

計畫編號：MOHW106-CDC-C-114-114118

衛生福利部疾病管制署 106 年委託科技研究計畫

計畫名稱：自動化輿情監測系統與非結構性資料分析模式建置

106 年度研究報告

執行機構：中華 R 軟體學會

計畫主持人：陳嫻如

研究人員：丘祐璋

執行期間：106 年 5 月 18 日至 106 年 12 月 31 日

研究經費：新臺幣 139 萬元整

本研究報告僅供參考，不代表本署意見，如對媒體發布研究成果應事先徵求本署同意

目 錄

壹、研究報告中文摘要.....	4
貳、研究報告英文摘要.....	5
參、研究報告本文.....	7
(一) 前言.....	7
(二) 材料與方法.....	10
1. 輿情系統架構.....	12
2. 資料處理.....	16
3. 資料分析.....	19
4. 系統環境.....	23
(三) 結果.....	30
1. 系統登入功能.....	30
2. 文章檢索功能.....	32
3. 進階探勘功能.....	38
4. 疫病同義辭典.....	44
5. Python 教育訓練.....	45
(四) 討論.....	46
1. Google Trends.....	48
2. CDC 疫情管理.....	49
3. Gphin.....	49
4. 本計畫輿情分析平台.....	50
5. 平台搜尋結果比較.....	50
6. 輿情監測系統對機關效益.....	56
(五) 結論與建議.....	59
(六) 重要研究成果及具體建議.....	60
(七) 參考文獻.....	61
肆、附錄.....	66
(一) 附錄一：疫病同義詞典.....	66

(二) 附錄二、教育訓練出席統計 75

壹、研究報告中文摘要

社群媒體的興起讓訊息的傳遞也變得更加容易。但網路言論除了傳播的便利外，也帶來潛在風險。部份人士為了自身的利益，恣意在網路上散佈假新聞。若民眾未查訊息的真偽便將該謠言分享到自身的社交網路之中，不但可能造成不必要的恐慌外，更可能會挑起民眾對執行機關的不信任感。

從公共治理的角度來看，要能遏止謠言傳播，必須先了解不實言論的起源與傳播方式。但至今網路訊息眾多，如用人工方式一一檢視網路言論，不但耗費大量人力、時間，更可能因為檢視速度緩慢，錯失宣導正確政令的最佳時機，因而我們希望能建立一個自動化輿情分析系統，以系統化的方式快速觀測網路對疫病相關的討論。

為了能夠廣泛、快速蒐集大量的疫病輿情，該系統首先蒐集了各大網路頻道對疫病的相關討論，接者將輿情資料儲存於非關聯式資料庫之中，當搜尋引擎索引所有輿情資料後，我們便能以關鍵字查詢到與疫病相關的最新輿情。此外，為了能夠讓疾管署署內同仁熟悉如何利用 Python 工具分析非結構化資料，我們另外舉辦了為期九天的 Python 教育訓練課程，教導同仁如何懂得利用 Python 處理、蒐集與分析非結構化資料。

為了凸顯輿情分析能帶來不同的分析維度，我們另外比較了 Google Trend, GPHIN, CDC 與輿情數據，希望透過不同數據的比較與詮釋，讓當責單位了解輿情分析的優勢與侷限，之後便能將輿情分析的成果做為未來防堵謠言散播的第一道防線。

中文關鍵詞(至少三個)：輿情監測、非結構性資料、文字探勘

貳、研究報告英文摘要

As more and more people are connected to the social network, one can quickly reach to others via the internet. But on the other hand, the internet also brings potential hazard. For example, if some people intent to spread the fake news on the internet, the rumor may likely terrify innocent crowd, and also lead people to question the capability of the official unit.

For the perspective of public governance, one can stop the rumor only if they know how they are broadcasted on the internet. However, as the amount of internet opinion is numerous; one can hardly digest all messages solely in person. Therefore, if the government can build up an opinion monitoring system to monitor and analyze online comments, the government can consequently immediately authenticate the truthfulness of the news. For this purpose, we come out a four-year plan to build up a platform that can monitor multilingual opinions.

To monitor vast opinions quickly, we deploy a distributed opinion monitoring system. This system first uses the web crawler to collect disease related reviews on all kinds channels. Then, the system stores the opinions into a Non-SQL database. After, the search engine index all non-structured data, one can query the most updated opinions through our SAAS service. Also, to let the member of CDC understand how to analyze unstructured data with Python programming language, we hold a nine-day training course, so they can understand how to process text data with python.

To emphasize the importance of opinion data, we compare the data collected from Google Trend, GPHIN, CDC and opinion reviews; we hope to let the official unit understand the advantage of opinion mining is best for stopping rumor rather than disease trend prediction.

Keywords: Text Mining, Unstructured data, Opinion Mining

參、研究報告本文

(一) 前言

伴隨著社群媒體、智慧裝置與雲端計算的興起，引領我們進入了資訊爆炸的時代。根據知名研究機構 IDC 2016 年對巨量資料的預測，巨量資料的需求從 2015 年到 2020 年的複合成長率將會以每年 22.6% 的成長率成長，而在 2020 年將會達到 589 億美元的規模，足以證明各界對巨量資料的需求與重視，如何利用巨量資料帶來的機遇，並從數據中挖掘出價值資訊，是現今各國政府所爭相關注的議題。

在談到巨量資料時，我們往往會以三個 V 描述巨量資料，分別是資料量(Volume)，資料累積速度(Velocity)與資料多樣性(Variety)。如何面臨巨量資料所帶來的三個挑戰，並從中挖掘出數據價值(Value)，便是公部門與私營企業在這巨量數據時代中必須面臨的挑戰。

公共衛生議題上的數據處理與分析並非新鮮事，長久以來各國政府就非常重視收集相關資料與建立疾病監測系統，例如法定傳染病通報系統，預防接種登錄等系統，這些系統所蒐集之數據一向是政府衛生單位的重要參考資料。但除了傳統的結構化資料分析外，國外使用巨量數據分析疾病的案例為疾病監測系統帶來新的刺激。例如：美國愛荷華大學研究團隊比對 2004 年三月至 2008 年五月由 Yahoo 搜尋引擎查詢「流感」相關訊息的搜尋量，發現可以較監測系統早偵測疫情流行。除此之外，知名的網路巨擘 Google 早在 2008 年全球禽流感爆發疫情之際時，便於知名 Nature 期刊發佈一用 Google 趨勢(Google Trends) 預測疫情的文章，該案例是從近三十億筆搜

尋關鍵字中，篩選出五千萬筆搜尋關鍵字，再與美國疾病管理局的流感案例進行比對，在分析過搜尋關鍵字與案例數量的相關性後，發現從關鍵字得到的分析結果與官方資料相當符合。雖然隔年 Google 即發現該預測方法失準，因而將該論文從知名期刊中撤下。雖然 Google 的分析方法並不能真正實際用在疫情預測上，但利用巨量資料作為參考的研究方法，卻開啟了巨量資料研究公共衛生的先端。

但除了利用預測準確度的思量以外，改從公共管理的角度出發，政府如何了解民眾對疫病與公共衛生的討論，掌握民眾是否有正確的防疫概念，並避免不實謠言在社群網路上任意散佈，是政府的重要責任之一。但是隨著社群網路與智慧裝置的發達，資訊累積的速度與傳播量也是每年俱增，如以傳統以人的方式一一檢測，除了浪費時間、物力以外，更可能會因為檢測速度太慢，導致疾管署可能會錯失第一手重大訊息，失去最佳能執行政令宣導的時機，如能建立第一線的輿情通報系統，疾管署便可以更即時聆聽民眾的心聲，即早發現政策實行上的缺失。

另外，隨著全球化的影響，疫病在各國間的傳染、蔓延，也讓各國政府的衛生單位也上緊發條，驚怕若防疫把關上如有所疏失，外國的疫病便會在國內蔓延開來，除了可能造成傷亡外，亦會給民眾帶來莫大的恐慌與對政府的不信任感。因此除了要掌握國內的相關討論外，如何知悉國外是否有大規模討論特定傳染病，亦會是防疫關卡建築不可缺失的一塊。

因應以上需求，我們因而建立了一即時輿情平台，希望能夠即時搜尋國內外各種對疫病的討論與相關新聞，方能讓疫病發生之際，疾管署能第一手

掌握消息，並待確認訊息真偽後，再擬訂相關政策或應對方針。若所掌握的訊息為謠言，便可以即時透過關謠專區，宣導正確概念，遏止謠言叢生；但若掌握的訊息可能性較高時，便可以比對內部數據，明辨是非後，便可以提出相關政策，擬訂對應方針。

但若單單只是研究外部的輿情資料，則會因為過多龐雜的訊息而讓人無法掌握真正的研究方向，因而必須與內部數據關聯後，才能讓數據產生價值，例如：比對外部資訊，知悉宣導是否有效，比對實際回報數據，了解網路所傳是否謠言。為了能讓疾管署內部的同仁能關聯外部輿情與內部資訊，我方因而提供疾管署一系列 Python 課程，希冀內部同仁能學習完整套 Python 知識後，便有能力利用手邊的開源工具，分析非結構化數據，讓輿情工具不只是工具而已，而是變成數據的提供來源，讓內部人員可以利用這套數據輔以工具，便可以從數據中挖掘出更多價值。

根據計畫的規劃下，我們這今年完成了以下目標：

- 建構一自動輿情監控平台該平台能夠利用使用網路爬蟲技術即時監控來自新聞、論壇與社群媒體等各疫情資訊來源的最新訊息
- 設立一疫情輿情相關的繁體中文同義字典，讓系統得以歸納包含同義關鍵字的輿情資訊。
- 系統之後必須能夠將輿情資訊儲存到一非關聯式資料庫中，供使用者能夠存取該輿情資訊
- 系統能將這些儲存於非關聯式數據庫中的輿情訊息匯入到搜索引擎之中，讓使用者得以透過搜索引擎的介面上下關鍵字查找與疾病管制相關

的所有輿情

- 系統需能夠搜集該輿情資訊等來源、評論數、按讚數、回文數等不同指標，讓執行單位得以用不同指標排序或檢視最具威脅性的謠言。
- 系統能夠將輿情資訊根據聲量、頻道來源、情緒指標、擴散路徑等資訊繪製視覺化圖表，讓使用者得以視覺化方式探索輿情資料。
- 舉辦 Python 教育訓練課程，提升疾管署對非結構化文字資料的處理與蒐集能力。第一年的課程包含：Python 基本教學、Python 網路爬蟲、Python 文字探勘等三大課程項目。

(二) 材料與方法

為了能夠觀測國內外疫病各大頻道相關輿情訊息，並從非結構化資料之中分析出價值資訊，我們建立了一個即時輿情觀測平台，並將該平台建構於雲端平台上，各部門的分析人員只須透過瀏覽器，便可以瀏覽並取用輿情分析數據，用做報表或決策依據。以下將詳列建立該輿情分析系統的架構與方法。

今年完成的資料來源如下：

臺灣新聞網

公視新聞網，三立新聞，東森新聞，蕃新聞，中廣新聞網，台灣大紀元，風向新聞，關鍵評論

臺灣社群

卡提諾論壇，拓峰網論壇，TT1069 男男貼圖交友區，基地-社團法人台灣基地協會

中國大陸國家網

中華人民共和國農業部，中華人民共和國中央人民政府門戶網站，中國疾病預防控制中心，中華人民共和國國家衛生和計劃生育委員會

中國大陸地方政府

吉林省衛生和計劃生育委員會，安徽省疾病預防控制中心，安徽省衛生和計劃生育委員會，北京市疾病預防控制中心，北京市衛生和計劃生育委員會官方門戶，浙江省疾病預防控制中心，浙江省衛生和計劃生育委員會門戶網站，廣東省疾病預防控制中心，廣東省人民政府应急管理辦公室，廣東省衛生廳，重慶市衛生和計劃生育委員會，重慶市疾病預防控制中心，福建省疾病預防控制中心，福建省衛生和計劃生育委員會，甘肅省衛生和計劃生育委員會，甘肅省疾病預防控制中心，廣西壯族自治區疾病預防控制中心，廣西衛生信息網，貴州省衛生和計劃生育委員會門戶網，貴州省疾病預防控制中心，湖北省衛生和計劃生育委員會，湖北省疾病預防控制中心，河北省衛生廳，河北省疾病預防控制中心，湖南省疾病預防控制中心網站，湖南省衛生和計劃生育委員會，河南省衛生和計劃生育委員會，江蘇省疾病預防控制中心專業版，江蘇省衛生和計劃生育委員會，江西省衛生和計劃生育委員會，江西省疾病預防控制中心，遼寧省衛生和計劃生育委員會，遼寧省疾病預防控制中心，雲南省衛生計生委首頁

新聞媒體源

cnYES 鉅亨網，央視網，彭博新聞社，MoneyDJ 理財網，民報 Taiwan People News，SOGI 手機王，經濟學人，NOWnews 今日新聞，民視 FTV，

