Data Structure:Beat Google

補YerMoore補習班搜尋系統

Final Project Proposal

Group 3:資管二 112306043卓珙奇 112306007朱廷翊
112306060劉冠昕 112306027廖偉翔 112306057鄭宇彤

一、主題與動機

主題:補**YerMoore:**補習班搜尋與評分系統

在當前的教育環境中，選擇合適的課後補習資源對於學生的學業發展和未來生涯發展具有深遠的影響。然而，選擇補習班的過程往往繁瑣且時間消耗，尤其是對於家長和學生來說，需從大量的補習班資訊中篩選出符合自己需求的機構，但不精確的資料和搜尋結果也讓尋找的過程十分煎熬。因此，開發一個可以有效搜尋補習班資訊並根據特定標準（如學科、城市、其他特別要求等）進行評分的系統，能夠顯著提升家長與學生對補習班選擇的效率和準確性。

我們建立一個基於網絡爬蟲技術的學校數據搜尋系統，能夠自動爬取並整合來自不同來源的學校數據，並根據不同指標（如補習班的學科、地點、權重關鍵字出現次數等）計算分數，幫助用戶做出更加明智的學校選擇。此系統可根據用戶需求篩選補習班，並提供具體的評分結果，進一步幫助學生和家長更快找到最適合的補習班。

主題更改:家教搜尋→補習班搜尋

因關於家教的大量資訊處於FB家教社團中，無法爬取，光靠爬取文字也無法完全正確的辨別學歷、以圖片專業證書真偽，提供正確資訊給使用者，因此本組決定將主題從家教搜尋系統更改為尋找有立案保證，如果出事有人找。也有一定公信力並已接受公評的補習班。

二、搜尋和評分方式

搜尋過程:

1. 使用關鍵字構建搜尋條件:

   系統會根據用戶提供的city(所在城市)、subject (欲補習科目)和 other (其他要求)字段來組合搜尋條件，並使用這些條件進行 Google 搜尋。搜尋結果會返回以該條件進行權重分數計算後排列的補習班列表。

2. 發送搜尋請求:

   系統構建好搜尋 URL(如"https://www.google.com/search?q=city subject other 補習班")後，使用 Jsoup 發送 HTTP 請求並解析返回的 HTML 頁面。

3. 解析搜尋結果:

   在搜尋結果中，系統解析出 h3 標籤的內容，這些標籤通常包含搜索結果的標題，並且抓取該標題的連結。

4. 生成補習班對象並評分:

   每個搜索結果會被轉換為 School 物件，並存儲其中，包括學校名稱、城市、科目、URL 和特殊選擇標記，再根據後端四項評分標準計算各網頁及其子頁面加權分數。

5. 返回結果:

   系統將排列後結果返回給前端，依分數由高至低進行展示。

評分過程:

1. 設定關鍵字及權重:

   系統設定了一組關鍵字(如短期衝刺、升學、學測、視聽等),每個關鍵字有不同的權重,用於計算補習班的評分。

2. 檢查城市和科目:

   如果補習班的 city 和 subject 與用戶選擇的匹配,則為補習班加上額外的分數。具體來說:

   - 城市匹配:如果補習班的城市與用戶提供的城市匹配,學校會獲得 35 分。
   - 科目匹配:如果補習班的科目與用戶提供的科目匹配,學校會獲得 35 分。

3. 計算關鍵字匹配分數:

   用戶在 other 欄位提供的關鍵字會與學校的網頁內容進行匹配,系統會計算關鍵字在補習班中出現的次數,並根據出現次數為補習班加分。

4. 加權關鍵字匹配:

   系統還根據預設的關鍵字(例如,名師,學測,個別輔導,數學,專業證照 等)在補習班網站內容中的出現次數進行額外加分。每個關鍵字的權重會影響分數的計算。

   - 例如,如果補習班網站中提到學測,且這個關鍵字的權重是 3,那麼就會為學校加上 count * weight 的分數。

5. 計算最終評分:

