

社群資料研究：關鍵字詞分析與趨勢追蹤

Social media data research: Keywords analysis and
trend tracking



行政法人

國家災害防救科技中心

National Science and Technology Center
for Disaster Reduction

NCDR 108-T04

社群資料研究：關鍵字詞分析與趨勢追蹤

Social media data research: Keywords analysis and
trend tracking

劉致灝、蔣佳峰



行政法人

國家災害防救科技中心

National Science and Technology Center
for Disaster Reduction

國家災害防救科技中心

中華民國 109 年 01 月

中文摘要

本研究以本中心於災害應變期間實行之社群網路蒐整流程為基礎，結合統計與文字探勘技術，從文章中初步分析熱門字詞，並透過機器學習找出特定領域之權威詞(Power Term)，且依照字詞類型加以分類。研究成果亦將整併至攀爬平台中，作為後續應變期間分析之使用，目前提供關鍵字過濾文章、地理詞擷取、文章主題分析走勢、權威詞分析等。本文以尼莎、海棠颱風為解說範例，說明地理詞擷取、災情敘述擷取結果以及聲量趨勢變化。

關鍵字：社群網路、統計分析、文字探勘、機器學習

ABSTRACT

This research is based on the process of collecting, analyzing and integrating the disaster information from social web sites. The topic combines statistics and text mining technology.

The first step is to analyze popular words from articles, and find "Power Terms" in a certain domain. According to the type of those keywords, we will categorize them by using machine learning. The research results will also be integrated into the platform for analysis during the period.

Keyword filtering, location related words extraction, article topic analysis, and power terms analysis are currently provided functions by the platform. We use Typhoon Nesat and Haitang as examples to illustrate the extraction of location related words, the extraction of disaster descriptions, and the changes of social volume.

Keywords: Social network, Statistics, Text mining, Machine learning

目錄

第一章 社群案例應用	3
第二章 社群攀爬平台系統架構與功能介紹	5
第三章 關鍵字詞分析與趨勢追蹤	8
關鍵字詞分析	9
關鍵字詞分析-統計分析	9
關鍵字詞分析-命名實體擷取	12
趨勢追蹤	14
第四章 尼莎海棠颱風事件案例說明	16
第五章 結論與建議	20
附錄 參考文獻	21

圖目錄

圖 1 攀爬平台系統架構	5
圖 2 攀爬平台查詢介面	6
圖 3 社群資料來源分布	7
圖 4 關鍵字詞分析-統計分析之簡要流程.....	10
圖 5 尼莎、海棠颱風關鍵字詞分析範例	12
圖 6 序列標記範例說明	13
圖 7 文章標記範例	14
圖 8 權威詞擷取範例	14
圖 9 尼莎、海棠颱風社群資料之關鍵字擷取結果.....	17
圖 10 尼莎、海棠颱風社群聲量變化趨勢	18
圖 11 屏東地區社群關注熱度變化	18
圖 12 南部地區社群關注熱度變化	19

表目錄

表 1 尼莎、海棠颱風關鍵詞擷取數量	16
--------------------------	----

前言

網際網路的發展加速資訊交換的效率，隨著時代的演進，人們從早期坐在座位上，使用桌上型電腦瀏覽網頁資訊，可謂 Web1.0 的時代。而後網路發展提供使用者互動的機制，單方向接收資訊的模式已然改變，漸漸地，討論區與論壇等交流平台開始被廣泛使用，成為交換訊息之 Web2.0 時代的代表作品。

除了傳遞訊息平台的演進，硬體的製造技術使得設備逐漸縮小化，使用者從笨重的桌上型電腦，到隨身帶著走的筆記型電腦，最後變成輕薄可放入口袋的智慧型手機，社群網路的演進與方便攜帶的硬體設備相輔相成之下，交換訊息的頻率越發迅速了。

藉由網路服務建立社交圈，同時網路服務也提供個性化聚合服務，讓使用者取得個人感興趣的資訊，並能分享個人化的資訊至社交圈中，形成當下最流行的社群網路平台，即是所謂的 Web3.0。

從網路科技服務的發展歷程反應出使用者對於科技使用的需求，同時也改變使用者以往在資訊傳遞的方式。現今網路服務的使用者習慣將網路上瀏覽到的資訊，藉由社群網路服務平台發佈或轉貼，如旅遊資訊、美食情報或政治議題等。這些資訊會被其他社交圈或是網民瀏覽到，並且不斷的被轉載或觀看，也讓資訊的傳遞速度變得非常的

快速。同樣地，當災害發生期間，民眾間最新災情的訊息也隨之往來於各大社交平台之間，若能善加利用這股能量，對於災情的取得、提供給指揮官於應變期間的參考資料，有助於現場情勢的掌握。

目前傳遞於社群網路的資訊多屬文字資料，遺憾的是，這些資料的真實性參差不齊，文章結構亦無統一格式，即自由文本(Free Text)，錯別字、標點符號的瑕疵等，使得文章的品質難以兼顧，在文本分析上實屬一大難題。但是，對於關鍵字詞的掌握，我們仍然可以透過簡單的統計方法加以計算，僅僅是在災害發生期間，統計頻率最高的用詞，就能顯露出平常少見的關鍵字，如同台灣人民適逢大選期間，討論話題偏向政治議題一般，當發生颱風、地震等災害事件之時，討論的話題不外乎「淹水災情」、「大樓倒塌」、「募集物資」等議題，平日討論的多元話題在特定事件發生後趨向於單一主題，在初步的關鍵字掌握上較為容易許多。

