世 新 大 學 管 理 學 院
資 訊 管 理 學 系
碩 士 論 文

PTT 社群網站網民意見之探勘

-以太陽花學運為例

A Case Study on Posts from PTT Boards

-The “318 SunFlower Student Movement”

指導教授： 劉育津 博士
研 究 生： 邱旻翰

中華民國 105 年 7 月

世新大學管理學院資訊管理學系碩士班

論文題目： PTT 社群網站網民意見之探勘-以太陽花學運為例

研 究 生： 邱旻翰

本文承蒙下列口試委員審議通過。

口試委員 _____

_____

_____

_____

指導教授 _____

中華民國 105 年 5 月 25 日

誌謝

時光飛逝，轉眼間兩年的研究所生涯即將邁入尾聲，而在這論文完成之際，首先要感謝我的指導教授劉育津老師，在研究過程中悉心的教導，使我得一一窺文字探勘領域的奧妙，並給予我許多的建議以及指導，使論文得以順利完成，讓我在學習過程中受益良多。也感謝口試委員方孝華老師、游聖瑾老師在口試以及論文上的指正與建議，使論文可以更加完善。

兩年的日子裡，在 Lab 的點點滴滴，不管是課業上的討論、沒有意義的閒聊，或者是與學長姐打不停的桌遊，讓我這兩年的碩班人生一直都很熱鬧，而今天即將畢業，回想起這段日子，想必會相當不捨。

感謝怡如，從大學開始到碩班這六年的相伴，雖然時有爭吵，但這六年沒有妳還真的不行；感謝雯棋，總是聆聽我各種抱怨，以及口試當天的各種協助，很高興在這段旅途當中遇見你們，我會永遠珍惜這個緣分。

最後，感謝默默支持與鼓勵我的家人，讓我可以無後顧之憂的專心寫論文，謹以此文獻給我所感謝的各位。

邱旻翰 謹致

世新大學資訊管理所

中華民國 105 年七月

摘要

在網路快速發展下，人們習慣在網路上傳遞訊息，而這些訊息背後所隱藏的資訊對於許多人而言非常重要，舉凡個人、企業亦或者是政府等，如何去處理這些大量的訊息就成了很重要的議題。而近年來社群網站的崛起，原先傳遞訊息的結構也趨於複雜，在社群網站上的訊息不一定會是具有結構性的，很可能只是單純一句話。因此許多研究開始針對非結構化的資訊進行探勘，找出其背後所隱含的主題以及針對主題的情緒語意。

本研究中以台大電子佈告欄(PTT)為資料集，並以 318 太陽花學運為案例，透過 Python 撈取相關資料，將其進行中文斷詞後取出議題中的關鍵字，並以 Latent Dirichlet allocation(LDA)找出文章相關的主題，本研究進一步依此主題特性進行分析與命名。此外並透過情緒語意分析針對主題進行分析，實驗結果顯示，在 PTT 中發言的網民大多數是支持學運的態度。

關鍵字： 318 太陽花學運、LDA、PTT、語意分析

Abstract

Nowadays, social network sites (SNSs) are popular platforms to discuss issues of all kinds. Since the “The Arab Spring” event being spread out and spawn revolutions via SNSs, the governments began to aware the importance of hearing messages from SNSs.

In this study, the forum “Hatepolitics” on National Taiwan University bulletin board (PTT) is used to explore the politic event “318 SunFlower student movement” which was happened on March 2014. The posts from March 2014 to March 2015 are gathered and further analyzed. We gain relevant information by Python, and get the keywords after Chinese tokenization, further find out the related topics via Latent Dirichlet allocation (LDA), and at last perfrom sentiment analysis to explore the topics.

The research reach conclusions by naming each topic appropriately; in addition, it can be concluded that people writing comments on “Hatepolitics” during the period tended to support the “318 SunFlower student movement”.

keywords： 318 student movement, LDA, PTT, sentiment analysis

目錄

表目錄

圖目錄

第一章　緒論

1.1 研究背景

近年來，隨著網路科技的進步，以及行動裝置的普及，社群媒體因而被廣泛的使用，且不分男女老少，人們利用這些社群媒體去取得新資訊。也導致比起在真實的世界中，人們更習慣在網路上去傳達訊息。在以往，人們由電視、廣播等方式取得資訊，而這些方式只能單向的進行傳遞，提供資訊的人無法得到回饋。而在現代，以網路、社群媒體的方式，可以雙向的進行交流，人們可以更直接的取得或者交換新資訊，而提供資訊者也能得到閱讀者的回饋。在過去，一則廣告需要播放一段時間後，才能夠去統計其效益，但在現在，人們只要在網路上看到這則廣告，只需一個點擊，對於廣告提供者就是一種效益的展現。

這些情況在社群媒體中更可以得到許多案例，企業、商家、個人、或政治人物，甚至是其他的媒體業者，不論何者都開始經營起社群媒體，在 Facebook 上我們總可以看到各企業、店家的粉絲專頁，在社群媒體中進行行銷，而民眾也可以直接的在上面給予回覆，表達對於這個企業產品的看法、或者是這間餐廳服務的意見；也能夠看到政府、政治人物在上面宣導政府法令，或是自身政見等等情形；而部分的電視媒體、平面媒體也開始走向網路社群媒體的經營。綜合以上情形可以得知社群媒體在現在已經是與大眾密不可分的重要工具。

藉由這些社群媒體，政府、企業都能夠直接且快速的蒐集到人民或者消費者的意見，而藉由這些意見，政府就能夠去調整原本政策的執行、新政策的制定；企業也能因而了解市場趨向，調整銷售方針，或者是設計新產品。相較於這種以社群媒體為主的雙向交流，以往以報章雜誌、電視、廣播為主的廣告效益就沒有那麼的顯著、即時。(黃亦筠，2012)

Facebook(臉書)、Plurk(噗浪)、Instagram...等，人們在這些社群媒體中去取得、傳遞資訊，而這些大量傳遞的資訊也造就了大量的資料，構成了所謂的「巨

量資料」。目前世界上累積的資料量相當龐大，Google 一天需計算超過 24PB 的資料[1]；Youtube 每分鐘被上傳的影片長度超過 100 小時，而每一天所上傳的影片需要花 15 年才看得完(楊穎鈞，2014)；全世界每分鐘有約 2 億封電子郵件被寄出，一天下來就會有超過 2500 億的信件量，而其中絕大部分都是垃圾郵件。這些大量的資料，在以往人們或許會因為當時的統計、規劃能力的欠缺，或者是設備不足，無法妥善利用、分析這些資料，而在現在，隨著電腦效能的提升，以及技術逐漸成熟，人們利用電腦去處理這些大量的資料，已不同於以往如此困難，這些取得、處理、並且分析的技術就是所謂的資料探勘，而對應的，這些資料就是「巨量資料」，利用資料探勘的技術，從大量的資料中，取出有利於使用者需求的部分，不論是市場分析、企業需求、科學技術研究，亦或者是影響人民生活密不可分的政治。

在全世界都可以看到將社群媒體應用於政治相關的社會事件中，而其中最早的例子，則是發生在阿拉伯世界的社會革命運動「阿拉伯之春[2]」。阿拉伯之春事件背景是在2008年金融危機後，在北非的突尼西亞，有位青年因長期失業，只好自行開設攤販，卻又遭到警察的刁難，為表達他的不滿，選擇了自焚。青年自焚的影片在社群媒體中大肆的被傳遞，也因而勾起了突尼西亞人民長期以來對政府的不滿，進而發生全國性的社會運動，推翻了當時的政權，其效應甚至擴展到周邊的埃及、利比亞等國家，甚至擴展到亞洲、歐洲部分地區。

而最先將社群媒體應用於選舉的則是在 2008 年美國總統選舉中的候選人

---

[1] 國立交通大學統計學研究所<<巨量資料帶來的契機與挑戰>>

網址 http://www.stat.nctu.edu.tw/data/super_pages.php?ID=data1

[2] 阿拉伯之春 Wikiedia 網址：

https://zh.wikipedia.org/wiki/%E9%98%BF%E6%8B%89%E4%BC%AF%E4%B9%8B%E6%98%A5

歐巴馬，他藉由成功的網路行銷手段，以新的選舉策略，成功的打贏了不被看好的選戰，在這次的選舉中，有著大量的年輕人，而這些年輕人中有 66%將票投給了歐巴馬，由此可以發現其全新的選舉策略的影響力。(邢光愷，2012)

除了以官方網站傳達政策，更在網站上架設工具，讓支持者得以在網站上直接與彼此溝通。而除了網站以外，歐巴馬還利用了各種的社群網站，像是 Facebook、Twitter、Youtube 等，激起了群眾對於本次選舉的熱度，也藉此讓選民間可以彼此交流，互相聯繫。

而最先將這種方式應用在台灣的選舉，則是去年縣市長選舉時的台北市長選戰，本次選舉中很多資訊都是由大量的數據去計算出來的，在競選團隊的背後，都有一個大數據的團隊，這些人負責去記算在 Facebook 粉絲專頁中的 PO 文留言，有多少人進行討論、點讚、分享，而這些計算的基礎都是來自於人民，以及社群媒體 (王泰俐，2013) 。

在這次選舉中，候選人不斷的以社群媒體為媒介，進行政策的宣導、針對對手的攻訐，以及各種選舉預測。而在民眾之間，對於候選人的討論熱度從其登記參選起，直到 MG149 事件爆發後達到巔峰，由下圖 1 可見，針對連勝文、柯文哲兩候選人的議題其高峰都是負面事件，這些訊息也是競選團隊可以去著墨的部分。