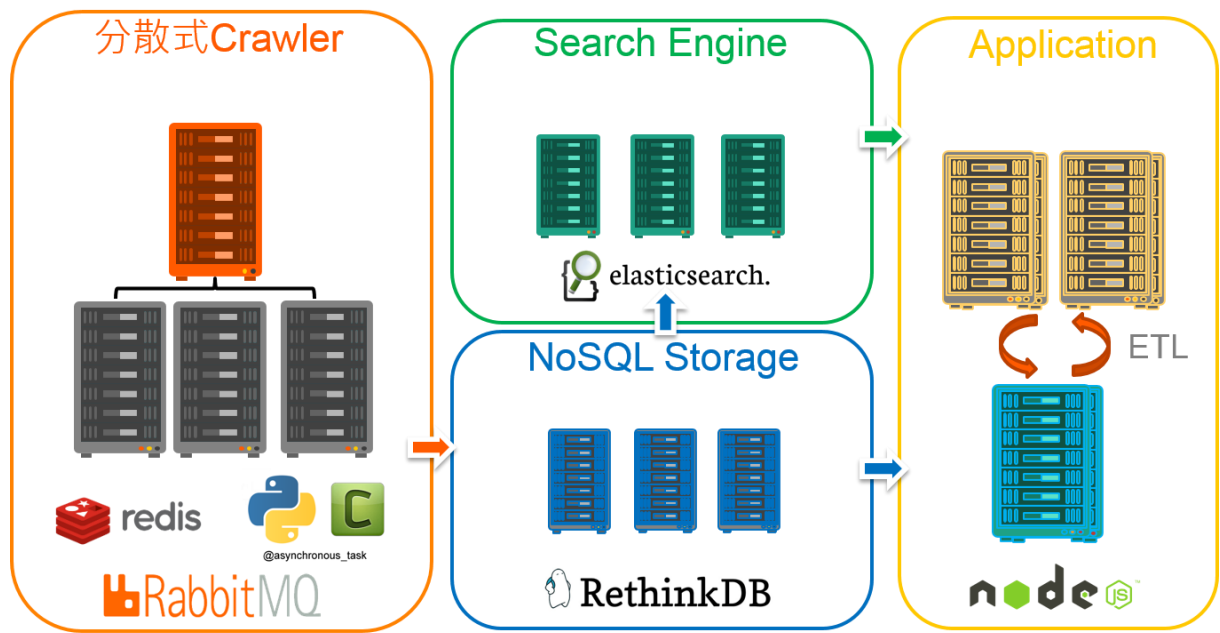
巴哈姆特電玩資訊站, PanSci 泛科學, 年代電視台, PeoPo 公民新聞, 自由時報電子報, TechNews 科技新報, 觀察者網, TVBS 官方網站, 東森電視, Yahoo, 环球网, 人民網, 美麗島電子報, 三立新聞網, 風向新聞, 上下游 News&Market 新聞市集, 風傳媒, 上報, 報導者, 大紀元, 壹電視 NextTV, 中央社即時新聞 CNANEWS, 華視全球資訊網, 中央網路報, 新浪, 中國時報, 新頭殼 newtalk, 中國評論新聞, 解放網, 中視, 農傳媒, 中廣, 端傳媒, 公視, 蕃薯藤, 卡卡洛普 Gamme, 聯合報, 台視, 聯合新聞網, 台灣醒報 Awakening News Networks, 關鍵評論網, 台灣環境資訊協會, 蘋果日報

社群媒體源

PTT、DCARD、Mobile01、Youtube、Facebook

1. 輿情系統架構

由於量大、積累速度快、格式龐雜的輿情訊息皆符合一般對巨量資料的認定，為了能夠讓系統能即時蒐集、儲存、索引、分析並呈現輿情分析結果，我們架構了一分散式平台架構，以期能以該架構迅速處理並分析海量的輿情訊息。系統架構如下：

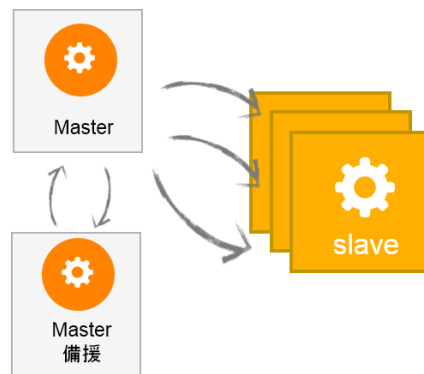


圖一、系統架構

以下將分為四個元件（分散式網路爬蟲、分散式儲存架構、搜尋引擎與應用程式）分別說明各元件的主要功能：

分散式網路爬蟲

本研究以 Python 建立分散式網路爬蟲，根據輿情資訊來源，系統將自動蒐集輿情資訊的主文、評論數、按讚數、回文數等不同指標資料。由於資料量龐大，為了能夠加速資料爬取的速度，將會架設分散式資料擷取模組爬取 PTT, Facebook、各大新聞媒體及其他疫病資訊源。為了能夠協調各分散式系統的運作，採用主從式架構，建立分散式爬蟲系統，概念如下圖。



圖二、主從式爬蟲架構

根據實機測試，每台機器可以每天抓取三十萬頁網頁，透過四台（一主三從）的部屬，將可每天抓取約 100 萬網頁的資料量。由於該架構屬於分散式架構，所有的工作統籌都由主節點分配，因此當如果有增加資料抓取的需求時，只需加機器便可擴增資料量的抓取。

除了抓取網頁內容外，必須使用網頁剖析程式方能抽取網頁中重要的資料，但往往目標網頁格式會有所變動，因此在抓取的過程之中，我們必須要保留網頁的原始檔，以免目標格式變更後，導致資料遺失。因此，系統會先在磁碟空間上預儲存一份資料源，而後再透過剖析程式，抽取重要資料。如果抓取資料頁面格式有變，該系統將會自動發出警示至權責人員信箱。

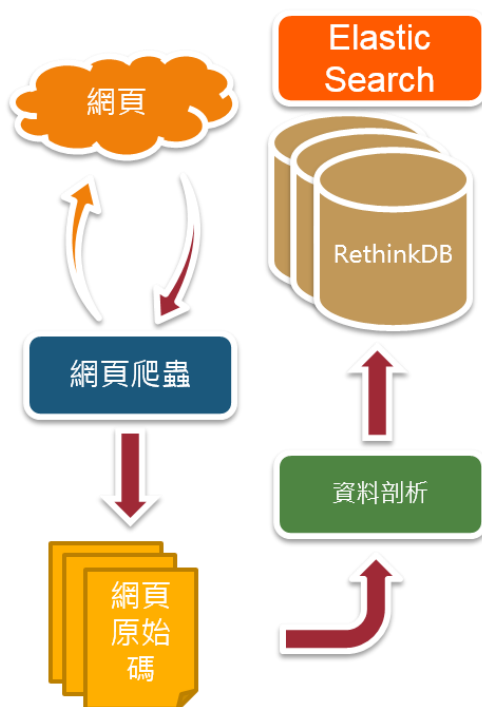
分散式儲存架構

為了因應抓取的資料屬於非結構化資料，不同來源資料要存儲的欄位可能不一致，因此我們採用了 RethinkDB 做為分散式資料儲存引擎，確保系統能以無綱要(Schema-Free)的 JSON 模型儲存大量的非結構化資料。由於該架構亦採用分散式架構，因此當使用者若需要加大儲存空間時，只需要

增添機器便可以增加系統儲存容量。

搜尋引擎

而為了能夠加速資料搜索的速度，我們在寫進 RethinkDB 資料庫的同時，寫入一份資料至 ElasticSearch，使用者之後便可以透過其搜尋功能篩選輿情資訊。流程圖如下圖。



圖三、爬蟲流程

由於搜尋引擎是以反向索引機制(Inverted Index)索引資料，當文章進入搜尋引擎後，系統將透過比對字典內的字詞，便可以將輸入文章分詞，並針對各詞建立索引，因此，使用者只須透過關鍵字，便可以快速查詢到跟關鍵字相符合的疫病相關文章。

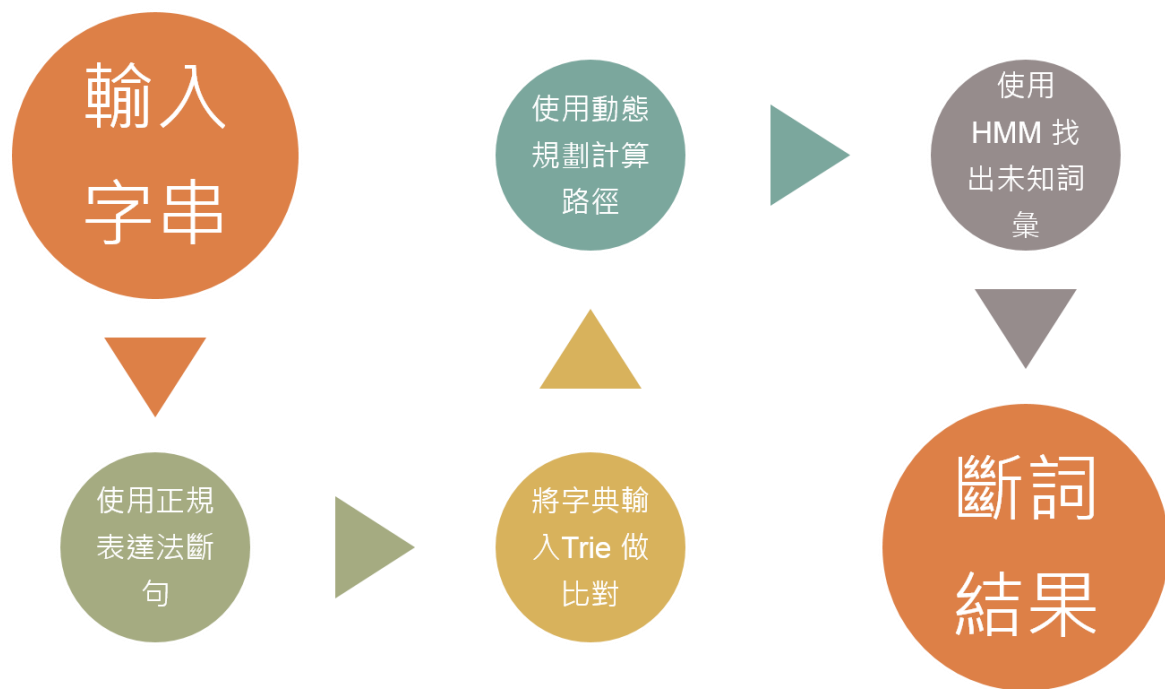
應用程式

為了能夠在雲服務上即時分析非結構化數據，我們在分析程式端佈署了三台虛擬機器，其中一台用做資料 ETL(Extract, Transform, Loading)，而應用及分析程式則佈署在另外兩台機器上。該應用程式端負責分詞，並透過文字探勘機制（分群、詞性標註、分類、語意分析、情緒判斷）處理及分析爬蟲所搜集到的輿情資料，之後可透過可視化儀錶板呈現文字探勘後的分析結果。

2. 資料處理

文章分詞

為了能夠支援中文資料斷詞，以及可以從中挑出重要的人、事、時、地、物等名詞，在本系統中使用 Python 的 Jieba 套件作為斷詞工具。該系統首先使用正規表達法斷句，接者使用首碼樹(Trie Tree) 等資料結構去生成句子成詞的所有可能組合，接者使用動態規劃找出組合中最大概率路徑，這個路徑就是基於詞頻的最可能分詞結果。系統只需要擴充辭典，便可以精確斷出中文詞彙，而使用者亦只需要擴充簡體中文詞典，便可以因應簡中斷詞需求。除了提供字串比對方式斷詞，系統亦採用隱馬可夫等統計模型以辨識文章中可能的新詞。



圖四、分詞過程

詞性標記

本系統的詞典除了字詞以外，另外有標記每個詞的詞性，當產生斷詞結果時，除了會斷出字詞外，亦可以選擇是否顯示該字詞的詞性。例如針對下列句子「很高興您對這個議題有興趣」斷句，則會產生下列結果：

很/副詞

高興/形容詞

您/代詞

對/介詞

這個/數量詞

議題/名詞

有/動詞


興趣/形容詞

詞庫建立

本計畫會利用網路爬蟲技術先到維基百科 (Wikipedia) 蒐集跟疫情相關的所有關鍵名詞，並利用維基志工標註的同義字詞 (粗體字) 建立同義字典 (請見範例圖)。

登革熱 [編輯]

維基百科，自由的百科全書

 維基百科中的醫療相關內容僅供參考，詳見**醫學聲明**。如需專業意見請諮詢專業人士。

登革熱 (英語：dengue fever)，也稱為**骨痛熱症、斷骨熱、天狗熱**，是一種由登革熱病毒引起的由蚊媒熱帶病^[1]。患者大約會在感染後3到14天後發作^[2]，症狀包括發熱、頭痛、肌肉和關節痛，還有典型性的**麻疹樣皮疹**^{[1][2]}。一般會於2至7天痊癒。少部分患者病情可進一步惡化，出現危及生命的登革出血熱，患者有**出血、血小板減少**和血漿蛋白滲出，或者進展為登革休克綜合徵，此時會出現致命性的**低血壓休克**^[2]。

圖五、維基百科同義字詞來源

另外，本系統將會利用新聞關鍵字(如下圖所示)每天更新系統中文字典，以利系統可以使用最新的中文詞典正確斷詞，產生精準的文字分析結果。

蚊子是登革熱的主要傳染病媒。(路透資料照)

疾管署統計，今年有5例本土登革熱、234例境外移入(含2例死亡)，近一個月境外移入感染國家以越南為多，今年越南登革熱已達12萬5000例，較去年同期增加逾40%。

本週末受鋒面影響，國內部分地區將有短暫陣雨，疾管署提醒，雨後積水容器有助病媒蚊生長，民眾要落實孳生源清除；前往登革熱流行地區，則須做好防蚊措施。

相關關鍵字：

健康醫療

登革熱

圖六、利用新聞關鍵字擴充詞典

3. 資料分析

詞頻分析

本系統可以將文章斷詞後，統計該篇文章所使用的詞彙頻率，並可透過統計時間、頻道、來源繪製關鍵字聲量圖表，並可以根據來源與頻道繪製統計圖，使用者便可以透過來源與頻道統計分析輿情發生的熱點。

關鍵字分析

為了判斷一個詞是不是關鍵詞，系統建立了 TFIDF 演算法，用來判斷文章中的關鍵字，判斷規則如下：如果某個詞比較少見，但是在這篇文章中多次出現，那麼該詞很反映了這篇文章的特性。

TFIDF 是由 TF 乘以 IDF 所計算出來，公式如下所示：

- TF (Term Frequency)

單詞在該文件的出現次數

單詞 w 在文檔 d 中出現的次數: $\text{count}(w, d)$

文檔 d 中總詞數: $\text{size}(d)$

$\text{tf}(w, d) = \text{count}(w, d) / \text{size}(d)$

- IDF (Inverse Document Frequency)

一個詞語普遍重要性的度量

設文檔總數為 n

設詞 w 所出現檔數 $\text{docs}(w, D)$

$\text{idf} = \log(n / \text{docs}(w, D))$

TFIDF 可用來評估該詞對於該文件的重要程度，假設單詞對文章的重要性越高，TF-IDF 值就越大。

為了能夠快速計算 TFIDF，系統將會維護一個 IDF 字典，之後如果使用者希望能夠快速從文字中摘要出關鍵詞時，便可以透過該字典快速摘出該文章關鍵字。系統便可以在文章下方顯示該關鍵字，提示使用者可以透過該關鍵字了解通篇文意。

相似度計算

為了能夠計算文章間或字詞間的相似度，本系統將會使用餘弦相似度 (Cosine Similarity) 計算相似度。為了加速系統的運算速度，系統將會從每篇文章中依 TF-IDF 各抽取 20 個關鍵字，並且將文章建立詞頻矩陣(以文章當列、字詞當欄位)，並且利用該關鍵字計算文章間或字詞間的相似度。系統將會自動會傳相似度最高的文章或字詞，讓使用者了解文章或字詞間的關聯性，以就近挑出最相近的文章或最相關的字詞。