除了簡單的統計分析外，自然語言處理(Natural Language Processing, NLP)也是當今常見的文字探勘(Text Mining)技術，本研究所適用的方法為條件隨機域(Conditional Random Field, CRF)，將文字抽象化為特徵矩陣，進而訓練學習模型，達到擷取特定領域的關鍵字。

第一章 社群案例應用

日本，與台灣相似地常遭受地震、颱風等天災影響，Takeshi Sakaki 等人建立時間推估與空間定位的辨識模型，以社群網站的文章數量、內容，推估地震事件的發生時間與地點，使用日本網民常用的推特(Twitter)作為主要資料來源，根據文章發文時間點與發文者的 GPS 座標之間的關係。

他們統計文章出現數量之變化，估算災害發生的時間點，並藉由地震等災害字詞獲取相關文章，建立以下三種特徵：統計特徵（文章字數、關鍵字出現於文章位置）、關鍵字特徵，以及查詢詞的前後詞特徵。依此做為分類器的特徵進行分類，並使用相關文章的 GPS 座標，以推估災害地點 (Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo ,2010)。

美國在社群網路也有相關災情判識之研究，Yury Kryvasheyeu 等人針對發生於美國 2012 年的珊迪颶風，亦利用 Twitter 當作資料來源，透過使用者的個人資料，從他們自身提供的住址資訊，結合事發當下發文者自身的 GPS 座標，對照發文內容是否與災情相關，根據其密集區域推估災害發生的地點。

同時，根據文章內容之情感描述，推論出災害發生的時間範圍。他們認為一般人會在社群網站上發表文章，通常是以分享喜悅

為目的，文章多半使用一些較開心、正向的字眼。然而在災害發生期間，文章內容則偏向敘述災情、報平安等，負面情緒的字眼亦會相對的增加。實驗過程觀察文章內容的平均情感值是否跌至負數，以此推估可能的發生時間(Kryvasheyev, Yury, et al. ,2015)。

第二章 社群攀爬平台系統架構與功能介紹

在本中心建置的社群攀爬平台中，透過資料攀爬程式 (WebHunter) 擷取各大社群網站、新聞等資料來源，除了將原始資料進行結構化倉儲外，亦透過資料分析模組 (ENLP) 進行各項資料分析應用，資料的分析結果與查詢皆以 API 進行操作，詳細如圖 1 所示。

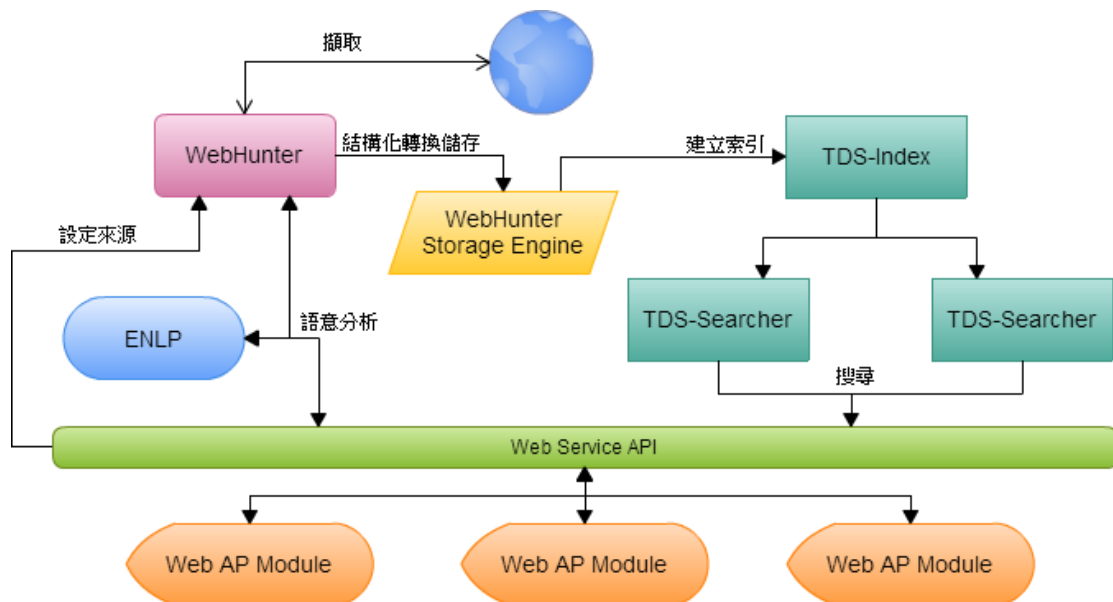


圖 1 攀爬平台系統架構

圖 2 為資料查詢介面，可根據時間選擇區間，選擇該時段發布之文章集，中央的來源列表可決定資料來源的種類，右側的主題列表則透過事先設定的正規表達式 (Regular Expression, RE) 初步篩選文章，大量減少非相關文章之數量，以颱風為例，當我們遇到颱風侵襲時，容易聯想到的關鍵字常為水災、淹水、豪雨等字眼，視當時災情決定適用之災害關鍵字集合，使用 AND(&)、OR(|)、NOT(!) 等

邏輯運算元組合，達到初步篩選之行為。

查詢條件

期間：
 最近3小時
 最近6小時
 2019/09
 2019/09/01 00:00:00
 2019/10/31 00:00:00
 可直接以輸入方式修改上方自訂查詢時間範圍

來源
 討論區
 社群網站
 facebook
 facebook-event
 facebook社團
 bbs
 新聞

主題：
 土石流聲量
 水災災情關鍵字
 放假測試
 測試測試
 花樣作嘢
 測試-梅姬
 0823測試
 測試之

關鍵字篩選： 請輸入搜尋關鍵字...
作者篩選： 請輸入作者關鍵字...