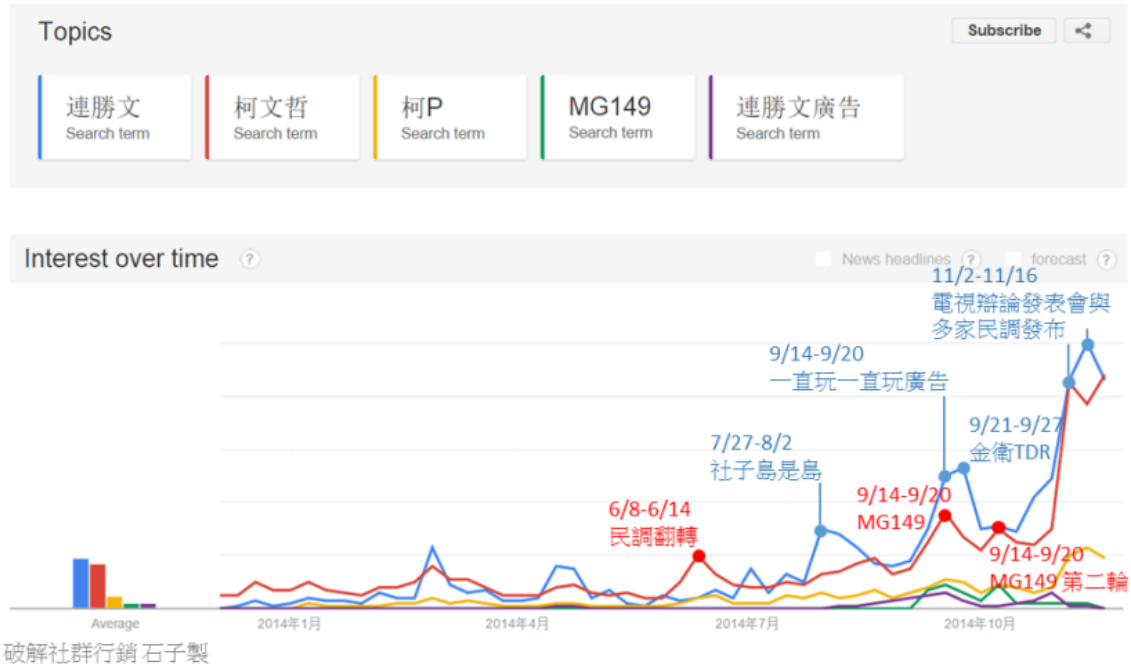

圖 1　2014 年台北市長選舉，柯文哲與連勝文 Google Trend 趨勢圖[3]

而不論是選舉，亦或者是社會運動，其反映的都是在民主社會中的民意，若政府能夠善用網路媒體去了解民意，對於政策的訂定與推行，也會較為順利，也可以減少民眾對於政府的反感，避免民眾抗爭。

然而，即便政府開始以社群媒體來了解、廣納民眾之意見，在台灣卻仍然發生許多人民對政府作為感到不滿而發生的社會事件，像是 2011 年的反媒體壟斷運動、2013 年的大埔事件、2015 年 3 月的阿帕契事件...等事件，而這些事件也大量的在網路上被進行討論。

本研究目的在於藉由蒐集網路媒體中，人們的發文以及他人對於這篇文章的回應，觀察文章討論的變化，試圖找出其關連，並找出其中關鍵字，對討論主題進行分析與研究，在案例部分選中在 2014 年發生的 318 學運，本事件主要

[3]圖片來源: https://taiwansmm.files.wordpress.com/2014/11/e68a95e5bdb1e789871.png

訴求草率通過的海峽兩岸貿易協定(以下稱服貿)須退回重審，而藉由社群媒體，人民達到訴求，服貿退回，且至今尚未生效，在事件發生期間，其議題的討論在 Facebook 部分的粉絲團以及 PTT 政治相關版塊中相當熱烈。

事件期間，在 Facebook 部分，除了政治人物各人粉絲專頁對於服貿的評判以外，亦出現了大量與事件相關的粉絲專頁、社團等，其主要訴求皆為退回服貿，反黑箱作業等等。在 PTT 的部分，在政治版、政治黑特版、八卦板等版塊中，長時間維持「紫爆[4]」的狀態。

Facebook 部分在洪綾君老師的「318 學運期間政治人物網路發言探究—大數據爬文技術之應用」(2015)，已經討論過，故本研究中以 PTT 為主，對前文所述之版塊，進行資料的爬文、斷詞、關鍵字搜尋...等動作。

1.2 研究動機

對政府而言，當有越來越多的民眾利用社群媒體進行討論與分享議題時，這些大量增加的民意資料，卻也可能造成政府無法真正理解民眾的意見。而隨著各種上網設施的普及，這些數位化的資料大量、快速且多元的發展，使的已往處理結構化資料的傳統分析方式失靈（Eaton et al., 2012）。在這些大量的資料中僅有約 15%為結構化的資料，剩下的 85%的非結構化資料可來就是來字前面提到的社群媒體中的文章、留言等等。而長期下來，這些資料還可能會以倍數成長，在這個情況下，就造就了巨量資料，以及分析這些資料的技術。

[4] 紫爆係指於單一版塊中，線上同時間 10 萬人

1.3 研究目的

由於社群媒體在公共事務影響力的增強，本研究希望藉由對 2014 年發生的太陽花學運社群網路留言意見進行分析，藉以了解以下幾個問題：

1. 本期間網民在 PTT 上主要談論的關鍵字詞有哪些？
2. 依所探勘出的關鍵字可將其歸類成哪些主題？
3. 這些主題的語意取向為何？

第二章　文獻探討

本研究欲探討在社群媒體中，網民對於議題討論之情形以及討論過程中情緒語意狀況之變化，本章將針對上述內容，查閱相關典籍，分為四節分別為網路輿情與社群媒體、文字探勘、主題群集分析以及情緒分析等部分進行介紹。

2.1 網路輿情與社群媒體

輿情是指在社會中，圍繞著某項事件的討論與變化中，民眾對於政府的作為而產生的政治態度，它是人民意見的表現，亦即民意是輿情的始源(MBA 百科)，而其傳播方式隨著科技的進步，進展到以網路方式傳播，政府得以使用不同的方式，在網路上不同的社群、網頁蒐集民意，做為政策規劃以及執行的依據。(廖洲棚，2014)

自從 Web2.0 網路技術的發展及普及後，世界各國皆開始應用 Web2.0 之技術來改善政府施政效能，除了在各政府部門網站外，也在一些非政府的網站與人民接觸(Chen, 2010)，並了解網路中的民意。而隨著 Web2.0 技術的發達，社群媒體對群眾的影響也日益擴增(呂建億，2015)，在 Web 2.0 技術下，臉書、部落格、推特、微博...等等「新媒體[5]」 (Kushin, 2010.)大量的產生，其共同具有的特性為開放性，在這些平台上，人們可以分享自己的意見、看法，並以其他社群媒體使用者之間維持連繫（Ellison et al., 2007），前文中提到，2008 年歐巴馬開啟了社群媒體的選戰，使用了大量的社群媒體進行傳播，成功的動員了年輕選民進行投票，也標榜了網路政治時代的到來 (Macnamara and Kenning，2011) 。

根據 Internet World Stats(2015)統計，截至 2015 年 11 月止，全球使用網路

[5]智庫百科–新媒體網址：
http://wiki.mbalib.com/zh-tw/%E6%96%B0%E5%AA%92%E4%BD%93

的人數總量已經達到了 33.6 億，在台灣有超過一千九百萬的上網人口，網路成為了在報章雜誌、廣播以及電視之後的第四媒體。其快速傳播的特性，就是他快速發展的主因，而社群媒體在網路世界中的發展更是不容小覷。台灣的臉書使用者高達了一千八百萬，約為上網人口的 91%[6]。以往人們藉由報章雜誌、電視去取得資訊，而現在人們使用社群媒體以及行動裝置，比起過去，他們可以更快的得到最新的資訊。不僅是個人，企業也跳脫了以往的行銷手段，走向了數位化、網路化的路線，根據 Advice Interactive Group 在 2015 年所做的統計(圖 2)。

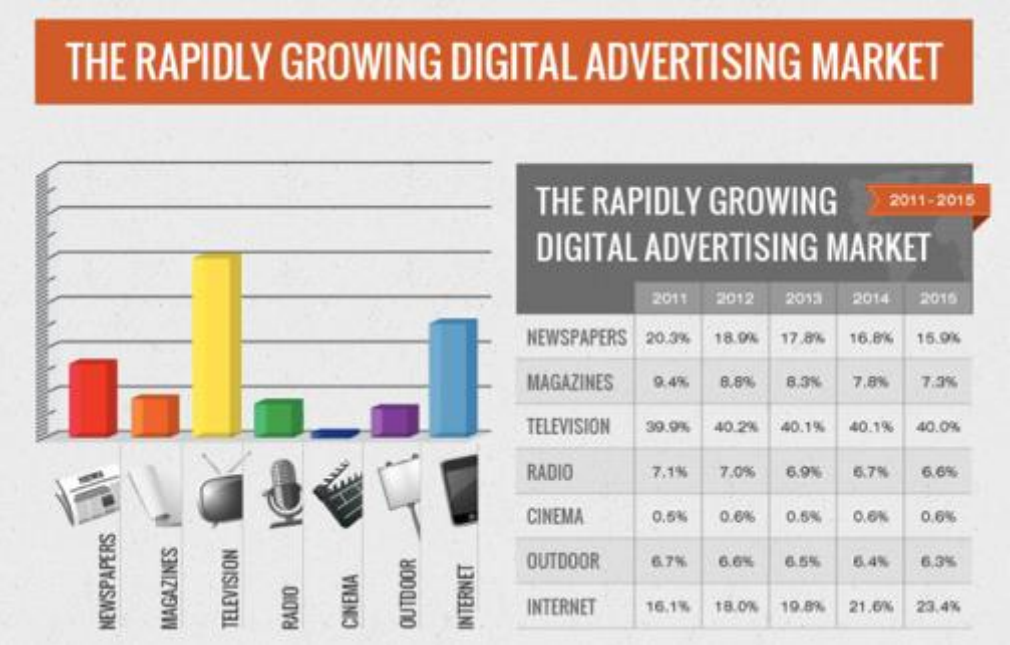

圖 2 傳統廣告與數位化廣告成長率

(圖片來源：Advice Interactive Group)

可以看到在 2011 年時企業在報章雜誌上的廣告有 29.7%，而且 2015 年只剩下 23.2%。電視、廣播的部分在 2011 年有 47%，在 2015 年時則有 46.6%，成長較緩慢，而在網路廣告的部分 2011 年只有 16.1%，在 2015 年則達到了 23.4%，

[6]網路人口統計 http://www.internetworldstats.com/stats.htm

成長較為明顯。根據尼爾森(Nielsen) [7]「對廣告的信任度報告」(2013)，相較於一般廣告(14%)，人們更相信社群媒體中其他消費者的推薦(78%)，也就是口碑行銷。這份報告也顯示出人們易受其他人或者是網路輿論的影響。

由此以上文獻可得知，社群媒體不僅僅是民眾了解資訊的平台，他同時也是人們可以發表評論、互相交流、互相影響的網路空間。

2.2 文字探勘

為了達到對前述網路輿情進行分析，需要對其資料做文字探勘，有別於資料探勘的對象資料較為結構化，文字探勘的對象長短不一、沒有規則，很可能只是生活中的一句話，或者是某項在某環境中的特定用語，這些辭彙可能會隨著網路使用者每天的更新而無法事先定義它，唯近年來文字探勘技術的進步，使的這些新辭彙，得以產生新的價值(李慶堂,2015)。

文字探勘被定義為一種編輯、組織及分析大量非結構化文件的過程，以達到使用者某項特定需求(Sullivan ,2001)，找出其中人事時地物間等關鍵字的關聯，並加以分類後呈現。其範圍包含了文字語意的分析、文章內容的截取、透過關鍵字分析文件、文件與討論議題的群集關係等。而其應用範圍廣泛（Cimiano, 2006），包含了資訊檢索、資訊擷取、自然語言處理、計算語言學...等。藉由找出資料中的關鍵字，加速檢索資料的速度。

