關鍵字搜索

專題編號：105 -CSIE -S022
執行期限：104 年第 1 學期至 105 年第 1 學期
指導教授：王正豪
專題參與人員：102590014 蘇育銘
102590031 簡屏軒
102590048 蔡中維

一、摘要

本專題為財經搜尋引擎與文本分析，以網站的方式呈現。內容擁有關鍵字尋找文章、TagCloud、圖表分析等功能。我們會用爬蟲連結到網頁擷取資料並儲存至資料庫，然後前端部分會用 Solr 將資料庫資料索引出來呈現於網頁介面，再運用文字視覺化讓使用者能更簡單輕鬆的了解文章內容，使用方法很簡單，使用者只要在搜尋欄內輸入想要查找的關鍵字，那就可以找出相關的財經文章、關鍵字排行…等等，可以自由選擇想要的功能。

二、緣由與目的

現代的資訊傳播快速，而在這麼資料爆炸的時代中「搜尋引擎」就顯得格外的重要，讓使用者能快速、方便並知道現今趨勢，而財經資訊更是注重分析與視覺化，所以我們以此目標來建構出我們的專題。

三、研究報告內容

(一) 使用技術方法

Python: 爬蟲框架 Scrapy，HTML, XML 來源資料 選擇及提取 的內置支援 提供了一系列在 spider 之間共用的可複用的篩檢程式(即 Item Loaders)，對智慧處理爬取資料提供了內置支援。 通過 feed 匯出 提供了多格式(JSON、CSV、XML)，多存儲後端(FTP、S3、本地檔案系統)的內置支援 提供了 mediapipeline，可以自動下載爬取到的資料中的圖片(或者其他資源)。

PHP: 我們使用 PHP 來製作標籤雲，那標籤雲或稱做文字雲是關鍵詞的視覺化描述，用於匯總用戶生成的標籤或一個網站的文字內容。標籤一般是獨立的詞彙，常常按字母順序排列，其重要程度又能通過改變字體大小或顏色來表現，所以標籤雲可以靈活地依照字序或熱門程度來檢索一個標籤。大多數標籤本身就是超連結，直接指向與標籤相聯的一系列條目。

Solr:是 Apache Lucene 專案的開源企業搜尋平臺。其主要功能包括全文檢索、命中標示、分面搜尋、動態聚類、資料庫整合，以及富文字（如 Word、PDF）的處理。Solr 是高度可延伸的，並提供了分散式搜尋和索引複製。Solr 是用 Java 編寫、執行在 Servlet 容器（如 Apache Tomcat 或 Jetty）的一個獨立的全文搜尋伺服器。Solr 採用了 Lucene Java 搜尋庫為核心的全文索引和搜尋，並具有類似 REST 的 HTTP/XML 和 JSON 的 API。 Solr 強大的外部配置功能使得無需進行 Java 編碼，便可對其進行調整以適應多種類型的應用程式。

Mysql: phpMyAdmin 是由 PHP 寫成的 MySQL 資料庫系統管理程式，讓管理者可用 Web 介面管理 MySQL 資料庫。藉由此 Web 介面可以成為一個簡易方式輸入繁雜 SQL 語法的較佳途徑，使用 phpMyAdmin 就可以方便的建立、修改、刪除資料庫及資料表。

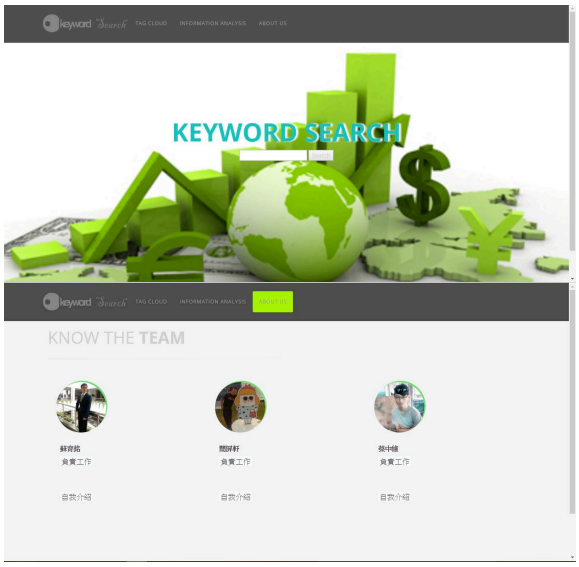
Html5:包含了 HTML、CSS、JavaScript 三方面的革新。在 HTML 方面，HTML5 增加了新的語意標記、多媒體標記、各種屬性和其他規範，也正式放棄了幾項舊標記。在 CSS 方面，它提供新的 CSS 語法和樣式；在程式方面，它也提供新的 JavaScript 程式開發介面。

D3.js: 是一套 Javascript 函式庫，包含一整組操縱畫圖很好用的輔助工具,還有很方便的資料操作模型。

(七) 參考文獻

參考文獻請參考 IEEE 或本校學位論文參考文獻格式規範撰寫，請依參照使用之順序，依次編號列出。

實驗結果



結論

完成這個專題讓我們明白了資料可視化的重要，光是政府開放了資料是沒有用的，還必須依靠資料可視化的處理，讓這些資料不會只是一格一格的數字，那樣的呈現方式閱讀性是微乎其微的。

最重要的是，我們還學到了完成一個作品所需要的能力與團隊合作的重要性。

參考文獻

[1] 基於詞頻統計的分析和討論:
http://libsearch.ntut.edu.tw:2337/eds/detail/detail?sid=ee7c3a82-148a-4873-a571-fa0bc89108c4%40sessionmgr120&vid=0&hid=114&bdata=Jmxhbmc9emgtdHcmc2l0ZT1lZHMtbGl2ZQ%3d%3d#AN=edsarl.10083715.201204.201207180015.201207180015.61.64&db=edsarl

[2]Python 後端網路爬蟲:
http://wiki.jikexueyuan.com/project/python-crawler-guide/summarize.html

[3]PHP 標籤雲技術:
http://www.111cn.net/phper/php/42844.htm

[4]Solr 詞彙統計
https://zh.wikipedia.org/wiki/Solr

[5]Mysql 資料庫架設與使用
http://www.codedata.com.tw/database/mysql-tutorial-getting-started/