- 【#生活】最雷停車位！行家一看地面「2孔蓋」急勸退一位網友正買房準備下訂金，...**

【主文】【#生活】最雷停車位！行家一看地面「2孔蓋」急勸退一位網友正買房準備下訂金，但是一看到停車位有點疑慮，「不曉得這2個孔蓋的作用是什麼」，其他剩下的停車位都靠牆壁邊緣或往下的坡道前，不方便停車。

共 0 則回文 | MSN 新聞 2019-10-28 09:45:00 | 共 0 次點閱 | 0.0 | 55.3
[Facebook 粉絲團](#) > [MSN 新聞](#) | [庫存內容](#) | [加入追蹤](#) | [本篇地理詞](#)
- 【#娛樂】認愛5個月就分！許維恩爆「被利用」呂銳突然在IG自曝「我和她選擇回歸...**

【主文】【#娛樂】認愛5個月就分！許維恩爆「被利用」呂銳突然在IG自曝「我和她選擇回歸到朋友的位置」，不少粉絲得知後，便湧入雙方各自的社群平台寫下安慰字句，但就有一名許維恩的友人疑似對呂銳不滿。

共 0 則回文 | MSN 新聞 2019-10-28 09:30:00 | 共 0 次點閱 | 17.6 | 23.4
[Facebook 粉絲團](#) > [MSN 新聞](#) | [庫存內容](#) | [加入追蹤](#) | [本篇地理詞](#)
- 【#香港】史上第2次！半島酒店庇護民眾港警挨批比皇軍凶殘香港民眾27日再度發...**

【主文】【#香港】史上第2次！半島酒店庇護民眾港警挨批比皇軍凶殘香港民眾27日再度發起梳士巴利花園「追究警暴」集會，下午3時許，示威者走出梳士巴利道，警民衝突隨即爆發，警方發出多枚催淚彈，造成半島酒店外民眾感到不適，逃到半島酒店避難，工作人員並未阻止，而後避難的民眾從其他通道離開

共 0 則回文 | MSN 新聞 2019-10-28 09:01:00 | 共 0 次點閱 | 19.6 | 52.2
[Facebook 粉絲團](#) > [MSN 新聞](#) | [庫存內容](#) | [加入追蹤](#) | [本篇地理詞](#)

圖 2 攀爬平台查詢介面

本研究使用的資料為攀爬平台擷取之社群攀爬資料，著重於尼莎、海棠颱風事件，詳細資料設定如下所示：

1. 時間：106 年 7 月 29 日至 7 月 31 日
2. 資料來源：PTT(八卦板、颱風板、各地方板等)、臉書(新聞媒體粉絲團、在地社團、熱門社團等)、噗浪搜尋(Plurk)
3. 關鍵字設定：「淹?水|溪水|暴漲|路樹|積水|暴漲|溢堤|淹水|水位|豪?雨|暴雨」與「((積水|淹到|淹|淹水)&(輪胎|層|膝蓋|小

腿|大腿|公尺|公分))|(雷雨|豪雨|飄雨|驟雨|風雨|有風有雨||滲
水|滲漏)」

時間取自發布尼莎颱風之陸上颱風警報開始，至解除海棠颱風之海上颱風警報為區間，根據本次事件為颱風事件，使用與淹水相關的關鍵字眼進行初步篩選，主要為淹水相關災情字詞，如淹水、豪雨災情、淹到腳踝、輪胎等程度詞，並抓取共計 9,355 筆留言，其中，前三名資料佔比為 Facebook 粉絲團、PTT、Facebook 社團，詳細如圖 3 所示，臉書粉絲團主要透過記者拍攝災情照片後由社團的小編上傳照片、發文，民眾瀏覽照片後再進行留言；PTT 與臉書社團皆為一般民眾之討論內容，品質與內容則相對較差。

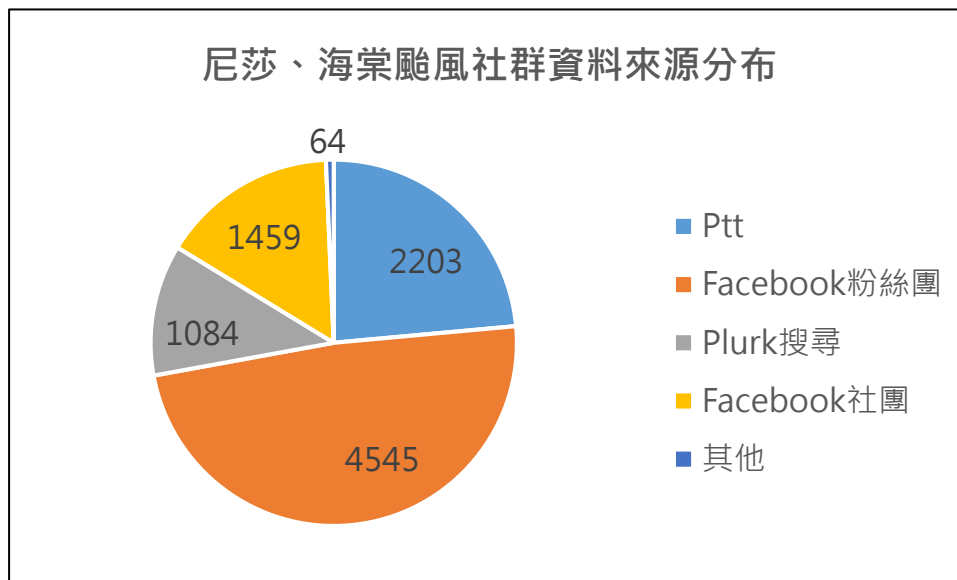


圖 3 社群資料來源分布

第三章 關鍵字詞分析與趨勢追蹤

關鍵字詞分析在災防領域的應用上，著重於社群網路討論的災情文章進行關鍵字擷取，試圖找出帶有時間性(時間相關詞類、發文時間等)、空間性的熱門字詞(如地名、知名地標等)，藉此推論何地或何地可能發生嚴重災情，比起透過電話進行通報的事件數量，一般民眾在討論各地災情相關文章，其資訊量必然遠超過前者，能否快速地在茫茫文海中，找出民眾熱烈討論的地點、災情描述，是為主要課題。

