Project 2 for the Biomedical Information Retrieval Course

資工所 P78101514 黃仁鴻

https://github.com/toonnyy8/ncku/tree/bir2021/hw2

系統環境

程式語言		JavaScript
開發環境		Node.js 16.10.0、esbuild 0.13.3、ts-node 10.4.0
函式庫	前後端	Vue 3.2.19、Koa 2.13.3、Koa-router 10.1.1
	文本處理	Compromise 13.11.4、porter-stemmer 0.9.1
	xml 解析	xml2js 0.4.23

文本前處理

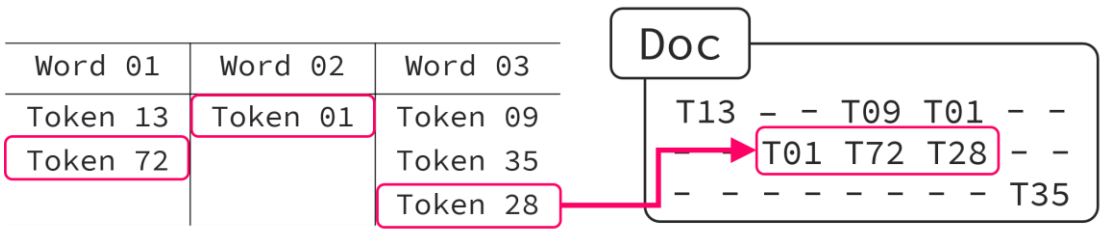
使用 compromise.js 的功能將文檔的標點符號移除以及拆分像是 don't, isn't 等縮寫詞，再將所有字元轉成小寫並依空白字元切割出 token 後，配合每個 token 在文檔出現的位置建構成 token table。建立 token table 時會分成有經過 porter stemmer 處理與沒有處理兩份 table。

Token Table 類型

```
interface TokenTable {
  [token: string]: {
    [didx: `${number}`]: {
      title?: Set<number>;
      [aidx: `${number}`]: Set<number>;
    };
  };
}
```

關鍵字組搜尋

在查找關鍵字時會先對其執行與 token table 相同的前處理，接著使用 Levenshtein Distance 從 table 中找出相似的 token 與這些 token 在文檔的位置。再來將關鍵字組替換為近似 token 後跟文檔進行比對，找尋其中是否存在 Levenshtein Distance 小於閾值的子序列，若存在就將此文檔回傳。



文檔中存在 Levenshtein Distance 低於閾值的子序列