文章分群

本系統可透過文章相似度，自動將同類文章彙整成一類。目前坊間的分群方法多倚賴 K-Means 做文章分群，但是若必須將特定文章硬分成 K 群實屬不合理，因此我們採用了社群偵測(Community Detection)演算法分群文章，該演算法的概念如下，首先計算文章的相似度，接者便利用相似度將同類文章建立關聯連線，接著便可以利用圖論的社群偵測找出高度相連的群體，便可自動產生多個主題的文章群。另外，除了將同類文章放置於同群外，系統並可從同群中自動整理出詞頻較高的關鍵字，供使用者參閱(如下圖)。

第 5 群

重要關鍵字: 川普

[新聞\]擋不住的風暴 川普主宰美總統選情](#)
[問卦\]川普贏了美國人會崩潰嗎?](#)
[Re: 討論\]川普 大危機!](#)
[新聞\]川普又贏 美國人恐爆出走潮](#)
[Re: 討論\]川普 大危機!](#)
[Re: 討論\]川普 大危機!](#)
[新聞\]共和黨「反川普」陣線成立 羅姆尼親上火](#)
[新聞\]擔心川普式外交 60共和黨大老發公開信](#)
[新聞\]共和黨前總統候選人羅姆尼稱川普是騙子](#)
[新聞\]川普致勝秘訣 把選舉當交易](#)
[Re: 新聞\]擔心川普式外交 60共和黨大老發公開信](#)
[討論\]川普當選美國總統對台灣未必是壞事](#)
[Re: 討論\]川普當選美國總統對台灣未必是壞事](#)
[Re: 討論\]川普 大危機!](#)
[Re: 討論\]川普當選美國總統對台灣未必是壞事](#)

第 6 群

重要關鍵字: 姊弟,恐嚇,大家,奶奶

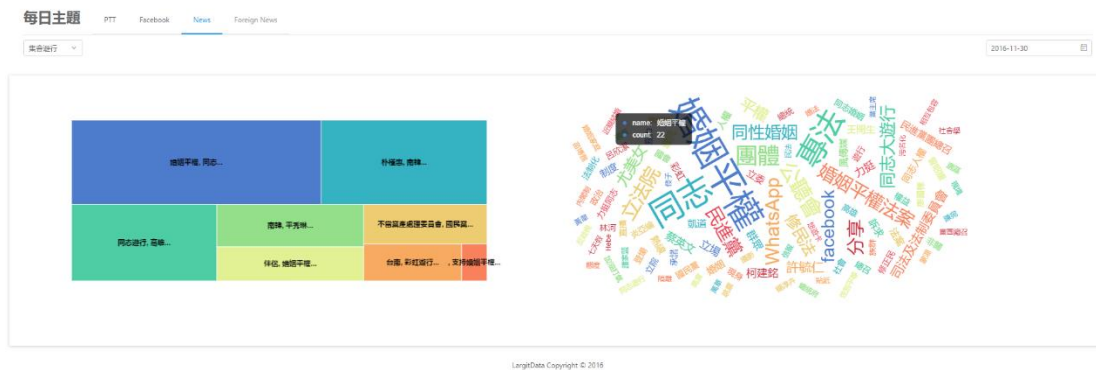
[新聞\]未完待續 三姊弟布丁風暴尚未完結](#)
[新聞\]三姊弟布丁要關了! 臉書PO告別文:從今](#)

圖七、文章分群示意範例

文章主題

本系統可以將文章分群後整理成樹狀圖，使用者可以透過該樹狀圖理解每個主題被討論的數量，並從該圖中瀏覽每個主題中的代表字詞。另外使

用戶可以點選每個主題，以觀看該主題的代表字，系統將以文字雲的方式呈列各主題的詞頻。



圖八、文章分群結果呈現

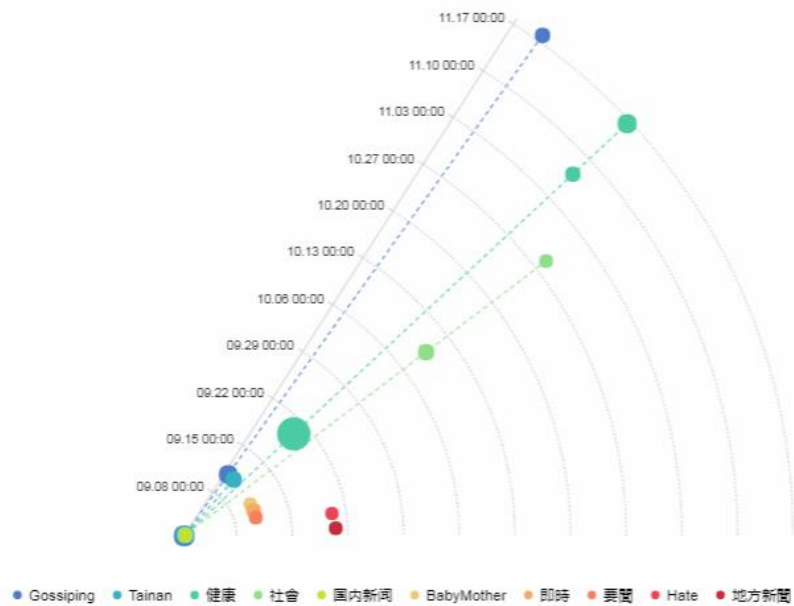
情緒分析

當使用者點擊切換情緒分析時，將會產生符合條件文章之正負情緒統計圖。文字的正面情緒與負面情緒，是我們取用文字中的形容詞後，根據文字出現文章的的正負面標籤以及文字出現的領域，計算出每個文字的正負面極性，若文字意表正面，則正面極性分數越高，但若文字意表負面，則負面極性分數越高。該圖則是透過計算該關鍵字所出現文章之正面極性減掉負面極性而得之，若相減結果為正，則將該文章列為正面情緒。反之，則列為負面情緒。

擴散分析

擴散分析報表可以讓使用者掌握輿情擴散程度與方向。該圖表將輿情來源表示成一雷達狀圖表；以時間區隔不同時間間隔；以圓圈大小顯示輿情聲量；並以圓圈顏色標註不同頻道來源，使用者便可以透過該雷達圖表關注哪些頻道有持續在討論疫情資訊，並可關注該輿情是否有擴散到其他媒體

源。



圖九、擴散分析圖表範例

4. 系統環境

在此節將說明，建立「即時輿情分析系統」所需之軟硬體技術。

軟體部分

1. NGINX 伺服器：

Apache 與 Nginx 伺服器是世界上最常見到的兩個網頁伺服器軟體。兩者綜合起來服務了超過網路上約 50% 的流量。這兩者不但皆可以服務大量的網路請求，並可協同多種程式語言（Java, Python, PHP）建構完整的網路應用程式。

在國內，大多數應用程式還是使用 Apache 所建構的，但 Apache 伺服器的缺點在於無法同時乘載成千上萬個 HTTP 請求時，而讓網頁應用程式能

同時承載成千上萬個請求的問題也被稱為 C10K 挑戰。為了解決這個問題，Igor Sysoev 於 2002 年發明了 Nginx 伺服器，並在 2004 年推出了第一個版本，希望透過非同步，事件驅動架構解決 C10K 問題。Nginx 後來逐漸成為網頁伺服器的主流之一，特點在於它使用較輕量的資源而且可以分散佈署在低端的硬體上，可以快速服務靜態的資源，並可以動態傳遞請求至其他應用程序。

在本規劃中，為了能讓「輿情分析系統」可以乘載大量的請求，故使用 Nginx 做為底層的網頁服務器。

2. RethinkDB：

網路輿情資料的特性是龐大且非結構化，為了能有效儲存符合這樣特性的資料，我們選用了非關聯式資料庫 RethinkDB 存放爬蟲蒐集到的輿情資料。RethinkDB 如同 MongoDB，可以使用 JSON 格式儲存資料，並可透過分散式的架構，使用多台機器乘載大量資料。但 RethinkDB 有一項其他 NoSQL 所沒有的特點，提供了主動推送更新(Push-Query)功能。

當以往前端頁面要更新呈現內容時，必須使用輪詢 (Polling) 的方式不停的對後端資料庫發出請求，以查詢最新的資料。如果輪詢 (Polling) 的頻率過於頻繁，將會對資料庫造成負擔，但若減少輪詢 (Polling) 的頻率，則實際資料跟資料庫中的內容會有落差。而 RethinkDB 的主動推送更新功能解決了這問題。

3. Ubuntu：

本系統將建構在 Ubuntu 16.04 上，Ubuntu 是以桌面應用為主的

GNU/Linux 作業系統，除了具有免費、開源的優點外，相較於其他 Linux 作業系統有下列好處：

1. 對使用者相對親善，Ubuntu 隱藏了許多 Linux 操作的細節，使用者可以透過圖形化介面操作作業系統，並透過其完備的套件管理工具安裝及管理各式應用程式
2. 具有完備的程式管理套件，使用者可透過套件管理工具安裝及管理各式應用程式
3. 擁有相對成熟的開源社群，使用者可以輕易透過社群尋求幫助
4. 具商業公司 Canonical 做技術支援，被回報的軟體缺陷，都會在下一個新版本被修正。

4. PostgreSQL

PostgreSQL 是可以媲美 Oracle 的開源資料庫，由全世界超過 1000 名貢獻者所維護。PostgreSQL 只提供單一版本，不像 MySQL 有區分社區版、商業版與企業版。基於自由的 BSD/MIT 許可，任何組織可以使用、複製、修改和重新發佈程式碼。

PostgreSQL 相較於其他資料庫而言，有相當高的可靠度、資料一致性、安全性與完整性，企業可以使用 PostgreSQL 打造相對穩健、安全的企業資料倉儲。另外，PostgreSQL 的文件非常完整，使用者可以從線上找到大量免費的線上手冊，或可以從歸檔文件中找尋舊版本的文件。

而除了資料庫本身功能外，多數企業使用者通常會將該開源產品是否容易維護納入考量之一，而由於 PostgreSQL 具有強大的社群及獲得些許商

業公司的支持，使用者還是可以透過社群或商業公司的管道獲得使用上的協助。

5. Python 程式語言：

Python 是一通用的直譯式、交互式、物件導向高階程式語言，由於其非常簡單易用，具有廣大的社群支持，還有完整的套件管理工具，Python 已經被廣泛運用在自動化腳本、物聯網及資料分析上。而由於 Python 的幾個套件可以大幅簡化資料 ETL(Extract, Transformation, Loading)的工作，並且可以透過其高階分析套件 Pandas 及 scikit-learn 進行敘述性統計與機器學習，因此在本專案上將會使用 Python 做為系統開發的主力工具。以下將簡介我們會使用到的套件：

1. Requests:

Requests 套件讓我們可以使用 Python 對遠端伺服器下達 REST (包含 GET, POST, PUT, DELETE)的操作，因此我們可以使用該 Requests 對網頁伺服器遞送 GET 與 POST 請求，蒐集伺服器回應資訊，進而取得網路輿情頁面原始碼。

2. Scrapy:

Scrapy 是一 Python 爬蟲框架，內建許多爬蟲所需的支援與函式庫，讓使用者可以輕鬆開發、維護一多線程(Multithread)的網路爬蟲。

3. BeautifulSoup4:

取得頁面原始碼後，為了能萃取重要資訊如：文章主題、作者、主

文、推文等資訊，我們可以使用 BeautifulSoup4 剖析頁面原始碼，並可透過 css selector 或 XPath 等操作抽取頁面中關鍵資訊。

4. Pandas:

為了能夠快速清理及分析資料，在本系統中使用了 Pandas(Python for Data Analysis)套件，Pandas 套件能將 csv, Json, html 等格式快速轉變成 DataFrame 格式，之後使用者可以透過類似 SQL 操作對資料作敘述性統計。

5. Jieba:

為了能夠對中文資料斷詞，以及可以從中挑出重要的人、事、時、地、物等名詞，在本系統中使用了 jieba 套件，該套件使用 Trie Tree 等資料結構去生成句子中文字所有可能組合，接者使用動態規劃找出組合中最大概率路徑，這個路徑就是基於詞頻的最可能分詞結果。而在新詞辨識部分則使用了隱馬可夫模型的 Viterbi 演算法。該套件是目前 Python 界最受歡迎的中文斷詞工具。

6. Scikit Learn:

在本研究中，會使用機器學習技術分群、分類輿情資訊。而 scikit learn 套件內建有回歸分析、資料分類、資料分群與降低維度等功能，使用者便可在整理好資訊後，使用機器學習套件快速分析研究資料。

7. django:

django 提供一 MTV 架構，讓使用者可以用模組化方式架構一完整

的網頁應用程式。另外，django 有提供一個完整的後台操作介面，使用者便不需要另外編寫管理介面，即可透過後端的管理介面操做資料庫。由於本計劃將大多使用 javascript 做前端頁面的呈現，因此本計劃將使用 django 搭建一個 API 服務，做為前端畫面與後端資料庫的接口，供前端頁面存取資料庫內容，並呈現到地圖上。

8. Networkx:

Networkx 是可以用來建構複雜網路模型的 Python 套件，除了可以透過該套件建構一網路模型外，該套件內建許多圖型探勘的演算法，使用者可以使用該套件找出子網路模組(module)或探勘圖型的重要性質。

6. Javascript 程式語言：

相較於使用 Python 開發伺服器端（後端）的 API，我們必須使用客戶端程式語言 Javascript 呈現頁面資訊，使用者才能與前端介面進行互動。而在本計劃中，為了讓使用者能享受互動體驗，並觀看相關動畫，我們將使用 Javascript 撰寫所有前端頁面的呈現內容(包含地圖座標的標記，前端互動的表單，圖表的呈現，前端動畫)，而我們除了使用原生的 Javascript 外，會搭配以下套件完成我們的頁面製作。

1. JQuery:

Jquery 是一跨平台的 Javascript 套件，主要的目的是為了簡化 Javascript 的撰寫，讓使用者可以更方便的透過 DOM 操作頁面元

素、建立動畫與處理非同步事件。使用 JQuery，我們便可以 Write Less & Do More 完成所有頁面製作。

2. Leaflet:

為了製作互動式的輿情地圖，我們除了將使用 Openstreetmap 提供的圖資外，我們可以使用 Leaflet 建構互動地圖。Leaflet 是由 Vladimir Agafonkin 所開發的免費且開源的 Javascript 套件，套件本身可以讓使用者在地圖上任意添加點、線、面等資訊，也支援主要作業系統與移動裝置，非常適合套用在互動地圖的設計。