針對處理後的資料，依據其特徵區分為許多群，使性質相似的資料被分為同一群，讓使用者能快速區分文件類別，並迅速找到需要的文件。另一方面，對所有文件的分布提供一個綜覽，以提升文件的搜尋效益，並自動建立文件的分類架構，辨識文件中的字詞與關聯性，以減少文件檢索和查詢的誤判（謝家蘭，2006）。

本研究針對文字探勘的處理，分為資料蒐集、斷詞、關鍵字萃取、主題分

[7] NIELSEN 廣告信任度，網址
http://www.adweek.com/socialtimes/files/2013/01/social-media-help-desk.jpg?red=at

析與情緒語意分析等步驟，主題分析與情緒分析於後續待做討論。以下先討論中文斷詞與關鍵字萃取。

2.2.1 中文斷詞

為找出在文件中的關鍵字，我們需要對文章進行斷詞的動作，英文的資料不需要進行斷詞的處理，但本研究中選用 PTT 版塊中資料為中文，故須對其進行斷詞動作。詞是語言學家所定義的「能夠獨立運用，具有完整語意的最小語言成分」(陳言熙，2006)。若是英文，則每個單字都可以為一個詞，具有自身的意義，且在使用習慣上，詞與詞之間都會有空白做為分隔，也就沒要斷詞的必要性[8] 。然而在中文中，詞和詞之間並無明顯分隔，故需要進行斷詞的動作。針對斷詞的部分，在本研究中使用 Python 中 Jieba 中文斷詞套件。

Jieba 是一套由中國百度的一位開發者所撰寫，並開放其原始碼供眾人使用、維護。Jieba 中文斷詞所使用的演算法是基於 Trie Tree 結構去生成句子中中文字所有可能成詞的情況，然後使用動態規劃算法來找出最大機率的路徑，這個路徑就是基於詞頻的最大斷詞結果。而對於辨識新詞（字典詞庫中不存在的詞）則使用了 HMM 模型（Hidden Markov Model）及 Viterbi 算法來進行辨識。[9]

2.2.2 關鍵字萃取

一般要由文章中找尋出關鍵字，須考慮到兩項因素：

i. 這個字(詞)在這篇文章中出現的頻率(Term-Frequency,TF)。
ii. 在所有的文章中，有幾篇文章出現這個字(詞)(Inverse-Document-Frequency,IDF)。

TF 是用來計算在文章、句子分數的最基本方式，若在文件中出現的頻率次

[8]如何使用 JIEBA 結巴中文分詞程式，https://github.com/fxsjy/jieba

[9] http://blog.fukuball.com/ru-he-shi-yong-jieba-jie-ba-zhong-wen-fen-ci-cheng-shi/

數越高，代表這個字(詞)越能夠代表這篇文章的主題，也就是關鍵字(Luhn,1957)。Jones(1972)則認為各關鍵字間的相對性權重，是由各關鍵字在所有的文件中是否出現的頻率倒數來決定，並提出了 IDF 的詞頻計算公式。

而為了改善詞頻的計算方式，Salton 與 Buckley(1988)結合了 TF 與 IDF 的技術，並發展出一種新的詞頻計算方式， TF-IDF (Term Frequency - Inverse Document Frequency)。TF-IDF 代表辭彙出現在文件中的頻率越高且出現在其他文件中的頻率越低的話，代表辭彙就越具有代表性，重要度也越高。

2.3 主題群集分析

斷詞後的辭彙，必定與文章間存在某種關係，該關係也就是「主題」，一般在撰寫文章時，會先思考文章之主題為何，再去根據主題，找出合適的詞語來表達對於這個主題自己的看法。

而對於這些詞語與文章主題之間的關係，在本研究中以隱含狄利克雷分布(LDA，Latent Dirichlet allocation)的模型，試圖找出辭彙與文章之間的主題分群。

該模型由 Blei 等(2003)提出，其優點在於它是一個非監督式模型，故每個主題都可以找到對應的詞與對主題進行描述，而其缺點則是當資料量不多時，效果較為不佳(陳冠瑜，2015)。LDA 模型被大量應用於資訊檢索、機器學習等相關領域，吳政毅(2015)將 LDA 應用於新聞主題萃取，藉 LDA 分群之主題以避免以往對於新聞主題總有模稜兩可的情形；蕭昱維(2014)利用 LDA 技術，試圖找出 Twitter 交談中背後的主題，並試圖改善 LDA 面對短文章中主題發掘的困難；邱怡菁(2015)以 LDA 對英文課程教材進行主題分析，並整理出其摘要，令學習者可以快速了解課程內容；而對於社群討論與情緒分析的部分，張日威(2014)透過 LDA，分析微網誌噗浪(Plurk)中的相關主題，並透過情感分析針對相關主題給予極性分類，讓使用者可以快速了解大眾對於相關主題的喜好程度。

LDA 可以在一系列的文本語料庫中產生離散數據集合概率模型，即它認為一篇文章是由多組詞構成的一個集合，詞與詞之間沒有先後順序的關係，且每篇文章中可能包含多個主題。它可將文章中的主題按照機率分布的形式推斷出每一主題所代表的詞語[10]，其模型如圖 3。

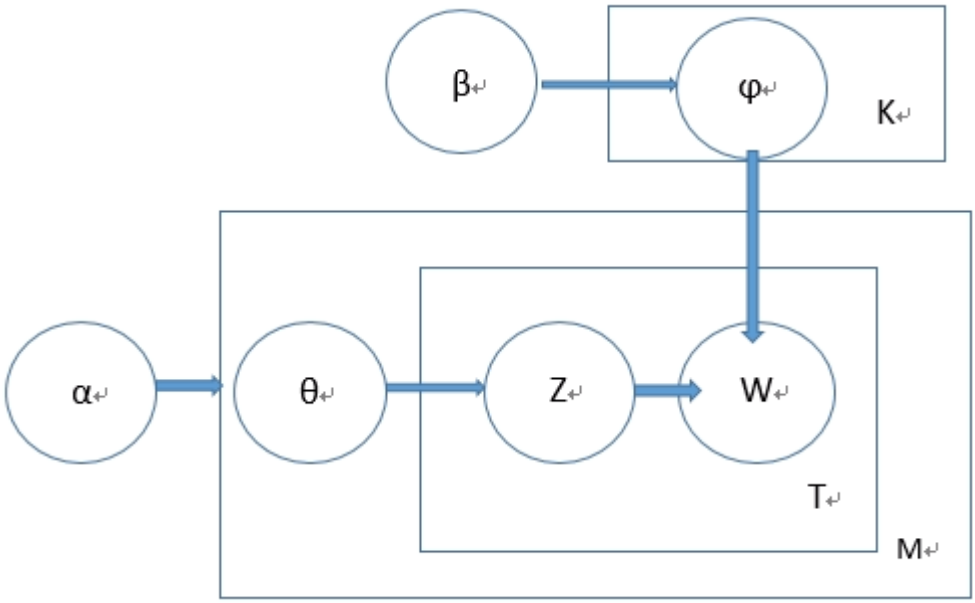

圖 3 LDA 模型

LDA 相關各符號意義如下表格 1

表格 1 LDA 相關符號表[11]

M：文件數量
K：主題數量
W：所有文章的所有辭彙
Z：所有主題
T：單篇文章的總詞數
N：所有文章的總詞數
α：狄利克雷參數，表達主題在文章中的分布

[10]通俗理解主題模型 LDA，http://www.csdn123.com/html/topnews201408/93/7793.htm

[11]隱含狄利克雷分布 LDA Wikipedia，https://zh.wikipedia.org/wiki/%E9%9A%90%E5%90%AB%E7%8B%84%E5%88%A9%E5%85%8B%E9%9B%B7%E5%88%86%E5%B8%83

β：狄利克雷參數，表達辭彙在主題中的分布
θ：文章與主題分布關係
ϕ：主題與辭彙分布關係

其模型生成方式如下：

1. 從狄利克雷分布 α 中取樣出文章 i 的主題分布 θ_i
2. 從主題的多項式分布 θ_i 取樣出文章 i 第 j 個辭彙的主題 $Z_{i,j}$
3. 從狄利克雷分布 β 取樣出主題 $Z_{i,j}$ 的辭彙分布 $\phi_{\approx I,j}$
4. 從辭彙的多項式分布 $\phi_{\approx I,j}$ 採樣得到最後生成辭彙 $W_{i,j}$

根據上述步驟，可得其模型公式如下：

$$p(w_i, z_i, \theta_i, \Phi|\alpha, \beta) = \prod_{j=1}^{N} p(\theta_i|\alpha)p(z_{i,j}|\theta_i)p(\Phi|\beta)p(w_{i,j}|\theta_{z_{i,j}})$$

當中 θ_i 為文章 i 的文章與主題的分布，ϕ_i 為主題 i 的主題與辭彙分布，$Z_{i,j}$ 為文章 i 中辭彙 j 的主題，$W_{i,j}$ 為文章 i 中辭彙 j，α 為主題中的文章分布，β 為辭彙在主題中的分布。

2.4 情緒分析

前文提到，詞是語言學家所定義的具有完整語意的最小語言單位，而每個辭彙又是由字所組成，字又是在中文書寫時的最小成份[12]，除了字本身帶有的情緒外還可能因為口語上的問題導致同樣的字可能會有不同的情緒，在一個句子中也可能因文字與其他文字間的關係產生不同的情緒(黃信華，2013)。

在本研究中會將前述斷詞結果後的辭彙進行情緒歸屬的判斷，而為了進行情緒歸屬的判斷，需要有健全的語意辭典作為輔助。目前常見的情情語意辭典有 NUTSD[13]以及知網(HowNet) (董振東，1988)。

---

[12] 單詞 Wikipedia，https://zh.wikipedia.org/wiki/%E5%96%AE%E8%A9%9E

[13] NLPLab，http://academiasinicanlplab.github.io/

NTUSD 是台灣大學自然語言處理實驗室由陳信希教授所建立的語意辭典，其主要架構為 General Inquirer (GI) 及 Chinese Network Sentiment Dictionary（CNSD），約包含正面情緒 2800 筆辭彙，負面情緒 8000 筆辭彙，共約 2 萬多筆辭彙。

中科院知網在 2007 年發布中英文字詞情感分析，其中包含中文正面情感詞語、中文負面情感詞語、中文正面評價詞語、中文負面評價詞語、中文程度級別詞語、中文主張詞語，共約九千多個詞(徐筱雁，2014)。

