計畫編號：MOHW107-CDC-C-114-124113

衛生福利部疾病管制署 107 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

107 年 度/全 程 研 究 報 告

執行機構：中華 R 軟體學會

計畫主持人：陳嬿如

研究人員：丘祐瑋

執行期間：107 年 1 月 1 日至 107 年 12 月 31 日

研究經費：新臺幣 壹仟貳佰零捌萬元 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..... 3
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 11
1. 輿情系統架構 11
2. 資料蒐集 16
3. 資料處理 25
4. 資料分析 30
5. 系統環境 36
(三)　結果 43
1. 擴展大陸輿情監測範圍 43
2. 擴增簡體中文字典 51
3. 簡繁搜索功能 55
4. 擴展英文媒體輿情監測範圍 56
5. 建立輿情篩選模型 59
6. R 語言教育訓練..... 60
(四)　討論 66
1. 國內疫病輿情監測與分析 66
2. H7N9 疫病輿情監測與分析..... 68
(五)　結論與建議 75
(六)　重要研究成果及具體建議 78
(七)　參考文獻 81

壹、研究報告中文摘要

網路世界的發達讓訊息的傳播力變得無遠弗屆，值得慶幸的是，如果現在要蒐集與一個疫病相關的新聞訊息，只需要在電腦前點點手指，就可以立即蒐集到大量的新聞與社群討論訊息，但值得憂心的是：1. 幾乎每分每秒都有新的輿情訊息產生，如何快速蒐集、消化這麼龐大的資訊量，並辨別真正有價值的資訊變得比以往困難許多。2. 在網路世界裡，任何訊息只要透過有心人士的傳播或操弄，這些訊息就可能快速在網路世界散播，今天如果被創造的訊息是帶有負面影響力的網路謠言，將造成人民不必要的恐慌與擔憂。因此，如何快速掌握、處理各訊息來源的疫病輿情，並能迅速過濾、通報具有價值的訊息，便是本研究需要克服的課題。

本計畫為四年期的研究計畫，延續了去年關於中文疫病的輿情蒐集計畫，本年度的計畫將以擴充輿情蒐集的廣度與深化輿情系統訊息的分析、通報與篩選機制做為本年度計畫主軸。在擴充輿情資料蒐集的廣度上面，除了增加繁中、簡中的資料蒐集頻道外，本年度另外蒐集了東協國家的官方英文媒體，擴充外國國家的資訊來源，希冀透過監控外文媒體的訊息，充分瞭解國外與疫病相關的輿情。在深化輿情系統的分析、通報與篩選機制上，本研究先解決了在中文分析上，簡繁字詞的歧異，與處理掉中文同義字詞的問題，確保系統能有效處理中文字詞資訊。在通報與篩選部分，本計畫在系統上建置了 Line 與 Email 等通知管道，使用者可以在設定關鍵字後，便可以透過資訊設備迅速收到與關鍵字相關的輿情訊息。為了能夠篩選出真正重要的情資，本研究亦加上機器學習功能，透過電腦自動篩選輿情，減少人

工閱讀的時間。

中文關鍵詞(至少三個)：

輿情分析、輿情監測、非結構性資料、文字探勘

貳、研究報告英文摘要

The booming of the internet now makes messages travel faster than ever before. The upside of this phenomenon is that now people can quickly gather messages from the internet with few clicks. However, the downsides are 1. Since new messages are coming out every second, how to collect and digest a vast amount of message, and identify essential information now becomes harder than ever before. Secondly, if anyone creates rumors on the internet, these rumors can now easily spread and cause unnecessary panic to the crowd. Therefore, how to collect and process online opinions from different sources, and try to screen valuable information from tons of online opinions is the goal of this study.

This research is a four-year project, which continues the Mandarin opinion mining research of the last year. The target of this plan is to expand the number of monitoring sources and to strengthen the functionality of current opinion mining system in the aspect of analysis, reporting, and screening.

Concerning monitoring sources expansion. Besides adding more Traditional Chinese and Simplified Chinese monitoring channels, we add English media sources from ASEAN countries. Hoping by monitoring the news from foreign online media, we can get to know how disease and news spread in overseas countries.

Regarding enhancing the functionality of current opinion monitoring system, this study first solved the differences of traditional and simplified Chinese,

as well as the problem of eliminating Chinese synonyms, to ensure that the system can effectively understand mandarin messages. In the notification and screening section, this study setup notification mechanism such as Line and Email on the system, system users can receive the public opinion information immediately once any related news is published. At last, in order to screen out the critical information, this study also utilizes machine learning techniques to screen out public opinions. With screening function, the system user can quickly spot on valuable information instead of reading tons of spamming messages.

Keywords: Text Mining, Unstructured data, Opinion Mining

参、研究報告本文

(一)　前言

疾病防治對任何國家都是不可忽視的重要問題。疾病管制署(下稱疾管署)作為台灣人民疾病防治的把關者，應隨時關注可能影響人民健康的疾病情報，以便在正確的時間作出正確的決定，有效防止疫病的爆發。

隨著全球化的趨勢，每個人接觸到的人與環境也不只侷限於一個國家或一個地區，因此傳染病的流行，往往也會從一個地區開始，漸漸擴散到其他區域。因此為了防疫的需求，除了要關注國內的新聞外，更要緊盯外國疫病的消息，以避免該疾病在國內引起大流行。

以往訊息的傳播方式有限，決策者只消透過閱讀報紙及觀看電視台，便可以掌握所有跟疾病防治的相關訊息，但近年來網路發達，訊息以前所未有的速度累積。如果只透過傳統方式蒐集資訊，速度與涵蓋範圍則都有所未逮。因此，如果能透過網路技術，快速且大量蒐集跟疫病相關的輿情資訊，便能以更即時的方式了解跟疾病防治相關的輿情。

透過了解網路輿情，除了能夠以言論數據的角度了解疾病的傳染模式外，疾管署亦可以利用輿情數據做為感測器，檢測流感等季節性流行病在人群中的傳播行為。若能透過輿情早期發現可能的流行病，便可以利用輿情作為早期預警，提醒人民應該為即將流行的流感做好準備，以避免疾病暴發，造成區域間的大流行。

之前關於疫病相關的輿情研究，大多數的研究都有利用社群媒體上的網路輿情檢測(Detect)或預測(Predict)季節性流感或豬流感等流感暴發

事件。而因為預測(Predict)與檢測(Detect)在定義上不同，因此研究方法也區分為對流感的檢測(Detect)與對流感的預測(Predict)。流感檢測(Detect)是指發現已經發生的流感病例的過程。另一方面，流感預測(Predict)則是討論如何收集資料來預測流感趨勢，因此“即時預測”一詞通常指的是即時預測流感病例的過程。

而在近期的研究中，網路輿情已經被證明在疫情爆發的早期階段，當缺乏可靠的監測數據時，社群網路可以成為預測傳染病爆發的可靠工具。喬治亞州立大學的研究人員在追蹤和分析 2014-2015 年西非伊波拉疫情和 2015 年韓國中東呼吸症候群冠狀病毒感染症疫情期間，公共衛生部門和知名媒體通過社交媒體或網站發佈的報告。研究人員發現這些數據可用於識別突發疫情期間暴露和傳播模式。而利用輿情建立早期的預警模式，將有助於疾管署評估流行病爆發之前所應採取的公共衛生防治策略與手段。

輿情除了可以扮演防疫所使用的第三方資料源，亦可以協助疾管署早期發現是否有不實的謠言在網路上傳播。近年來因為社群媒體的發達，現在人人都可以利用近乎零成本的方式在社群媒體上發佈各式訊息。而雖然網路是免費的，但是廣告生態與政治利益，卻可以讓一些不肖分子可以藉由假新聞獲利或煽動群眾。因此如何快速偵測，並遏止假新聞的傳播，便是現在所有政府部門皆須關心的重大議題。但假新聞的產生來源相當多變，可能來自於個人的玩笑、一個組織的運作甚至到傳統媒體的斷章取義，因此尚沒有一套系統化的方式能斷定一則訊息必然是假新聞。但透過系統蒐集各式不同的資料源，從中交叉比對資料的來源，再加以確認訊息的真偽，便成為目

前找出假新聞的直接方法．如疾管署能夠透過輿情系統及早發現可能的假消息，並適時闢謠，便可以達公共治理之效，有效跟民眾宣導正確的疾病防治概念。

因應分析以及闢謠的需求，我們在這四年期的計畫中，搭建一即時與情分析平台，希望能夠透過即時搜尋國內外各種對疫病的討論與相關新聞，方能讓疫病於爆發早期時，疾管署便能第一手掌握消息。如能利用該輿情資訊建立監測與預測模式，便能利用該輿情預測疫情爆發的可能性，以達早期預防之效。另外，若能掌握各個消息來源，疾管署則可以在謠言廣泛散佈之前，先行闢謠消毒，對民眾宣導正確的防疫概念，避免民眾受到謠言的影響，導致不必要的損失。藉由這兩方面雙管齊下，疾管署便可以利用該輿情系統，達公共治理之效。

近年來深度學習與資料科學的興起，引領大家對人工智慧擁有多所期待，而若疾管署的同仁皆有資料分析的基本底子，則在面對內部大量的數據與外在大量的輿情數據時，方有能力處理與關聯，從中找出對既有業務有益處的重要資訊。為了能讓內部同仁具有資料處理的知識，我方即提供疾管署一系列 R 語言課程，希冀內部同仁能學習 R 語言的基礎、繪圖功能、資料處理功能、統計分析與機器學習等相關知識後，便有能力利用 R 語言，分析身邊的數據，更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：

- 整合簡體中文的資料來源，蒐集所有中國政府的地方各級城市的官方疫情訊息，擴展系統的監測範圍，當有謠言產生時，可以透過該資料源驗

證訊息真偽性。

- 增加簡體中文的詞典，讓系統得以處理簡體中文的斷詞，與識別文章中的名詞，提供資要的可讀性。
- 增加簡繁中文對譯的功能，讓使用者得以下繁體中文便能檢索簡體與繁體中文的資料。
- 增加兩岸三地語言的中文同義字典，讓監測系統得以歸納簡繁中文同義輿情資訊。
- 增加主要國家官方英文媒體的疫情訊息來源，補足資訊來源只有中文語系的不足，讓系統得以用英語監控世界各地的輿情訊息，強化系統的監控能力。
- 我們使用貝氏網路建立一輿情篩選模型，再將模型建立至即時警示系統，讓系統得以在收到相關輿情後，篩選出適當的輿情，並警示給權責人員，使權責人員第一時間收到最相關輿情資訊，而能過快速濾掉不相關的輿情。
- 舉辦 R 語言分析課程，提升疾管署內部人員對非結構化文字資料的處理與蒐集能力，提升同仁的數據分析技巧。課程內容包含：R 語言簡介、R 語言 ETL、R 語言的儲存與資料探索實務、R 語言與機器學習。

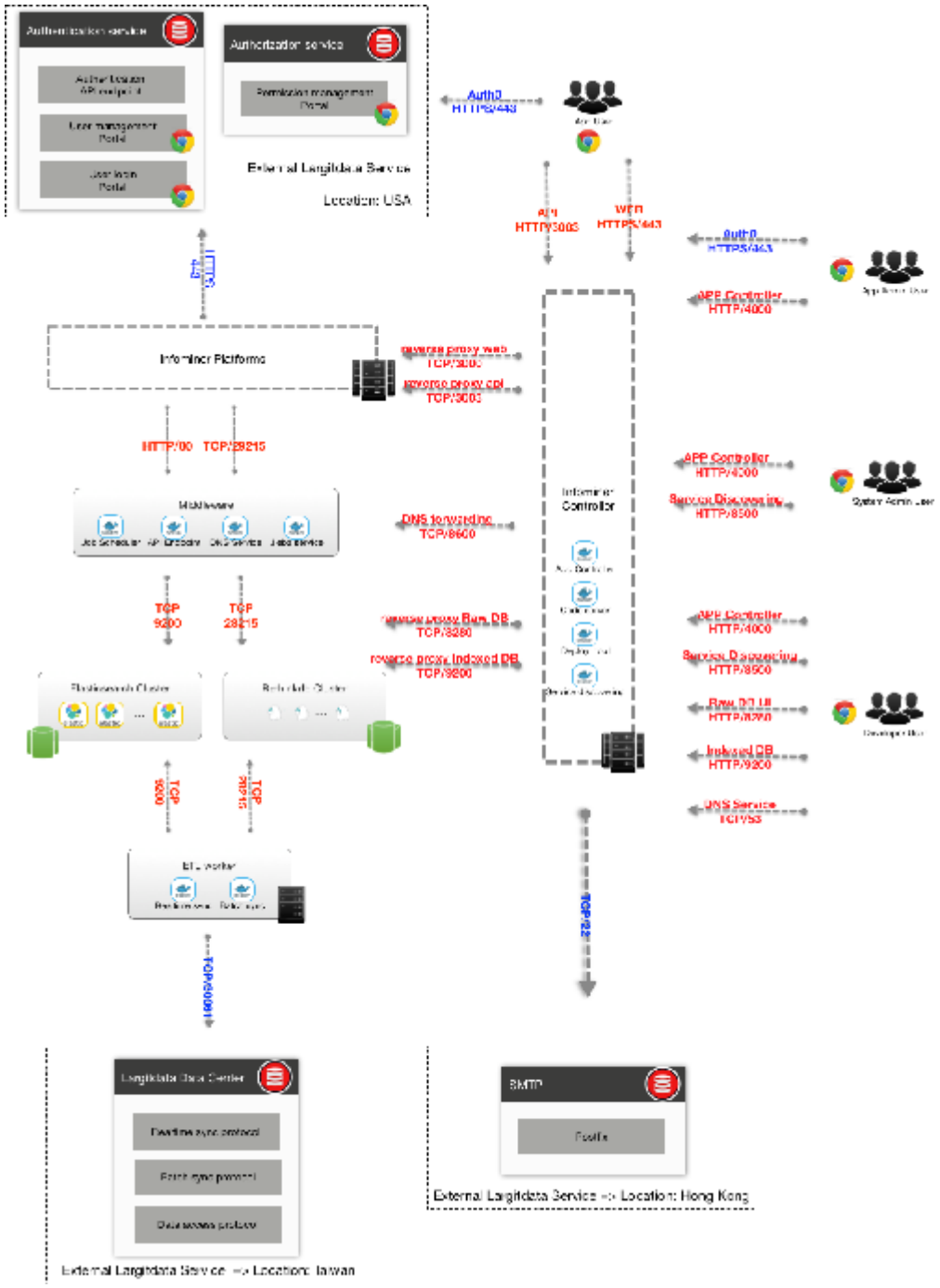
（二） 材料與方法

為了能夠快速蒐集國內外疫病各大頻道相關輿情訊息，並從非結構化的輿情資料之中分析出價值資訊，我們在第一年的計劃中建立了一個輿情觀測平台。並將該平台建構於雲端平台(Google Cloud Platform)上。透過該雲端平台，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。