在時間推移下，一起災害事件應有它的生命週期，如同災害管理的四個階段：減災、備災、應變、復原，災害事件從產生、處理、最後結束的過程，同時亦會反映在民眾討論的議題上，關鍵字隨著時間的趨勢變化也能顯示出民眾目前關心的議題為何。

在事件發生初期，民眾討論的主題在於什麼地方有災情、災情敘述為何，熱心民眾會回報最新的災情資訊以求支援，到了事件中后期，整體事件情勢獲得控制後，零星的地方災情回報開始減少，主要嚴重受災區域持續受到關注，緊繃的心情開始逐漸放鬆，民眾的閒言閒語也相繼產生，到了事件末期階段，則開始討論責任歸屬、損失評估，從事件的討論逐漸轉移至執政者對於事件應變的效率評價、或討論經濟的損失、以往歷史事件的比較等。

本章節將針對本研究在關鍵字詞分析與趨勢追蹤之方法進行說明與案例示範。

關鍵字詞分析

關鍵字詞分析主要分為兩個主軸：1. 利用統計分析計算字詞頻率的初步分析，藉此快速找出熱門關鍵字詞，此作法較為快速，但無法指定特定領域字詞，2.命名實體辨識(Named Entity Recognition, NER)，透過觀察文章特徵的機器學習，將文章結構抽象化，從中擷取特定片段，搜尋特定領域之關鍵字詞，這種作法較為費力，需標記大量訓練資料、設計合適特徵。

關鍵字詞分析-統計分析

人類能閱讀文章並找出關鍵字句，主要為能理解文章中各單詞的意義(Meaning)，對於程式來說，文章中的字詞本身不具有任何意義性，只是由不同的字元組合的句子，這些詞義難以將其數據化，故無法被程式所「理解」。

在這前提下，簡單的做法是：無視每個字詞的意義，所謂熱門的關鍵字詞，照常理推論應是當下被熱烈討論、提及的字眼，所反映出的應該是異常高的曝光率，那麼，我們可以僅計算各詞類自身的出現在單一文章的詞類頻率(Term Frequency)，表示該詞類被提及的重要程度；計算該字詞在各個文章的出現的文本頻率(Document

Frequency)，表示該字詞的應用廣泛程度，透過統計分析的方式，找出具有代表性的關鍵詞類。

作為文章的前處理，考慮到來自社群網站的文章多屬自由文本 (Free Text)，即書寫文章的格式毫無規定，容易出現不必要之冗言贅字，故須進行初步的清除作業，先把每一個句子進行斷詞處理，後將多餘的停用字詞(Stopwords)移除。

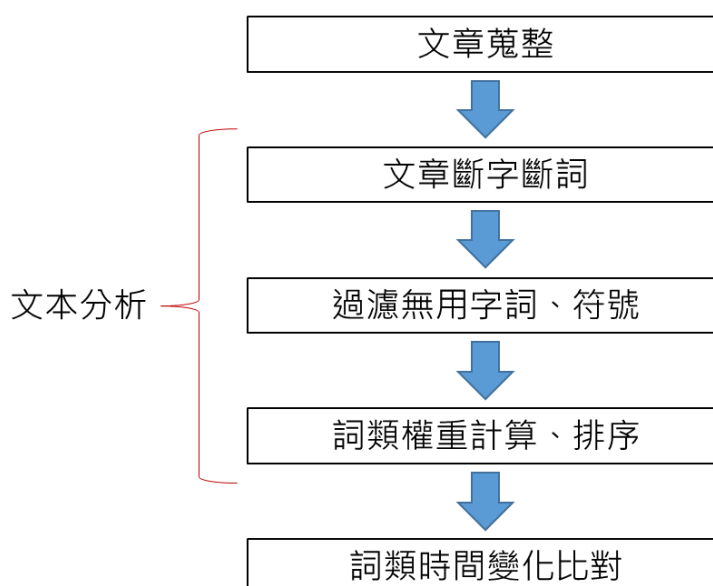


圖 4 關鍵字詞分析-統計分析之簡要流程

以統計分析的方式，作為直觀的權重計算方法共有三個：(1)計算特定詞類的總出現頻率(Term Frequency, T_f) (2)計算特定詞類出現於每一篇文章頻率(Document Frequency, D_f) (3)結合前述兩種計算方法，出現頻率與文章頻率倒數之乘積(TF-IDF)。

我們以尼莎海棠颱風為範例進行說明：尼莎與海棠颱風於 2017 年 7 月 26 日至 7 月 31 日影響台灣本島，本次使用共 1000 筆資料，

主要資料來源來自臉書，由圖 5 所示，可以得知使用 TF-IDF 方法的前 20 名關鍵字，其內容較難以顯示與尼莎、海棠颱風相關之關鍵字詞，相對的，單單使用 TF 或 DF 所得到的關鍵字，內容較為接近災害發生期間，民眾談論颱風災害相關的詞語。