目前中文情緒辭典之建置多半都是由大型資料庫中蒐集辭彙，再以人工方式，標記其情緒以及極性(張日威，2014)。

而對於情緒分辨的方式，近年來採用一些機器學習技術，其中 SVM 是最好的分類為中文文本情感分類方式(Tan & Zhang,2008)，以 SVM 和 SeCeVa 改善情感辭彙和語義結構訊息的應用，可以將識別率從 40.3%提升到 68.15%(黃祖菁,2012) ，另外在資料探勘演算法中也可以做情緒分類例如：統計方法、類神經網路、決策樹等(黃承龍等，2004)，而在中文短句之情緒分類的研究中提及三種方式為關鍵字偵測法、學習偵測法、混合方法(孫瑛澤等，2010)

第三章　研究方法

本研究在實際操作上是以 318 學運為案例，而首先要面對的問題就是要決定分析哪一些社群媒體上的留言。在有關 318 學運的事件中，最常被提到的社群媒體為 Facebook 及 PTT，Facebook 部分如前文所提，在洪綾君老師的研究中已提及，故本研究選用台大電子布告欄(PTT)為研究對象。

3.1 研究架構

在架構部分分為資料蒐集、文字探勘兩部分，如下圖 4：

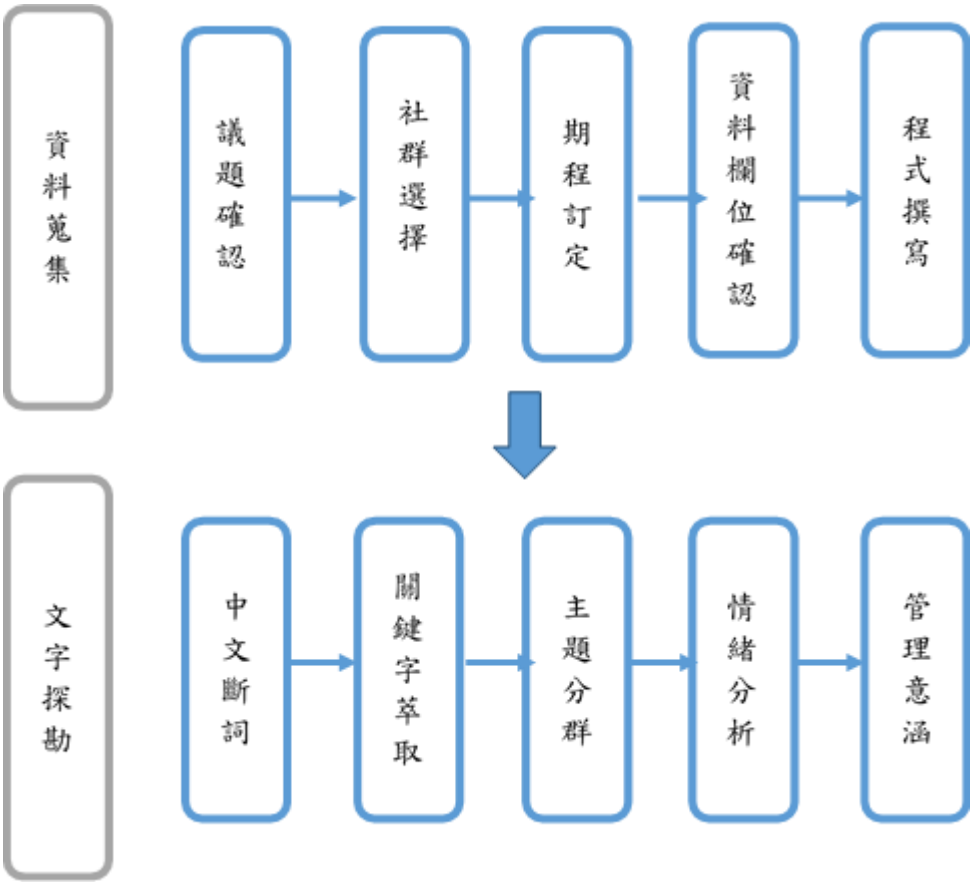

圖 4 研究架構

第一部分為資料蒐集，目的在確認議題、社群、以及討論的時間範圍，和最後的程式撰寫以爬取資料。

第二部分為文字探勘，分為中文斷詞、關鍵字萃取、主題分群、情緒語意分析，以及最後的管理意涵解說。

3.2 研究方法

3.2.1 資料蒐集

首先，在資料蒐集部分，分為五個步驟：

1. 議題確認

針對台灣各項社會議題使用社群媒體者之中進行選擇，舉凡洪仲丘事件、1129 選戰、太陽花學運...等，而其中太陽花學運更是由民眾主動發起，且成功影響政府決策者，故選用此議題做為討論議題。

2. 社群選擇

觀察與議題相關之討論多出現於哪些社群媒體中，發現多半在於 Facebook 與 PTT 中，因 Facebook 部份於洪老師之研究中已使用過，故本研究中選用 PTT 做為目標社群媒體。

而與本議題相關之討論多出現於 PTT 中 HatePolitics(政治黑特版)與 Politics(政治版)兩政治相關版塊中，故後續將對兩版塊之文章進行爬取。

3. 期程訂定

針對議題之時間範圍進行設定，以便後續資料擷取作業之進行。

本研究議題發生於 2014 年 3 月 18 日，故在本研究中也將期程訂為 2014 年 3 月 1 起至 2015 年 3 月 18 日止。

4. 資料欄位確認

在資料內容部分，因 PTT 中文章格式不一，我們對 PTT 文章中標題、作者、發文時間、內文、回應…等部分欄位進行撈取。

5. 程式撰寫

本研究資料收集之程式以 Python 進行撰寫，針對 PTT 與議題相關之版塊進行爬文的動作，這部分分為兩個階段，先對文章頁面進行爬取(一頁有 20 篇文章)，這部分會包含各文章網址，接著對每一篇文章進行剖析，將其中欲撈取欄位的標籤取出，並寫入 Excel 中，相關欄位如下。

針對文章，我們所抓取的欄位如下表格 2

表格 2 文章欄位

a_ID	b_作者	c_標題	d_日期	e_ip	f_內文	numg	numb	numn	numall

針對文章之回文，我們所抓取的欄位如下表格 3

表格 3 回文欄位

index	a_ID	num	狀態	留言者	留言內容	留言時間

於爬文過程中，因 PTT 格式不一，亦或者有亂碼、特殊符號等情形出現，導致偶有程式中斷情形，故後續以分段方式進行資料爬取，縮小爬取範圍並找出有問題的文章，加以處理，或者予以排除，以確保資料正確性。以下為程式流程圖(圖 5)

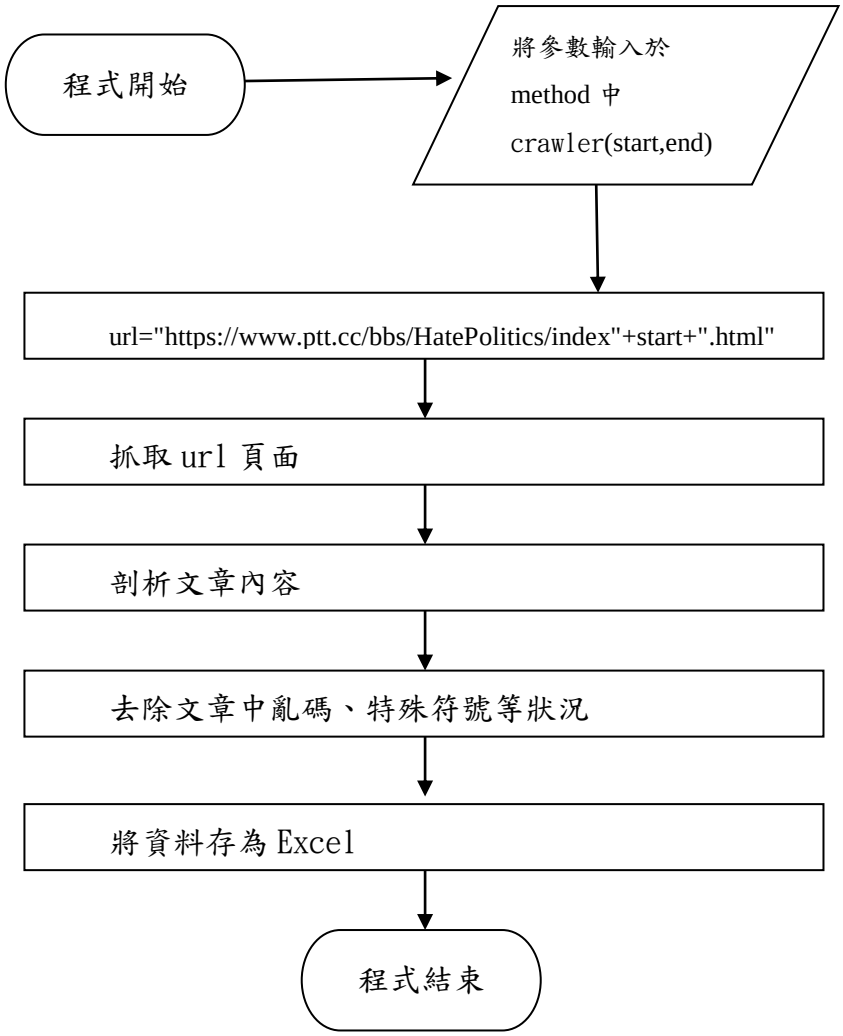

圖 5 資料撈取程式流程

下就上述流程，進行部分程式碼解說(表 4、5)。

表格 4 擷取文章頁面程式碼

```
def crawler(start,end):
    page = start; times = end-start+1; g_id = 0;
    for a in range(times):
        resp = requests.get(
        url="http://www.ptt.cc/bbs/hatepolitics/index"+str(page)+".html")
        soup = BeautifulSoup(resp.text, "html.parser")
        for tag in soup.find_all("div","r-ent"):
            try:
                link = str(tag.find_all("a"))
                link = link.split("\"")
                link = "http://www.ptt.cc"+link[1]
                g_id = g_id+1
                parseGos(link,g_id)
            except:
                pass
```

```
        sleep(0.2)
        page += 1
```

表格 4 之程式碼為針對文章列表所執行之函式 crawler，首先為該函式設定起始值 start 與結束頁面 end，該函式由兩個迴圈組成，在第一個迴圈會藉由 Python 中 BeautifulSoup 的套件對文章列表進行剖析。如前文所述，每一文章列表的頁面皆含有 20 篇文章，該函式會剖析出頁面中每篇文章的連結。在第二個迴圈則會將剖析出的連結交由 parseGos 函式進行針對文章內容的剖析以及相關的除錯等動作。