3. Chart.js:

Chart.js 是個以 HTML5 的 Canvas 為基礎的圖表插件(Plugin)，可以用圖表視覺化呈現數據，支援動畫效果，且可以運作在所有支持 HTML5 的瀏覽器上。

7. ElasticSearch :

ElasticSearch 是建構在 Apache Lucene 的開源搜尋引擎，但比較起其前輩 Lucene，Elasticsearch 安裝簡單、以 JSON 做為模型，使用者可以透過 JSON 建立模型、索引資料、查詢資料，以及可以佈署在分散式架構等優點，讓 ElasticSearch 成為當前最熱門的開源搜尋引擎。

8. Redis :

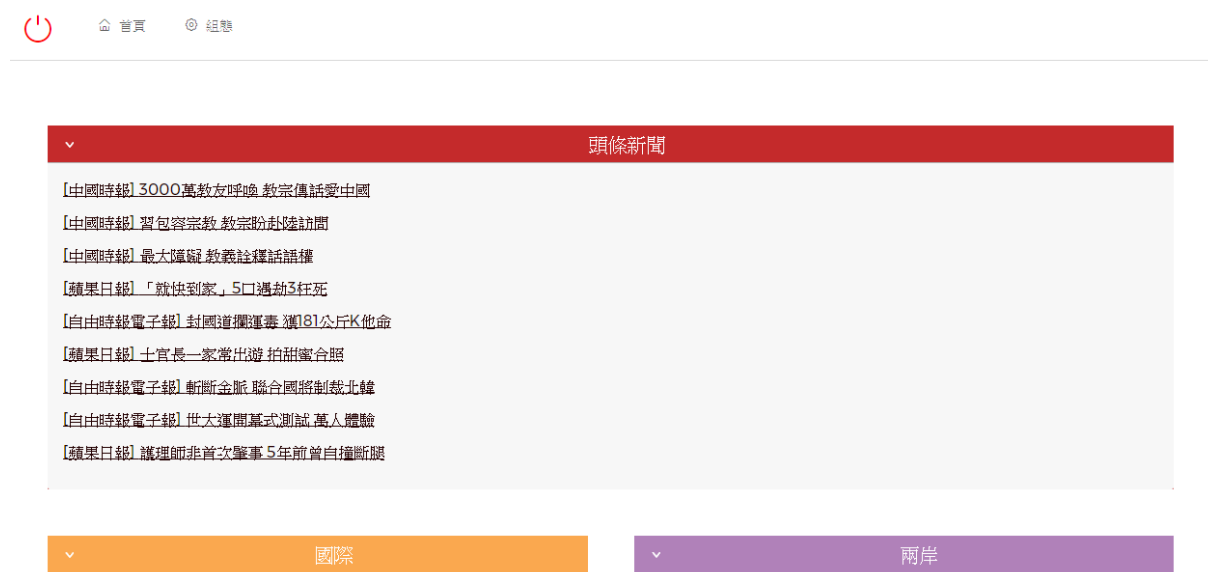
Redis 是一存放在記憶體中的非關聯式資料庫。不同於另外一個記憶

體的儲存方案 Memcached，其可存取較大物件與支援 SQL 語法，讓他成為記憶體資料庫的最佳選擇。因為該資料庫存放在記憶體中，因此可快速吞吐輸入與輸出資料，通常被用作儲存使用者的會話與購物車資訊，另外可以用來當作分散式叢集的佇列架構。

(三) 結果

1. 系統登入功能

輿情系統是一雲端服務，使用者只須要帳號與密碼，便可以順利登入系統，並開始取用各式分析模組。首先，在瀏覽器(支援 Chrome, Firefox 以及 IE 10 以)上輸入服務網址後進入雲端平台。



圖十、系統主頁儀表板

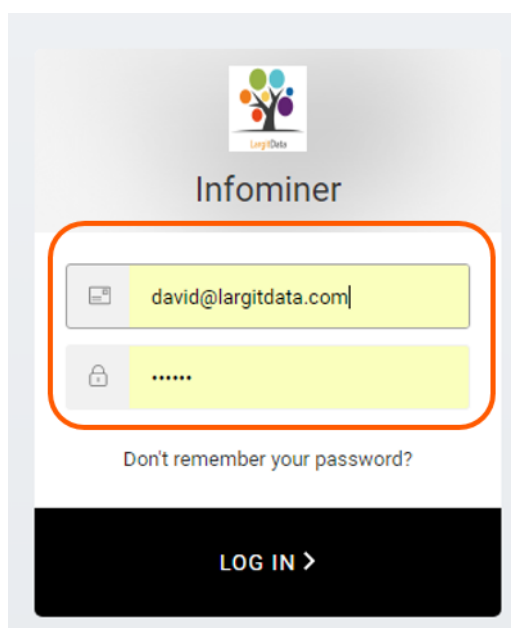
使用者便可以看到儀表板，呈列瀏覽頭條新聞、國際、兩岸、社會、財經、政治、地方等七大新聞版塊最新新聞資訊。接者，使用者便可以在左上角看

到系統登入按鈕。



圖十一、系統登入按鈕

點選登入按鈕以後，便可以進入到輸入帳密的頁面。



圖十二、帳號密碼輸入畫面

等待輸入正確帳號與密碼之後，便可以順利進入系統主畫面，並可以開始使用系統所提供的各項模組。



圖十三、操作系統畫面

2. 文章檢索功能

關鍵字群組設定

使用者可以設定關鍵字群組，並在每個群組下使用 AND, OR, NOT 等條件設定關鍵字篩選條件，只要透過設定關鍵字、自訂欲觀測的關鍵議題，就能隨時監控您所關心地網路輿情與訊息，讓您不再錯過第一手消息！



圖十四、關鍵字群組設定

動態搜尋功能

除了設定關鍵字群組外，使用者可以在首頁上方或搜尋文章功能列中，輸入文章關鍵字做即時資料搜尋，系統便會即時根據使用者輸入關鍵字呈現搜尋結果。



日期	標題	來源	版區	評論數	讀文數	噓文數	正向情緒	負向情緒	回覆數	操作
2017/11/12 15:09	[問卦] 台南高雄一直地震真的適合遷都?	ptt	Gossiping	29	7	11	1	1	0	檢視文章 檢視回應 分享
高雄台南最近一直地震 還有海水倒灌問題，地層下陷問題 登革熱 ，天氣爆熱爆濕等問題 真的適合當首都？還是只是想炒地皮？圖利自己？位置太差 也證明了，台中線拿的比台南高雄少 發展卻屢打台南高雄 台南高雄真的適合遷都？...										
2017/11/12 00:32	[創作] 邊錄記者事件簿之上吊紅衣女屍 (4)	ptt	marvel	12	7	0	0	3	0	檢視文章 檢視回應 分享
編？還是韓鄭或蚊子？因為想傳染 登革熱 ，所以被人一巴掌打死嗎？連交配都來不及交配，我一命嗚呼？」「閉嘴！你給我閉嘴！」她衝進浴室，火冒三丈地站在我的面前大叫，我睜睜地看著她，一種...										
2017/11/11 23:47	[火大] 被叮	ptt	Hate	8	0	0	1	1	0	檢視文章 檢視回應 分享

圖十五、文章列表

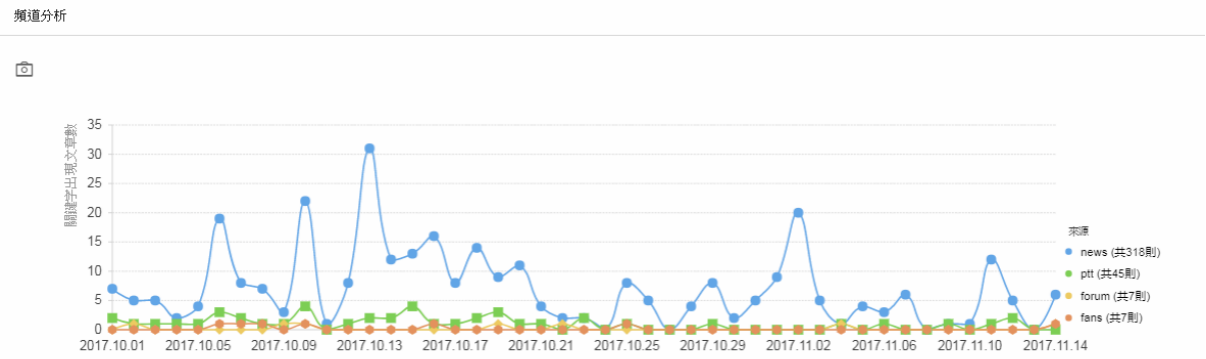
輿情聲量統計與媒體來源統計

完成欲觀測之關鍵字群組後，便可開始對這些設定好的關鍵字群的監控，每次登入系統首頁即能觀看這些議題的最新監測內容，包括所有有關該議題的輿情聲量統計圖(不同時間點的輿情數量折線圖)



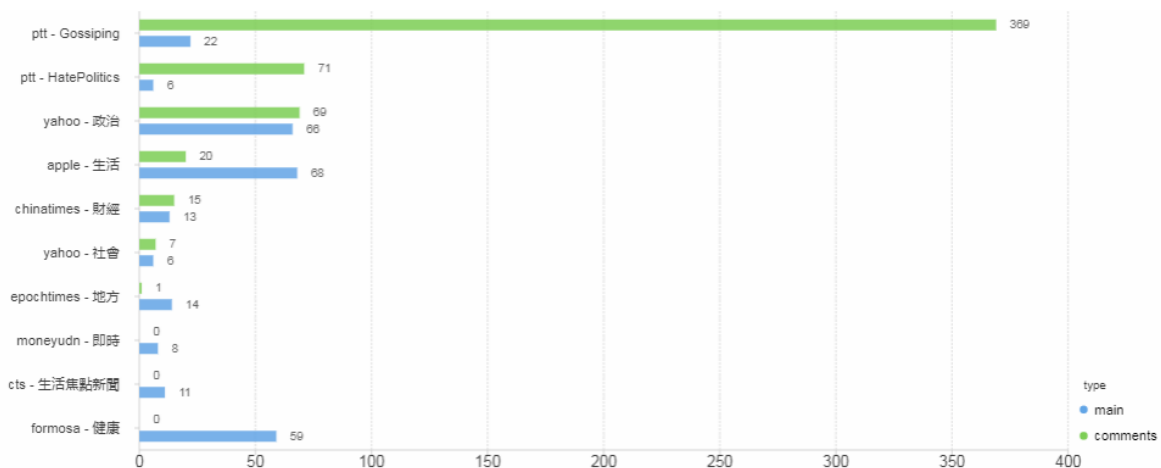
圖十六、文章聲量分析

與媒體來源統計圖(資料來自各來源的折線圖)。



圖十七、文章來源分析

與頻道分析統計圖(資料來自各頻道的主回文統計圖)。



圖十八、頻道分析

表列搜尋結果

完成欲觀測之關鍵字群組後，便可開始對這些設定好的關鍵字群的監控，每次登入系統首頁即能觀看這些議題的最新監測內容文章，包含文章標題、內文、來源頻道、發佈時間、推文數、閱覽數、情緒指標及作者等資訊；使用者亦可以點擊該標題，連結到原文位置。



圖十九、搜尋結果

當使用者點選後檢視文章時，可以瀏覽系統保存在資料庫中的原文內容。

新頭殼newtalk台北市今年登革熱境外移入病例數為全國第一，衛生局呼籲，若出現相關症狀應立即就醫！雖然時序已進入冬季，但境外移入登革熱病仍存在風險，台北市衛生局呼籲，在旅遊返台後如果出現發燒、倦怠、頭痛、肌肉痛、關節痛及後眼窩痛等症狀，應立即就醫，並且主動告知醫生旅遊史，避免延遲通報讓病毒傳播範圍擴大。衛生局指出，台北市今年截至11月13日，登革熱病例數已累計58例，皆為境外移入，高居全國之冠，感染源以東南亞國家為主，其中以馬來西亞12例最多，其次分別為菲律賓、越南各8例。台北市進一步發現部分個案於流行地區旅遊返台，出現發燒、倦怠、頭痛、肌肉痛、關節痛及後眼窩痛等症狀，輕忽症狀且自行服藥未即時就醫，使疫情隱藏期超過3日，甚至延遲通報達9日之多。依據台北市106年登革熱確診個案通報前就醫次數分析，由機場入境篩檢被通報計29名(占50%)、就醫1次被通報計20名(占34.5%)、就醫2次8名(占13.8%)，就醫3次1名(占1.7%)，資料顯示仍有15.5%的個案未被及時通報。台北市衛生局疾管科科长陳少卿表示，境外移入病例可能引發本土登革熱的風險，對市民健康及社區防疫造成極大的威脅，旅遊期間或回國後如出現發燒、頭痛、後眼窩痛、肌肉痛、關節痛、噁心及出疹等症狀，應儘速就醫，並主動告知醫師旅遊、活動史等；醫院與診所方面，應將旅遊史納入問診標準作業流程，以利早期診斷及治療，降低社區傳播風險。衛生局呼籲，國人前往登革熱流行地區旅遊、探親或經商時，應落實個人防蚊措施，於身體裸露處使用經衛福部核可的防蚊藥劑，並儘量穿著淺色長袖衣褲，以防遭病媒蚊叮咬。此外，提醒民眾平時應落實居家戶內外孳生源清除，除了避免病毒於社區傳播外，也能保護同住親友的健康。

關閉

圖二十、文章原文

當使用者點選後檢視回應時，可以瀏覽系統保存在資料庫中的推文內容。

1796789693953777:	2017/11/14 11:34
749730615229505: 這就是開放東南亞免簽的後果！	2017/11/14 11:35
10207815884714425: 北市不意外。	2017/11/14 11:46
1654562237948187: 哈哈 臺南登革熱市長也把蚊子帶去台北了 台北人幫QQ	2017/11/14 12:52

關閉

圖二十一、推文內容

儲存搜尋結果

此外，使用者可點擊下載檔案，以 CSV 打包符合查詢條件之文章內容。當點擊下載檔案按鈕後，系統會執行背景工作，將打包資料上載至檔案暫存區儲存搜尋結果，使用者可以等待資料打包完成後，前往檔案暫存區查看資料。

日期	標題	來源	版區	評論數	譯文數	噓文數	正向情緒	負向情緒	回覆數	操作
2017/11/14 15:20	新北登革熱群聚警報解除 赴東南亞仍要小心...	apple	生活	0	0	0	0	1	0	檢視文章 檢視回應

新北市先前發生本土登革熱群聚事件，共有7人遭感染，疾病管制署今表示，此波感染觀察至本月6日止，已無新發個案，研判警報解除；該署提醒，目前南部地區氣溫仍高，還是病媒蚊活躍季節，且登革熱境外移入國家.....以越南、泰國等東南亞國家為主，民眾仍要提高警覺。新北市鶯歌區自9月底開始出現有本土登革熱群聚，感染者從一開始的2人增到了7人，所幸至本月初已無新增病例，研判疫情已獲控制。疾管署副署長羅一鈞說，經分析此波.....登革熱群聚感染病毒為越南株，推測是境外移入病例導致，且該處有竹林、菜園等積水處，恐是造成疫情傳播的關鍵。羅一鈞說，截至上週今年共有295例登革熱境外移入病例，其中91例來自越南，提醒民眾赴該國要格外...