比較三者顯示的內容，與考慮到公式的計算方法，我們可以知道：「淹水」的 TF 值為 725，而「淹水」的 DF 值為 495，TF-IDF 值則為 $725/495=1.46$ ，屬於相當低的統計結果，理應不會出現在 TF-IDF 的前 20 名中，主要因為災害發生期間，民眾討論熱度大量集中於災情上，導致相關詞頻(TF)大幅提升，這些詞類散布在各文章的比例也隨之升高，進而造成 DF 值也隨之提升，TF-IDF 值也相對降低了。

	TF-IDF		TF		DF
住屋	11.4	淹水	725	淹水	495
事實	9.67	颱風	283	颱風	100
居住	8.5	屏東	152	屏東	89
救助	7.4	海棠	152	海棠	59
費用	7	尼莎	110	地區	50
地下道	7	造成	107	尼莎	48
月台	7	佳冬	107	造成	47
魏明谷	6.5	林邊	80	政府	44
伍婉華	6	屏東縣	79	嚴重	42
鄉焰	6	居民	68	災情	41
理賠	6	嚴重	66	屏東縣	39
洪水	6	政府	64	恆春	38
水氣	6	民眾	61	大雨	37
五河	6	災情	58	林邊	36
校園	5	住屋	57	台灣	36
芭樂	5	豪雨	54	影響	34
投保	5	地方	53	民眾	34
前天	5	積水	53	問題	32
爸爸	5	公分	51	東港	32
林揆	5	台灣	51	積水	32

圖 5 尼莎、海棠颱風關鍵字詞分析範例

關鍵字詞分析-命名實體擷取

前述的方法乃忽略詞義，僅透過統計方式進行篩選熱門詞類，命名實體擷取(NER)則透過人為定義的方式，進行額外的特徵賦予，也就是用人工的方式替不具意義的詞類，強行賦予一些能夠被運算的特徵，例如常見前、後綴字詞、詞性、句首(尾)、英數字、標點符號等，利用事先建立的詞庫進行比對，為字詞增加特徵，作為機器學習的特徵使用。

針對文字處理的模式，目前主要透過人為賦予的方式，針對詞類提供人造特徵、標記欲辨識的答案，將文章抽象化成數據後，提

供程式學習文章的撰寫模式，我們將此行為稱之命名實體擷取 (Named Entity Recognition, NER)。

NER 的實作目前主要透過序列標記(Sequence Labeling)達成，必須先行將文章進行標記化(Labeling)，將每一個句子拆解成一個個標記，概念如圖 6 所示。

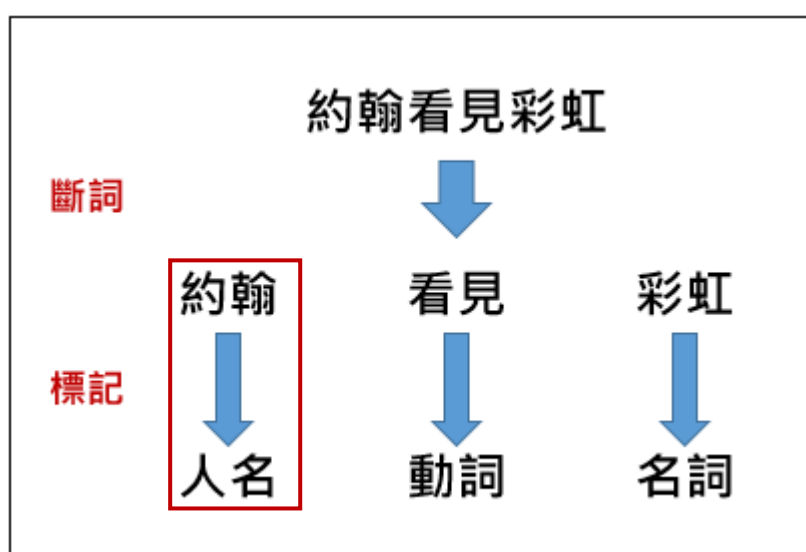


圖 6 序列標記範例說明

約翰這個字被標記為人名，代表我們想找的目標，我們可以藉由標記後的結果，判斷命名實體的位置。序列標記模型基於訓練資料原文、人工的標記結果，配合訓練句子本身的句型建立，常見的作法有隱性馬可夫鏈 (Hidden Markov Model, HMM) 及條件隨機域 (Conditional Random Field, CRF)。

本研究使用的 CRF 是一種模式識別及機器學習的建模方法，由 John Lafferty 等人於 2001 年提出，用於分析序列資料，如自然語言或生物序列。CRF 是一種無向性的機率圖形模型，針對給定的句子，我

們考慮相鄰的字與字之間，其標記結果是否有關係性，希望能從中找出最佳的標記組合，透過訓練資料轉換成的編碼，觀察出已知上下文的關係，並建立出一致的解釋。對於一個字來說有 5 種標記(BIESO)可能，我們的目標為從一個句子可能產生的所有標記組合中，找出機率最有可能的配對。

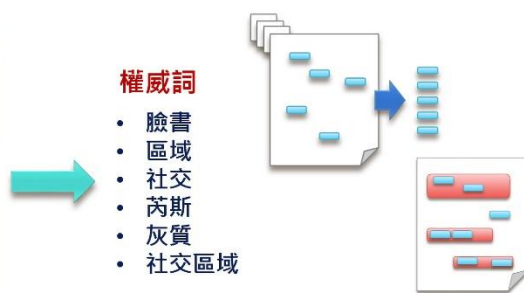
透過機器學習技術，針對非結構化文件內既有**特殊屬性的人事時地物**詞彙進行擷取，建立特殊屬性模型。