表格 5 擷取文章剖析程式碼

```
def parseGos(link , g_id):
    global stupidid
    resp = requests.get(url=str(link),cookies={"over18":"1"})
    soup = BeautifulSoup(resp.text,"html.parser")
    try:author=soup.find(id="main-container").contents[1].contents[0].contents[1].string.replace('
', '').replace(", ")
    except:
    try: title =soup.find(id="main-container").contents[1].contents[2].contents[1].string.replace('
', '').replace(",")
    except: title ="Invalid Title"
    try: date = soup.find(id="main-container").contents[1].contents[3].contents[1].string
    except: date = "date is wrong"
    try:
        ip = soup.find(text=re.compile("※ 發信站:"))
        ip = re.search("[0-9]*\.[0-9]*\.[0-9]*\.[0-9]*",str(ip)).group()
    except:    ip = "ip is not find"
    try:
        a = str(soup.find(id="main-container").contents[1])
        a = a.split("</div>")
        a = a[4].split("<span class=\"f2\">※ 發信站: 批踢踢實業坊(ptt.cc),")
        content = a[0].replace(' ', '').replace(",").replace('\n', '').replace('\t', '')
    except:
```

表格 5 為針對文章頁面所執行之函式，該函式接收 crawler 函式所解析出之網址後，會對文章進行剖析，同樣藉由 BeautifulSoup 套件，找出頁面中與資料相對應的標籤，並將其取出後加以除錯，以便後續儲存動作。

3.2.2 文字探勘

於文字探勘部分，同樣分為五個步驟：

1. 中文斷詞

文字探勘中重要的第一個步驟就是斷詞，不同於英文辭彙於詞與詞間有空白做為分隔，中文是以一個句子做為分隔，故詞與詞的分隔在中文中顯得格外模糊，故須對其進行斷詞動作。

本研究中斷詞同樣使用 Python 進行撰寫，並使用 Jieba(結巴)中的 posseg 套件(林冠宇，2014、彭桂香，2014)，在得到斷詞結果後，同樣寫入 Excel 中。

以下就斷詞部分，以部分程式碼進行解說(表格 6)

表格 6 擷取中文斷詞程式碼

```
import jieba
import jieba.posseg as pseg
import csv
jieba.set_dictionary('dict.txt')
f = open('hatecontent2000.csv', 'r')
i=1
s=list()
for row in csv.reader(f):
    if len(row):
        line = row[0].strip()
        words = pseg.cut(line)
        for word in words:
            ss=[str(i), word.word, word.flag]
            s.append(ss)
    i=i+1
f.close()
```

因本研究中所使用之 Jieba 套件所提供之辭典為簡體版本，以致斷詞時恐有所疏漏，而在 Jieba 中提供了自行增加辭典的功能，故在程式執行

前，須以 set_dictionary 函式匯入繁體辭典。匯入辭典後，再將 PTT 資料中內文(content)欄位餵入本程式中，利用 Jieba 中的 posseg 套件當中的 Cut 語法，即可依據辭典，將資料進行斷詞，藉以獲得斷詞結果，並寫入到外部檔案中，以備後續使用。

2. 關鍵字萃取

斷詞後的第二步驟即為關鍵字的萃取，在本研究中使用的同樣是以 Python 去撰寫的程式進行，在 Jieba 中同樣提供了關鍵字萃取的功能，亦可以對斷詞後的結果進行關鍵字的萃取。在得到關鍵字後同樣寫入 Excel 中。

以下就關鍵字萃取部分，以部分程式碼進行解說(表格 7)

表格 7 擷取關鍵字萃取程式碼

```
import jieba
import jieba.analyse
fr = open('Book1.csv', 'r')
f=fr.read()
tags = jieba.analyse.extract_tags(f, topK=1000, withWeight=True)
for wd, weight in tags:
    ss = [wd, weight]
    s.append(ss)
fr.close()
```

同樣利用 Python 中的 Jieba，並利用當中的 analyse 套件，利用其中 extract_tags 的方法，可以由匯入之資料集中取得關鍵字，並可提供萃取出之關鍵字在資料集中的權重為何，另外，該套件也提供了取 K 個數量關鍵字的功能(topK)。

3. 主題分群

針對取出的關鍵字進行分群，在本研究中主要以 LDA 模型為主，LDA 可將文章中的主題按照機率分布的形式，進而去推斷出各主題所代表的詞語，而在本研究中實作部分以 R 語言中 LDA 套件進行主題分群的動作。

首先需先將文章斷詞為 LDA 所能接受的格式(這部分以 Python 進行)，

接著將該檔案與關鍵字的檔案，餵入程式中，LDA 的套件可以設定預計要分幾個主題、以及其分配的 aplha、eta 參數等。

在本研究中也使用了 R 中的 LDAvis 套件，該套件可以將主題分群的結果以圖形化的方式呈現，其結果顯示出我們所設定欲分類之主題數 K，以及各主題對應到之前 30 個關鍵字。

以下就主題分群部分，以部分程式碼進行解說(表格 8)。

表格 8 擷取 R 語言 LDA 程式碼

```
#計算 LDA 模型
K <- 10          #預計分幾個主題
G <- 1000
alpha <- 0.05    #D 分配的參數 alpha
eta <- 0.1       #D 分配的參數 eta

fit <- lda.collapsed.gibbs.sampler(documents = documents, K = K, vocab = vocab, num.iterations = G, alpha = alpha, eta = eta, initial = NULL, burnin = 0, compute.log.likelihood = TRUE)
theta <- t(apply(fit$document_sums + alpha, 2, function(x) x/sum(x)))
phi <- t(apply(t(fit$topics) + eta, 2, function(x) x/sum(x)))
      json <- createJSON(phi = phi, theta = theta,
                                 doc.length = doc.length, vocab = vocab,
                                 term.frequency = term.frequency)
serVis(json, out.dir = 'C:/Users/user/Documents/vis_hate0414',)
```

先將斷詞後的資料集，以及取出之 1000 個關鍵字讀入程式中，比對後將對應到的資料放入 vocab 變數中，接著設定 LDA 模型的各項參數，在本研究就，針對所取之關鍵字，分為十個主題做討論，故將變數 K 設為 10，alpha、eta 則是在建立 LDA 模型時必備之參數。在建立模型後再以 LDAvis 套件，對模型進行視覺化的呈現，該套件會藉由 theta 及 phi 兩參數，去建立 json 檔，再以瀏覽器開啟 serVis 函式產生之 html 檔，即可看到視覺化的模型，該模型會顯示圖每個主題中，對應到之前 30 個重要的關鍵字，在 json 檔中亦可看到整個主題中所分配到的關鍵字。

4. 情緒語意分析

依據上一步驟所得之各主題的關鍵字，計算出各主題所代表之情緒。

在這部分分為兩個步驟，首先是建立情緒詞庫，接著才是計算情緒分數。

在建立詞庫的部分，本研究中選用台大語言實驗室陳信希教授的NTUSD 詞庫以及中科院知網情緒辭典，因選用兩份辭典，故須將其整合，並將重複辭彙刪除，有問題的辭彙去除，以便後續正負情緒計算之用，其中中科院知網情緒辭典為簡體辭典，故須先進行編碼的轉換後將其轉為繁體，才可使用。有不少研究也是將兩份辭典結合後在使用(陳昱年，2013、張日威，2014)。

在計算分數的部分，將分群後主題中的關鍵字，與整理後的辭典，以Access 加以進行查詢，並統計出其分數。在分數的部分目前分為加權前，以及加權後的分數。

加權前的分數即為該主題中辭彙與情緒辭典中相符合之字數，而加權後的分數則是再乘上該字彙在主題中出現的次數。

第四章　研究結果

本研究將以研究方法所述為基礎，利用相關技術將 PTT 中取得之資料進行斷詞、取關鍵字、主題分群以及情緒語意分析等動作，試圖找出在學運期間網民討論議題之變化，以及針對主題討論的情緒變化，並依分析之結果，建立一個資料探勘的流程，便於未來得以應用於其他議題中。

以下第一節會說明 PTT 資料內容，並呈現資料蒐集之結果，第二節會逐步說明文字探勘由中文斷詞至關鍵字萃取，再分為四個模型說明主題分群至情緒語意分析的結果，並以圖形化方式呈現主題分群之結果。

4.1 資料蒐集

在本研究中會針對 PTT 政治相關版塊進行資料的抓取，在本研究中以 Python 程式語言，在 Python 中 Ipython Notebook 版本的環境中撰寫爬蟲程式進行資料爬取的動作。圖 6 所呈現之畫面為 PTT 中 HatePolitics 版(政治黑特版)中的文章列表頁面， 於 PTT 中，每一頁列表皆會有 20 篇文章，包含其編號、標題、發文作者以及發文日期等； 圖 7 則為 PTT 文章頁面，該頁面包含文章標題、作者、發文時間、文章內容、發文 IP 以及他人之回應等資料(圖 6、圖 7 為閱讀方便，色調皆經反轉) 。

圖 6 PTT 文章列表(色調經反轉)

圖 7 PTT 文章頁面(色調經反轉)

4.1.1 資料蒐集成果

本研究中資料部分因 PTT 刪文機制[14]，在研究期間回顧原始資料，發現有部分文章已被刪除，是以在本研究中將分為 4 個模型做解說，下表格 9 為各模型資料之時間範圍。

表格 9 資料模型

模型	刪文前後	資料範圍
A	後	2014 年 3 月起至 2015 年 3 月
B	後	2014 年 3 月起至 2014 年 11 月
C	前	2014 年 11 月起至 2015 年 3 月
D	後	2014 年 11 月起至 2015 年 3 月

標中 A 模型為發現刪文的所有資料，模型 B 則是以 2014 年 3 月學運期間起至 2014 年 11 月六都選戰前，資料同樣是刪文過後之資料，此範圍希望能專注探討學運期間討論情行。模型 C 為刪文前之 2014 年 11 月至 2015 年 3 月學運滿周年，模型 D 則是刪文後之 2014 年 11 月至 2015 年 3 月學運滿周年，後續將探討 C、D 模型之間的差異。

以下將呈現各模型資料蒐集之成果，表格 10 為 PTT 資料於 Excel 儲存的畫面。

表格 10 擷取 PTT 資料 Excel 畫面

b_作者	c_標題	d_日期	f_內文
tonemay(booksin)	[討論]認為洪案司法不公且自己不是 D 能兒請進	Fri Mar 7 19:19:41 2014	有誰能夠回答我三個問題，我就願意賜予你「非 D 能兒」.....