圖二十二、下載文章內容

瀏覽可下載檔案列表，點選下方的下載連結：Link，便可以開始下載 CSV 檔案

[Ptt]: 腸病毒

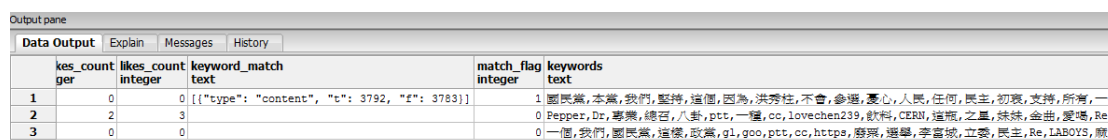
狀態(重新下載): 完成
建立時間: 2017-08-24 10:59:20
更新時間: 2017-08-24 10:59:21
日期區間: 2017-08-17 - 2017-08-24
下載連結: [Link](#)

圖二十三、下載資料連結

3. 進階探勘功能

產生文章關鍵字

當爬蟲蒐集到新文章時，系統會根據 TF-IDF 演算法，自動提取每篇蒐集文章中重要的關鍵字組（一至多個關鍵字），並將關鍵字清單增添在該筆資料的 keywords 欄位之中，使用者可以隨後取用該關鍵字作進一步的分析



	kes_count	likes_count	keyword_match	match_flag	keywords
	ger	integer	text	integer	text
1	0	0	[{"type": "content", "t": 3792, "f": 3783}]	1	國民黨,本黨,我們,堅持,這個,因為,洪秀柱,不會,參選,憂心,人民,任何,民主,初選,支持,所有,一
2	2	3		0	Pepper,Dr,專案,總召,八卦,ptt,一種,cc,lovechen239,飲料,CERN,這瓶,之真,妹妹,金曲,愛喝,Re,
3	0	0		0	一個,我們,國民黨,這樣,政黨,gl,geo,ptt,cc,https,廢票,選舉,李富城,立委,民主,Re,LABOYS,麻

圖二十四、文章關鍵清單分析

告警服務

系統提供即時警示系統，讓使用者可以透過 EMAIL 接受即時輿情警訊以及分析報表，以快速掌握每天發生的重要輿情。使用者可以透過警示設定畫面設定不同頻道（新聞、社群、媒體頻道）的警示時間區間、門檻、門檻依據欄位以及接受警示之使用者 EMAIL。



示警條件設定

* 示警名稱: 登革熱

* 信箱: david@largitdata.com

* 關鍵字群組: 登革熱

關鍵字: 登革熱

* 頻道: All

進階設定:

取消 確定

圖二十五、警示設定畫面範例

根據不同功能，本系統提供兩種不同的警示類型：

分群警示

自動摘要每天發生的議題，系統會根據各頻道當天所發生的新聞及輿情相似度，將同類輿情分至同一主題，並透過關鍵字抽取引擎，將同一群文章的關鍵字詞抽取出來，讓使用者透過群組關鍵字便可掌握每天不同主題輿情的輪廓。每個主題下面將會表列同一主題的相關文章，使用者可以點擊連結以進一步連結到原文。

重要關鍵字: 帽子 同性戀 愛滋

[Re: \[問卦\] 同性戀為什麼要打人恐同症的帽子](#)

[\[問卦\] 同性戀為什麼要打人恐同症的帽子](#)

[Re: \[問卦\] 同性戀為什麼要打人恐同症的帽子](#)

[Re: \[問卦\] 同性戀為什麼要打人恐同症的帽子](#)

第4群

重要關鍵字: 氣候 經濟部長 能源 台東 缺電 安全 考量 節能 李世光 需求面 經濟部 危機 需量 17 台灣 法定程序 核一 鐵軌

[\[新聞\] 經濟部長：核一不是「重啟」是「併聯」](#)

[\[新聞\] 「廢核沒有錢變」——李世光：用盡手段仍缺](#)

[\[新聞\] 李世光：極端氣候讓核一併聯發聲](#)

[\[新聞\] 經濟部長：核一不是「重啟」是「併聯」](#)

第5群

重要關鍵字: 外交 立法院 太平島 主權 中國 委員會 部署 安全 國家 海軍 建設 馮世寬 國防 江啟臣 國防部長 陳亭妃 大陸 安

全局 台北 軍事 外交部 軍艦 所需 內運 備詢 南海 綠委 海巡署 周美伍 國防部 通報 行政院 立委 部份

[\[新聞\] 應急備戰 國防部長：1個月內運4萬枚彈](#)

[\[新聞\] 應急備戰 國防部長：1個月內運4萬枚彈](#)

[\[新聞\] 應急備戰 國防部長：1個月內運4萬枚彈](#)

第6群

重要關鍵字: 老師 傳學

[\[問卦\] 會計學和數值學論難](#)

[Re: \[問卦\] 會計學和數值學論難](#)

[Re: \[問卦\] 會計學和數值學論難](#)

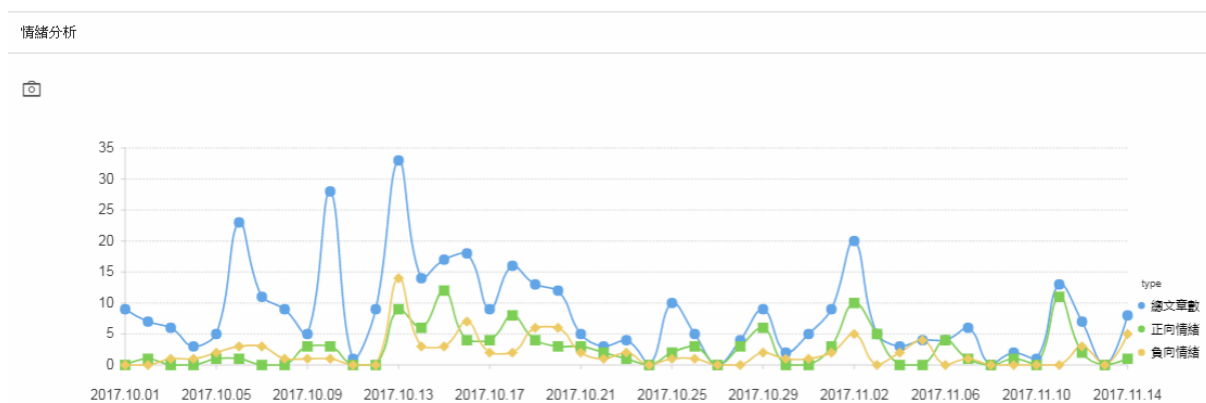
圖二十六、分群警示範例

關鍵字警示

若使用者有在系統設置相關關鍵字群組，則當出現與關鍵字條件相符的輿情時，系統便會自動發信告警使用者是否有符合關鍵字的文章出現，使用者可以點擊連結以連到原始文章。

情緒分析

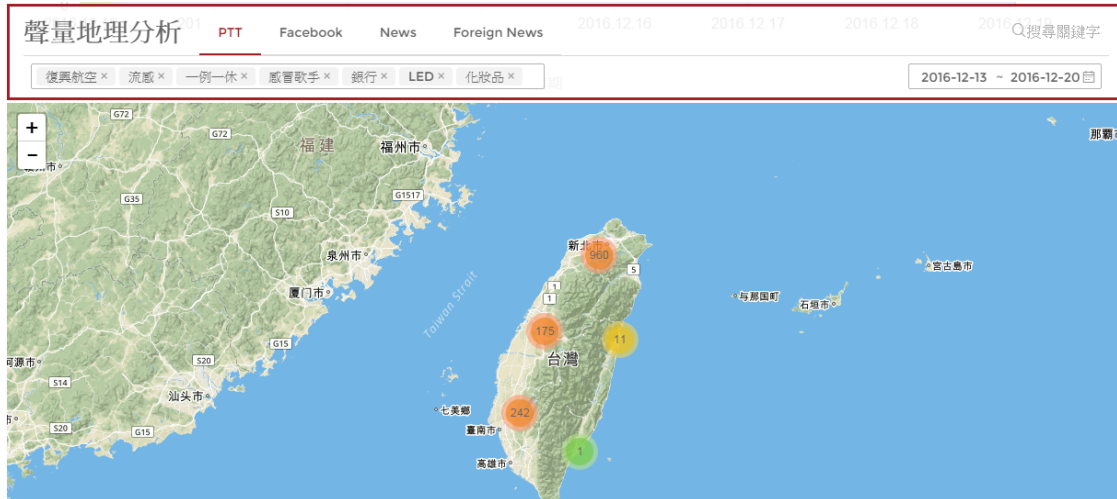
當使用者點擊切換情緒分析時，將會產生符合條件文章之正負情緒統計圖。文字的正面情緒與負面情緒，是我們取用文字中的形容詞後，根據文字出現文章的的正負面標籤以及文字出現的領域，計算出每個文字的正負面極性，若文字意表正面，則正面極性分數越高，但若文字意表負面，則負面極性分數越高。該圖則是透過計算該關鍵字所出現文章之正面極性減掉負面極性而得之，若相減結果為正，則將該文章列為正面情緒。反之，則列為負面情緒。該圖則同時表列所有文章數，讓使用者得以比較在同一議題上，一般群眾對該議題意表正面亦或負面。



圖二十七、使用者情緒分析

議題分群

而為了能夠更快速掌握輿情主題為何，本系統提供了摘要系統，供使用者可以根據文字相似度，分群議題事件，並抽取該議題的重要關鍵字，讓使用者快速了解每天有哪些重要的議題正在發生。



圖三十、地理聲量資訊呈現

意見領袖分析

輿情分析平台可以快速搜尋討論該關鍵字社群領袖，找尋各頻道中討論該關鍵字主題最多的使用者，讓您在該議題中快速挖掘出主要發文的意見領袖！系統會統計每個使用者的發文數，文章被按讚數、文章被噓數、回文數、正面情緒分數、負面情緒分數等，並綜合這些指標後，計算出領袖指數。藉由領袖指數的排序功能，我們即可快速找到討論該議題的可能領袖。

Author	Doc Count	Positive	Negative	Comments	Likes	Dislikes	Leader Score
TrueSpace	2	0	0	4	3	1	18
poplove700	2	2	4	0	0	0	16

圖三十一、意見領袖

點選使用者名稱後，便可以表列使用者的歷史 PO 文，我們即可透過觀察使用的歷史 PO 文，了解該使用者對議題的見解。

日期	標題	來源	版區	評論數	讚文數	噓文數	正向情緒	負向情緒	回覆數	操作
2017/09/28 2:12	[爆卦] 新品種的政治變色龍	ptt	Gossiping	0	0	0	0	3	0	檢視文章 檢視回應
2017/09/28 10:12	[新聞] 請中信：新品種的政治變色龍(王浩威)	ptt	Gossiping	0	0	0	2	1	0	檢視文章 檢視回應

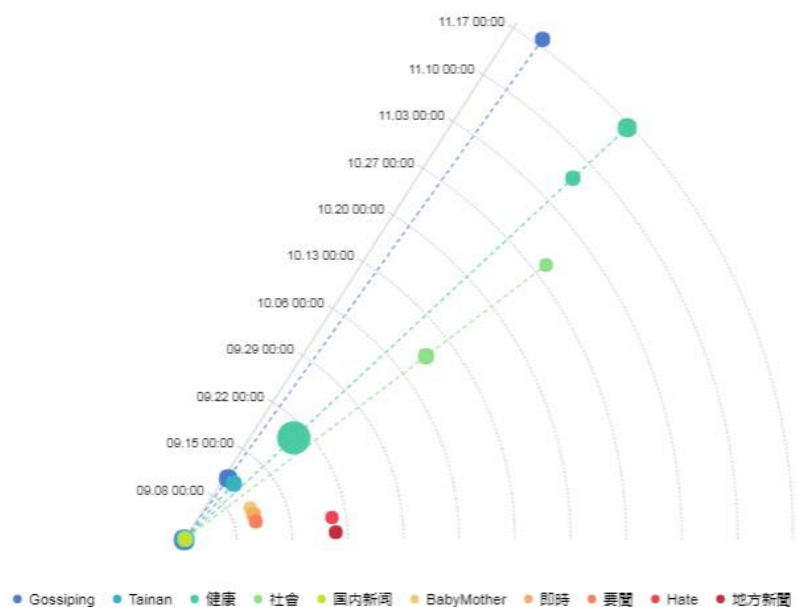
< 1 >

圖三十二、意見領袖發文內容

擴散分析

擴散分析報表可以讓使用者掌握輿情擴散程度與方向。透過擴散分析圖，我們可以發現有哪些頻道有持續關注特定事件。另外，可以從點的散佈程度觀測是否跟該議題相關的討論有擴散到其他頻道去。舉例來說，當我們以登革熱做為搜尋關鍵字時，便可以透過該擴散分析圖表發現即使今年台南並未有任何登革熱的病例，但 PTT 的台南版持續有民眾發佈跟登革熱議