   綜合考慮以上因素,最終得出每個補習班的評分。評分是根據匹配條件和關鍵字出現的頻次計算的,分數越高,學校的評價就愈高,同時高於100分的搜尋結果會在前端特別標示成特特佳選擇,供使用者快速參考。

三、系統架構設計

(一)前後端敘述

- 前端界面:提供使用者設定其所在地區、需求科目及其他需求條件。需要顯示家教的本資訊及其對應的分數，並提供排序、過濾等功能。
- 後端服務:處理補習班數據的儲存與管理，根據使用者選擇的條件計算補習班的評分，並將結果傳送至前端進行顯示。

(二)Class設計圖

1.前端

1. **search.html**
    - 搜尋頁面模板，使用 HTML 與 CSS 設計輸入表單，包含選擇地區、科目及自訂條件的欄位。
    - 支援 RWD，確保不同裝置上的使用體驗一致。
    - 將地區和科目使用下拉式選單顯示，供使用者方便且快速選擇，並設有“其他”選項供使用者輸入非預設選項。
    - 將其他條件設為輸入區，使用者可自行輸入所需關鍵字，並用,隔開(如 學測,台北車站,高中,台大,……)，系統就會將以上關鍵字列入權重計算。
2. **searchResults.html**
    - 搜尋結果頁面模板，動態渲染補習班清單，顯示學校名稱、網址及分數。
    - 特佳選擇用醒目文字標記顯示。
3. **styles.css**
    - 提供統一的樣式支持，設計簡潔且易於閱讀的視覺介面。
    - 包含陰影、背景顏色及過渡效果以增強使用者體驗。

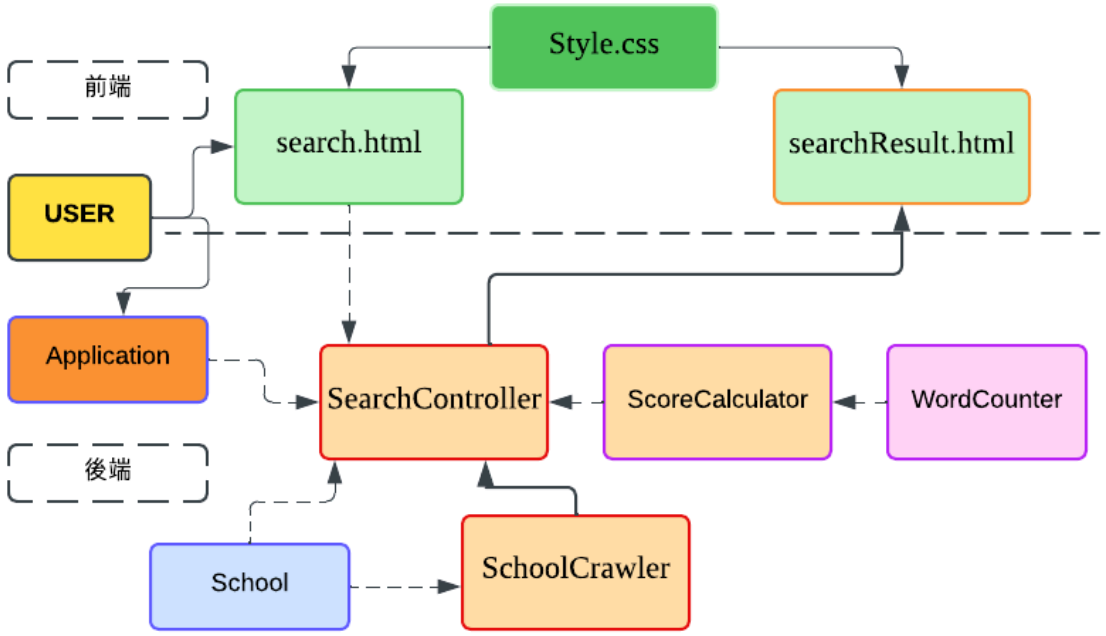
2.後端

1. **Application.java**
    - Spring Boot 啟動類，初始化並使整個應用程式開始運行。
2. **SearchController**
    - 負責處理搜尋請求，根據使用者條件(地區、科目等)調用爬蟲及分數計算邏輯。