[14] PTT 刪文機制:於單一版塊文章超過 20000 篇，且文章未被鎖定(M)，文章會由舊聞開始刪除至低於 20000 篇

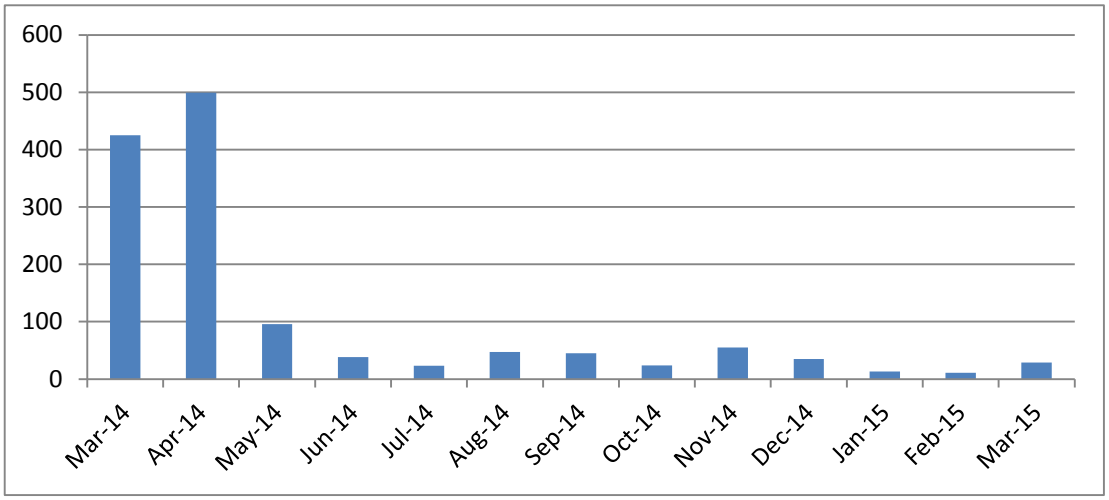

圖 8 PTT 資料蒐集成果(A 模型)

圖 8 為 A 模型資料蒐集之結果，可以看到文章數量於事件發生期間數量遠超過其他月份，而當中 2014 年 11 月則因為適逢六都選戰，使得文章數量相對其餘月份較為多一些。本模型將探討研究範圍中所有資料。

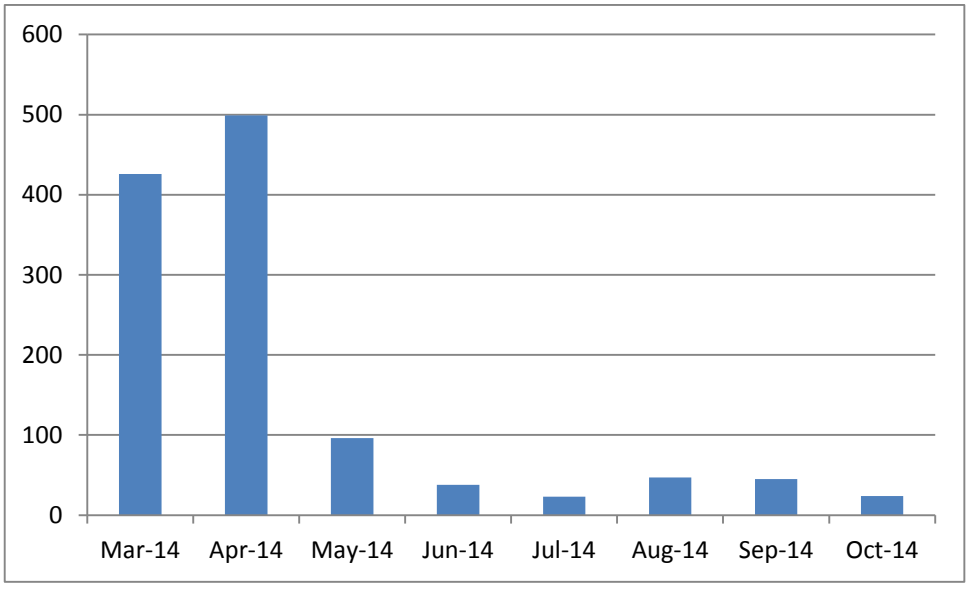

圖 9 PTT 資料蒐集成果(B 模型)

圖 9 為 B 模型資料蒐集之結果，資料區間為 2014 年 3 月至 2014 年 11 月，本模型在第一次爬取資料時，文章已被 PTT 刪文機制刪文過，故本區間資料與 A 模型 3 月至 11 月相同，試圖探討學運期間之情形。

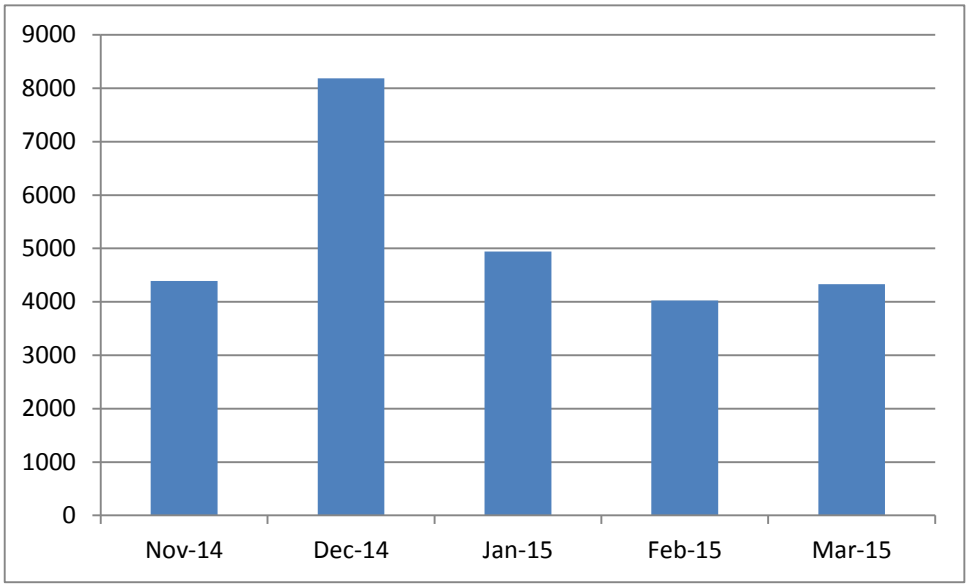

圖 10 PTT 資料蒐集成果(C 模型)

圖 10 為 C 模型資料蒐集之結果，資料區間為 2014 年 11 月至 2015 年 3 月，本模型資料為刪文前爬取，故數量上相對多於其他模型。

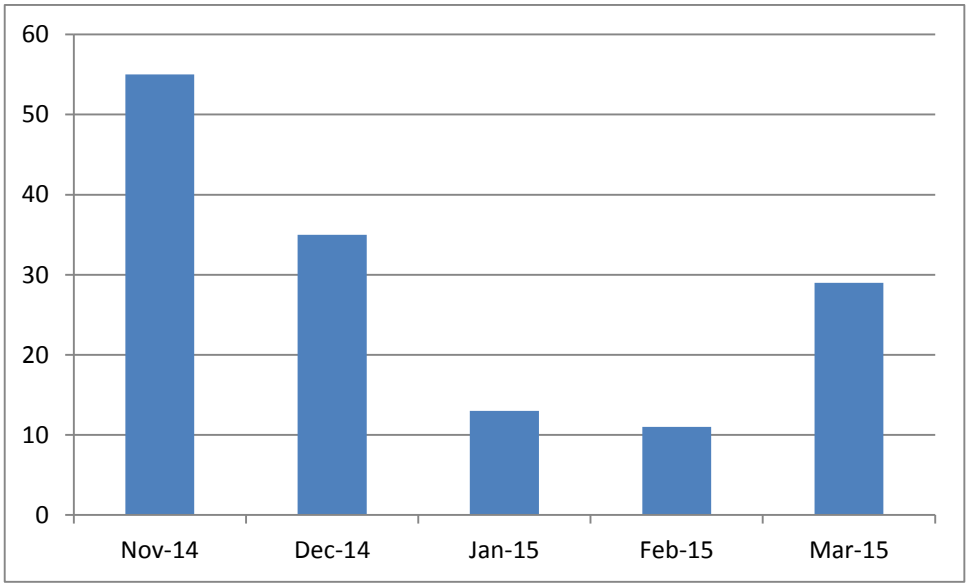

圖 11 PTT 資料蒐集成果(D 模型)

圖 11 為 D 模型資料蒐集之結果，資料區間為 2014 年 11 月至 2015 年 3 月，本模型資料為刪文後重新爬取之資料，後續將比較 C 模型與 D 模型之差異。

4.2 文字探勘

4.2.1 中文斷詞

本研究中中文斷詞部分同樣使用 Python 進行撰寫，斷詞結果舉例如表格 11。

表格 11 斷詞結果舉例

原文	斷詞結果			
有誰能夠回答我三個問題，我就願意賜予你「非 D 能兒」的頭銜。1.洪案的被告犯了哪些罪？請提出罪名與法條。(30%)2.你手上有哪些證據可以證明被告犯了上述罪名？(30%)(溫情提醒：請不要拿名嘴說過的話、新聞來當證據)3.如果你是法官，會怎麼判？(30%)送分題：你覺得自己是不是 D 能兒？(10%)--哈哈哈～--<SPANCLASS="F2">※發信站：批踢踢實業坊(PTT.CC)</SPAN> ◆ FROM:114.36.153.221<DIVCLASS="PUSH"><SPANCLASS="F1HLPUSH-TAG">→</SPAN><SPANCLASS="F3HLPUSH-USERID">VALEPIY</SPAN><SPANCLASS="F3PUSH-CONTENT">: 無視亨</SPAN><SPANCLASS="PUSH-IPDATETIME">03/0717:41</SPAN>	有 誰 能夠 回答 我 三個 問題 我 就 願意 賜予 你 「 非 D 能兒 」 的 頭銜 洪案 的 被告 犯 了 哪些	罪 ？ 請 提出 罪名 與 法條 你 手上 有 哪些 證據 可以 證明 被告 犯 了 上述 罪名 溫情 提醒 請 不要 拿 名嘴	說 過 的話 新聞 來 當 證據 如果 你 是 法官 ， 會 怎麼 判 送 分題 你 覺得 自己 是不 是 D 能兒 哈哈	spancl ass = 哪些 罪 ？ 請 F2 ※ 發信 站 批 踢踢 實業 坊 (ptt . cc span > ◆ From :

4.2.2 關鍵字萃取

本研究關鍵字萃取亦使用 Python 進行，針對四組模型視其資料量，分別取出不同數量的關鍵字，下表格 12 為四組模型與取出關鍵字數量。

表格 13則為取出關鍵字之舉例，在Python中同時可以計算關鍵字的權重，其理論為 TF-IDF，依據斷詞結果出現頻率進行關鍵字的取得，其中產生部分關鍵字為無意義之英文(如 PTT 格式中的 class、span 等)，故在後續有將無意義之英文關鍵字予以排除。