題相關的文章，代表當地民眾對該議題的重視程度。



圖三十三、擴散分析圖表

4. 疫病同義辭典

於分析文字之前，如何能夠正確將文章分詞，並有校歸納同義詞才能提高文字分析的準確度。但是由於疫病的相關字詞眾多，我們無法一一篩選歸納同義詞，所幸維基百科上的志工在編寫條目時，除了原本字詞外，會用粗體字標示同義詞與縮寫，因而有黃純敏等學者提出利用維基百科建構中文縮寫詞與同義詞庫等構想[31]，於是我們便利用該構想，先使用網路爬蟲維基百科先將所有疫病的條目抓取下來，並剖析出粗體字位置(網頁標籤為的區域)，我們便可利用該同義字詞資訊建構同義字典，同義字典內容可以見附錄一。

5. Python 教育訓練

為提升疾管署對非結構化文字資料的處理與蒐集能力。我們提供疾管署包含：Python 基本教學、Python 網路爬蟲、Python 文字探勘等三大課程項目。課程時間如下：

表一、Python 教育訓練日程表

日期	課程名稱
106/07/05	Python 簡介與環境設定/物件與資料結構
106/07/19	運算式與陳述/方法與函式
106/07/26	物件導向程式設計/模組、套件
106/08/18	使用爬蟲採集資料清洗、整理採集資料（上）
106/08/23	清洗、整理採集資料（下），使用資料庫儲存與管理資料（上）
106/09/20	使用資料庫儲存與管理資料（下），爬蟲實戰
106/09/27	文字斷詞、自訂詞庫文本分類、分群（上）
106/10/11	文本分類、分群（下）
106/10/18	新詞探勘，語意分析（關聯字詞、同義詞）

透過九天的教學內容，我們討論了如何使用 Python 處理資料、爬取資料與分析資料的觀念，希冀疾管署同仁能透過課程內容理解該如何使用手邊 Python 開源工具，輔以豐富的第三方 Python 套件，能立刻分析手邊的非結構化資料。實際上課出席狀況則可以參考附錄二。

經統計，上課出席率約為 56.9%，而到課堂後期，作業繳交人數也逐次下降。經探討原因。歸咎於該教育訓練多偏重編程實務，對從未接受過編程訓練的同仁而言，課程難度偏高，如果沒有搭配到實際任務，很容易會因為編程門檻過高而放棄學習。建議貴署明年可以提出實際的分析任務，我們可以搭配任務提供諮詢，並舉辦工作坊讓同仁利用 Python 解決實際分析任務，貴署同仁便就會因為參與實際解題的過程而產生成就感，不但可引發同仁對編程的興趣，進而可增強學習意願，讓該課程不會淪為鴨子聽雷，而實際能在署內發揮成效。

(四) 討論

於建立輿情分析平台後，為了能夠凸顯輿情分析平台的特點，我們評比了 Google Trends, GPHIN, CDC 的通報紀錄與本計畫的輿情分析平台，得出以下比較表。

表二、各分析平台比較表

	Google Trends	CDC 疫情管理	Gphin	本計畫輿情分析平台
來源	網民在 google search 的搜尋關鍵字	國內各大醫院診所就診紀錄 國外已發生之疫情	全世界衛生相關 官方網站 全世界主流媒體	國內各大新聞媒體 指定中國官方新聞媒體 國內知名社群平台 1. Facebook 2. PTT

				<ul style="list-style-type: none"> 3. Dcard 4. Mobile01 5. Youtube
功能項目	<ul style="list-style-type: none"> 1. 被搜尋次數 2. 相關搜尋 3. 各地區搜尋熱度 (eg. 就登革熱和流感而言，中南部 > 北部) 4. 兩關鍵字熱度比較 	<ul style="list-style-type: none"> 1. 各疾病每日就診資訊 	<ul style="list-style-type: none"> 1. 主動示警通知 2. 專業人工審查 	<ul style="list-style-type: none"> 1. 關鍵字搜尋 2. 主動示警通知 (email) 3. 聲量地理分析
小結	<ul style="list-style-type: none"> 1. 顯示搜尋熱度並不一定能代表疫情 	<ul style="list-style-type: none"> 1. 實際就診紀錄 	<ul style="list-style-type: none"> 1. 約 1000 個來源 2. 涵蓋各國語言 	<ul style="list-style-type: none"> 1. 可深入社群現場 2. 可自由設定多組關鍵字，持續監控 3. 擷取內文、留

			3. 專職監控疾病相關 4. 語意分析 5. 反饋機制, 降低假警報發生 (False Alarm) 6. 每 15 分鐘更新	言, highlight 關鍵字 4. 每 5 分鐘更新 5. 摘要輿情內容 6. 情緒分析
--	--	--	--	---

1. Google Trends

Google Trends 搜集全世界網民在 google search 的搜尋關鍵字，但卻未包含社群的討論 eg. ptt, fb 的文章或留言，所以可以從 Google Trends 的數據窺探民眾目前的搜尋喜好，但確不能分辨關心的群眾來自於哪些來源。該平台的特色以下包含四個項目：

- 被搜尋次數
- 相關搜尋
- 各地區搜尋熱度 (ex. 就登革熱和流感而言，中南部 > 北部)
- 兩關鍵字熱度比較

為了解 Google Trends 的數據是否能用在疫情預測上，我們比較了 CDC 公佈的登革熱就診數據與 Google Trends 數據，發現搜尋熱度高低並不一定能代表疫情嚴重程度。經統計，截至 2017.11.13 為止，臺北市為登革熱確診病例最多的城市。但 google trend 卻顯示高雄市為關鍵字：『登革熱』

搜尋熱度最高的城市



圖三十四、2017 台灣各縣市關鍵字「登革熱」之搜尋熱度



圖三十五、2017 台灣各縣市關鍵字「登革熱」之就診數量

2. CDC 疫情管理

疾管署的通報系統涵蓋國內各大醫院診所就診紀錄以及已發生之疫情資訊，因為為實際就診之數量，所以可以用做對照組資料集，用以評比各數據平台的分析準確度。

3. Gphin

GPHIN 平台搜集全世界官方衛生相關網站與媒體資料，該平台內有許

多專業人工審查，因而可以將低誤報的機會。搭配主動示警功能，可以扮演一個早期預警的角色。由於涵蓋多個國家的資料來源，包含多語系的解讀，加上反饋機制，是信度跟校度最佳的防疫資訊平台。但該平台較未搜集其他社群媒體的訊息，無法從該平台了解是否有疫情謠言在坊間流傳。

4. 本計畫輿情分析平台

本計畫平台蒐集國內各大媒體新聞及國內知名社群(包含 Facebook, PTT, DCARD, Mobile01 平台)以及本計畫所指定的網站，而本平台提供良好的搜尋介面及分析平台，使用者可以依自己的觀察與需求任意以關鍵字查找及分析資料，若有即時訊息的需求，使用者可以訂閱示警內容，當一有不實謠言散佈時，系統可以即時警示使用者。該系統的優勢是因為有包含一般社群媒體，使用者可以用關鍵字以多角度去理解訊息的真偽，並可深入社群現場，了解謠言的起因，並能持續透過訂閱示警追蹤事件，是用以了解訊息傳遞脈絡，提高公共治理的輿情分析工具。

5. 平台搜尋結果比較

本計畫以登革熱為例，比較 Google Trend，實際疫情數據與本計畫輿情平台所監測到的訊息，試論述疫情與輿情的相關性，以及以數據說明這三數據來源在本質上的區別。

首先，我們於 Google Trends 上搜尋 2017 年的登革熱搜尋趨勢，發現 2017 年，民眾有持續關注登革熱議題，議題聲量集中在五月與八月（圖三十六）。



圖三十六、2017『登革熱』輿情聲量

而若以地區別排序搜尋量，則發現登革熱的搜尋量集中在於高雄與台南一帶，可見圖三十七。



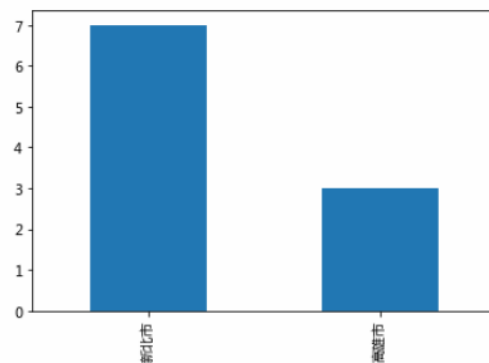
圖三十七、2017 台灣各縣市關鍵字『登革熱』之搜尋排行

但實際查詢疾管署所開放之「登革熱 1998 年起每日確定病例統計」(<https://data.gov.tw/dataset/21025>)，則可發現 2017 年的發病月份集

中在八月與十月（圖三十八），而發病地點則集中在新北市與高雄市，證明 Google Trends 上面的搜尋數據不足以支持用搜尋熱度可以預測疫情的推論。



圖三十八、2017 年登革熱病例數統計



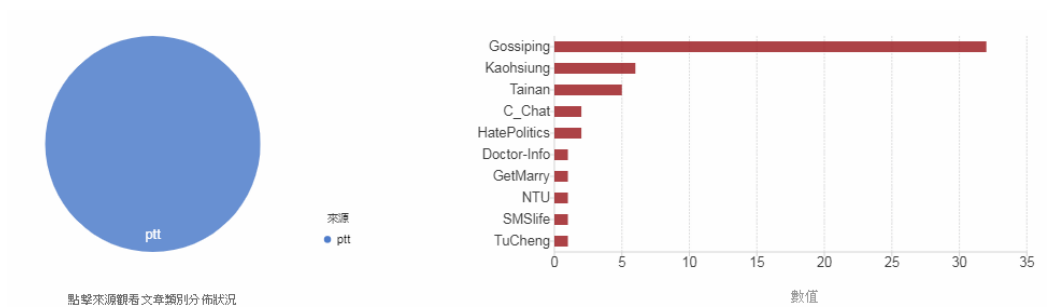
圖三十九、2017 年台灣各縣市登革熱病例統計數

而使用本計畫平台觀測輿情討論數量，則發現輿情討論聲量的高點出現在 2017/7/27，當天最多討論的新聞則為「高雄搶頭香 出現首例本土登革熱」（<http://www.chinatimes.com/realtimenews/20170727004952-260405>）。



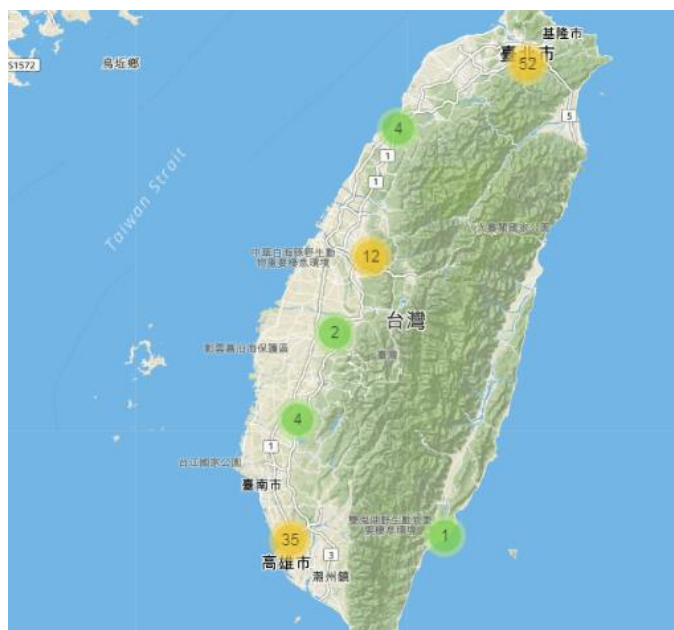
圖四十、分析輿情聲量高點

另外，經檢視輿情來源後，發現討論亦集中在台南與高雄相關的討論版，代表該輿情所監測到的聲量變化與 Google Trends 所觀測到的搜尋量有一定的相關度。



圖四十一、聲量頻道來源分析

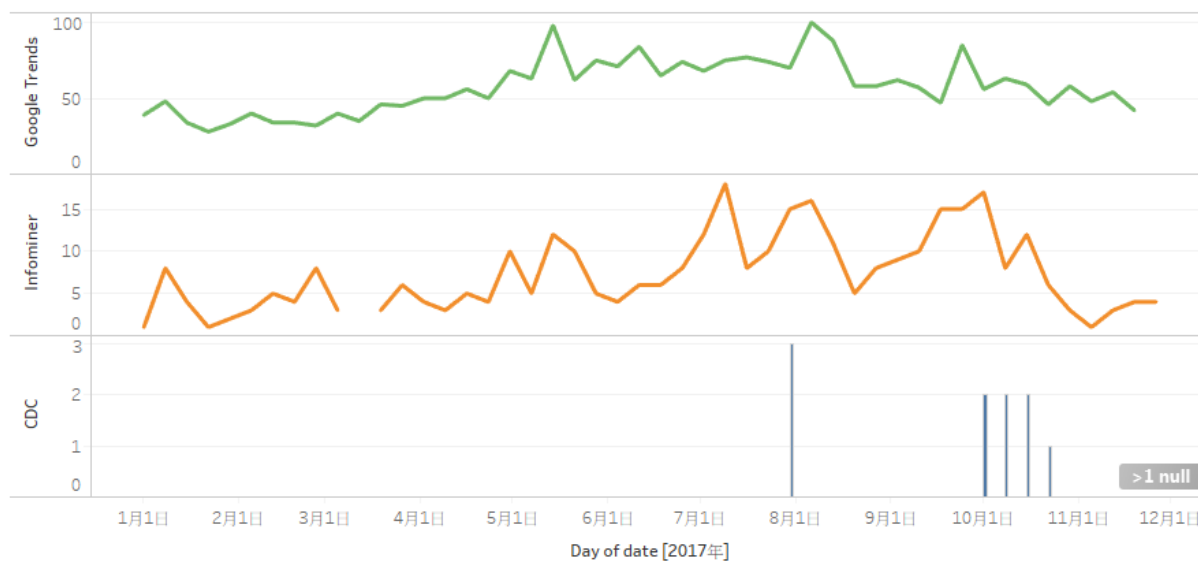
我們更進階的追蹤所有討論的來源 IP，並將該 IP 轉換為實際地址，發現疫情討論的相關輿情除北台灣外，的確集中於南台灣一帶。代表該地區民眾相當關心疫情的發展，但關心跟疫情是否發生並無任何顯著相關，意味者若是使用輿情用來判斷疫情是否會發生並不合理。



圖四十二、根據發言地理位置統計輿情數量

最後為了驗證上述論點，我們將 Google Trends、實際疫情發生數據與輿情觀測數目表列在同一張統計圖（圖四十三），從圖中可發現無論是 Google Trends 或是輿情觀測平台，討論登革熱的聲量皆發生在實際疫情發生後，證明 Google Trends 與輿情資料無法當作疫情預測的先行指標。