姓名	王大明
組織名	家樂福
地址(點)	台北市忠孝東路四段
銀行帳戶	088888888888888
身分證號	E123456789
車牌	1234-AB
EMAIL	123@gmail.com
日期	2015年5月12日
金額	新台幣300元
物品(刑)	金飾、壓制器、電源線、狗鍊、空彈殼、海洛英
事件(刑)	行車不慎、毆打、踹開、詐騙、遭竊、攻擊、催討

亦可透過標記訓練建立特殊屬性模型如水災辨識模型

圖 7 文章標記範例



採用無詞庫斷詞法，應用字詞的位置、次數、上下文之關係，自動找出文件內重要的關鍵詞

圖 8 權威詞擷取範例

趨勢追蹤

趨勢追蹤在社群網路上根據時間或空間上的推移，進而觀察事件走勢或民眾討論主題變化之行為，具空間特性的地點類關鍵字可

透漏該區域可能出現嚴重災情。

我們可以由前述所獲取的熱門關鍵字作為參考，觀察不同時間下的頻率變化，推估其影響時間、嚴重程度；根據不同時間的熱門關鍵字分布，推估目前民眾的談論主題。

以颱風或地震為例：根據災害當下所獲取的地點類關鍵字，觀察其數量變化，在出現頻率急遽提升的時間點，有較高的機率是處於受到災害衝擊的 1~2 小時之時間區間內，在災害應變的過程中，民眾討論的主題也隨著事件處理演進，逐漸由災區地點討論，轉而進行災後檢討，反映的關鍵字描述也會由地點類、災情敘述類，開始提及執政者、地方政府、經濟損失等相關字眼，這些民眾的輿論分析皆可如此從中觀察之。

第四章 尼莎海棠颱風事件案例說明

本次案例使用尼莎與海棠颱風災害事件作為說明，尼莎颱風影響台灣時間為 106 年 7 月 28 日自宜蘭蘇澳登陸，沿途影響屏東南端沿海地區造成豪大雨，22 時從苗栗竹南出海，海棠颱風則在 7 月 30 日登陸屏東楓港，再一次影響屏東地區，後轉西北方向前進，7 月 31 日從彰化芳苑出海，夾帶的西南氣流帶來南部地區豪大雨災情。

本次測試資料為 PTT、臉書粉絲團等社群資料，於 7 月 29 日至 7 月 31 日共計 9355 筆資料做為測試，分別以地理詞、災情敘述等項目進行擷取，共擷取出 5485 筆地理詞、3036 筆災情敘述詞，詳細如表 1 所示。

表 1 尼莎、海棠颱風關鍵詞擷取數量

文章數量	地理詞擷取數	災情敘述擷取數
9355(筆)	5485(筆)	3036(筆)
	地理詞類別數	災情敘述類別數
	1006(種)	55(種)

將這些詞彙進行詞頻(TF 值)排序後如下表所示，屏東、高雄、台南、嘉義等南部地區為本次受災之主要區域，宜蘭、台北、基隆等北部區域，則為尼莎颱風登陸時影響之區域，其餘地區顯示兩個颱風移動路徑或受災區域。

災情敘述的部分則多以淹水成災等詞彙有關，尤其民眾在討論災情時僅訴說該區域「淹水」而已，此詞被提及 1543 次，超過總詞頻之一半，相關的災情：停電、土石流、道路中斷、倒塌等災情。

地點相關詞彙	頻率	災情相關詞彙	頻率
屏東	344	淹水	1543
高雄	311	停電	78
台南	235	土石流	51
台灣	225	坍方	40
臺灣	129	落石	39
嘉義	104	道路無法通行	13
宜蘭	100	停水	11
恆春	77	溪水暴漲	9
台北	72	路樹倒壓死人	9
東港	68	水深約3公分	7
馬祖	68	道路封閉	6
屏東縣	66	路樹倒塌	5
南投縣	66	路面淹水	4
基隆	64	道路淹水	3
仁德	63	樹木倒塌	3
彰化	62	路面積水	3
苗栗	62	路面有較高積水	1
台南市	55	水深約1公分	1
佳冬	55	屋頂連同椽木塌毀	1

圖 9 尼莎、海棠颱風社群資料之關鍵字擷取結果

聲量的趨勢追蹤部分則可以對照本中心之「2017 尼莎暨海棠颱風災害報告」之時間點進行比較，民眾討論聲量分別在整起事件的開始與結束為最高討論熱度，分別為 7 月 29 日 14 時發布尼莎颱風海上颱風警報，與 7 月 31 日 8 時解除海棠颱風海上警報作為兩個討論熱點。