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：

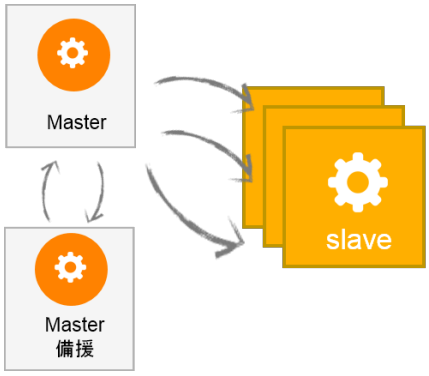


圖一、系統架構

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。由於系統必須先針對網頁建立反向索引，而反向索引是根據詞庫內的字詞所建立，如果有

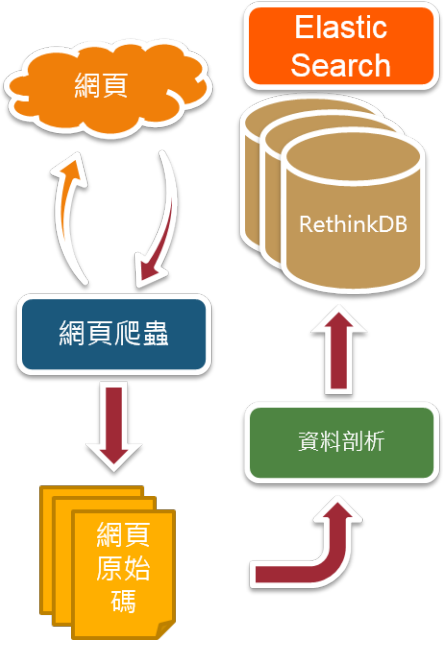
新增字詞，則資料便必須重建索引，方能讓使用者可以根據新字詞檢索文字內容，也因如此，我們必須要保留原始網頁內容，以供後續檢索用。因此，系統會先在磁碟空間上預儲存一份資料源，而後再透過剖析程式，抽取重要資料。如果抓取資料頁面格式有變，該系統將會自動發出警示至權責人員信箱。

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、資料索引機制

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多台虛擬機器，部分機器用做資料 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，之後可透過網頁介面呈現文字探勘後的分析

結果。

2. 資料蒐集

在 106 年度的計畫中，為了能夠廣泛搜索國內外的疫情資訊，因此已佈建部分網路爬蟲蒐集國內各來源的輿情資訊。於 107 年度為了能擴展輿情系統的既有能力，於是今年整合更多繁體、簡體中文與英文媒體的疫情訊息來源。由於輿情系統所蒐集之輿情訊息來源為非結構化資料，因此如果沒有將資料做過一定的清整，我們將無法對資料進行統計分析，甚至探勘語意內容．由於全世界的網頁廣泛，如要全部蒐集下來，將是一個很龐大的工，因此我們選定了多數民眾關心的社群媒體當做資料蒐集目標，以期能夠更精準處理跟分析輿情訊息。以下將資料粗分為:台灣新聞、中國新聞、中國疾管、論壇、部落格、影音資料與內容農場等，107 年的資料來源可見下表：

表一、台灣新聞網

資料來源	來源連結
Yahoo	yahoo.com
財訊	wealth.com.tw
台灣好新聞	taiwanhot.net
報導者	twreporter.org
風傳媒	storm.mg
yam蕃薯藤	yam.com
關鍵評論網	thenewslens.com
三立新聞網	setn.com

上報	upmedia.mg
NOWnews 今日新聞	nownews.com
聯合報	udn.com
PeoPo 公民新聞	peopo.org
民報 Taiwan People News	peoplenews.tw
端傳媒	theinitium.com
台視	ttv.com.tw
TechNews 科技新報	technews.tw
聯合新聞網	udn.com
TVBS官方網站	tvbs.com.tw
新頭殼 newtalk	newtalk.tw
鏡週刊	mirrormedia.mg
上下游News&Market新聞市集	newsmarket.com.tw
公開資訊觀測站	twse.com.tw
新新聞	new7.com.tw
MoneyDJ理財網	moneydj.com
PanSci 泛科學	pansci.asia
公視	pts.org.tw
line	line.me
風向新聞	kairos.news
自由時報電子報	ltn.com.tw
HiNet	hinet.net
民視地方	ftv.com.tw
美麗島電子報	my-formosa.com
中國評論新聞	crntt.com
遠見雜誌	gvm.com.tw
東森電視	ebc.net.tw
年代電視台	eracom.com.tw
民視ＦＴＶ	ftv.com.tw
卡卡洛普 Gamme	gamme.com.tw
cnYES 鉅亨網	cnyes.com

Ettoday新聞雲	ettoday.net
中時電子報	chinatimes.com
台灣大紀元	epochtimes.com.tw
公民行動	civilmedia.tw
中視	ctv.com.tw
中國時報	chinatimes.com
今周刊	businesstoday.com.tw
台灣環境資訊協會	e-info.org.tw
Cheers快樂工作人雜誌	cheers.com.tw
中央社	cna.com.tw
Duda	dopost.com
中天快點TV	ctitv.com.tw
天下雜誌	cw.com.tw
中央網路報	cdnews.com.tw
中天 必PO TV	ctitv.com.tw
華視全球資訊網	cts.com.tw
康健雜誌	commonhealth.com.tw
中廣	bcc.com.tw
爆料公社	bc3ts.com
中央通信社CNA	cna.com.tw
朝日新聞中文網	asahichinese-f.com
蘋果日報	appledaily.com
台灣醒報 Awakening News Networks	anntw.com
農傳媒	agriharvest.tw
商業周刊	businessweekly.com.tw

表二、中國新聞網

資料來源	來源連結
輔仁文誌	vjmedia.com.hk
奇摩	yahoo.com

中國台灣網	taiwan.cn
成報	singpao.com.hk
人民网	people.com.cn
星島頭條網	stheadline.com
新浪	sina.com.cn
華爾街日報(中國)	wsj.com
新明日報	shinmin.sg
香港電台網站	rthk.hk
新華網	xinhuanet.com
熱血時報	passiontimes.hk
852郵報	post852.com
now新聞	now.com
香港文匯網	wenweipo.com
澎湃	thepaper.cn
日本經濟新聞	nikkei.com
東方日報	on.cc
你好台灣網	hellotw.com
立場新聞	thestandnews.com
香港政府新聞網	news.gov.hk
晴報	ulifestyle.com.hk
無綫新聞	tvb.com
大公網	takungpao.com.hk
樂古	nakuz.com
印尼星洲日報	sinchew.com.my
明報香港新聞網	mingpao.com
环球网	huanqiu.com
ＨＫ Cable TV	i-cable.com
881903	881903.com
華夏經緯網	huaxia.com
香港獨立媒體網	inmediahk.net
新城電台	metroradio.com.hk

國度復興報(香港)	krt.com.hk
謎米香港	memehk.com
公教報	kkp.org.hk
解放网	jfdaily.com
香港商報	hkcd.com
香港經濟日報	hket.com
香港01	hk01.com
ＦＴ中文網	ftchinese.com
信報	hkej.com
香港蘋果新聞	appledaily.com
观察者网	guancha.cn
多維新聞網	dwnews.com
中国新闻网	chinanews.com
2000fun遊戲資訊網	2000fun.com
央视网	cctv.com
中國台灣網	taiwan.cn
8頻道新聞	channel8news.sg
博訊新聞網	boxun.com
海峽兩岸關係協會	arats.com.cn
巴士的報	bastillepost.com
am730	am730.com.hk

表三、中國疾管

資料來源	來源連結
浙江省疾病預防控制中心	cdc.zj.cn
浙江省衛生和計劃生育委員會	zjwjw.gov.cn
天津政務網	tj.gov.cn
雲南疾控資訊網	yncdc.cn
西藏自治區疾病預防控制中心	tibetcdc.cn
西藏自治区卫生和计划生育委员会	xzwsjsw.cn

四川省衛生和計劃生育委員會	scwst.gov.cn
山西省疾病預防控制中心	sxcdc.cn
西藏自治區人民政府	xizang.gov.cn
寧夏回族自治區衛生和計劃生育委員會	nxws.gov.cn
吉林省人民政府	jl.gov.cn
內蒙古自治區政府門戶網站	nmg.gov.cn
上海市衛生和計劃生育委員會	wsjsw.gov.cn
內蒙古自治區綜合疾病預防控制中心	nmcdc.com.cn
山東省疾病預防控制中心	sdcdc.cn
山西省衛生和計劃生育委員會	sxwsjs.gov.cn
中華人民共和國國家衛生健康委員會	nhfpc.gov.cn
中華人民共和國農業部	moa.gov.cn
山東省衛生和計劃生育委員會	sdwsjs.gov.cn
青海省衛生和計劃生育委員會	qhwjw.gov.cn
陝西省疾病預防控制中心	sxcdc.com
遼寧省國土資源局	lndoh.gov.cn
遼寧省疾病預防控制中心	lncdc.com
寧夏疾病預防控制中心	nxcdc.org
河南省衛生和計劃生育委員會	hnwsjsw.gov.cn
湖北省衛生和計劃生育委員會	hbwsjs.gov.cn
甘肅省衛生和計劃生育委員會	gsws.gov.cn
廣東省衛生和計劃生育委員會	gdwst.gov.cn
江蘇省疾病預防控制中心	jshealth.com
湖南省人民政府	hunan.gov.cn
廣東省人民政府應急管理辦公室	gdemo.gov.cn
江西省疾病预防控制中心	jxcdc.cn
中华人民共和国中央政府网	sousuo.gov.cn
江西省卫生和计划生育委员会	jxwst.gov.cn
湖北疾控	hbcdc.cn
湖南省疾病預防控制中心	hncdc.cn
江蘇省衛生和計劃生育委員會	jswst.gov.cn

海南省人民政府	hainan.gov.cn
貴州省疾病防預控制中心	gzscdc.org
黑龍江省疾病預防控制中心	hljcdc.org
黑龍江省衛生和計劃生育委員會	hljhfpc.gov.cn
貴州省衛生和計劃生育委員會	gzhfpc.gov.cn
廣西壯族自區疾病預防控制中心	gxcdc.com
廣西壯族自治區衛生和計劃生育委員會	gxhfpc.gov.cn
河北省卫生和计划生育委员会	hebwst.gov.cn
重慶市衛生和計畫生育委員會	cqwsjsw.gov.cn
福建省衛生和計劃生育委員會	fjhfpc.gov.cn
甘肅省疾病預防控制中心	gscdc.net
福建省疾病預防控制中心	fjcdc.com.cn
重慶市疾病預防控制中心	cqcdc.org
廣東省疾病預防控制中心	cdcp.org.cn
中國疾病預防控制中心	chinacdc.cn
安徽省衛生和計劃生育委員會	ahwjw.gov.cn
北京市衛生和計劃生育委員會	bjchfp.gov.cn
安徽省疾病預防控制中心	ahcdc.cn
北京市疾病預防控制中心	bjcdc.org
天津市疾病預防控制中心	cdctj.com.cn

表四、外國新聞

資料來源	來源連結
天主教香港教區	catholic.org.hk
The Standard 英文虎報	thestandard.com.hk
South China Morning Post	scmp.com
The Wall Street Journal	wsj.com
Washington Post	washingtonpost.com
South China Morning Post	scmp.com
Viet Nam News	vietnamnews.vn

RISING STAR	indiatimes.com
紐約時報	nytimes.com
sanook	sanook.com
The Indian EXPRESS	indianexpress.com
inquirer	inquirer.net
Los Angeles Time	latimes.com
The Myanmar Times	mmtimes.com
香港ポスト	hkpost.com.hk
PIA	pia.gov.ph
ＨＫＦＰ	hongkongfp.com
經濟學人	economist.com
ai Gon GP Daily	saigon-gpdaily.com.vn
BH online	bharian.com.my
彭博新聞社	bloomberg.com
中國日報網	chinadaily.com.cn
BBC	bbc.com
NHK	nhk.or.jp
安塔拉新聞	antaranews.com
ministry of health	moh.gov.sg
From the Desk of the Director-General of Health Malaysia	kpkesihatan.com
Kementerian Kesehatan	depkes.go.id
MInistry of health	cdcmoh.gov.kh

表五、PTT、臉書、論壇、部落格、影音頻道

資料來源	來源連結
Youtube	youtube.com
壹電視 NextTV	nexttv.com.tw
香港地下天文台	wxhk.org
ＵＷＡＴＳ	uwants.com

TVB討論區	tvb.com
ＰＣＤＶＤ數位科技討論區	pcdvd.com.tw
SOGI手機王	sogi.com.tw
吹水台	lihkg.com
MIUI米柚論壇	miui.com
香港信用卡優惠網	hongkongcard.com
Mobile01	mobile01.com
香港高登	hkgolden.com
捷克論壇	jkforum.net
巴哈姆特電玩資訊站	gamer.com.tw
DCARD	dcard.tw
2000Fun遊戲資訊網	2000fun.com
香港討論區	discuss.com.hk
伊莉討論區	eyny.com
中國留學社	chinaeducenter.com
癮科技	cool3c.com
卡提諾論壇	ck101.com
親子王國	baby-kingdom.com
udn部落格	udn.com
隨意窩	xuite.net
痞客邦	pixnet.net
樂多日誌	roodo.com
Medium	medium.com
批踢踢實業坊	ptt.cc
臉書	facebook.com

表六、內容農場

資料來源	來源連結
TEEP亮新聞	teepr.com
GreatDaily	twgreatdaily.com

壹讀	read01.com
生活知識+	lifeonea.com
ptt01	ptt01.cc
Joy啦	joylah.co
壹A新聞	lifeonea.co
KKNews	kknews.cc
GETIT01	getit01.com
香港玩樂網	brotherhood.space
bomb01	bomb01.com
coco01	coco01.today
美麗日報	bldaily.com
Buzzhand	buzzhand.com

