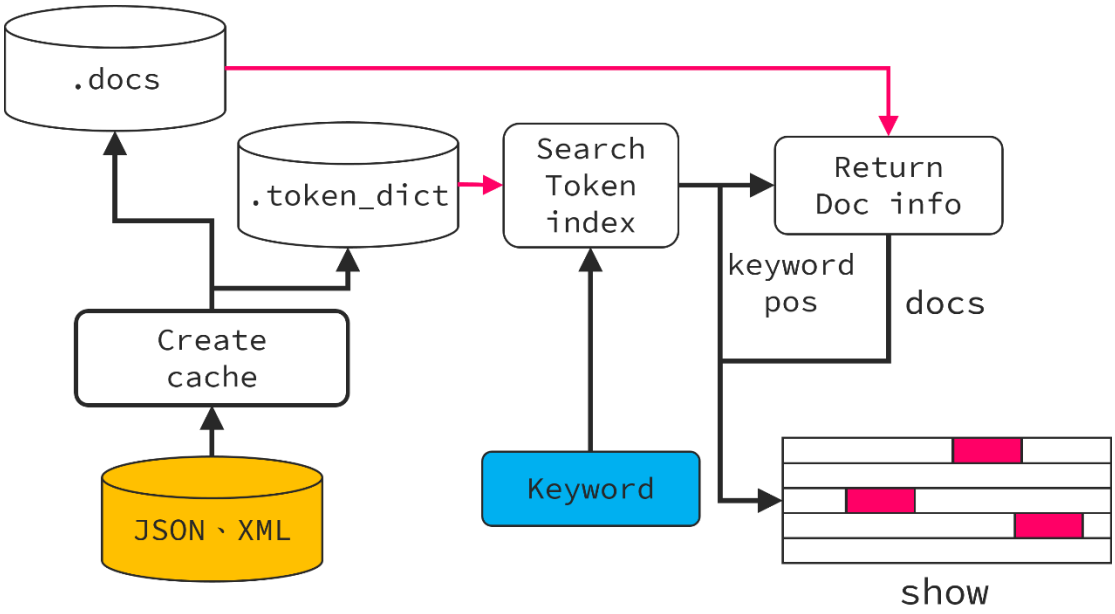

Figure 1 系統架構圖

系統環境

程式語言		JavaScript
開發環境		Node.js 16.10.0、esbuild 0.13.3
函式庫	前後端	Vue 3.2.19、Koa 2.13.3、Koa-router 10.1.1
	文本切割	Compromise 13.11.4、Compromise-sentences 0.3.0
	xml 解析	xml2js 0.4.23

文本分析

文檔的 word、sentence 數量資訊目前是由 compromise.js 這個函式庫輔助所得出的，其中 word 會去除標點符號並拆分像是 don't, isn't 等縮寫詞後才進行數量統計。

文本前處理

系統架構由 server 與 client 兩區塊所組成。在 server 啟動之後會檢查是否存在 .doc 與 .token_dict 這兩個檔案，若不存在的話會讀取資料夾內的 JSON 與 XML 檔，並使用各文檔的「標題」與「內文」組成 .doc，接著對 .doc 的 word 進行分析後建構文字（自動轉為小寫）的索引檔 .token_dict。

關鍵字搜尋

在查找關鍵字時會利用是先建立的 .token_dict 來查詢關鍵字出現的文檔 ID、段落與字符數，並從 .doc 中取得對應的文檔後，將這些資訊一同回傳給 client。