表格 12 模型與關鍵字數

模型	關鍵字數量	去除英文後的關鍵字數量
A	1000	832
B	500	389
C	4000	3604
D	300	250

表格 13 取出關鍵字之舉例

關鍵字	權重	關鍵字	權重
民進黨	0.026417	中國	0.020673
獲得	0.025132	現在	0.02003
學生	0.025057	服貿	0.019501
大陸	0.024679	恭喜	0.018706
國民黨	0.022449	還是	0.018027

4.2.3 主題分群結果

本研究中主題分群以 R 語言實作，針對取出之關鍵字進行分群，其理論依據為 LDA，以下分為四個模型，分別說明主題分群之結果，並以圖形化方式呈現。

1. 模型 A

圖 12 為模型 A 所有資料之分群結果(Default)，針對本模型，分為 10 個群組(含所有資料集共有 11 個)，該群組之關鍵字為在所有分群結果中，較常出現者，可猜測若該關鍵字出現於後續各主題中，可能較無意義，亦或者是較無法代表該主題。

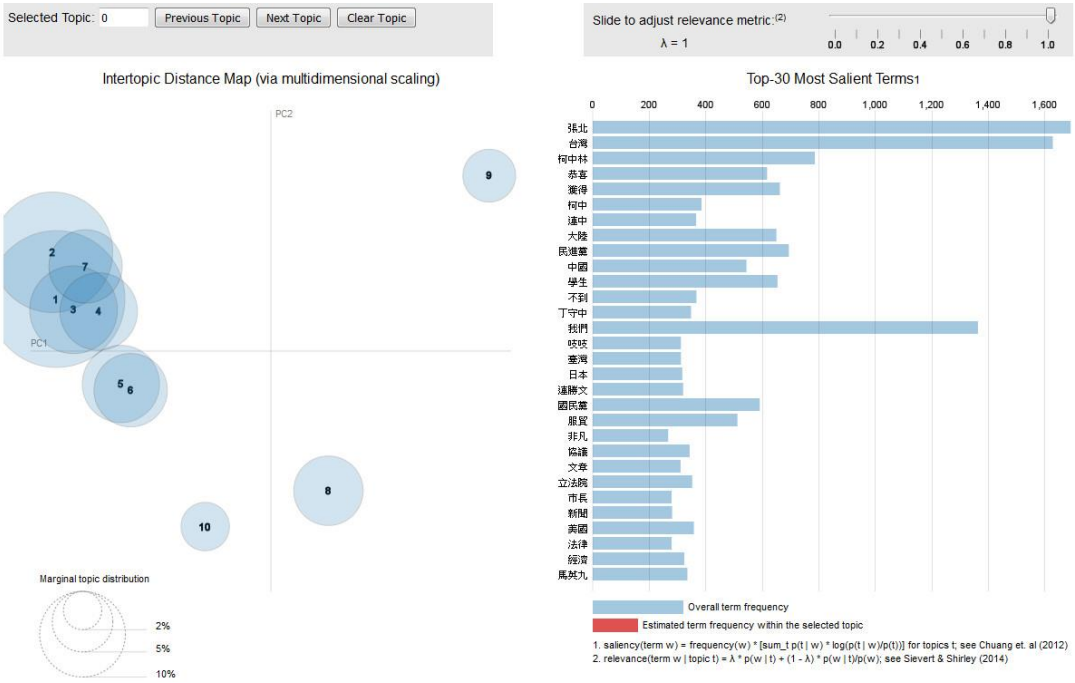

圖 12 模型 A 整體資料集 LDA 圖像化及前 30 個關鍵字

表格 14 則為在全體關鍵字這個群組中，較為重要的前 30 個關鍵字。

表格 14 模型 A 整體資料集前 30 個關鍵字列表

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
張北	1691	學生	654	柯中	385	協議	343	臺灣	312
台灣	1628	大陸	650	不到	367	馬英九	335	文章	311
我們	1363	恭喜	617	連中	366	經濟	324	新聞	281
柯中林	786	國民黨	591	美國	358	連勝文	320	市長	279
民進黨	694	中國	544	立法院	352	日本	317	法律	279
獲得	662	服貿	512	丁守中	348	吱吱	312	非凡	267

下圖 13 為模型 A 中，分群結果為主題 1 的關鍵字。

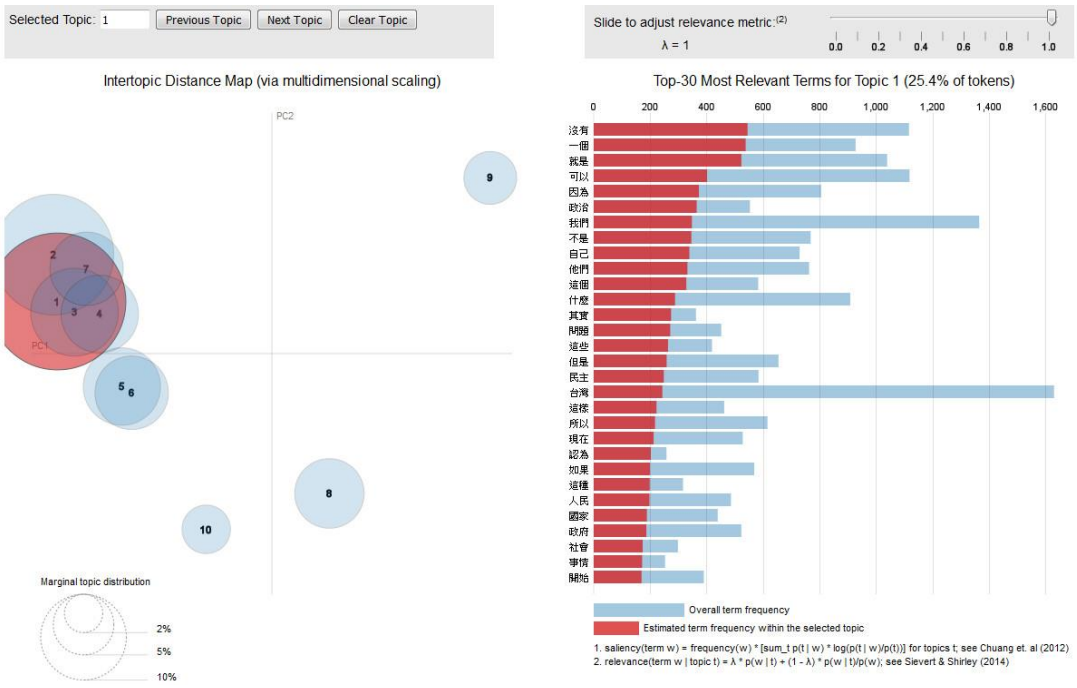

圖 13 模型 A 主題 1 之 LDA 圖像化及前 30 個關鍵字

表格 15 是在模型 A 的主題 1 中，較為重要的前 30 個關鍵字，觀察本群組與 Default 群組的差異，僅「我們」與「台灣」兩詞相同，其餘關鍵字多為討論國家、政府、政治、人民、民主相關辭彙，猜測應為討論台灣民主政治議題，本研究中將此主題命名為「民主」。

表格 15 模型 A 主題 1 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
沒有	544.43	~~我們~~	347.4	其實	273.76	這樣	222.01	人民	197.13
一個	537.47	不是	345.41	問題	270.77	所以	217.04	國家	188.18
就是	522.54	自己	338.44	這些	262.81	現在	212.06	政府	186.19
可以	401.14	他們	331.48	但是	257.84	認為	202.11	社會	173.25
因為	372.28	這個	327.5	民主	247.89	如果	199.13	事情	171.26
政治	364.32	什麼	287.69	~~台灣~~	242.91	這種	198.13	開始	169.27

關鍵字加註橫槓者表示其出現於 Default 群組中

下圖 14 為模型 A 中，分群結果為主題 2 的關鍵字。

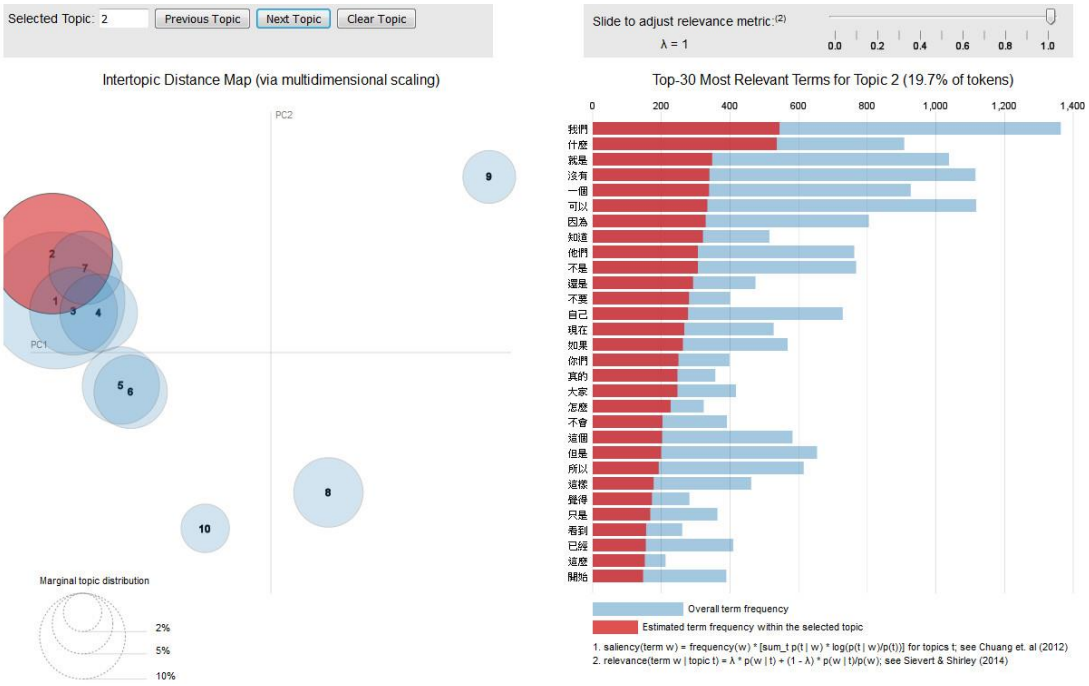

圖 14 模型 A 主題 2 之 LDA 圖像化及前 30 個關鍵字

表格 16 是在模型 A 的主題 2 中，較為重要的前 30 個關鍵字，觀察本群組與 Default 群組的差異，僅有「我們」一詞相同，但其餘辭彙並無特殊意義，猜測本群組分群之辭彙多無意義，亦或者是介系詞、主詞，並無特殊意義，故本研究針對此主題不作命名。