3. 資料處理

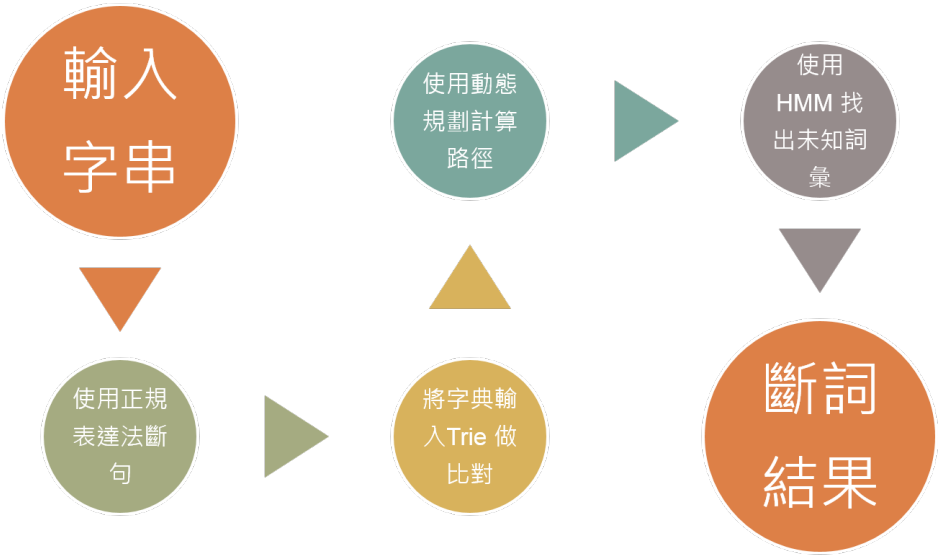
由於本計畫除了需要能夠同時支援中文與英文輿情資料處理能力，以下將會針對中文與英文資料分別列出本計畫所使用到的工具以及方法。

文章分詞

英文文字並不存在斷詞問題，只需要透過文字間的空白，即可完成文字斷詞。如以下範例句：'Social Media and Fake News in the Election' 可以根據字詞中的空白斷為'Social', 'Media', 'and', 'Fake', 'News', 'in', 'the', 'Election'等文字。但中文斷詞比較繁複，如以下範例句子: '流感疫苗打不打' 即無法單純使用空白斷詞。

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，便可以因應繁簡中斷詞需求。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，確保即使在字典中的字詞雖不完整，亦可以透過該套件進行中文斷詞。



圖四、Jieba 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「對雞蛋過敏，仍可施打流感疫苗」斷句，則會產生下列結果：

對(介詞)　雞蛋(名詞)　過敏(動詞)，仍(副詞)　可(副詞)　施打(動詞)　流感(名詞)　疫苗(名詞)

英文資料則是透過 NLTK 的 Wordnet 提供的詞與標記功能，透過該功能標記各英文字詞的詞性。

詞幹提取

英文資料雖然沒有分詞問題，但卻有詞幹提取的問題，例如，要識別字符串「cats」、「catlike」和「catty」是基於詞根「cat」；「stemmer」、「stemming」和「stemmed」是基於詞根「stem」。

詞幹提取是去除詞綴得到詞根的過程，並得到單詞最一般的寫法。透過詞幹提取算法，便可以簡化詞 「fishing」、「fished」、「fish」和「fisher」為同一個詞根「fish」。

詞庫建立

由於 Jieba 會使用到字串比對的方式進行斷詞，因此必須先準備一個夠大的詞庫，Jieba 方能準確地進行斷詞，並標示出各字詞的詞性，以利之後的分析。

為了能夠建構基本的繁體中文字典，本計畫使用了三個資料來源：

1. 萌典

萌典共收錄教育部《重編國語辭典修訂本》、《臺灣閩南語常用詞辭典》及《臺灣客家語常用詞辭典》十六萬筆國語、兩萬筆臺語、一萬四千筆客語條目。將作為斷詞用詞庫的基底。

2. 維基百科

考量到萌典收納字詞的數量可能稍嫌不足，因此本計畫亦利用網路爬蟲技術到維基百科(Wikipedia)蒐集所有台灣正體與大陸簡體等條目資料，藉以增加字詞的數量。

條目 討論 台灣正體 汉 漢 閱讀 編輯 檢視歷史

中文維基百科Facebook粉絲專頁正式上線，邀請大家一同關注。

登革熱 [編輯]

維基百科，自由的百科全書
（重新導向自 登革熱）

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（英语：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]。

登革病毒由黑斑蚊屬的幾種蚊子傳播，主要是埃及斑蚊（*A. aegypti*）[1]。登革熱病毒有五型[7]；感染後對同型病毒可獲得終身免疫，但對異型病毒免疫力維持時間較短。且感染異型病毒會增加嚴重併發症的風險[1]，目前的篩檢方式包含偵測血液中是否存有對抗病毒或其RNA的抗體[2]。

圖五、台灣正體條目

条目 讨论 大陆简体 汉 漢 阅读 编辑 查看历史

十一月是维基百科亚洲月， 现在就报名参加吧！

台湾知识种子计划志工召募中，请参看计划页面、WSOTK粉丝团

登革热 [编辑]

维基百科，自由的百科全书

维基百科中的医疗相关内容仅供参考，详见医学声明。如需专业意见请咨询专业人士。

登革热（法语：La fièvre de la dengue；英语：dengue fever），也称为**骨痛热症**、**断骨热**、**天狗热**，是一种由登革热病毒引起的由蚊媒热带病[1]。患者大约会在感染后3到14天后发作[2]，症状包括发热、头痛、肌肉和关节痛，还有典型性的麻疹样皮疹[1][2]。一般会于2至7天痊愈。少部分患者病情可进一步恶化，出现危及生命的登革出血热，患者有出血、血小板减少和血浆蛋白渗出，或者进展为登革休克综合征，此时会出现致命性的低血压休克[2]。

登革病毒由黑斑蚊属的几种蚊子传播，主要是埃及斑蚊（*A. aegypti*）[1]。登革热病毒有五型[7]；感染后对同型病毒可获得终身免疫，但对异型病毒免疫力维持时间较短。且感染异型病毒会增加严重并发症的风险[1]，目前的筛检方式包含侦测血液中是否存有对抗病毒或其RNA的抗体[2]。

圖六、大陸簡體條目

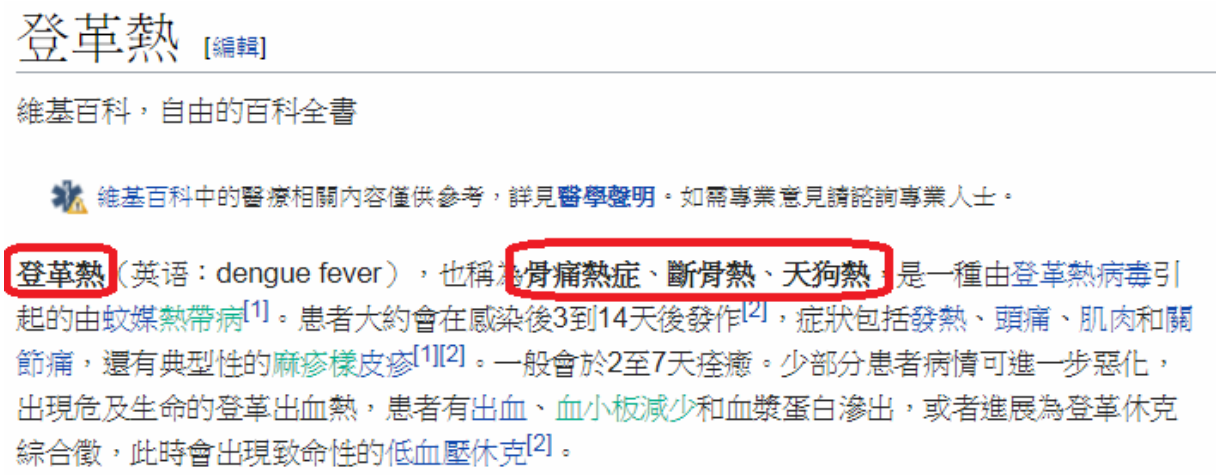
3. 新聞關鍵字

另外，為了應付每天被創造出來的新詞（例如：左流右肺），本系統將會透過網路爬蟲抓取最新新聞關鍵字(如下圖所示)，並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。

圖七、利用新聞關鍵字擴充詞典

同義詞典

中英文同樣會有同義詞的處理問題，所幸英文資料較為完備，可透過 wordnet 提供的字典即可以處理掉同義詞，但同義字的中文資源相較起來稀少許多，因此我們另外使用網路爬蟲爬取維基百科條目中的粗體字（維基志工標註的同義字詞），嘗試建立同義字典。



圖八、維基百科同義字詞來源

中文簡繁轉換

在增加簡體資料的檢索來源後，便會面臨到該如何處理兩岸三地字詞有差異的問題，因此我們即利用了維基百科擁有簡體及繁體中文字詞的特性，蒐集繁中與簡中中文詞典，並產生對應同義詞典，讓系統在檢索跟分析輿情資料時無礙。

4. 資料分析

由於輿情系統採集到的皆為非結構化的文字資料，相較於結構化資料，分析的過程需要較多的資料預處理，方能使用文字探勘技術嘗試從文字中擷取出價值資訊。以下將詳列我們在本系統中採用的分析方法：

詞頻分析

本系統可以將蒐集到的進行斷詞後，產生出每篇文章的關鍵字詞，皆者我們便可以利用各字詞，統計該篇文章所使用的詞彙頻率。利用詞彙頻率，我們亦可透過統計時間、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: count(w, d)

文檔 d 中總詞數: size(d)

tf(w,d) = count(w, d) / 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 docs(w, D)

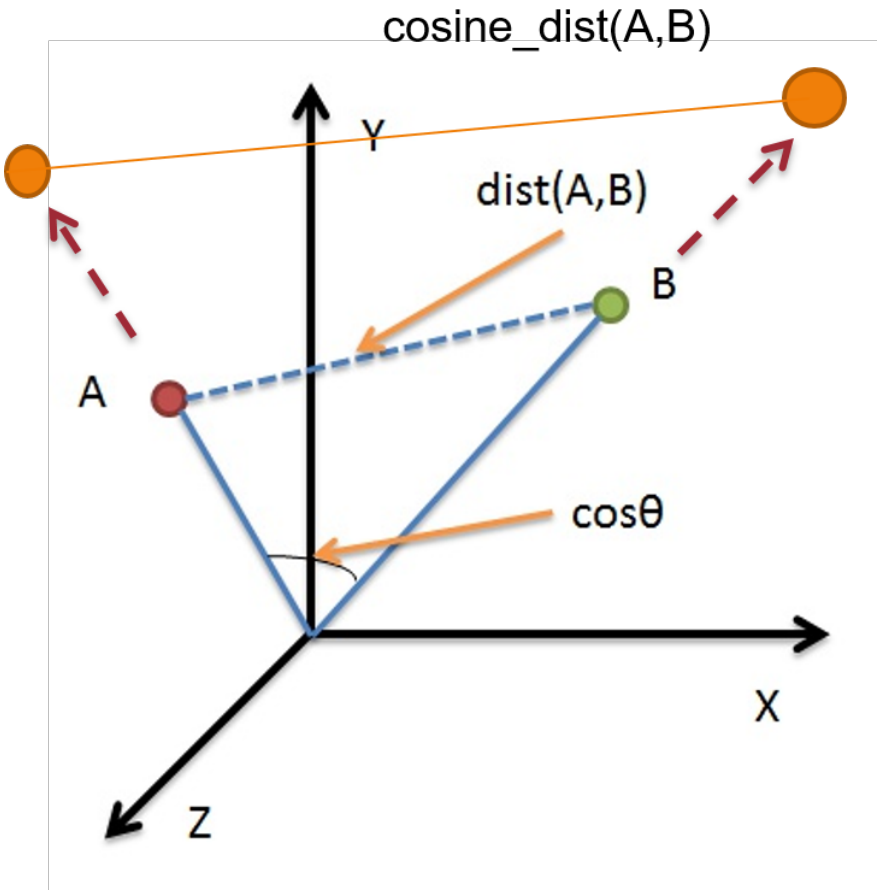
  idf = log(n / docs(w, D))

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。



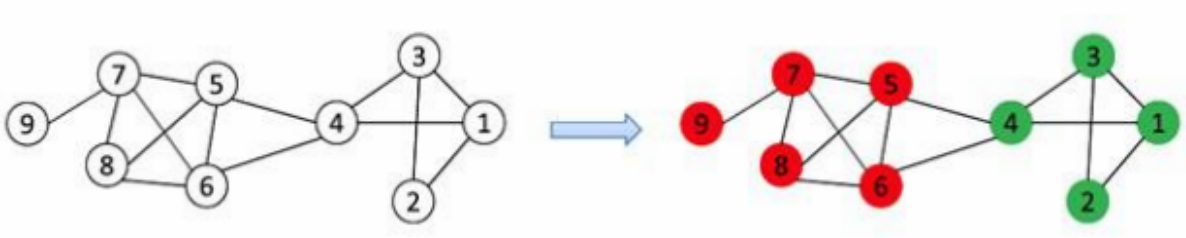
圖九、字詞相似度計算

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群。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

本系統可以將文章分群後整理成不同主題，每個主題下收納多篇文章，

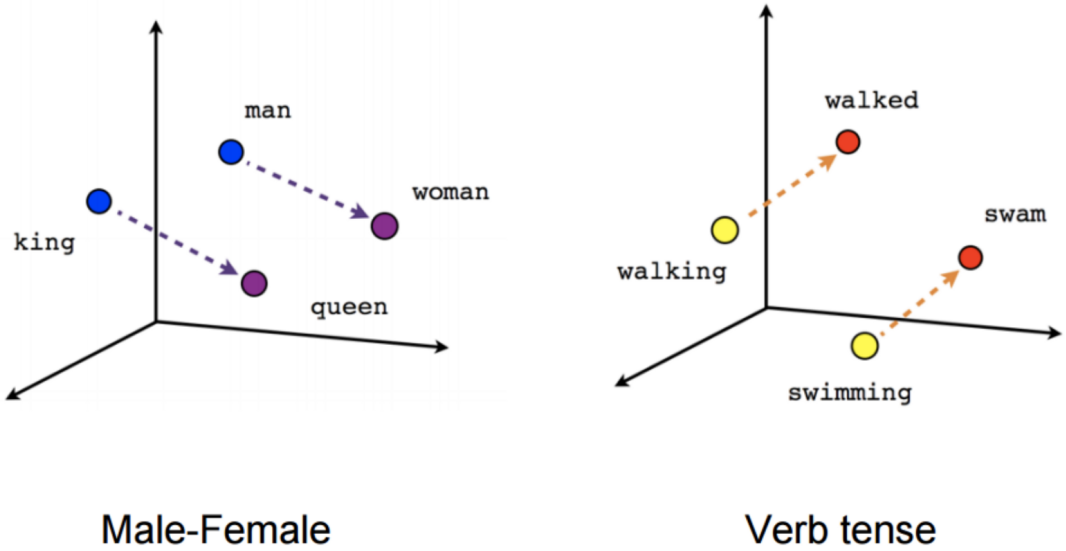
使用者只要閱讀文章標題，便可以了解該主題的原意。每個主題下都有詳列跟該主題相關的關鍵字詞，使用者只須點選該關鍵字詞，便可以分析該關鍵字詞的聲量、熱詞並分析詞與詞之間的關聯。



