CE6155 Homework1 Report

student ID：112522087　　name：黃懷萱

Can answer in Chinese or English(可以使用中文或英文回答)

1. (15%) Please record the performance statistics of the web crawler, for example: how long does it take to crawl 100 web pages, how long does it take to crawl 500 web pages? Then, how do you speed up the process, and what is the resulting performance after speeding up?

在進行網頁爬蟲的過程中，我使用了兩種不同的方法來搜尋網站：動態查詢（利用 Selenium）和靜態查詢（利用 BeautifulSoup）。最初，當我嘗試搜尋「數學系」相關的網站時，我注意到僅使用靜態查詢會遇到一個問題：網頁內容因為沒有等待 JavaScript 渲染完成而導致不完整，這使得網頁看起來內容一致。為了解決這個問題，我起初選擇了動態查詢的方法。

隨後，當我將搜索深度調整為 2 時，我發現許多數學系的網站存在著 404 錯誤，顯示未維護。這促使我轉向另一個搜索目標：「學習與教學研究所」。在這一階段，我發現目標網站是靜態網站，且內容已經完成渲染。因此，我決定改用 BeautifulSoup 進行查詢，這大大提高了查詢速度。

另外在原本的程式碼中沒有設定 error_list 這導致錯誤的網址也會重複查詢，也因此我多補了一個 error_list 來嘗試提速。

2. (15%) Based on the URLs you have crawled, analyze the structure and layout of the website, and draw an approximate sitemap of the website.

我最後使用的深度搜尋為 2，共有 159 個搜尋結果，且有 2367 延伸節點。很可惜的是因為延伸 Links 的數量太多了，因此會導致畫出來的 sitemap 節點圖會非常的密集，無法看清楚。

圖 1 Sitemap v1

另外我嘗試參考 Github 中的 sitemap-visualization-tool 的方法，這方法主要是嘗試透過 Google 的預設方法使用/sitemap.xml 來獲取網站的 XML 資料時，但當網站沒有提供 sitemap.xml 檔案時，這種方法就無法使用。所以我試圖繪製「國立中央大學學習與教學研究所」網站時就遇到這樣的情況，該網站沒有提供 sitemap.xml，我無法直接獲取所需的 XML 資料。

為了解決無法順利繪製 XML 的這個問題，我又額外參考 GitHub 中的 python-sitemap，這個套件可以透過 domain 去爬取生成相關的 XML 檔案，但是他不提供視覺化的部分，所以我同時結合上述方法實現目標。

在成功結合這兩種方法後，我終於能夠繪製出網站的結構圖。不過，由於最終生成的圖片尺寸過大，這裡就不直接展示了。參考附檔「sitemap_graph_3_layer.pdf」，可惜的是由於「國立中央大學學習與教學研究所」

的網站在設計上沒有考慮到提供 sitemap.xml 的支持，從該網站擷取出來的資料字串非常長，這對於繪製過程和結果的整潔度造成了一定的影響。

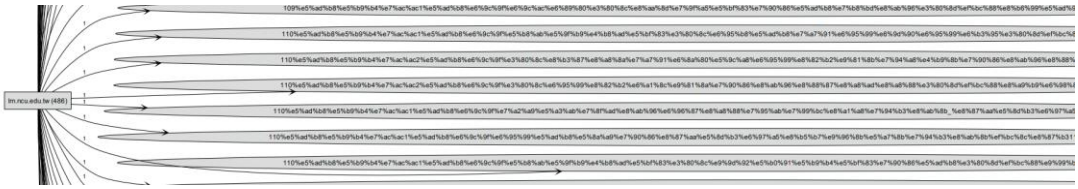

圖 2 Sitemap v2 by “sitemap-visualization-tool”

3. (20%) Please compare the differences in query results using different query methods. Please attempt to use embedding models for vector search.