- 連結前後端，提供 /search 和 /search/results 兩個路由。(/search: 處理搜尋介面 search.html，/search/results: 接收使用者輸入的地區、科目及其他條件，將搜尋結果傳遞至 searchResults.html 顯示)
- 使用 Trie 資料結構高效儲存和檢索搜尋結果。
- 過濾不必要的網站資料（如含廣告或不相關內容的網站）。
- 進行子頁面爬取及算分，因大部分補習班網頁子頁面深度不深故只爬取兩層，除子頁面城市和科目外之分數加到補習班總分。
- 標註總分高於100的補習班為特佳選擇(specialChoice=true)。

3. **SchoolCrawler**
   - 基於 Jsoup 爬取補習班相關網站資訊，將結果整理為 School 類型列表。
   - 用city+subject+other補習班進行搜尋，再以Google 搜尋之結果，爬取網站標題和連結。
4. **ScoreCalculator**
   - 根據使用者輸入條件及預設關鍵字計算每個補習班的分數。
   - 權重分數依據是否符合所選地區和科目，以及關鍵字出現次數和匹配程度計算。
   - 用node定義使用者輸入的自定義關鍵字，依出現次數增加額外分數。
5. **School**
   - 資料類，描述補習班的基本屬性，包括名稱、地區、科目、網址、分數和是否為特加選擇。
6. **WordCounter**
   - 用於計算目標網站內關鍵字出現次數，使用Boyer-Moore 字符串匹配算法。

Class Diagram



SearchController

Fields:

- - schoolCrawler
- - scoreCalculator
- - schoolTrie

Methods:

- + searchForm
- + searchResults

Properties:

- **功能**：處理用戶的搜尋請求，將結果傳回給前端。

SchoolCrawler

Methods:

- + fetchSchools

Properties:

- 負責從網頁上爬取學校資料。

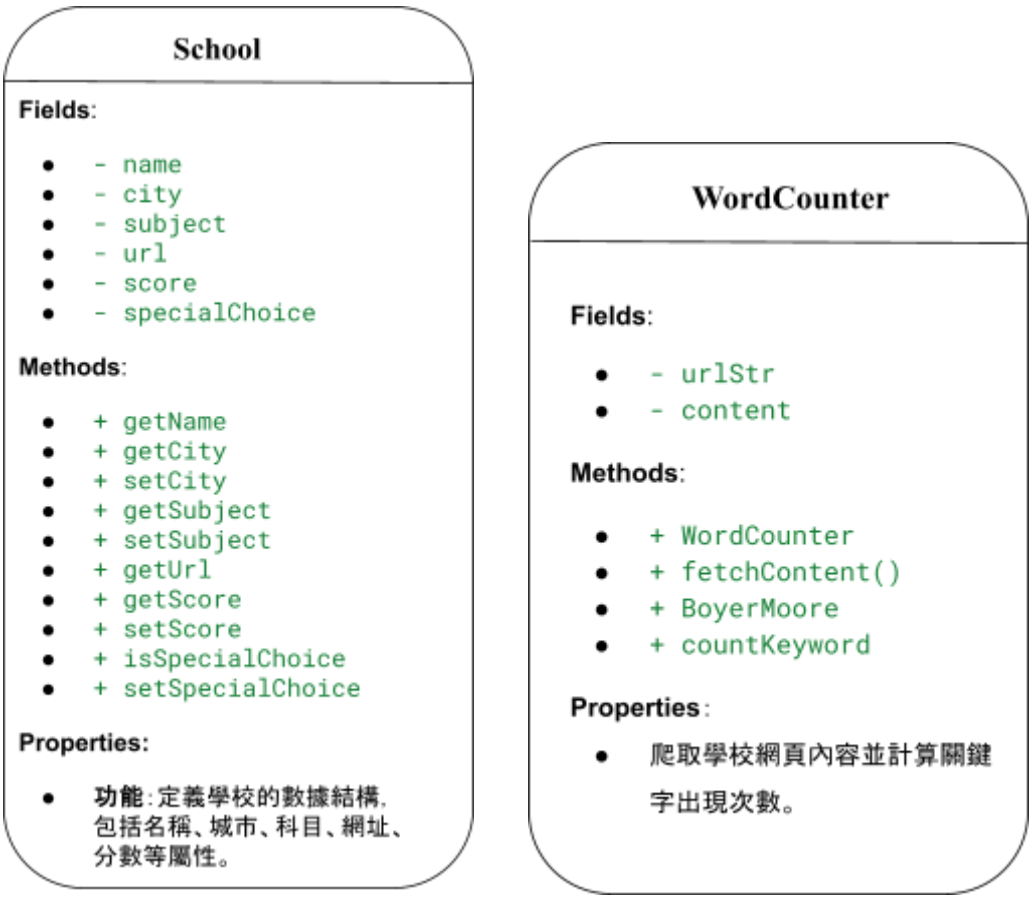
ScoreCalculator

Methods:

- + calculateScores

Properties:

- 根據特定關鍵字和條件計算學校分數。



四、專案目標與系統特點

1. 以演算法優化搜尋品質
   本專案的主要目標之一是確保搜尋與排序的效率。在處理大量補習班數據時，如何平衡運算速度與精度是一大挑戰。為此，我們使用@Map,tries(字典樹)、nodes、PriorityQueue、HashSet處理資訊並以BoyerMoore進行較高效的關鍵字比對。後續將持續測試並優化演算法以應對更大規模的數據。
2. 資料準確性
   爬蟲程式的設計需要確保從多個網站及其子頁面抓取的數據，故需要設定多項有參考價值的關鍵字及合理權重，以過濾無效或重複數據。同時，透過清理有不必要的關鍵字的網頁(如人力銀行、徵才、業配等網站)，進一步提升結果的可靠性。

3. 使用者體驗
   使用者界面的設計以簡單直觀為目標，採用下拉式選單與額外輸入欄位，讓使用者快速設定需求，並設定輸入提示讓使用者參考。搜尋結果頁面則通過分數排序與特佳選擇標記，幫助使用者快速找到合適的補習班。也在搜尋過程或網頁無法爬取時也有設定錯誤處理功能，跳過無法被爬取的網頁，並在搜尋結果不理想時給出建議標語。

五、遭遇挑戰與科技困難

1. 演算法效率問題
   系統需要處理大量數據並支持多重條件篩選。目前採用 Trie 儲存結構和加權排序演算法，但在處理更高複雜度的條件時，運算效率還有待加強，時常出現運作太久的問題。未來如要進一步發展，可能需引入更有效率的演算法或進一步改良現有排序演算法的技術。
2. 資料準確性問題
   補習班網站中可能存在以出現多次"名師"、"頂大”來洗版提升被搜尋到次數的案例，影響爬取、計分和排序結果。這個問題將影響結果的可靠性，因此需透過多次測試才能取得各項因素之配分平衡與關鍵字和其權重設定。
3. 特殊需求處理的靈活性
   為滿足使用者具體需求，系統提供了靈活且客製化的搜尋條件。然而，條件過於細化可能導致結果稀少。為此，我們加入提示字樣，引導使用者調整條件以獲得更多相關結果，並在搜尋不出結果時會顯示引導標示。

六、分工表

組員	負責工作
卓珙奇	主題發想、程式、簡報製作、系統展示、Word編輯
朱廷翊	主題發想及更改、程式、簡報製作、系統展示
鄭宇彤	Proposal、Word編輯:主題與動機、搜尋和評分方式 、系統架構設計
劉冠昕	Proposal、Word編輯:class diagram
廖偉翔	Proposal、Word編輯:任務和系統特點、遭遇挑戰與科技困難