圖十、社群偵測法

關聯字詞分析

本系統亦有透過深度學習技術 word2vec 學習字詞之間的關聯性，word2vec 為一非監督式的深度學習模型，旨在可以利用前後文的詞向量，計算字詞之間的關聯度，使用者可以透過該關聯字詞，理解字與字之間的關係。

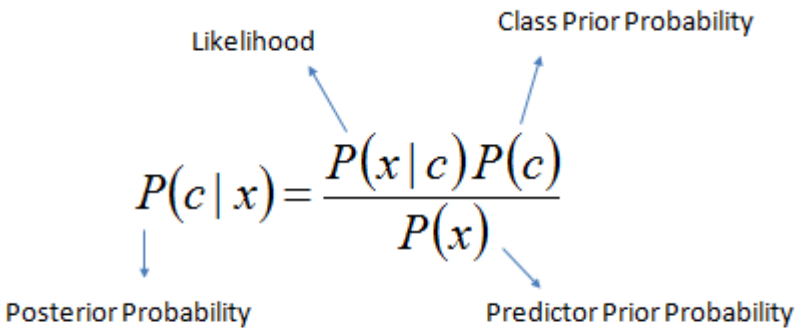


圖十一、word2vec

在建立 Word2Vec 的過程中，我們即會訓練出文字的詞向量，而詞向量除了可以減少文字的表示空間亦可以捕捉詞與詞之間的關係，在未來的系統中，我們將會應用詞向量進行文章分群與分類，讓系統能夠更準確的探勘出文字背後的資訊。

文章分類

為了避免使用者收到太多不相關的文章，本系統將會使用機器學習的分類方法學習使用者的標準，讓系統自動分類文章。在本研究中，我們將採用貝式學習方法作為文章分類器。貝式演算法採用機率模型，會計算每個字詞對到該類別的條件機率，因此當收集到新文章時，系統會自動將文章斷詞，並根據各字詞對應到該類別的機率做乘積，並以對應機率的高低，決定文章類別。



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

圖十二、貝氏分類法

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。

5. 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-leran 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 Javasciprt 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.　擴展大陸輿情監測範圍

原計畫目標

為了能夠擴展系統的監測範圍，我們將整合簡體中文的資料來源，目標蒐集所有中國政府的地方各級城市共 40 個網站(如山西省疾病预防控制中心、陕西省卫生计生委 、陕西省疾病预防控制中心、西藏自治区卫计委、西藏自治区人民政府、海南省疾病预防控制中心等網站的疫情訊息，以備有謠言產生時，可以透過該資料源驗證訊息真偽性。

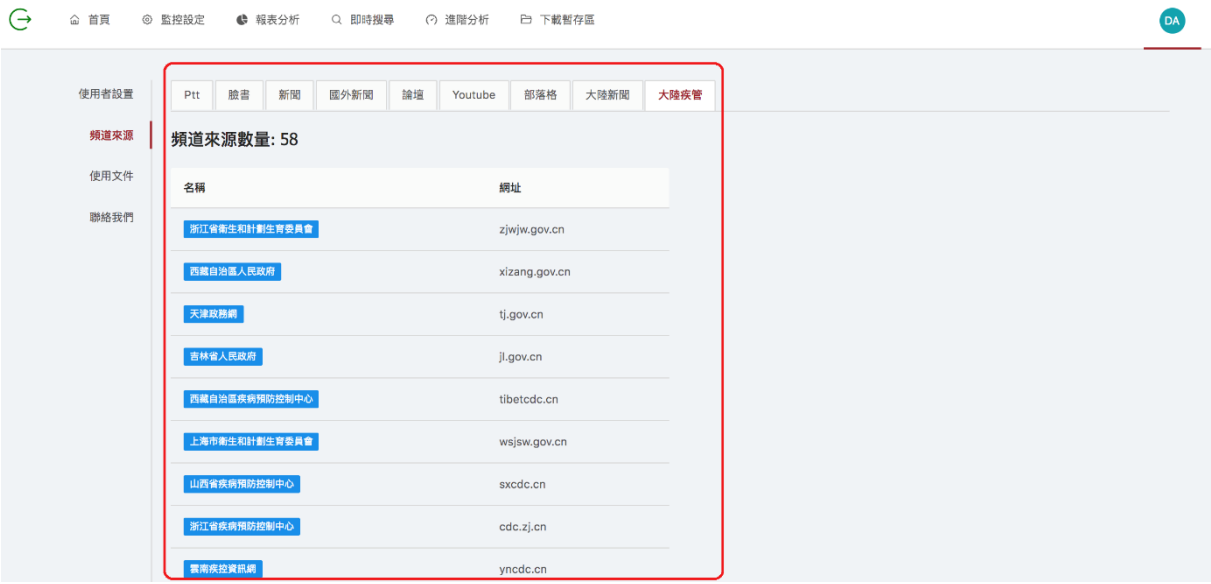
本研究成果

關於疾病管制相關的資料源，目前已增設 58 大陸相關來源資料，使用者可以透過輿情系統的頻道源列表，找到目前來源。

使用者可以先進到輿情系統中，點選右上角的個人設置(紅色框框標註的綠色圓圈圖樣)

圖十三、進入個人設置

點擊綠色圓圈以後，下方將出現使用者設置、頻道來源、使用文件與聯絡我們等選單畫面，在此點選頻道來源等畫面，右方將出現頻道來源。點選「大陸疾管」後，下方將列出跟中國疾病相關資料來源網站共 5 8 個。



圖十四、檢視「大陸疾管」來源

表七、大陸疾管網站名稱與連結

	名稱	連結
0	浙江省衛生和計劃生育委員會	zjwjw.gov.cn
1	西藏自治區人民政府	xizang.gov.cn
2	天津政務網	tj.gov.cn
3	吉林省人民政府	jl.gov.cn
4	西藏自治區疾病預防控制中心	tibetcdc.cn
5	上海市衛生和計劃生育委員會	wsjsw.gov.cn
6	山西省疾病預防控制中心	sxcdc.cn
7	浙江省疾病預防控制中心	cdc.zj.cn
8	雲南疾控資訊網	yncdc.cn
9	山東省疾病預防控制中心	sdcdc.cn
10	西藏自治区卫生和计划生育委员会	xzwsjsw.cn
11	青海省衛生和計劃生育委員會	qhwjw.gov.cn
12	中華人民共和國國家衛生健康委員	nhfpc.gov.cn

	會	
13	四川省衛生和計劃生育委員會	scwst.gov.cn
14	遼寧省疾病預防控制中心	lncdc.com
15	山西省衛生和計劃生育委員會	sxwsjs.gov.cn
16	內蒙古自治區綜合疾病預防控制中心	nmcdc.com.cn
17	寧夏回族自治區衛生和計劃生育委員會	nxws.gov.cn
18	陝西省疾病預防控制中心	sxcdc.com
19	遼寧省國土資源局	lndoh.gov.cn
20	湖北省衛生和計劃生育委員會	hbwsjs.gov.cn
21	山東省衛生和計劃生育委員會	sdwsjs.gov.cn
22	河南省衛生和計劃生育委員會	hnwsjsw.gov.cn
23	江西省卫生和计划生育委员会	jxwst.gov.cn
24	寧夏疾病預防控制中心	nxcdc.org
25	廣東省衛生和計劃生育委員會	gdwst.gov.cn
26	內蒙古自治區政府門戶網站	nmg.gov.cn
27	甘肅省衛生和計劃生育委員會	gsws.gov.cn
28	湖南省疾病預防控制中心	hncdc.cn
29	廣東省人民政府應急管理辦公室	gdemo.gov.cn
30	江蘇省疾病預防控制中心	jshealth.com
31	江西省疾病预防控制中心	jxcdc.cn
32	黑龍江省疾病預防控制中心	hljcdc.org
33	江蘇省衛生和計劃生育委員會	jswst.gov.cn
34	中華人民共和國農業部	moa.gov.cn
35	湖南省人民政府	hunan.gov.cn
36	中华人民共和国中央政府网	sousuo.gov.cn
37	河北省卫生和计划生育委员会	hebwst.gov.cn
38	黑龍江省衛生和計劃生育委員會	hljhfpc.gov.cn
39	貴州省衛生和計劃生育委員會	gzhfpc.gov.cn
40	貴州省疾病防預控制中心	gzscdc.org

41	廣西壯族自治區衛生和計劃生育委員會	gxhfpc.gov.cn
42	重慶市衛生和計畫生育委員會	cqwsjsw.gov.cn
43	湖北疾控	hbcdc.cn
44	廣西壯族自區疾病預防控制中心	gxcdc.com
45	福建省疾病預防控制中心	fjcdc.com.cn
46	海南省人民政府	hainan.gov.cn
47	福建省衛生和計劃生育委員會	fjhfpc.gov.cn
48	甘肅省疾病預防控制中心	gscdc.net
49	廣東省疾病預防控制中心	cdcp.org.cn
50	重慶市疾病預防控制中心	cqcdc.org
51	安徽省衛生和計劃生育委員會	ahwjw.gov.cn
52	安徽省疾病預防控制中心	ahcdc.cn
53	北京市衛生和計劃生育委員會	bjchfp.gov.cn
54	天津市疾病預防控制中心	cdctj.com.cn
55	中國疾病預防控制中心	chinacdc.cn
56	北京市疾病預防控制中心	bjcdc.org
57	石家庄市疾病预防控制中心	hbcdc.cn

另外，如果使用者將畫面切換到「大陸新聞」，將可以看到有 54 個資料來源。

圖十五、檢視「大陸新聞」來源

以下表列目前「大陸新聞」來源共 54 筆來源網站名稱與連結：

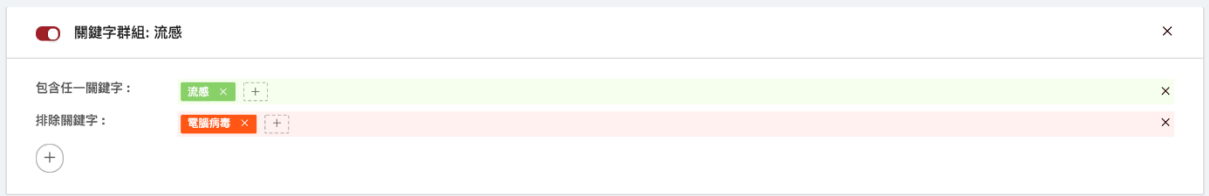
表八、大陸新聞網站名稱與連結

	名稱	連結
1	輔仁文誌	vjmedia.com.hk
2	奇摩	yahoo.com
3	中國台灣網	taiwan.cn
4	成報	singpao.com.hk
5	人民网	people.com.cn
6	星島頭條網	stheadline.com
7	新浪	sina.com.cn
8	華爾街日報(中國)	wsj.com
9	新明日報	shinmin.sg
10	香港電台網站	rthk.hk

11	新華網	xinhuanet.com
12	熱血時報	passiontimes.hk
13	852郵報	post852.com
14	now新聞	now.com
15	香港文匯網	wenweipo.com
16	澎湃	thepaper.cn
17	日本經濟新聞	nikkei.com
18	東方日報	on.cc
19	你好台灣網	hellotw.com
20	立場新聞	thestandnews.com
21	香港政府新聞網	news.gov.hk
22	晴報	ulifestyle.com.hk
23	無綫新聞	tvb.com
24	大公網	takungpao.com.hk
25	樂古	nakuz.com
26	印尼星洲日報	sinchew.com.my
27	明報香港新聞網	mingpao.com
28	环球网	huanqiu.com
29	ＨＫ Cable TV	i-cable.com
30	881903	881903.com
31	華夏經緯網	huaxia.com
32	香港獨立媒體網	inmediahk.net
33	新城電台	metroradio.com.hk
34	國度復興報(香港)	krt.com.hk
35	謎米香港	memehk.com
36	公教報	kkp.org.hk
37	解放网	jfdaily.com
38	香港商報	hkcd.com
39	香港經濟日報	hket.com
40	香港01	hk01.com
41	ＦＴ中文網	ftchinese.com

42	信報	hkej.com
43	香港蘋果新聞	appledaily.com
44	观察者网	guancha.cn
45	多維新聞網	dwnews.com
46	中国新闻网	chinanews.com
47	2000fun遊戲資訊網	2000fun.com
48	央视网	cctv.com
49	中國台灣網	taiwan.cn
50	8頻道新聞	channel8news.sg
51	博訊新聞網	boxun.com
52	海峽兩岸關係協會	arats.com.cn
53	巴士的報	bastillepost.com
54	am730	am730.com.hk

由於已經將跟大陸相關共 112 個網站加入到監測範圍中，使用者便可以在系統上鍵入關鍵字，即可搜尋到跟中國相關的疫病資訊。例如：使用者可以在關鍵字中建立「流感」關鍵字群組，接者在群組中設定流感（並排除掉電腦病毒等字樣），如下圖所示：