我設立了一個假設空間來進行搜尋。假設我需要找某位教授的資料，在這裡我預設使用「張立杰教授」作為關鍵字，希望可以直接找到該教授的個人介紹頁面，比如「http://lrn.ncu.edu.tw/lc-chang/」。

嘗試了兩種方法來改善搜尋效果。基本查詢和向量搜尋。第一種方法我使用了 Elasticsearch 的字符串匹配功能，這只是一個基本查詢方式。這種方法直接對字符串進行比對，簡單但有時可能不太準確。第二種方法，我將 Elasticsearch 中的數據轉換成 Text Embedding 並存在一個新欄位 vector 中，這個欄位的類型是設定為 dense_vector。Text Embedding 的維度必須根據 Embedding Model 的輸出大小來確定，如果維度設定錯誤，會導致程式碼報錯。在進行搜尋時，我會將查詢語句 encoding 成 vector，然後用它來與資料庫中的向量欄位進行匹配，這樣可以進行基於向量的搜尋。

```
hp_mapping = {
    "properties": {
        "url": {
            "type": "text"
        },
        "title": {
            "type": "text"
        },
        "depth": {
            "type": "integer"
        },
        "content": {
            "type": "text"
        },
        "links": {
            "type": "text"
        },
        "vector": {
            "type": "dense_vector",
            "dims": 768,
        }
    }
}
```

```
搜尋語句: 張立杰教授 搜尋欄位: content
未使用 Embedding 搜尋結果:
URL: http://lrn.ncu.edu.tw/research_areas/
URL: http://lrn.ncu.edu.tw/?p=1761
@ 使用 Embedding 搜尋結果:
URL: http://lrn.ncu.edu.tw/lc-chang/
URL: http://lrn.ncu.edu.tw/zh/lc-chang/
```

圖 3 Elasticsearch 根據 Embedding Model 所需的欄位設定(左)

圖 4 分別使用一般搜尋和 Embedding Model 的向量搜尋差異，使用 Embedding Model 的結果剛好符合我的假設空間(右)

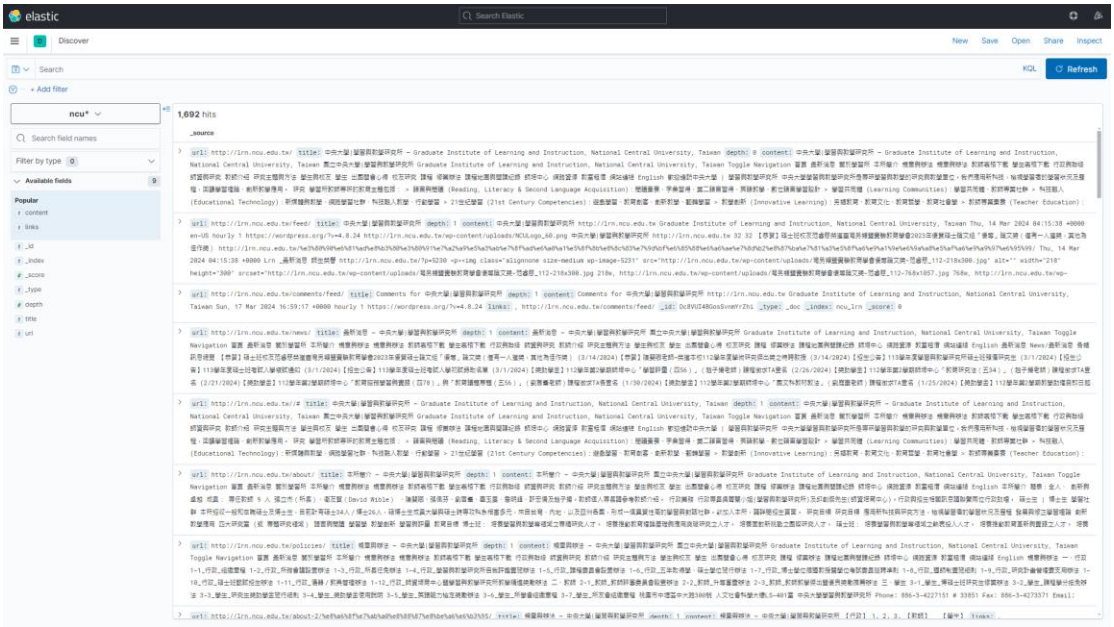

圖 5 Kibana 的搜尋結果