監測平台比較表

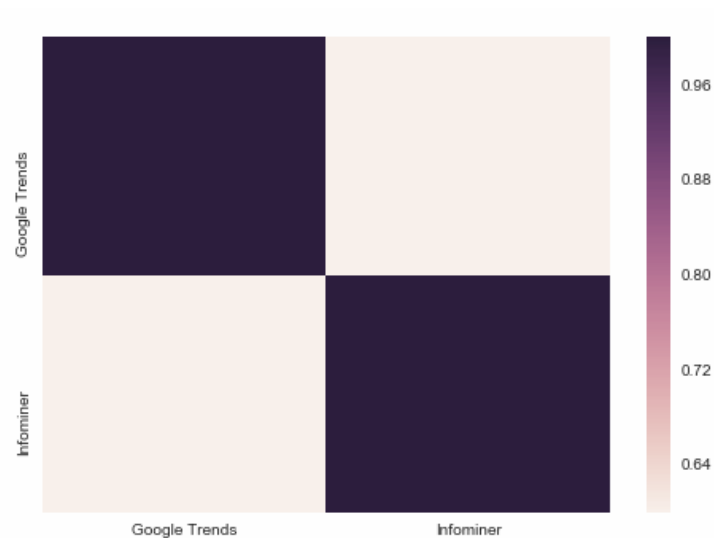


圖四十三、比較 Google Trends、輿情平台與疾管署通報疫情病例

但比較 Google Trends 與輿情監測平台，發現兩者之間的數據似乎有存在一定相關性，因此我們便使用了相關係數檢視兩者的相關性，結果發現兩者的相關係數為 0.6，代表 Google Trends 數據與輿情監測平台所觀測到的輿情聲量為中度相關(相關係數介於 0.6 ~ 0.8 之間)。

表三、Google Trends 與輿情觀測平台相關係數

	Google Trends	Infominer
Google Trends	1.0	0.6
Infominer	0.6	1.0



圖四十四、Google Trends 與輿情觀測平台相關性

但 Google Trends 統計的基準為網路民眾實際的搜尋量，而輿情觀測平台統計的是網民的發文數量，因此可以理解兩者之間並不會完全相關。

6. 輿情監測系統對機關效益

疾管機關除了關注疫情是否正在發生或擴散外，亦須了解民眾是否對疫病與防疫知識有正確認知。為了能夠掌握疫情討論的方向，避免網路上微不足道的訊息轉變為公關危機，關注網路輿情便是公部門提升公共治理品質的方法之一。然而以往在關注網路輿情時，多藉重於人力監測，但人力監測會衍生以下四個問題：1. 輿情搜集不夠全面、2. 蒐集速度不夠即時、3. 分析結果不夠準確、4. 訊息再利用不便利。

因此與其使用人力監測疫情相關輿情訊息外，透過系統自動化的搜集、清理、分析、視覺化，方能讓機關第一時間掌握重要輿情。但是，人

類語言博大精深，電腦只能用最簡單的統計方法量化輿情數據，確不能正確解讀字裡行間所蘊含的意義，因此，除了單靠輿情系統外，必需透過專家以質化方式解讀輿情訊息，方能從輿情訊息中正確解讀民眾的想法與輿情的走向。因此，我們認為此輿情系統可以成為署內一強而有力的工具，以協助署內同仁快速了解目前正在延燒的輿情方向；但工具本身還是要搭配署內人員的專業知識，對輿情數據提供質化分析，雙方相輔相成，方能有效提升公共治理品質。我們因此將輿情系統與人工之間的差異表列成以下比較表。

表四、輿情系統與傳統人工比較表

	傳統人工	輿情系統
資料蒐集廣度	只能單靠 Google 搜尋相關訊息，每日的搜尋量有限，也沒辦法一直關注各地輿情訊息。	可以搜集全世界上萬頻道的輿情訊息，掌握各地以及各頻道疫情相關輿情。
資料蒐集速度	人的搜尋速度緩慢，亦無法 24 小時不間斷關注各頻道的訊息來源，很可能會因為一時不察而漏接重要訊息。	透過分散式爬蟲 24 X 7 不間斷搜集各地輿情訊息，並可透過即時警訊寄送通知到權責人員信箱，確保最新訊息不漏接。

<p>輿情量化分析</p>	<p>以人工方式彙整各地訊息不便，如以手動方式整理不同來源的資訊，會導致分析緩慢，以致錯失分析黃金時間。</p>	<p>可透過分析工具快速產生、彙整大、雜、亂的輿情訊息，並可透過內建工具快速視覺化呈現輿情訊息，使用者便可以透過快速透過圖表及統計數字隨時掌握發燒議題。</p>
<p>資料利用程度</p>	<p>會因為不同人的資料留存習慣而導致歷史資料難以被搜尋或是被交換，讓歷史資料分析變得困難重重。</p>	<p>系統可以留存資料，提高資料的可利用率。讓權責人員之後還可以查閱過往資料。另外使用者可以將重要輿情保存下來，以便分析歷史輿情之用。</p>
<p>輿情質化分析</p>	<p>人類在解讀文字的能力尚樂勝當前的電腦技術，透過專家的解讀，才能發現輿情背後的真實意涵。</p>	<p>中文語意博大精深，電腦技術目前只能提供部分量化分析功能，尚無法有效理解文字背後所帶的深層蘊意。</p>

(五) 結論與建議

雖然目前已存在許多官方以及公共防疫監測平台，能夠透過監測實際數據，以及驗證新聞真實性來理解疫情的傳播方式，但本計畫希冀以一社群輿情的角度來重新理解疫情分析與公共治理的關聯性。因此本計畫平台同時搜集了官方媒體、新聞媒體以及社群媒體的資訊，希望能夠同時關注並比較三者的關聯，以釐清坊間關於疫病的傳言的真實性。而因為本平台已特別收納所有台灣知名的社群平台，因而會較國外知名平台更加有適地性質，能提供本地網民對疫病的詮釋分析。

本計畫所建立的平台，除了可以即時監測在網路上散播的訊息以外，亦將所有來源的資料整理成統一的綱要模型，因而可在輿情資料集套上一致的分析，讓使用者可以從聲量、來源、頻道、情緒等不同角度，理解關於疾病的輿情發生成因與熱點。讓權責人員可以綜觀當前的謠言散佈狀態與網民的討論方向。而因為所有統計資料亦有繪製成視覺化報表，透過圖表詮釋，權責人員便可以快速抓住輿情發展的大趨勢。

然而此分析成果目前都只有圍繞在公輿情上，尚未能仔細與實際發生的疫病資訊做相比對，了解民眾對當前疫病的看法是否有過度恐慌或過度臆測。未來應該可以建立一統一分析平台，彙整輿情與實際疫情發生的統計數據，將內外數據做一比對，便可以利用該資訊做為未來公共治理的決策依據。

另外，為了能夠觀測世界各國疫情相關輿情，本計畫將致力於發展多語系版本，提供多國語言的支援，並涵蓋更多國家的疫情來源，讓權責人

員可以更盡速掌握疫情相關輿情，並期望讓此輿情平台未來能扮演國際防疫平台的重要資料源之一。

(六) 重要研究成果及具體建議

根據計畫的規劃下，我們這今年完成了以下成果：

- 建構一自動輿情監控平台該平台能夠利用使用網路爬蟲技術即時監控來自新聞、論壇與社群媒體等各疫情資訊來源的最新訊息
- 設立一疫情輿情相關的繁體中文同義字典，讓系統得以歸納包含同義關鍵字的輿情資訊。
- 系統之後必須能夠將輿情資訊儲存到一非關聯式資料庫中，供使用者能夠存取該輿情資訊
- 系統能將這些儲存於非關聯式數據庫中的輿情訊息匯入到搜索引擎之中，讓使用者得以透過搜索引擎的介面上下關鍵字查找與疾病管制相關的所有輿情
- 系統需能夠搜集該輿情資訊等來源、評論數、按讚數、回文數等不同指標，讓執行單位得以用不同指標排序或檢視最具威脅性的謠言。
- 系統能夠將輿情資訊根據聲量、頻道來源、情緒指標、擴散路徑等資訊繪製視覺化圖表，讓使用者得以視覺化方式探索輿情資料。
- 舉辦 Python 教育訓練課程，提升疾管署對非結構化文字資料的處理與蒐集能力。第一年的課程包含：Python 基本教學、Python 網路爬蟲、Python 文字探勘等三大課程項目。

於第一年的任務中，我們已經建立好該輿情服務平台，唯需要增添更多相關資料來源，並設定關鍵字，權責人員便可以在第一時間利用系統蒐整疫情相關輿情。而由於本系統亦有建立即時警示系統，相信如果相關人員有妥善設置警示系統，便可以第一手掌握最新疫情輿情訊息。而在此年計畫中，因為投入較多心力在系統的建置，尚未有機會接觸第一線的觀測人員，了解他們該如何使用該平台，協助他們蒐集最新相關輿情。未來將在第二三年的計畫中陸續安排訪談，了解他們該如何使用該平台協助他們觀測最新疫情相關輿情。

本計畫在第一年的目標中，已增添許多中文輿情來源網站，並增加對中文語言的搜集、處理、分析與視覺化等功能，但防疫並非只是國內任務，有鑒於我們國家國民亦常前往東南亞國家或其他歐美國家交流與旅遊，如何快速掌握當地是否亦有疫情爆發，亦是國內主管機關單位所必須關注的議題，因此我們會在未來計畫致力於發展多語系版本，提供中文、英文、東南亞各國語世界其他各國語言的支援，並涵蓋更多國家的疫情來源，致力成為國際防疫平台的重要資料源之一。另外，為了能夠更理解輿情與疫病之間的關係，未來建議貴署建一分析平台，同時納入輿情、官方媒體與實際疫病統計數據，三者交叉比較，希冀能產生一些具公信力指標做為公共治理的指標。

(七) 參考文獻

- [1] V. Mayer-Schonberger, K. Cukier, 大數據, 林俊宏, 譯者) 台北市: 天下文化 (2013).
- [2] 陳亭愷, 國道計程電子收費實施後之網路輿情文本情感分析研究, 運輸管理學系碩士班, 淡江大學, 2015, pp. 122.
- [3] 鄒函升, 新聞輿情與民意偵測追蹤之研究—大資料之研究取向, 資訊管理研究所, 國立政治大學, 2013, pp. 73.
- [4] 林婉茹, 利用全民健康保險資料庫建置門診類流感監測系統, 醫務管理學研究所, 長榮大學, 2009, pp. 94.
- [5] 柳姚仁, 運用 Facebook 公開資料監測類流感疫情, 資訊管理學系碩士在職專班, 淡江大學, 2015, pp. 43.
- [6] 林世然, 大數據分析應用於登革熱疫情趨勢之研究, 電機工程系博碩士班, 國立高雄應用科技大學, 2015, pp. 70.
- [7] 吳和生, 莊人祥, 張筱玲, 我國傳染病監測系統簡介, 學校衛生護理雜誌 (2010) 51-58.
- [8] K.J. Henning, What is syndromic surveillance?, *Morbidity and Mortality Weekly Report* (2004) 7-11.
- [9] R. Heffernan, F. Mostashari, D. Das, A. Karpati, M. Kulldorff, D. Weiss, Syndromic surveillance in public health practice, *New York City, Emerg Infect Dis* 10 (2004) 858-864.
- [10] L. Steiner-Sichel, J. Greenko, R. Heffernan, M. Layton, D. Weiss, Field investigations of emergency department syndromic surveillance signals—New York City, *Morbidity and Mortality Weekly Report* (2004) 184-189.
- [11] R.C. Jones, M. Liberatore, J.R. Fernandez, S.I. Gerber, Use of a prospective space-time scan statistic to prioritize shigellosis case investigations in an urban jurisdiction, *Public health reports* (2006) 133-139.
- [12] M. Kulldorff, R. Heffernan, J. Hartman, R. Assunção, F. Mostashari, A space-time permutation scan statistic for disease outbreak detection, *PLoS Med* 2 (2005) e59.
- [13] H. Chen, R.H. Chiang, V.C. Storey, *Business Intelligence and Analytics: From Big Data to Big Impact*, *MIS quarterly* 36 (2012) 1165-1188.
- [14] P. Zikopoulos, C. Eaton, *Understanding big data: Analytics for enterprise class hadoop and streaming data*, McGraw-Hill Osborne Media 2011.
- [15] 蕭元哲, 葉上葆, 電子化政府之使用行為分析: 檔案分析法, *Electronic Commerce Studies* 1 (2003) 207-224.

- [16] H. Chen, M.-T. Lin, 陳祥, 林明童, 我國 [電子化政府整合型入口網站] 使用者行為分析, 圖書館學與資訊科學 28 (2002).
- [17] J. Caverlee, L. Liu, D. Buttler, Probe, cluster, and discover: Focused extraction of qa-pagelets from the deep web, Data Engineering, 2004. Proceedings. 20th International Conference on, IEEE, 2004, pp. 103-114.
- [18] Z. Nie, S. Kambhampati, A frequency-based approach for mining coverage statistics in data integration, Data Engineering, 2004. Proceedings. 20th International Conference on, IEEE, 2004, pp. 387-398.
- [19] O. Etzioni, The World-Wide Web: quagmire or gold mine?, Communications of the ACM 39 (1996) 65-68.
- [20] Y. Zhang, Y. Zhang, The Study on the Governmental Tactics of Persuasion of Network Public Sentiment, 2013 International Conference on Public Management (ICPM-2013), Atlantis Press, 2013.
- [21] R. Cooley, B. Mobasher, J. Srivastava, Data preparation for mining world wide web browsing patterns, Knowledge and information systems 1 (1999) 5-32.
- [22] 廖洲棚, 陳敦源, 蕭乃沂, 廖興中, 運用巨量資料實踐良善治理: 網路民意導入政府決策分析之可行性研究, 國家發展委員會, 2014, pp. 95.
- [23] S. Shindelar, Big Data and the Government Agency, Public Manager 43 (2014) 52.
- [24] United Nations Global Pulse, Big Data for Development: Challenges & Opportunities, New York, 2012.
- [25] 林蔚君, 詹雅慧, 巨量資料分析與 GIS 應用價值創造, 國土資訊系統通訊 88 (2013) 12-20.
- [26] M.A. Pirog, Data will drive innovation in public policy and management research in the next decade, Journal of Policy Analysis and Management 33 (2014) 537-543.
- [27] W.-Y. Ma, K.-J. Chen, Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff, Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17, Association for Computational Linguistics, 2003, pp. 168-171.
- [28] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Advances in knowledge discovery and data mining, (1996).
- [29] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to data mining, Pearson Addison Wesley

Boston2006.