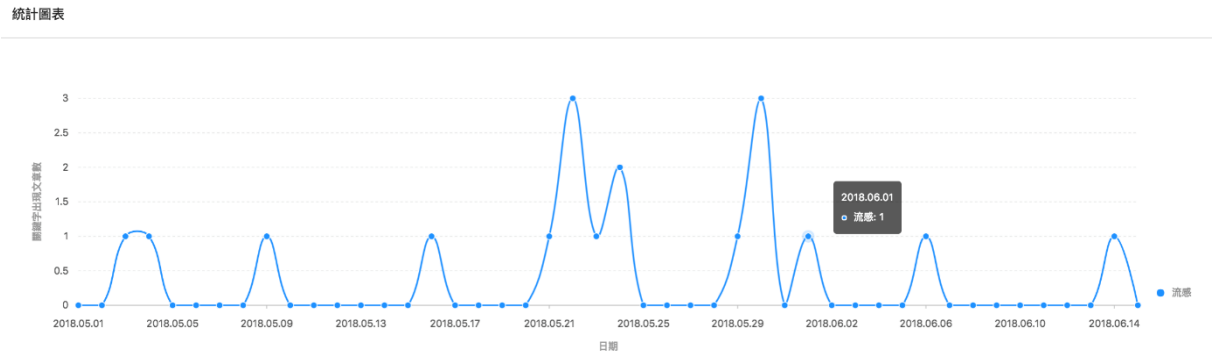


圖十六、設定關鍵字組

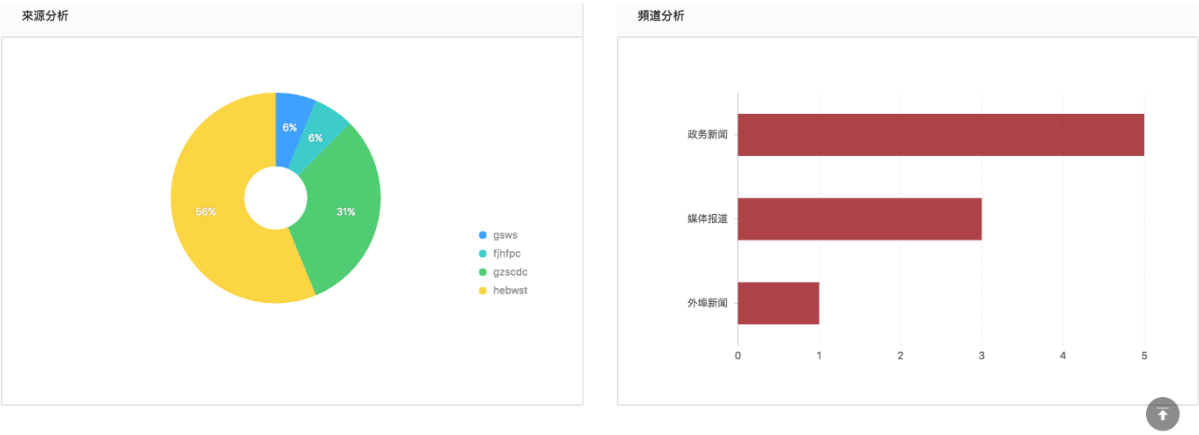
接者，將系統切換至關鍵字報表，使用者便可以在報表中，瀏覽跟流感

相關的輿情聲量、頻道分析、來源分析、熱詞分析及文章表列。

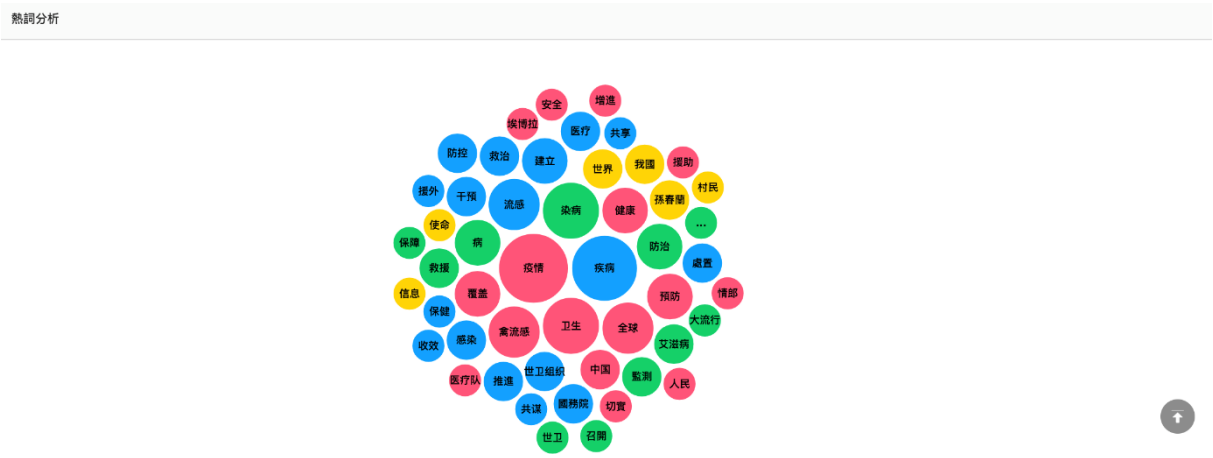
圖十七、設定關鍵字、觀測範圍與時間區間



圖十八、觀測輿情聲量



圖十九、分析輿情連結來源與輿情頻道來源



圖二十、熱詞分析

圖二十一、文章列表

2. 擴增簡體中文字典

原計畫目標

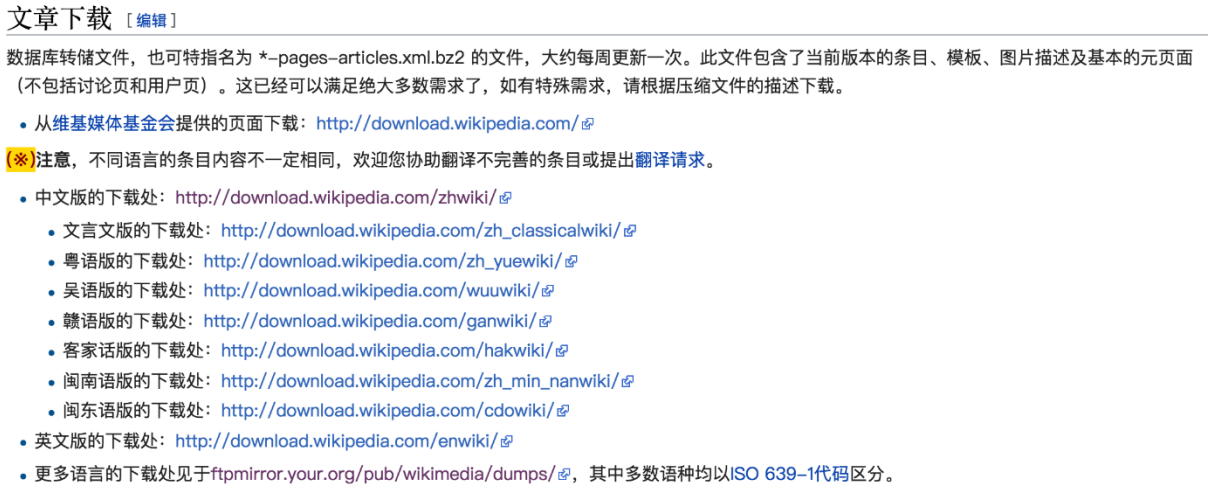
增加簡體中文的詞典， 並收納至少 5 萬個簡體中文詞彙， 讓系統得以

處理簡體中文的斷詞問題，與識別簡體文章中的專有名詞。

本研究成果

在增加簡體資料的檢索來源後，便會面臨到該如何處理兩岸三地字詞有差異的問題，因此我們即利用了維基百科擁有簡體及繁體中文字詞的特性，蒐集繁中與簡中中文詞典，並產生對應同義詞典，讓系統在檢索跟分析輿情資料時無礙。

建立簡體字的詞典方法如下，首先我們搜尋到維基百科的數據庫下載頁面（https://zh.wikipedia.org/zh-cn/Wikipedia:%E6%95%B0%E6%8D%AE%E5%BA%93%E4%B8%8B%E8%BD%BD），接者便找到中文版下載處，我們便可以打包下載維基百科的所有簡體字詞，並將字詞匯入到我們的詞庫之中.


文章下载 [编辑]

数据库转储文件，也可特指名为 *-pages-articles.xml.bz2 的文件，大约每周更新一次。此文件包含了当前版本的条目、模板、图片描述及基本的元页面（不包括讨论页和用户页）。这已经可以满足绝大多数需求了，如有特殊需求，请根据压缩文件的描述下载。

- 从维基媒体基金会提供的页面下载：http://download.wikipedia.com/

(※)注意，不同语言的条目内容不一定相同，欢迎您协助翻译不完善的条目或提出翻译请求。

- 中文版的下载处：http://download.wikipedia.com/zhwiki/
  - 文言文版的下载处：http://download.wikipedia.com/zh_classicalwiki/
  - 粤语版的下载处：http://download.wikipedia.com/zh_yuewiki/
  - 吴语版的下载处：http://download.wikipedia.com/wuuwiki/
  - 赣语版的下载处：http://download.wikipedia.com/ganwiki/
  - 客家话版的下载处：http://download.wikipedia.com/hakwiki/
  - 闽南语版的下载处：http://download.wikipedia.com/zh_min_nanwiki/
  - 闽东语版的下载处：http://download.wikipedia.com/cdowiki/
- 英文版的下载处：http://download.wikipedia.com/enwiki/
- 更多语言的下载处见于ftpmirror.your.org/pub/wikimedia/dumps/，其中多数语种均以ISO 639-1代码区分。

圖二十二、維基百科數據庫下載頁面

但為了能夠在搜尋條目時，能夠簡繁並用，並同時分析簡繁相關輿情資料，我們也建立了簡體與繁體的同義詞典，方法如下：首先，我們會到維基百科的頁面，鍵入疾病關鍵字，如愛滋病，我們便會進到愛滋病的內容頁面。

圖二十三、愛滋病條目頁面

圖二十四、語言切換選單

接者利用維基百科可以找到語言切換選項，並將語言切換成大陸簡體，此時就會跑出艾滋病的條目，如此可知艾滋病與愛滋病為同義詞，我

們便可以將同義詞典鍵入到同義詞庫中.

圖二十五、艾滋病的維基條目

根據上述方法，我們目前已經建立了五十萬條以上的簡體詞庫，以及三萬六千條的同義詞典，詞典範例如下：

- 淋巴癌/lymphoma/淋巴瘤
- e 肝/戊肝/急性病毒性戊型肝炎/急性病毒性 e 型肝炎
- 疱疹/HSV-2/HHV-1/人类单纯疱疹病毒/单纯疱疹病毒/HSV-1
- 腸病毒感染併發重症/echovirus/腸病毒/coxsackievirus groups a and b/enterovirus/肠病毒感染并发重症/肠道病毒/肠病毒/腸病毒感染併發重症
- 陰道滴蟲/trichomoniasis/滴蟲炎/滴虫性阴道炎
- 薩斯病/sars/嚴重急性呼吸道症候群/严重急性呼吸道综合征/煞斯病/萨斯病/非典/沙斯病/嚴重急性呼吸道症候群

- 尿滯留/尿液滯留/urinary retention/尿瀦留/renal retention
- 妊娠糖尿病/GDM/gestational diabetes mellitus/妊娠期糖尿病
- 福馬林/Formalin/福尔马林

於建立該同義與簡體中文詞典後，系統便可以同時分析與檢索簡體與繁體中文的疫病輿情資料. 由於本研究是利用維基百科建立同義詞典，可能會造成誤判的結果，因此在未來的研究中，會量測檢索的 Recall 與 Precision，作為檢視系統確度的評估標準。

3. 簡繁搜索功能

原計畫目標

在輿情系統搜尋列增加簡繁中文對譯的功能，讓使用者得以下繁體中文便能同時檢索簡體與繁體中文的資料。

本研究成果

雖然系統中建立有簡體字典以及同義詞典，但如果系統使用兩套方法分析文字輿情，將會造成系統很大的處理負擔，因此我們在處理簡體字時，會先將系統所蒐集到的簡體資料透過 OpenCC（開放中文轉換 Open Chinese Convert）翻譯成繁體中文，然後另外利用同義詞典將專有名詞從簡體翻譯成繁體字詞(例如：將愛滋病翻譯成艾滋病)，如此一來，我們便可以在系統中直接使用繁體字詞處理跟檢索簡體中文資料.

範例如下：首先進到輿情系統中的即時搜尋.

圖二十六、進入即時搜尋畫面

於畫面中鍵入愛滋病，即發現在中國新聞的區塊下，系統可以直接搜索出包含艾滋病的相關條目。

圖二十七、於即時搜尋搜尋愛滋病條目

4. 擴展英文媒體輿情監測範圍

原計畫目標

增加主要國家官方英文媒體的疫情訊息來源，補足資訊來源只有中文語系的不足，讓系統得以用英語監控世界各地的輿情訊息，強化系統的監控能力。

本研究成果

本研究為了廣納對海外國家的疫病監測範圍，因而除了監測繁體與簡體中文資訊外，亦納入東協國家等的英文媒體資料，希冀能透過監測海外的疫病輿情，了解其他國家的傳染情況，以期能利用該資訊做為早期預警·

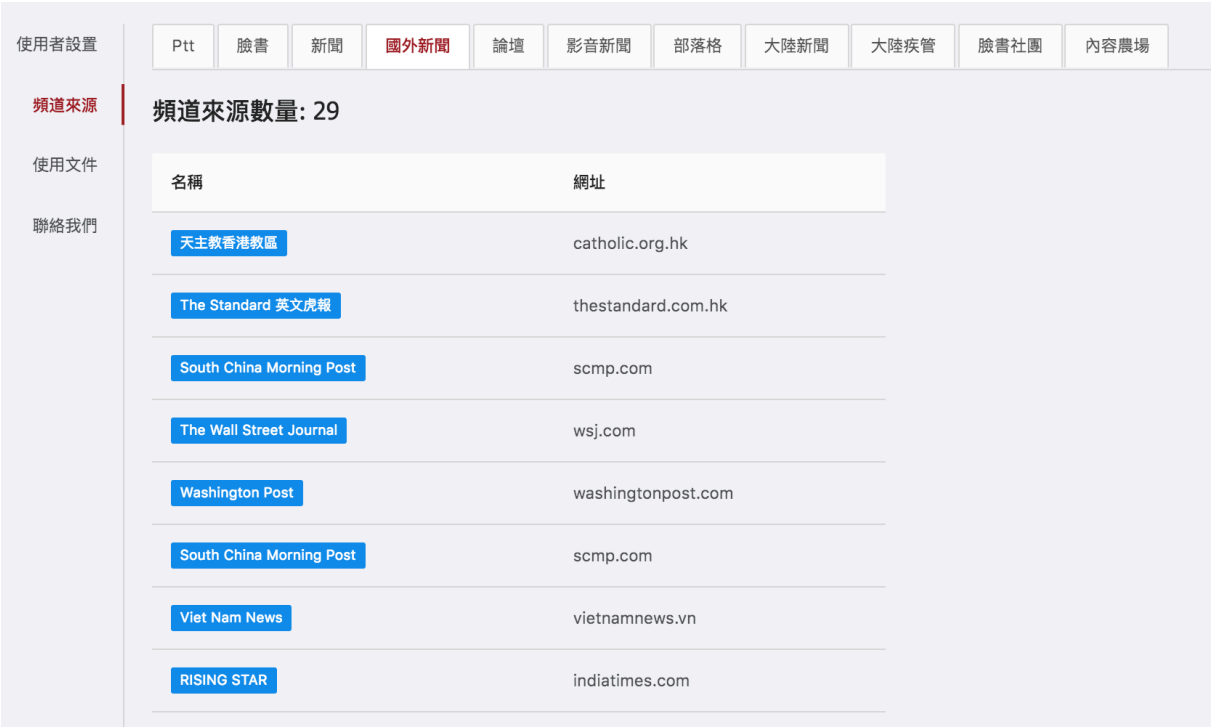
目前已增設 25 個外媒來源，使用者可以透過輿情系統的頻道源列表，

找到目前來源。

使用者可以先進到輿情系統中，點選右上角的個人設置（紅色框框標註的綠色圓圈圖樣）

圖二十八、進入個人設置

點擊綠色圓圈以後，下方將出現使用者設置、頻道來源、使用文件與聯絡我們等選單畫面，在此點選頻道來源等畫面，右方將出現頻道來源。點選「國外新聞」後，下方將列出資料來源網站共 29 個。



