Crawl Curation

專題編號：107-CSIE-S009
執行期限：106年第1學期至107年第1學期
指導教授：劉建宏 教授
專題參與人員： 104820003 曾立嚴
104820022 施逢怡
104820020 李家森

一、摘要

我們收集網路新聞電子報上的內文。分析不同類型的討論議題，將最近熱門關鍵字、可能之主題與代表文章可視化在網站上。(請參閱圖片 2~7)

關鍵詞：網路爬蟲(Web crawler)、關鍵字比對(Keyword matching)、文字分析(Text analytics)、雲端運算、容器技術(Container)

二、緣由與目的

現今新聞媒體影響力無遠弗屆，橫跨政治、經濟與社會文化等層面，並且也是民眾對於社會發生之事務相關訊息取得來源。但在審視各種特定議題相關大量新聞和評論之後，難以直接判別議題走向和分析媒體所關注的重點。此實務專題之目的為擷取新聞媒體中之資訊，進而統計分析，快速了解特定議題的重點。

三、研究報告內容

(一)使用技術方法及工具說明

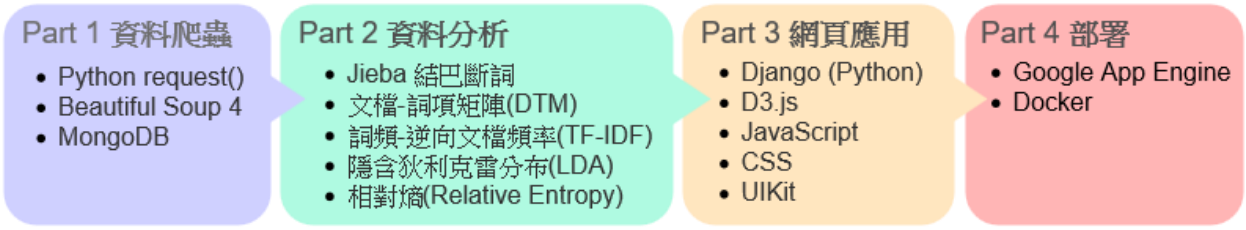
1、 新聞爬取：使用 Python 的 requests 模組到 Web server 取得特定頁面，以取得各家新聞之文章，並依爬取時間先後順序儲存貼文資訊。[1]

2、 頁面分析:以 Beautiful Soup 4來分析頁面，根據各種標籤與屬性，從 HTML 頁面裡過濾出想要的標籤。

3、 斷詞系統(Word Segmentation)：文本分析前我們必須將文章中的句子切割為具有意義的單詞。然而，有別於英文，中文的句子中詞是連在一起書寫的，因此，我們借助 Jieba、Genius 工具包進行初步切割再進行調整。

4、 文檔-詞頻矩陣(Document-Term Matrix)：將資料進行分析前，我們必須將她轉換為電腦能理解的語言。我們利用 one-hot-code 的方式，將文本中的單詞依序編號，每篇文章則轉換為一個陣列，陣列中對應欄位分別記錄各單詞出現次數。形成n × m的矩陣(其中 n:文本數，m:詞彙數)。

5、 詞頻-逆向文檔頻率(TF-IDF)：在文章中常常會存在一些「常見詞」，比例極高卻難以對文章進行特徵區隔，所以應將低其權重。相反的，我們關心的是僅在少數文章中經常出現的詞彙。常用的計算方式就是在文檔詞頻矩陣(TF)的基礎上乘上(IDF)進行權數調整。



圖片 1 研究流程圖

6、 隱含狄利克雷分佈(Latent Dirichlet Allocation, LDA)：對於結構化的文本數據，我們希望能歸類為數個主題。我們採用 LDA 主題模型，透過無監督式(機器)學習的方式，產生主題(詞)集合，以及這些主題中每個詞出現的概率是多少。

7、 相對熵(Relative Entropy)：主題模型對主題的描述往往不夠直觀，因此我們希望能在該主題下找到最具代表性的文章。我們透過相對熵，比對主題模型、文本中詞彙的機率分佈相似度。來找出最具代表性的文章。