[30] Ginsberg, Jeremy; Mohebbi, Matthew H.; Patel, Rajan S.; Brammer, Lynnette; Smolinski, Mark S.; Brilliant, Larry (19 February 2009). "Detecting influenza epidemics using search engine query data". *Nature*. 457 (7232): 1012–1014. PMID 19020500. doi:10.1038/nature07634.

[31] 黃純敏, 李亞哲, & 陳柏宏. (2015). 以維基百科為基礎之中文縮寫詞與同義詞庫建構. 資訊管理學報, 22(2), 117-140.

肆、附錄

(一)附錄一：疫病同義詞典

	疾管署法定傳染病	同義字
0	天花	天花
1	嚴重急性呼吸道症候群	嚴重急性呼吸道症候群, 严重急性呼吸道综合征, SARS, 沙斯病, 煞斯病, 沙士, 非典, 薩斯病, 萨斯病
2	鼠疫	鼠疫, 黑死病
3	狂犬病	疯狗症, 瘋狗症, 狂犬病
4	登革熱	断骨热, 斷骨熱, 骨痛热症, 骨痛熱症, 天狗熱, 天狗热, 登革热, 登革熱, 熱出血症, 热出血症, 蚊症, 花花公子熱, 花花公子热
5	德國麻疹	風疹, 风疹, 三日麻疹, 德国麻疹, 德國麻疹
6	霍亂	虎列刺, 虎列拉, 落吐症, 霍痢拉, 癘螺痧, 瘧螺痧, 霍亂, 霍乱
7	流行性斑疹傷寒	流行性斑疹傷寒, 流行性斑疹伤寒
8	白喉	白喉
9	流行性腦脊髓膜炎	流行性腦膜炎, 流行性腦膜炎, 流行性腦脊髓膜炎, 流行性腦脊髓膜炎, 流脑, 流腦
10	西尼羅熱	西尼羅熱, 西尼罗热, 西尼罗河病毒, 西尼羅

		河病毒
11	傷寒	伤寒热, 傷寒熱, 腸熱症, 肠热症, Typhoid Fever, 濕溫傷寒, 湿温伤寒, 腸傷寒, 肠伤寒, 傷寒, 伤寒, 腸窒扶斯, 肠窒扶斯
12	副傷寒	副傷寒, 副伤寒
13	小兒麻痺症/急性無力肢體麻痺	小兒麻痺症, 小儿麻痹症, 脊髓灰质炎, 脊髓 灰質炎, 急性灰白髓炎
14	桿菌性痢疾	志贺杆菌病, 志賀桿菌病, 细菌性痢疾, 桿菌 性痢疾, 杆菌性痢疾
15	阿米巴性痢疾	阿米巴痢疾, 阿米巴性痢疾
16	瘧疾	發瘧子, 发疟子, 瘧疾, 疟疾, 打擺子, 打摆子, 冷熱病, 冷熱病, 羅馬熱, 罗马热
17	麻疹	麻疹, 癩疹
18	急性病毒性 A 型肝炎	甲型肝炎, A 型肝炎, 急性病毒性 A 型肝炎, 甲肝
19	漢他病毒症候群	漢他病毒症候群, 汉他病毒症候群, 汉他病, 漢他病, 腎綜合征出血熱, 腎綜合徵出血熱, 流行性出血熱, 流行性出血熱
20	多重抗藥性結核病	多重抗藥性結核病, MDR-TB
21	屈公病	屈公熱, 屈公热, 屈公病, 契昆根亞熱, 契昆根

		亞熱, 基孔肯雅熱, 基孔肯雅症
22	炭疽病	炭疽病
23	先天性梅毒	先天性梅毒, 梅毒, 廣東瘡, 广东瘡, 楊梅瘡, 楊梅瘡
24	腸病毒感染併發重症	腸病毒感染併發重症, 腸病毒感染并發重症, Enterovirus, Coxsackievirus groups A and B, Echovirus, 腸道病毒, 肠道病毒, 腸病毒, 腸病毒,
25	結核病	TB, 結核病, 结核病
26	人類免疫缺乏病毒感染	後天免疫缺乏症候群, 人類免疫缺乏病毒感染, AIDS, 愛滋病, 艾滋病, 获得性免疫缺陷综合征, 获得性免疫缺陷综合征, 人类免疫缺乏病毒感染
27	漢生病	麻瘋病, 韓森氏病, 癩瘋病, 癩病, 麻瘋病, 癩瘋病, 癩病, 癩病, 漢生病, 汉生病, 韓森氏病, 癩風, 癩病
28	百日咳	頓咳, 顿咳, 百日咳
29	新生兒破傷風	新生兒破傷風, 新生儿破伤风
30	破傷風	四六風, 四六风, 七日風, 七日風, 臍帶風, 脐带风, 破傷風, 破伤风

31	急性病毒性 B 型肝炎	急性病毒性乙型肝炎, 急性病毒性 B 型肝炎, B 肝, 乙肝
32	急性病毒性 C 型肝炎	急性病毒性丙型肝炎, 急性病毒性 C 型肝炎, C 肝, 丙肝
33	急性病毒性 D 型肝炎	急性病毒性丁型肝炎, 急性病毒性 D 型肝炎, D 肝, 丁肝
34	急性病毒性 E 型肝炎	急性病毒性戊型肝炎, 急性病毒性 E 型肝炎, E 肝, 戊肝
35	流行性腮腺炎	流行性腮腺炎, 耳下腺炎, 痄腮
36	梅毒	廣東瘡, 广东疮, 楊梅瘡, 杨梅疮, 梅毒
37	淋病	淋病, 白濁, 白浊
38	侵襲性 b 型嗜血桿菌 感染症	侵襲性 B 型嗜血桿菌感染症, 侵袭性 B 型嗜血桿菌感染症, 侵襲性 b 型嗜血桿菌感染症, 侵袭性 b 型嗜血桿菌感染症, 嗜血桿菌, 拜菲爾氏菌, 嗜血桿菌, 流感嗜血桿菌, 費佛氏桿菌, 流感嗜血桿菌, 流感桿菌, 费佛氏桿菌, 拜菲爾氏菌, 流感桿菌,
39	退伍軍人病	軍團病, 军团病, 退伍軍人病, 退伍軍人症, 退伍軍人症, 退伍軍人病
40	先天性德國麻疹症候 群	先天性德國麻疹症候群, 先天性風疹症候群, 先天性风疹症候群, 先天性德国麻疹症候群

41	日本腦炎	日本腦炎, 流行性乙型腦炎, 流行性乙型腦炎, 日本腦炎
42	流感併發重症	流感併發重症, 流感併發重症
43	肉毒桿菌中毒	肉毒桿菌中毒, 肉毒桿菌中毒, 肉毒桿菌感染, 肉毒桿菌感染
44	庫賈氏病	Creutzfeldt-Jakob disease, 庫茲菲德-雅各氏症, 庫雅氏症, 庫雅氏症, 庫賈氏病, 庫賈氏病, 庫賈氏症, 庫賈氏症, 庫茲菲德-雅各氏症, 克雅二氏病, 克雅二氏症
45	鈎端螺旋體病	鈎體病, 鈎端螺旋體病, 鈎體病, 細螺旋體症, 鈎端螺旋體病, 細螺旋體症
46	萊姆病	萊姆疏螺旋體病, 萊姆病, 萊姆疏螺旋體病, 萊姆病
47	類鼻疽	惠特莫爾氏病, 奈特克里夫園丁病, 惠特莫爾氏病, 奈特克里夫園丁病, 類鼻疽, 類鼻疽, 類鼻疽伯克氏菌, 類鼻疽伯克氏菌, 類鼻疽假單胞菌, 類鼻疽假單胞菌
48	地方性斑疹傷寒	地方性斑疹傷寒, 地方性斑疹傷寒, 斑疹傷寒, 斑疹傷寒, 鼠型斑疹傷寒, 鼠型斑疹傷寒
49	Q熱	Q型流感, 寇熱, Q熱, 羊流感, Q熱, Goat flu, 寇熱, Q fever

50	水痘併發症	水痘併發症, 水痘併发症
51	恙蟲病	叢林斑疹傷寒, 沙虱熱, 沙虱熱, 恙蟲病, 恙虫病, 丛林斑疹伤寒
52	兔熱病	兔熱病, 兔熱病, 野兔病, 鹿蠅熱, 大原病, 法蘭西斯氏病, 野兔病, 鹿蠅熱, 大原病, 法蘭西斯氏病
53	疱疹 B 病毒感染症	疱疹 B 病毒感染症, 猴疱疹病毒, 猴疱疹 B 病毒, B 疱疹病毒, Herpes B virus, herpesvirus simiae, herpes simiae, monkey B virus, Cercopithecine herpesvirus 1
54	弓形蟲感染症	弓形蟲感染症, 弓形虫感染症, 弓蟲症, 弓虫症
55	布氏桿菌病	波狀熱, 馬耳他熱, 布氏桿菌病, 地中海弛張熱, 波浪熱, 馬耳他熱, 布魯氏菌病, 地中海弛張熱, 布魯氏菌病, 波浪熱, 波狀熱
56	新型 A 型流感	豬流感, 墨西哥流感, 豬流感, 人類豬 (型) 流感, H1N1 新型流感, 甲型 H1N1 流感, 豬源流感, 新型 A 型流感, 新流感, swine flu, A 型 H1N1 流感, 2009 年 H1N1 流感, 甲型流感病毒 H1N1 亞型, 北美流感, 豬源流感, 人類豬

		(型) 流感, 新型甲型 H1N1 流感, Influenza A virus subtype H1N1, A 型流感病毒 H1N1 亞型, 新型 A 型 H1N1 流感
57	茲卡病毒感染症	茲卡病毒感染症, 寨卡病毒感染症, 寨卡熱, 茲卡病毒, 茲卡病毒, 寨卡病毒
58	中東呼吸症候群冠狀病毒感染症	中東呼吸症候群冠狀病毒感染症, 中東呼吸症候群, 中东呼吸综合征, MERS
59	黃熱病	la fièvre jaune, 黄杰克, Yellow Fever, Yellow Jack, Yellow Plague, 黄热病, 黃熱病, 黄傑克, 黑嘔, 黑呕, 美洲瘟疫
60	伊波拉病毒感染	伊波拉病毒感染, 伊波拉出血熱, 埃博拉出血熱, 伊波拉病毒病, 埃博拉病毒病, 伊波拉, 埃博拉, 伊波拉病毒屬, 伊波拉病毒, 埃博拉病毒
61	拉薩熱	拉萨出血熱, 拉萨熱, 拉薩出血熱, 拉薩熱, Lassa fever, LHF, Lassa hemorrhagic fever
62	馬堡病毒出血熱	馬堡病毒出血熱, 马尔堡出血熱, 馬爾病毒, 馬爾堡病毒, 马尔堡病毒, 馬堡病毒, 马尔病毒, 马堡病毒
63	裂谷熱	里夫谷熱, 里夫谷熱, 裂谷熱, 裂谷熱

64	貓抓病	貓抓病, 猫抓病
65	NDM-1 腸道菌感染症	NDM-1 肠道菌感染症, NDM-1 腸道菌感染症, 多重抗藥性腸道菌感染症, 多重抗药性肠道菌感染症
66	發熱伴血小板減少綜合症	發熱伴血小板減少綜合徵, 蜱蟲病, 发热伴血小板減少綜合症, 發熱伴血小板減少綜合症, 蜱虫病
67	常見腸道寄生蟲病簡介	常見腸道寄生蟲病簡介
68	中華肝吸蟲感染症	中華肝吸蟲感染症
69	旋毛蟲感染症	旋毛蟲感染症, 旋毛虫症, 旋毛蟲, 豬肉蟲, 旋毛虫, 猪肉虫
70	肺吸蟲感染症	卫氏肺吸虫, 衛氏肺吸蟲, 肺吸蟲感染症, 衛氏並殖吸蟲, 卫氏并殖吸虫
71	廣東住血線蟲感染症	廣東住血線蟲感染症, 廣州管圓線蟲病, 广州管圓线虫
72	鸚鵡熱	鸚鵡熱, 鸚鵡热
73	病毒性腸胃炎	肠胃型感冒, 病毒性腸胃炎, 病毒性肠胃炎, 腸胃型感冒
74	沙門氏菌感染症	沙门氏菌感染症, 沙门氏杆菌病, 沙門氏桿菌

		病, 沙門氏菌感染症
75	疥瘡感染症	疥疮, 疥瘡, 疥瘡感染症
76	頭蝨感染症	头虱感染, 頭蝨感染, 頭蝨感染症, 头虱感染 症
77	李斯特菌症	李斯特菌症, 李斯特菌, 单核细胞增生性李斯 特菌, 單核細胞增生性李斯特菌, 李氏菌
78	隱球菌症	隱球菌症, 隐球菌病, 隱球菌病
79	CRE 抗藥性檢測	CRE 抗藥性檢測, CRE 抗药性检测, 抗碳青 霉烯類腸桿菌, Carbapenem resistant enterobacteriaceae
80	VISA/VRSA 抗藥性檢 測	VISA 抗藥性檢測, VISA 抗药性检测, VRSA 抗 藥性檢測, VRSA 抗药性检测, 萬古黴素敏感 性減低金黄色葡萄球菌, vancomycin- intermediate S. aureus, VRSA, 抗萬古黴素 金黄色葡萄球菌, vancomycin-resistant S. aureus
81	肺囊蟲肺炎	肺囊虫肺炎, 肺囊蟲肺炎
82	淋巴絲蟲病	淋巴丝虫病, 淋巴絲蟲病, 象皮病

(二)附錄二、教育訓練出席統計

節次	一	二	三	四	五	六	七	八	九	總和
應到 (人)	38	38	38	37	37	37	37	36	36	334
實到 (人)	33	28	24	18	21	20	20	16	10	190
出席率 (%)	86.8 %	73.7 %	63.2 %	48.6 %	56.8 %	54.1 %	54.1 %	44.4 %	27.8 %	56.9 %

作業繳交統計

次數	一	二	三	四	五	六	七
繳交人數 (人)	25	18	13	9	6	3	3