圖二十九、檢視「國外新聞」來源

表九、國外新聞網站名稱與連結

	名稱	連結
1	天主教香港教區	catholic.org.hk
2	The Standard 英文虎報	thestandard.com.hk
3	South China Morning Post	scmp.com
4	The Wall Street Journal	wsj.com
5	Washington Post	washingtonpost.com
6	South China Morning Post	scmp.com
7	Viet Nam News	vietnamnews.vn
8	RISING STAR	indiatimes.com
9	紐約時報	nytimes.com
10	sanook	sanook.com
11	The Indian EXPRESS	indianexpress.com
12	inquirer	inquirer.net
13	Los Angeles Time	latimes.com
14	The Myanmar Times	mmtimes.com
15	香港ポスト	hkpost.com.hk
16	PIA	pia.gov.ph
17	ＨＫＦＰ	hongkongfp.com
18	經濟學人	economist.com
19	ai Gon GP Daily	saigon-gpdaily.com.vn
20	BH online	bharian.com.my
21	彭博新聞社	bloomberg.com
22	中國日報網	chinadaily.com.cn
23	BBC	bbc.com
24	NHK	nhk.or.jp
25	安塔拉新聞	antaranews.com

26	ministry of health	moh.gov.sg
27	From the Desk of the Director-General of Health Malaysia	kpkesihatan.com
28	Kementerian Kesehatan	depkes.go.id
29	MInistry of health	cdcmoh.gov.kh

5. 建立輿情篩選模型

原計畫目標

我們使用貝氏網路建立一輿情篩選模型，再將模型建立至即時警示系統，讓系統得以在收到相關輿情後，篩選出適當的輿情，並警示給權責人員，使權責人員第一時間收到最相關輿情資訊，而能過快速濾掉不相關的輿情。

本研究成果

為了能夠在輿情訊息一出來，便可以發布警示至權責人員的信箱，系統有建置 Email 以及 Line 等通知機制，以確保當訊息一產生，便可以將訊息推送到權責人員的手持裝置，以達警示之效。但有時有有很多不相關的新聞也會因為關鍵字相符之故而被當成警示推送，為了能避免這樣的情事發生，我們增添了一使用貝式分類法建置的輿情篩選的機制。

圖三十、輿情篩選功能

當使用者在閱覽文章時，可以點選過濾此類文章，系統將會將該新聞當作訓練資料集，訓練模型能夠過濾此類文章，之後便能透過該模型適時排除掉不相關的資訊。

6. R 語言教育訓練

原計畫目標

為了強化疾管署內部人員的資料分析能力，該年會再加開 4 門 R 語言分析課程，提升同仁的數據分析技巧。課程內容包含：R 語言簡介、R 語言與資料 ETL、R 語言的儲存與資料探索實務、R 語言與機器學習等共 12 天，72 小時的課程。

本研究成果

為了因應同仁對 R 在視覺化與統計分析的需求，今年將課程修改為：R

語言簡介、R 語言與資料視覺化、R 語言與統計分析、R 語言與機器學習等課程。

為什麼要學習 R 語言

《哈佛商業評論》說「Data Scientist，數據科學家」是二十一世紀最性感的職業。根據 104 人力銀行預測 2018 年，前五大資料經濟職務需求，其中就有三個是資料分析相關職務，包括資料工程師、數據分析師與資料科學家。其求職者需要具備資料處理(ETL)工具開發經驗、熟悉 R 語言、Python、SQL、建置 Hadoop 或 Spark 平台經驗等等。R 語言是近年在資料分析領域中不可或缺的一項工具，R 語言的自由性與原始碼開放等特性讓許多資料分析人員在進行資料分析時廣泛的使用。

圖片資料來源：104人力銀行

圖三十一、前五大資料經濟職務需求

從資料取得的前期過程(ETL)，到中期的資料分析(Data mining)，到最後的資料視覺化，與機器學習，都可以輕鬆使用 R 語言進行。

課程目的在於降低初學者的學習門檻，也期望能減少專業使用者的程式撰寫時間。課程內容從 R 軟體基本概念與資料整理/組織切入，著重資料的理解與統計/機器學習等基本觀念的建立，以整合式開發環境 RStudio 進行實機操作。

為了能讓同仁充分吸收課程內容，上課時間規劃為每次三小時，授課日期 107 年 3 月 7 日至 107 年 7 月 11 日。

表十、課程課綱

課綱	課程內容	課程起始日	課程結束日	時數
R 語言簡介	簡介什麼是 R 語言、R 語言相關的資料結構以及 R 語言的基本操作，旨在讓署內同仁能夠使用 R 語言處理日常的數據資料。	2018/3/7	2018/4/11	21
R 語言與資料視覺化	簡介 R 語言的基本繪圖功能、ggplot2 及 plotly，旨在讓署內同仁能夠使用 R 語言的繪圖功能探索數據並能利用	2017/5/8	2017/5/16	9

	圖表述說數據背後的故事。			
R 語言與統計分析	簡介什麼是機率、敘述性統計以及推論性統計，旨在讓同仁能夠使用 R 做統計分析，了解如何利用統計方法見微知著。	2018/5/22	2018/6/5	15
R 語言與機器學習	比較機器學習與統計上的差異，並簡介該如何使用 R 實作迴歸分析、分類、分群與降低維度，旨在讓同仁如何利用機器學習方法挖掘數據背後的價值。	2018/6/6	2018/7/11	27

透過 12 天的教學內容，我們討論了如何使用 R 處理資料、整理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的 R 開源工具，輔以豐富的第三方 R 套件，能立刻分析手邊資料。

於課程結束後，亦舉辦了期末考，考試題目涵蓋所有課堂上教學的內容，並使用同仁一般在業務上會用到的資料做為測驗資料。測驗題目可連結到以下網頁：（http://rpubs.com/ywchiu/Exam20180903）題目如下圖所示：

疾管署R語言期末考題

David Chiu

9/3/2018

1. 請使用R 實作, 並以".R"檔繳交可運行之程式碼,以做為評分參考
2. 本期末考為Open Book, Open Computer 測驗,除抄襲其他人外, 所有公開資源都可以參考
3. 考試時間為一星期 (2018/09/03 12:00:00 ~ 2018/09/11 23:59:59)
4. 實作完期末考者，可以將.R 檔寄至 david@largitdata.com , 並在信件主旨上標注[繳交CDC R語言期末考]以做後續評分
5. 考試以60分為標準, 超過60分方算及格

基礎 R 語言 與資料視覺化 (12題共60分)

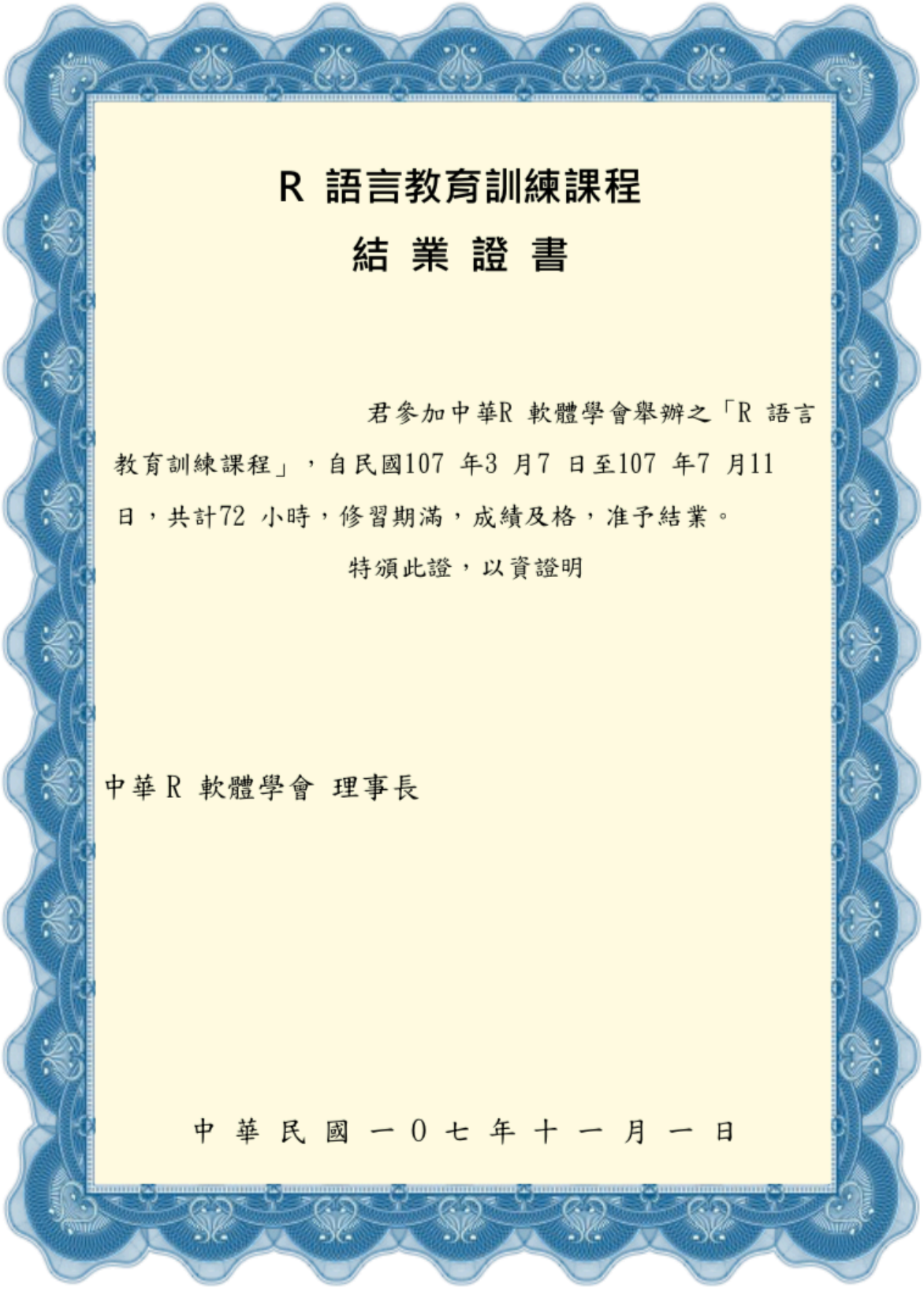
從疫情中心的開放資料網站蒐集到「登革熱近12個月每日確定病例統計」，資料如下:

```
library(readr)
dataset <- read_csv('https://raw.githubusercontent.com/ywchiu/cdc_course/master/data/Dengue_Daily_last12m.csv')
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   發病日 = col_date(format = ""),
##   個案研判日 = col_date(format = ""),
##   通報日 = col_date(format = ""),
##   確定病例數 = col_integer()
## )
```

圖三十二、期末考考題

經評量後統計考生資訊，共有 20 人參與考試，20 人全數通過考試. 為了嘉獎通過的同仁，亦頒發了中華 R 軟體學會的訓練課程結業證書，以肯定同仁們的學習成果以及資料分析能力。



R 語言教育訓練課程

結 業 證 書

君參加中華R 軟體學會舉辦之「R 語言教育訓練課程」，自民國107 年3 月7 日至107 年7 月11 日，共計72 小時，修習期滿，成績及格，准予結業。

特頒此證，以資證明

中華R 軟體學會 理事長

中 華 民 國 一 0 七 年 十 一 月 一 日

圖三十三、結業證書張樣

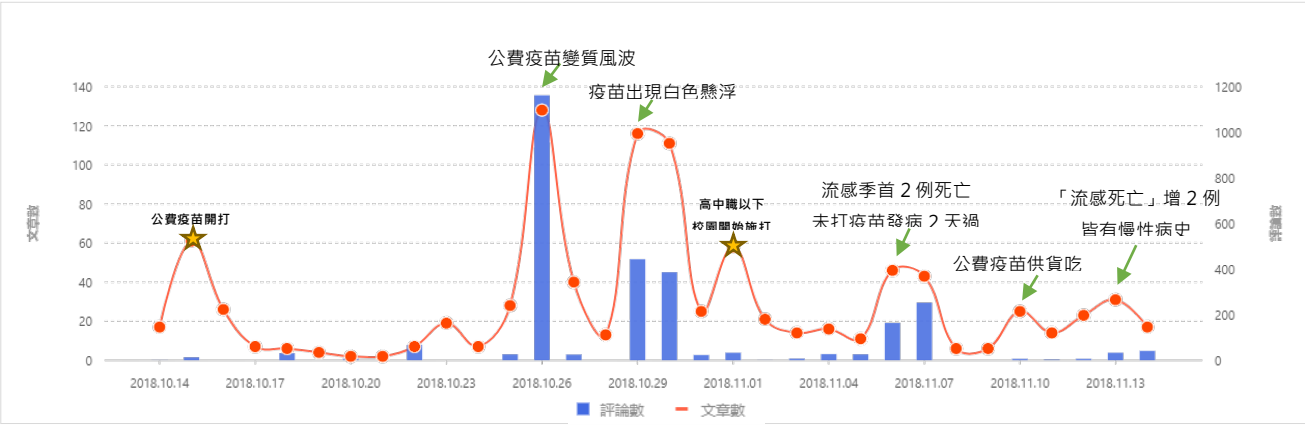
（四） 討論

1. 國內疫病輿情監測與分析

由於本計畫所建立的輿情分析平台廣納國內大部分的輿情資訊，因而分析人員可以輕易使用本系統，聆聽民眾心聲，即時發現公關危機，亦可以使用該系統建立輿情分析報表，分析事件始末。以下以最近最火紅的公費疫苗作為範例：