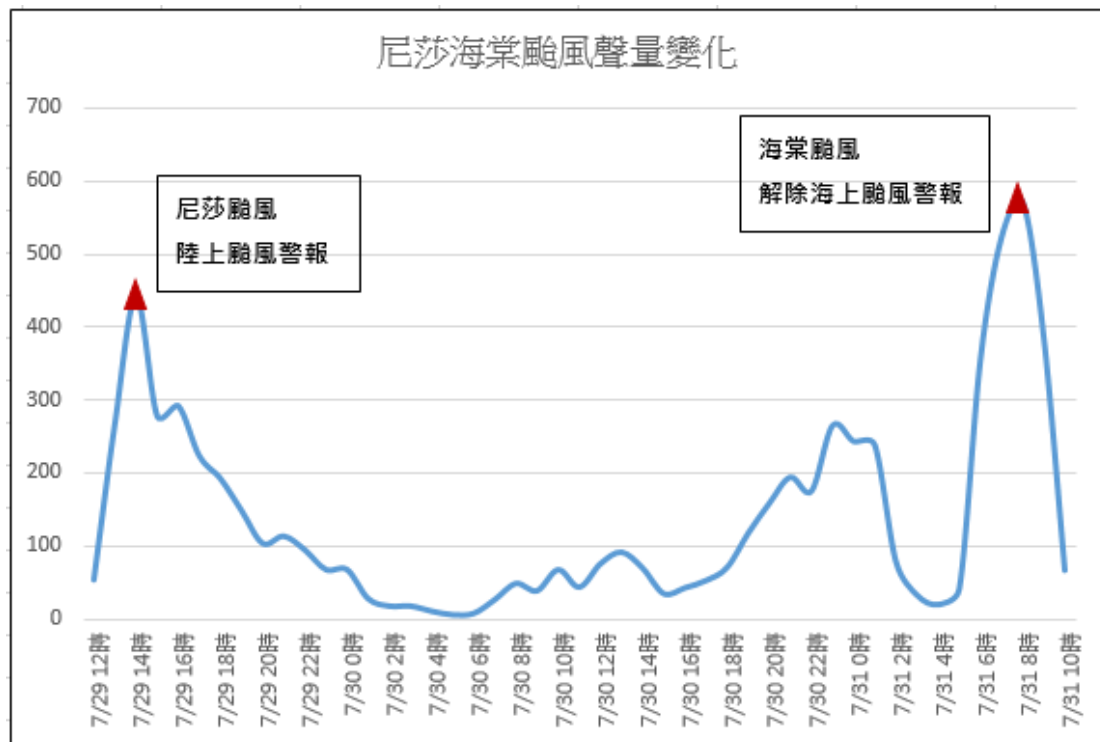


圖 10 尼莎、海棠颱風社群聲量變化趨勢

若將觀察目標轉移至受災嚴重地區時，考慮到民眾關注的地區與時間的變化，我們由前述所得到的關鍵字列表，從中選取被熱烈提到的地名：屏東、高雄、台南、佳冬、林邊等，我們單獨針對這些關鍵字被提及的頻率對照其時間的話可以由以下兩張圖片所示：