8、 MongoDB：我們使用 mLab 作為系統的資料庫，mlab 提供 MongoDB Hosting 服務，使開發者無需自行架設 DB Server。此種資料庫為 Non-Relational 資料庫，相較關聯式資料庫更能夠有效率地存取龐大資料，並且可單獨擴充單一資料的資料表中之欄位。[2]

9、 數據可視化：使用 D3.js 使數據可視化，D3.js 運用 JavaScript 存取 DOM 介面，依據資料來源(JSON)動態在 HTML 上繪圖。

10、利用 Django 網頁應用框架開發，處理頁面動態顯示與各式 Http Request 與 Post Method，另外使用 UIKit 作為前端框架開發響應式之網頁。

11、Google Compute Engine Custom Runtime：為 Google 的 PaaS 服務，提供 Ubuntu 的 VM (Virtual Machine) 環境進行雲端運算。為了環境的靈活度，我們使用 Custom Runtime，即可使用客製的 Docker 容器技術(Container)。我們在這樣的 Infrastructure 下，部署程式與網頁。[3]

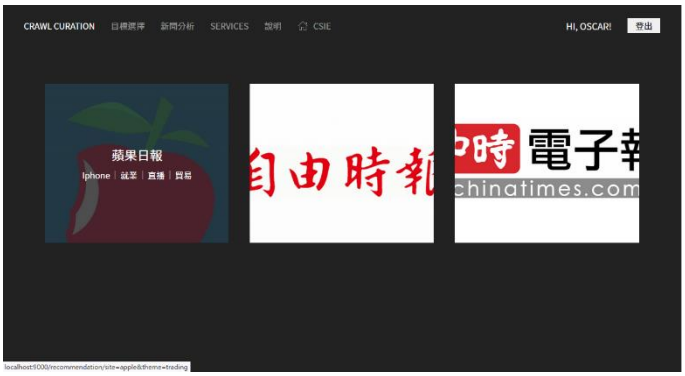
12、Firebase Authentication：Google 提供的帳戶登入驗證資料庫，使用者密碼等資訊被加密保存，管理員亦無法得知用戶之密碼，增加系統安全性。

(二)研究架構流程

如圖1，首先透過伺服器端爬取特定新聞內文。接著進行中文斷詞、資料處理、分析並透過模型歸納。最後進行視覺化處理，並顯示圖表於前端網頁。

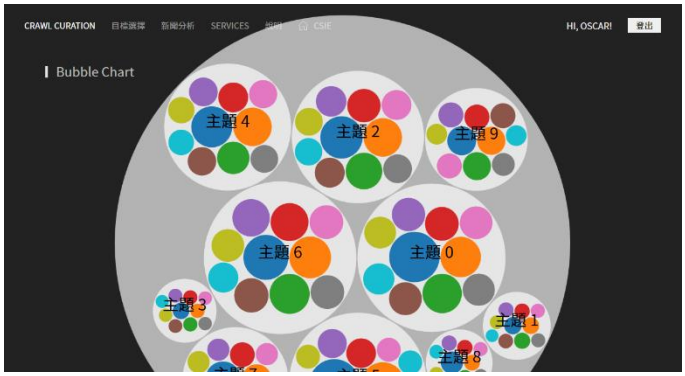
(三)專題成果

圖片 2 首頁



圖片 3 選擇目標網站分類

圖片4 主題分析結果頁面



圖片5 氣泡圖呈現分析結果

圖片6 氣泡圖可互動放大

圖片7 分析解果呈現詞雲

四、參考文獻

[1] Ryth Mitchell「網路擷取｜使用Python」, 歐萊禮出版社, 2017/09

[2] 郭遠威，**高效經營 BigData: MongoDB 資料庫系統管理與開發手札**，臺北：深石數位，2017，第1-6至1-8頁。

[3] Massimiliano Pippi, 2015, *Python for Google App Engine*, Birmingham: Packt Publishing, pp. 114-172.

[4] 文淵閣工作室,「Python 架站特訓班—Django 最強實戰」,2017.08