因應流感季的到來，衛生福利部疾管署宣布 10/15（一）起開放特定民眾1免費接種公費疫苗，避免流感找上門。然而，原是政府單位的一番美意，為何最終卻成了流感季未到，民眾怒火先到的窘境？

近一個月以來，公費疫苗總共引起 956 則網路聲量（圖 1），除疫苗開放施打日引起較多聲量外，引起最多聲量的原因莫過於十月底頻傳的疫苗問題。先是 10/26（五）發現巴斯德流感疫苗變色，一個週末過後，10/29

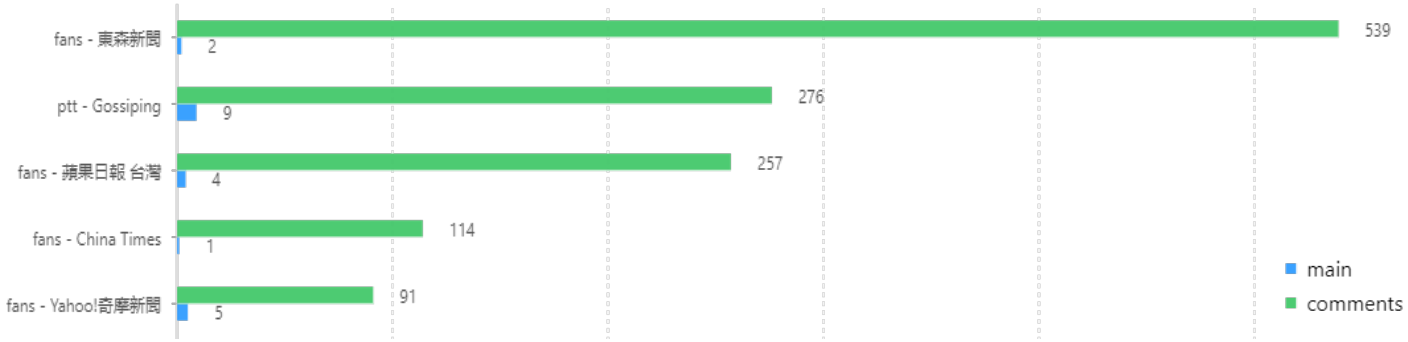


[1] 開放國內 50 歲以上成人，滿 6 個月以上至國小入學前幼兒、孕婦，及 6 個月內嬰兒之父母、幼保、安養機構人員、醫事及衛生等單位防疫相關人員，禽畜業者以及具有潛在疾病者，免費至各醫療院所施打。

圖三十四、疫苗議題的網路聲量

（一）流感疫苗內又出現白色懸浮物，接連異狀讓民眾氣得直跳腳，導致兩天單日聲量數都突破百則，更分別帶起各社群的討論熱度。

公費疫苗頻出包導致網路話題不斷，紛紛在各大網路社群引起留言討論，其中引發最多討論的平台是東森新聞以及蘋果日報的臉書粉絲專頁，PTT 的看板 Gossiping 也帶起不少討論度（圖 2）。



圖三十五、疫苗議題的網路討論度

面對接連發生的疫苗問題，網友在不同時間反應也不同，第一次（26 日）爆發變質疑慮時雖引起大量的批評聲浪，但仍有網友護航，表示還沒查清原因不用太驚慌。然而，第二次（29 日）發現異狀後，可以發現網友留言砲火更猛烈，直指政府的不是，更沒有網友願意跳出來表示抱持樂觀態度。直到首度出現未打疫苗死亡的憾事，部分網友才極力支持施打疫苗才有保障，不過仍有不少網友擔心疫苗的安全性（表 1）。

表十一、疫苗議題的網友反應

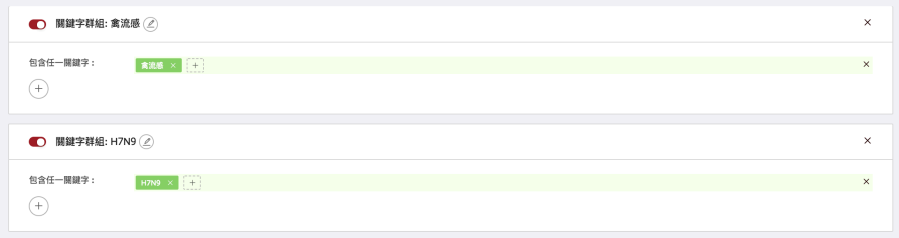
時間	話題	網友留言反應
10.26	公費疫苗變褐色	目前變色原因都還未查出，這麼驚慌失挫沒必要啦
		我昨天就施打了！你才給我爆這個！！
10.29－10.30	公費疫苗驚見白色懸浮物	不會太扯嗎？開放了，等有問題才回收，為什麼不一開始檢驗好沒問題在開放施打？
		最近蔡政府在冬眠嗎？ 台鐵出包，疫苗也出包，可不可以告訴我們什麼是安全的？
11.06-11.07	流感季首 2 例未接疫苗死亡	一定要打疫苗啊
		疫苗問題那麼多，想打但好怕

雖然事件爆發後，巴斯德疫苗立刻送原廠檢驗，11/1（四）也針對國光白色漂浮物疫苗澄清，證明是針筒碎片而非疫苗變質，但政府把關不嚴，處處螺絲拴不緊的既定印象恐怕還是很難抹去。政策美意加分不成，還讓民眾多日人心惶惶，此次公費疫苗接種大概也只剩下扣分的作用了！

為了能推動有感施政，從輿情了解民眾心聲，擬定政策方向，方能讓政策更貼近民心，讓機關單位受到民眾支持。因此建議未來疾管署可以利用該系統，檢視每次重大事件民眾的反應與輿論，並建立一套回饋機制，從系統蒐集民意，回饋疾管署做政策擬定，再將政策推動下去，蒐集民眾反應。便可以讓疾管署透過系統性的方法擬定有感政策。

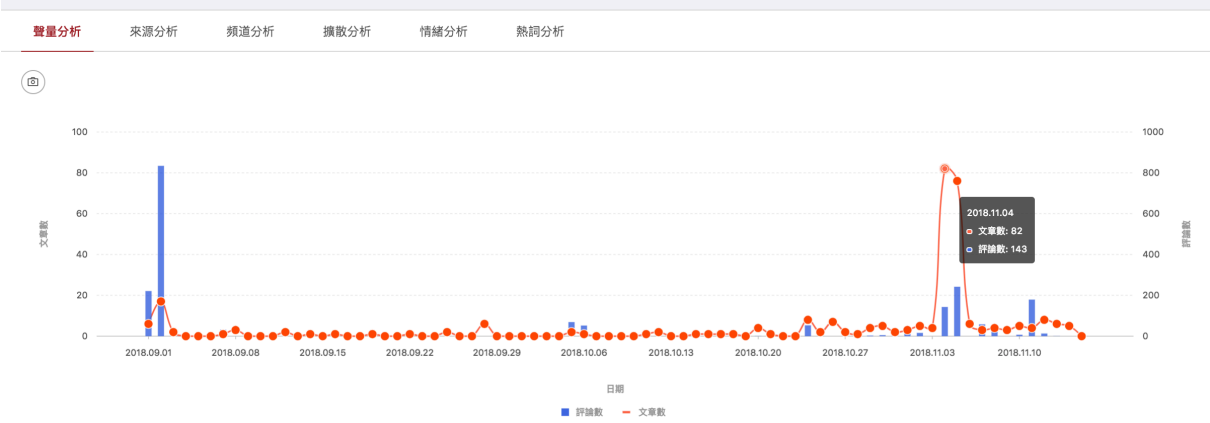
2. H7N9 疫病輿情監測與分析

日前 H7N9 人傳人的訊息鬧得沸沸揚揚，因此，我們將以 H7N9 做關鍵字，以了解系統的監測能力，並觀測訊息傳播的路徑與時序。我們首先設定 H7N9、禽流感作為觀測關鍵字組。



圖三十六、設定輿情關鍵字

接者，便可以切換到全頻道報表觀測事件發生的變化時序。於全頻道報表的設定下，我們將時間先切換成 9/1 ~ 11/15，我們便可以拉出這一段時間的聲量表。



圖三十七、H7N9 聲量圖表

從聲量圖表中，我們可以發現主文量最高峰(82 則)落在 11/4，但討論量最高(834 則)落在 9/2。

首先我們檢視 9/2 最主要的輿論是：「豬瘟、羊炭疽後驚爆 中國證實首例人感染 H7N4 禽流感」這篇新聞在 PTT 引發不少迴響，多半網友評論「支那真是全世界的禍害」，充分反映對對岸帶來傳染病的十足厭惡。

圖三十八、豬瘟、羊炭疽後驚爆 中國證實首例人感染 H7N4 禽流感

另外，我們檢視 10/4 的聲量高峰，發現當天討論最多的議題是「日本研究發現 H7N9 病毒可經飛沫傳播」。

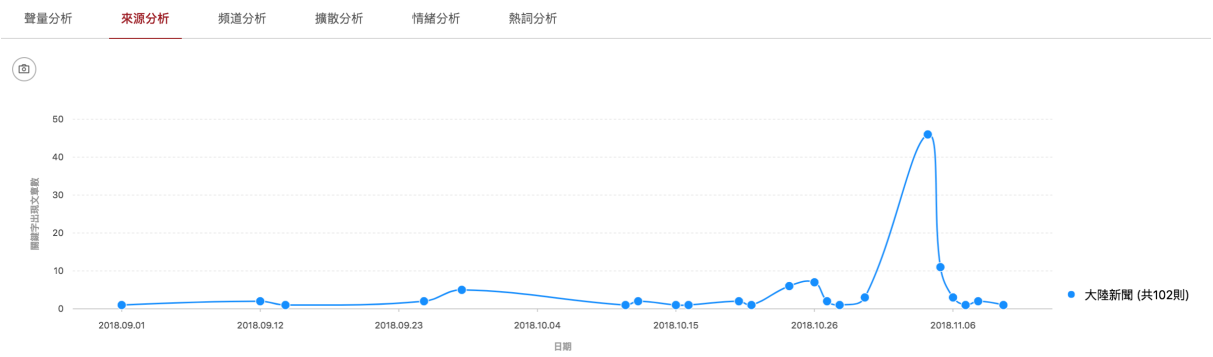
圖三十九、日本研究發現 H7N9 病毒可經飛沫傳播

多數傳媒都以公平公正的角度報導這篇發現，唯發現民視新聞卻在該臉書粉絲頁留下以下回文：

圖四十、民視新聞的臉書貼文

雖好像只是將事實陳述的方式重新排序，但這篇貼文完全沒有帶真實的新聞報導連結，敘述方式亦可能誤導聽眾誤以為 H7N9 現正在中國大爆發，實在有誤導聽眾之嫌疑。

為了能更深入了解到底是哪些資料來源對 H7N9 有諸多討論，我們即使用來源分析的功能，試圖找出聲量主要來源，從線圖中，便可以發現大陸新聞一共有 91 則對 H7N9 的相關討論，而在 10/4 就一共有 41 則相關的討論。移除掉中國新聞外的資料源後，重新檢視中國新聞源中 H7N9 的相關輿情，發現 10/4 的輿情亦集中討論日本大學的發現。

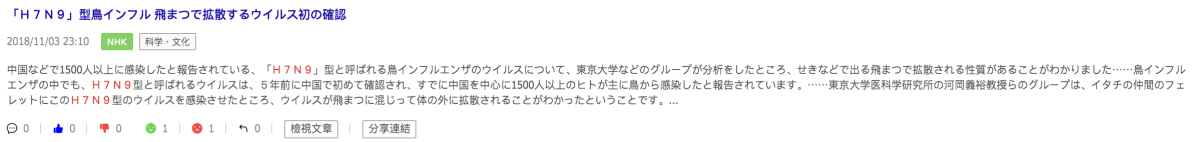


圖四十一、中國新聞對 H7N9 的聲量討論

但在線圖中，我們亦發現有另外兩波小山峰落在 9/28 與 10/24。檢視這兩天的輿情，可以發現 9/28 的討論主題為「家禽及時接種疫苗 減人染 H7N9 風險」，10/24 的討論主題為「廣西疫檢作假：3 只雞代表 60 只送檢 險引恐慌」代表中國各省級政府平時即對 H7N9 有採取一定的預防措施。

除了中國相關的新聞來源，我們亦有蒐集各計生委的發布新聞於大陸疾管專區，我們試圖搜尋一下是否各級政府有對 H7N9 的相關宣導，但卻發現除了定期的疫情匯報外，並無留下其他相關討論。

由於這次引發討論的來源，是來自日本新聞，因此我們分析外國的輿情資訊，試圖找出是否有相關報導。即發現 NHK 在 11/03 晚上 23:10 即有發布相關訊息：「Ｈ７Ｎ９」型鳥インフル 飛まつで拡散するウイルス初の確認，因此可以確定源頭資訊確實是來自於日本官方可信的媒體。

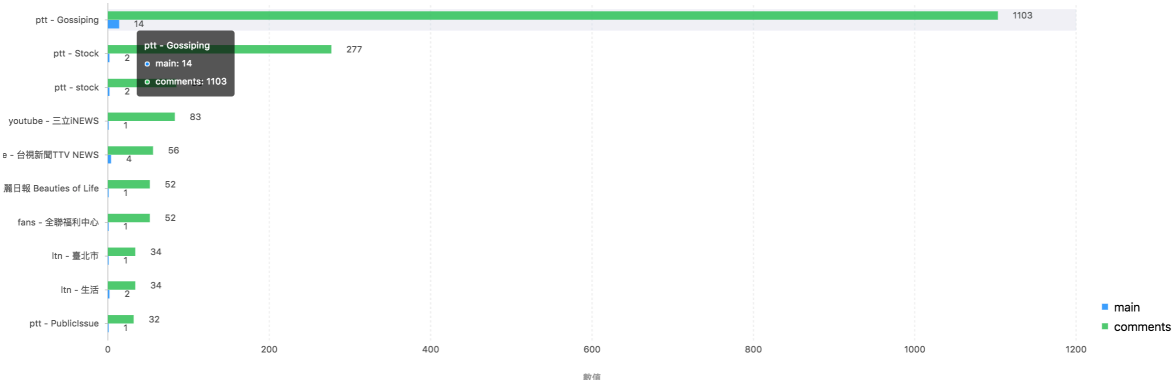


圖四十二、NHK 的討論新聞

除了查證資料的來源外，系統中亦有開闢一內容農場專區，我們可以從該專區尋找是否有 H7N9 相關的謠言。即發現內容農場「壹讀」擁有最多跟 H7N9 相關的輿情。其中以議題「100 年前，流感奪去 5000 萬人的生命，今天很有可能再發生」 最為聳動，文章的一開頭即號稱該研究來自約翰霍普金斯大學，嘗試增添該謠言的真實性，如該消息被透過 Line 或臉書或其他社群平台廣佈出去，將可能會引起許多不必要的恐慌。