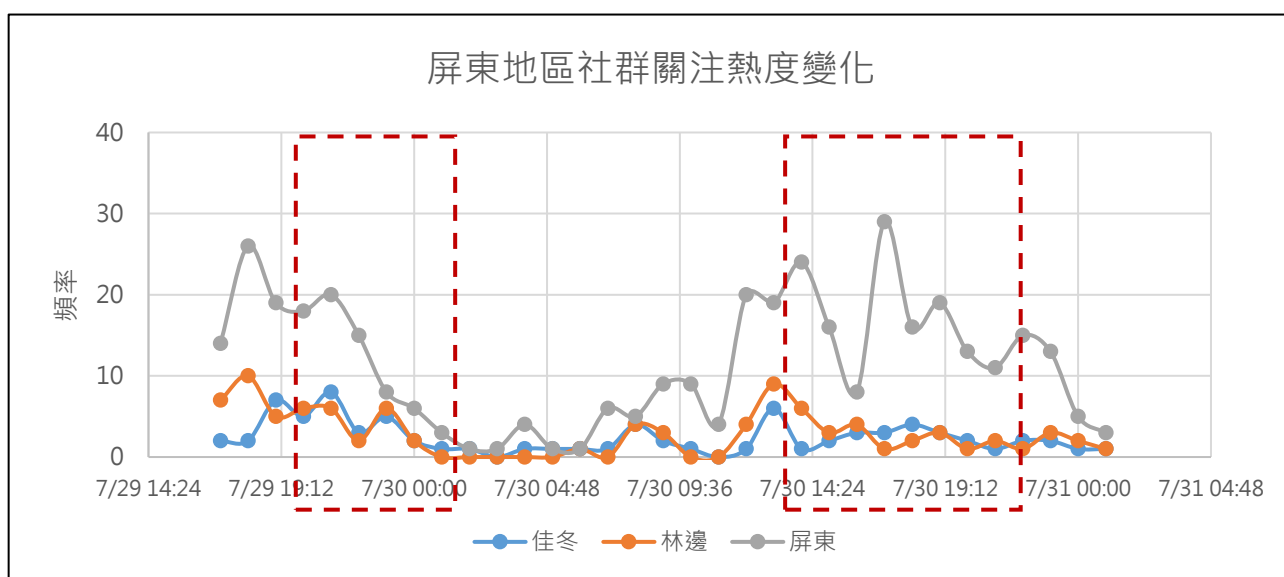


圖 11 屏東地區社群關注熱度變化

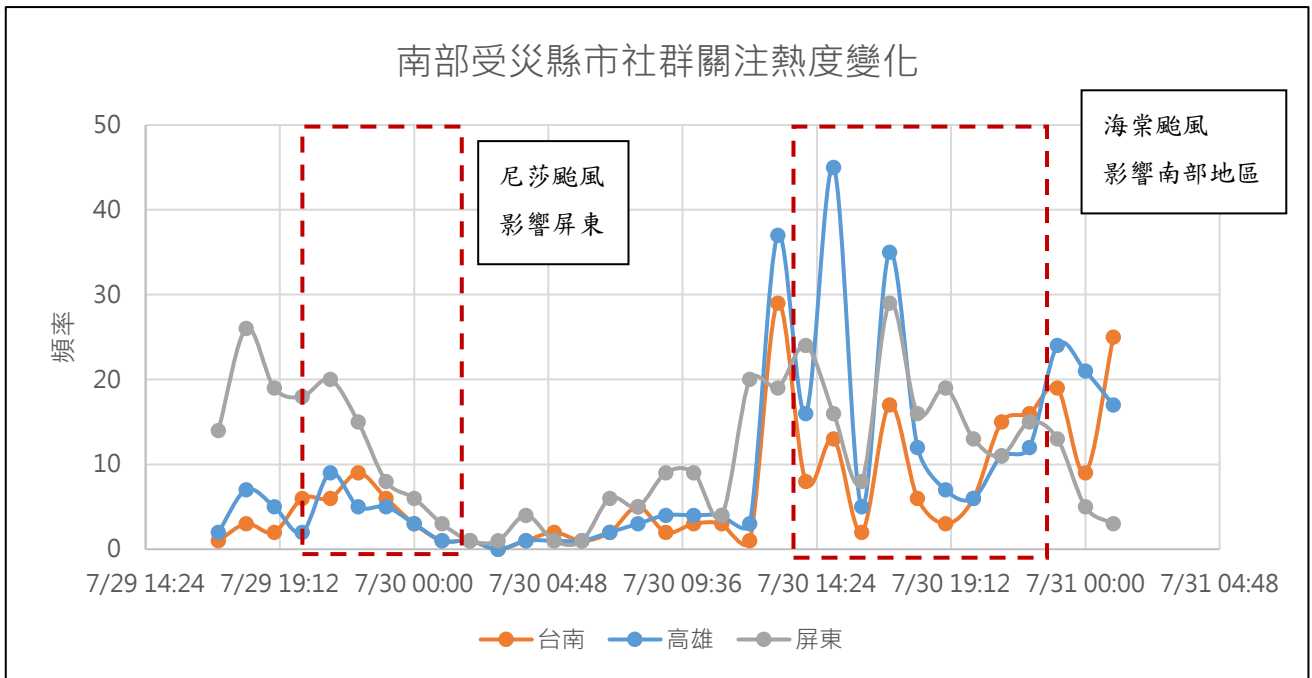


圖 12 南部地區社群關注熱度變化

圖 11 為屏東地區的社群關注變化，根據本中心之災害事件簿，尼莎颱風在 7 月 29 日 14 時至 17 時夾帶的西南氣流，造成造成屏東地區降下短延時強降雨，社群資料亦於 17 時開始有討論聲量

圖 12 則顯示受兩個颱風侵襲之時，尼莎先影響屏東地區後，海棠颱風於 30 日隨後造成南台灣降下大量豪雨，台南、高雄、屏東皆開始產生討論聲量，其時間區間與實際颱風侵時間相近。

第五章 結論與建議

社群網站在災害發生期間，能提供大量的災情資訊，然而，這些「資料」需經過先期的處理才能成為有用的「資訊」。

資料的初步過濾可以利用常見的災害關鍵字達成，其過濾的集合可藉由簡單的詞頻計算，快速地找出熱門的關鍵詞語，這些單詞可提供一定的地理、災情敘述等資訊。除了一般的統計分析外，自然語言處理可深化關鍵字的分析，藉由命名實體辨識模型，可找出地理詞、災情描述詞等，辨識程度更為精確，專一性也相對顯著。

在災害應變期間，這些來自民眾討論的熱門關鍵字，地理詞彙可能透漏目前災情較多的地區，災情敘述則顯示其嚴重性，在災情的掌握上應能提供一定的情報，聲量的高低趨勢則能顯示事件的嚴重程度、歷時長度或發生時間。

雖然社群網路可提供大量資訊，其正確性與精準性仍為一大課題，在本研究的未來工作，需考慮如何驗證這些「流言」是否屬實，如何將這些地理詞精準地定位在地圖上，對於進一步的災情掌握為主要難題所在。

附錄 參考文獻

- [1] Blunsom, Phil. "Hidden markov models." *Lecture notes, August 15* (2004): 18-19.
- [2] Chou, Chien-Lung, Chia-Hui Chang, and Ya-Yun Huang. "Boosted Web Named Entity Recognition via Tri-Training." *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 16.2 (2016): 10
- [3] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013.
- [4] Graves, Alex, et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.
- [5] Graves, Alex. "Generating sequences with recurrent neural networks." *arXiv preprint arXiv:1308.0850* (2013)
- [6] Huang, C.-M., Chan, E., & Hyder, A. A. (2010). Web 2.0 and internet social networking: A new tool for disaster management?-lessons from taiwan. *BMC medical informatics and decision making*, 10(1), 57.
- [7] Kryvasheyeu, Yury, et al. "Performance of social network sensors during Hurricane Sandy." *PLoS one* 10.2 (2015): e0117288.
- [8] Kryvasheyeu, Yury, et al. "Rapid assessment of disaster damage using social media activity." *Science advances* 2.3 (2016): e1500779.
- [9] Lafferty, John, Andrew McCallum, and Fernando Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001): 282-289.

- [10] Ma, Xuezhe, and Eduard Hovy. "End-to-end sequence labeling via bi-directional lstm-cnns-crf." *arXiv preprint arXiv:1603.01354* (2016).
- [11] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *EMNLP*. Vol. 14. 2014.
- [12] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.
- [13] Y. Y. Huang, C.H. Chung, "A Tool for Web NER Model Generation Based on Google Snippets," *Proceedings of the 27th Conference on Computational Linguistics and Speech Processing*, pp. 148–163, ROCLING, 2015.

社群資料研究：關鍵字詞分析與趨勢追蹤

發行人：陳宏宇

出版機關：國家災害防救科技中心

地址：新北市新店區北新路三段 200 號 9 樓

電話：02-8195-8600

報告完成日期：中華民國 108 年 12 月

出版年月：中華民國 109 年 01 月

版 次：第一版

非賣品

地址：23143新北市新店區北新路三段200號9樓

電話： ++886-2-8195-8600

傳真： ++886-2-8912-7766

網址： <http://www.ncdr.nat.gov.tw>