圖四十三、壹讀所散布的網路謠言

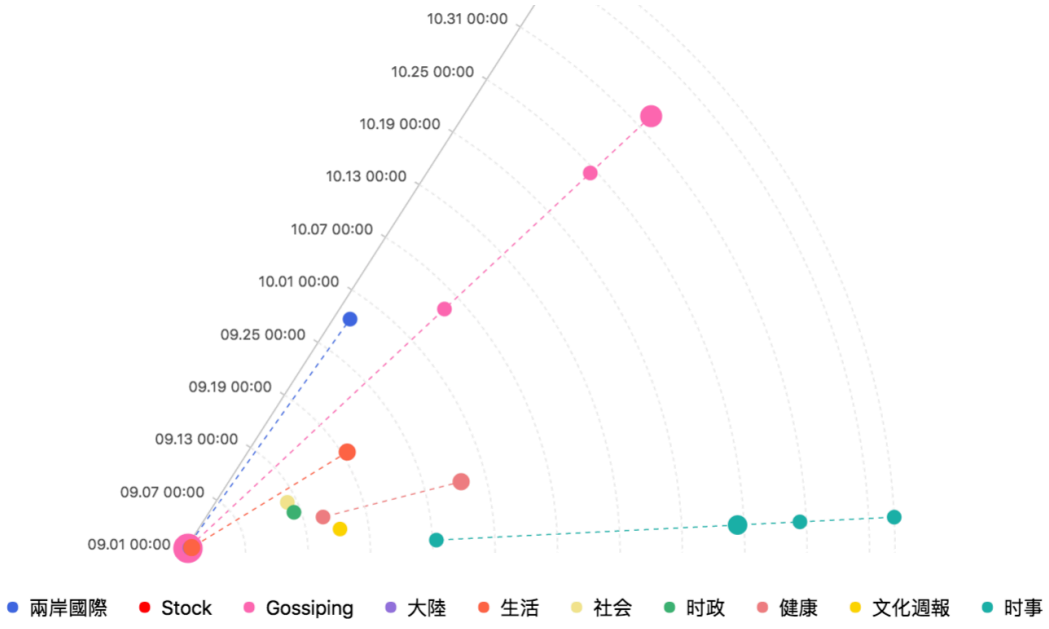
除了分析各新聞來源對 H7N9 的報導外，使用者可以善用頻道分析，分析相關新聞都是在哪個來源的哪個頻道被討論。



圖四十四、頻道來源分析

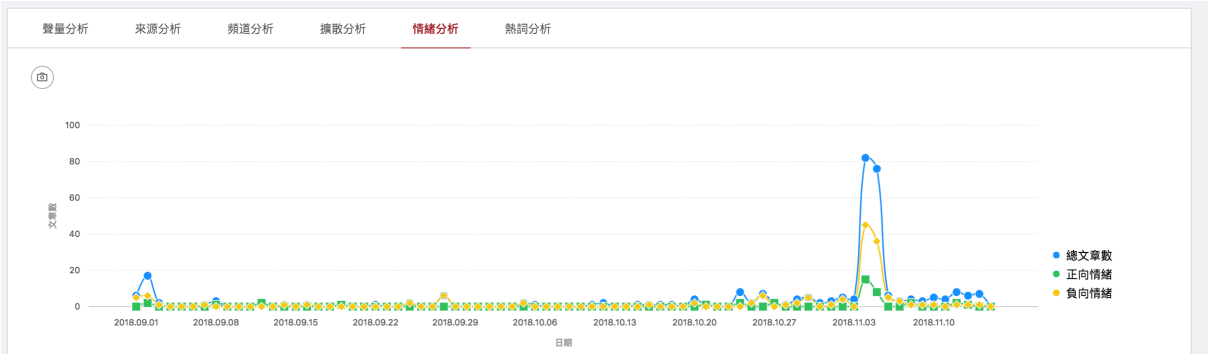
另外，使用者可以透過擴散分析，可以觀察相關新聞是否被廣泛的

被不同新聞源報導。如果擴散圖越分散，代表討論的來源越多，也就代表該輿情有不分板塊被廣泛的討論。



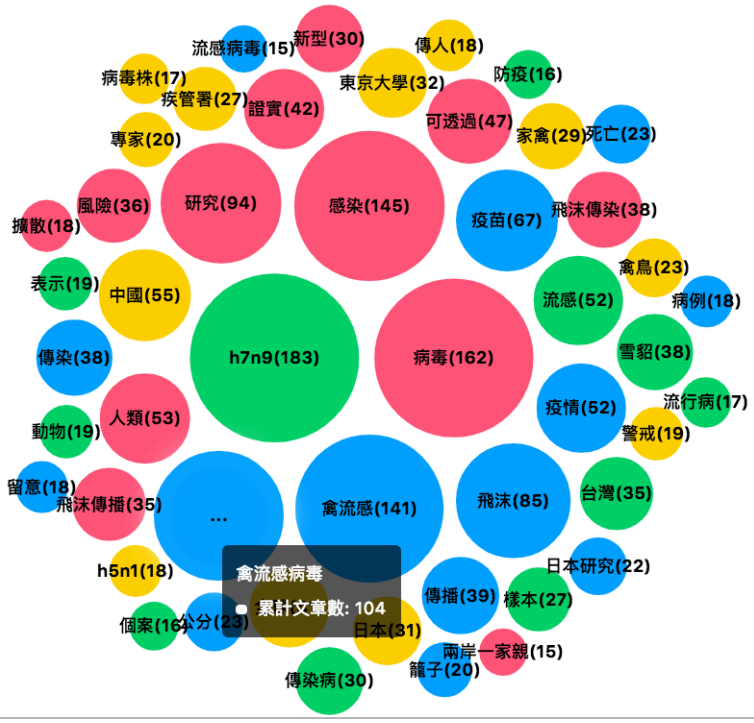
圖四十五、擴散分析

如要了解網民對該事件的看法，可以將分析功能切換到情緒分析。從圖中可以發現在 11/4 號當天，負面情緒明顯高於正面情緒，代表多數輿情為疫情傳染的負面消息。



圖四十六、情緒分析

此外，為了能關聯 H7N9 與其他的關聯詞，亦可以使用熱詞分析功能，便能探索出 H7N9 跟禽流感病毒與飛沫等字詞有相關性，可以以這些關鍵字詞作關鍵字，繼續探索是否有其他關聯新聞。



圖四十七、熱詞分析

（五） 結論與建議

從公關事件探查與新聞傳播及真偽性質確認等案例，可以發現輿情在公共治理扮演的角色越來越重要，因為網路的傳播性與渲染性非常強大，這些輿情訊息輕則影響公部門的聲譽，重則會造成社會動盪，因此在公共治理上，監測、分析、回應輿情的三步驟，必須是公部門不得不重視的課題。

在公費疫苗的案例中，可以發現民眾之所以接種公費疫苗其實就是出於對公部門的信任，但最近適逢選舉與普悠瑪事件等多個公共重大事件，

當該事件被聯想成政府的整體表現，民眾會把對既有政府的不滿加諸到這事件上，放大對該事件的反應，引發民眾強烈的不信任感。而雖然在第一時間，公部門雖然有對變色一事即時澄清，部分民眾也因為這樣的澄清，而提出相信政府的論調，但一旦引發第二次的疑慮，即會造成負面輿論高漲的情事。因此公部門在應對、澄清、把關、回應都需極端小心，以避免流感未到，負面輿論先行延燒。

在從 H7N9 例子的案例中，可以發現透過交叉比對，國內新聞、大陸官方頻道、大陸新聞、外國新聞與內容農場，便可以發現一事件新聞傳播的順序，便可以從各家媒體報導的一致性查證訊息的真偽。

往往在辨別訊息真偽時，消息來源是一個重要的判斷依據，以往我們會因為新聞記者站在第四權的角色，為新聞的真實性扮演審核與驗證的動作，但如今因為立場還有商業利益，有時候會發布些荒腔走板的內容。但所幸一經交叉比對後，即可發現民視新聞誤導人民以為 H7N9 正在中國大傳染的嫌疑，若民眾未能察覺該新聞確實有問題，往往可能會誤信貼在臉書的內容，而產生錯誤的結論。

另外，從訊息中亦可以發現內容農場也無所不用其極，對同一題材借題發揮，希望能透過聳動的消息，吸引眾人點擊，以獲取廣告利益，但若經無知或有心的民眾轉貼該訊息，而讓這樣的訊息在社交媒體上任意傳播，則有可能造成不必要的誤會與恐慌。

疾管署扮演具有公權力的角色，如能及早探查出這些假新聞，並適時處理，便可以有效遏止假新聞的傳播，還給民眾一個乾淨的視聽，避免錯

誤的疫病觀念引發社會動盪。

(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：

- 整合簡體中文的資料來源，蒐集所有中國政府的地方各級城市的官方疫情訊息，擴展系統的監測範圍，當有謠言產生時，可以透過該資料源驗證訊息真偽性。
- 增加簡體中文的詞典，讓系統得以處理簡體中文的斷詞，與識別文章中的名詞，提供資要的可讀性。
- 增加簡繁中文對譯的功能，讓使用者得以下繁體中文便能檢索簡體與繁體中文的資料。
- 增加兩岸三地語言的中文同義字典，讓監測系統得以歸納簡繁中文同義輿情資訊。
- 增加主要國家官方英文媒體的疫情訊息來源，補足資訊來源只有中文語系的不足，讓系統得以用英語監控世界各地的輿情訊息，強化系統的監控能力。
- 我們使用貝氏網路建立一輿情篩選模型，再將模型建立至即時警示系統，讓系統得以在收到相關輿情後，篩選出適當的輿情，並警示給權責人員，使權責人員第一時間收到最相關輿情資訊，而能過快速濾掉不相關的輿情。
- 舉辦 R 語言分析課程，提升疾管署內部人員對非結構化文字資料的處理與蒐集能力，提升同仁的數據分析技巧。課程內容包含：R 語言簡介、R 語言 ETL、R 語言的儲存與資料探索實務、R 語言與機器學習。

本計畫在第二年的目標中，已增添許多繁體中文與簡體中文輿情來源網站，並增加對系統對簡體中文的搜集、處理、分析的能力。此外，為了強化對國際疫情輿情的探測，今年亦增添了許多東協的英文官方媒體，希冀可以透過監測英文官方媒體，及早偵測出傳染病爆發的可能性。而透過幾個案例分析，已可知道透過輿情系統，政府可以了解公關事件發生的始末與辨別新聞的真實性，若能結合輿情分析與公共治理，及時理解民眾的需求與抱怨，並給予正確的回應，方能提升民眾對公部門的信任。此外，雖然疾管署扮演著防範假訊息散播的重要角色，但如果沒有民眾自發性的配合，也很難讓良善的政策能落實下去，因此未來可以利用系統偵搜出一些意見領袖，讓這些意見領袖在疾管署內發生，此時便能達到事半功倍之效，有效推廣良善政策。

經這兩年的開發下，雖然已經能透過系統，掌握區域性的輿情訊息，但離全球事件的探測卻尚有一段距離，因此在未來兩年的計畫中，希冀能藉由全球性的服務，如 Google、百度、Twitter 等平台擴增輿情的監控範圍，並納入區域性的分析功能，之後再搭配更多視覺化的分析，能夠讓疾管署以視覺化地圖快速掌握全世界的疫病資訊，再搭配既有的通知服務，方能確保隨時掌握不遺漏任何跟疫病相關的輿情，提升防疫能力。

等到確認疾管署可以運用該系統，有效的探勘非文字資料，以探勘出具威脅的疫病訊息源，之後便可以結合疾管署系統與資料，發展出一全面性的疫情預警與防堵平台，讓權責人員能透過自動且系統化的方式，防堵疫情，保護國民健康。

（七） 參考文獻

[1] Allcott H, Gentzkow M. Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives. 2017.

[2] FLETCHER, Richard, et al. Measuring the reach of “fake news” and online disinformation in Europe. Reuters Institute Factsheet, 2018.

[3] Georgia State University. "Study uses social media, internet to forecast disease outbreaks." ScienceDaily. ScienceDaily, 19 January 2017.

[4] Iafusco D, Ingenito N, Prisco F. The chatline as a communication and educational tool in adolescents with insulin-dependent diabetes: preliminary observations. Diabetes Care. 2000;23: 1853–1853.

[5] McIver D, Brownstein J. Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. PLoS Comput Biol. 2014;10: 1–8.

[6] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457: 1012–1014. pmid:19020500

[7] Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping Review on Search Queries and Social Media for Disease Surveillance: A Chronology of Innovation. J Med Internet Res. 2013;15: e147. pmid:23896182

[8] PHILLIPS, Lawrence, et al. Using social media to predict the future: a systematic literature review. arXiv preprint arXiv:1706.06134, 2017.

[9] ALESSA, Ali; FAEZIPOUR, Miad. A review of influenza detection and prediction through social networking sites. Theoretical Biology and Medical Modelling, 2018, 15.1: 2.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.
[15] K.J. Henning, What is syndromic surveillance?, Morbidity and Mortality Weekly Report (2004) 7-11.
[16] 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.
[17] 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.
[18] 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.
[19] M. Kulldorff, R. Heffernan, J. Hartman, R. Assunçao, F. Mostashari, A space–time permutation scan statistic for disease outbreak detection, PLoS Med 2 (2005) e59.
[20] H. Chen, R.H. Chiang, V.C. Storey, Business Intelligence and Analytics: From Big Data to Big Impact, MIS quarterly 36 (2012) 1165-1188.
[21] P. Zikopoulos, C. Eaton, Understanding big data: Analytics for enterprise class hadoop and streaming data, McGraw-Hill Osborne Media2011.
[22] 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.
[23] 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).
[24] 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.
[25] 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.
[26] O. Etzioni, The World-Wide Web: quagmire or gold mine?, Communications of the ACM 39 (1996) 65-68.
[27] 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.
[28] R. Cooley, B. Mobasher, J. Srivastava, Data preparation for mining world wide web browsing patterns, Knowledge and information systems 1 (1999) 5-32.
[29] 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.
[30] S. Shindelar, Big Data and the Government Agency, Public Manager 43 (2014) 52.
[31] United Nations Global Pulse, Big Data for Development: Challenges & Opportunities, New York, 2012.
[32] 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.
[33] 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.
[34] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Advances in knowledge discovery and data mining, (1996).
[35] 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.