表格 16 模型 A 主題 2 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~我們~~	544.27	因為	329.18	自己	277.63	怎麼	227.08	覺得	172.57
什麼	536.34	知道	321.25	現在	266.73	不會	203.29	只是	167.61
就是	348.01	不是	306.38	如果	262.77	這個	202.3	看到	155.72
沒有	340.08	他們	306.38	你們	249.88	但是	199.33	已經	154.73
一個	339.09	還是	292.5	大家	246.91	所以	192.39	這麼	151.75
可以	334.13	不要	280.61	真的	246.91	這樣	177.52	開始	146.8

下圖 15 為模型 A 中，分群結果為主題 3 的關鍵字。

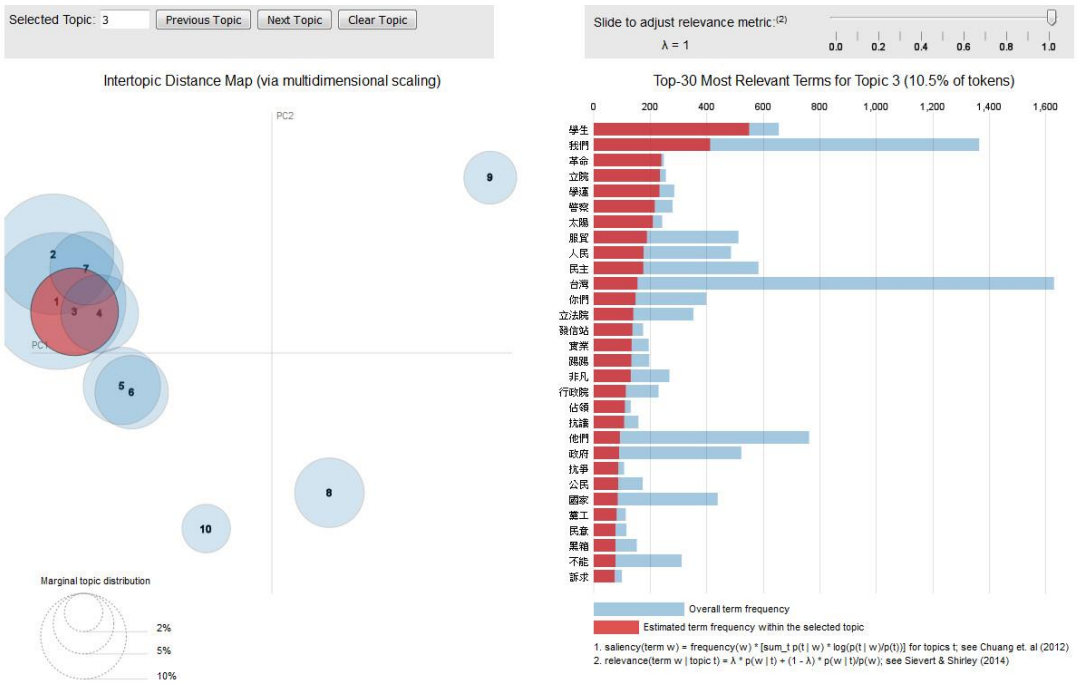

圖 15 模型 A 主題 3 之 LDA 圖像化及前 30 個關鍵字

表格 17 是在模型 A 的主題 3 中，較為重要的前 30 個關鍵字，觀察本群組與 Default 群組的差異，相同的辭彙有「學生」、「我們」、「服貿」、「台灣」、「立法院」、「非凡」，其餘辭彙則多為討論學運、立法院、抗爭、革命等，猜測本主題應在討論學運期間，學生占領立法院，也就是太陽花學運，故本研究將此主題命名為「學運」。

表格 17 模型 A 主題 3 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~學生~~	549.87	太陽	208.97	~~立法院~~	140	佔領	110.45	國家	84.83
~~我們~~	411.93	~~服貿~~	188.28	發信站	137.05	抗議	107.49	黨工	80.889
革命	239.51	人民	176.46	實業	134.09	他們	92.712	不能	76.948
立院	234.59	民主	175.47	踢踢	133.11	政府	89.756	民意	76.948
學運	232.62	~~台灣~~	154.78	~~非凡~~	131.14	公民	86.8	黑箱	76.948
警察	215.87	你們	147.89	行政院	113.4	抗爭	86.8	訴求	73.992

下圖 16 為模型 A 中，分群結果為主題 4 的關鍵字。

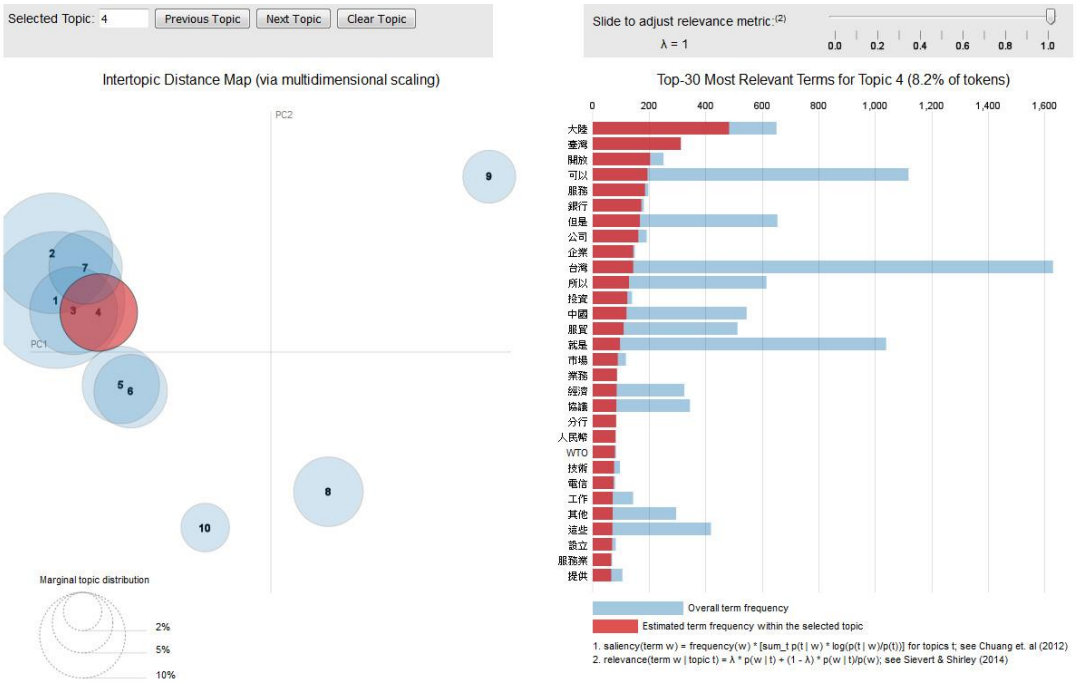

圖 16 模型 A 主題 4 之 LDA 圖像化及前 30 個關鍵字

表格 18 是在模型 A 的主題 4 中，較為重要的前 30 個關鍵字，觀察本群組與 Default 群組的差異，相同辭彙有「大陸」、「臺灣」、「台灣」、「中國」...等，其餘辭彙為服務業、工作、開放、市場、WTO 等，本主題應為討論開放服貿後導致服務業與市場受到衝擊，本研究中將此主題命名為「服貿」。

表格 18 模型 A 主題 4 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~大陸~~	482.49	但是	167.2	~~中國~~	119.45	~~協議~~	83.648	工作	70.718
~~臺灣~~	311.42	公司	161.23	~~服貿~~	109.51	分行	82.654	其他	70.718
開放	203	~~台灣~~	143.33	就是	96.578	人民幣	80.664	這些	69.723
可以	194.05	企業	143.33	市場	88.621	WTO	78.675	設立	68.729
服務	185.1	所以	128.41	業務	85.637	技術	75.691	可能	65.745
銀行	172.17	投資	122.44	經濟	84.643	電信	73.702	服務業	65.745

下圖 17 為模型 A 中，分群結果為主題 5 的關鍵字。

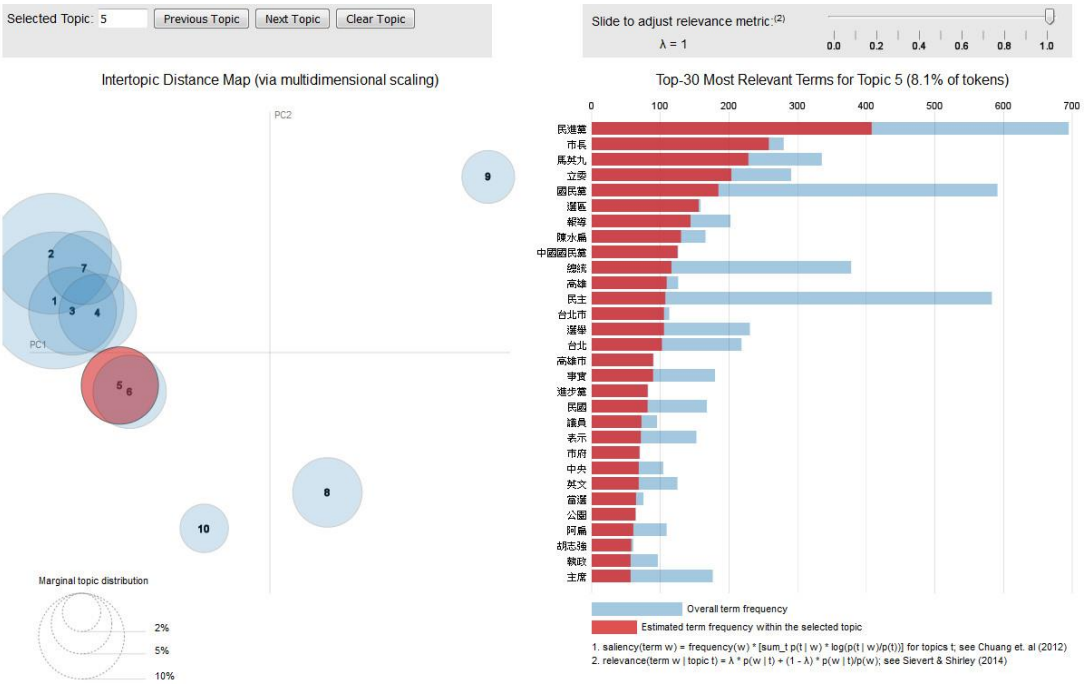

圖 17 模型 A 主題 5 之 LDA 圖像化及前 30 個關鍵字

表格 19 是在模型 A 的主題 5 中，較為重要的前 30 個關鍵字，觀察本群組與 Default 群組的差異，並扣除相同辭彙後，其餘辭彙多為政黨、選舉、台北、高雄、市長、議員等，猜測本主題應為討論 11 月的選戰，本研究中將此主題命名為「選舉」。

表格 19 模型 A 主題 5 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~民進黨~~	407.85	報導	143.95	台北市	105.26	進步黨	81.451	當選	64.585
~~市長~~	258.04	陳水扁	130.06	選舉	105.26	議員	72.522	公園	63.593
~~馬英九~~	228.28	中國國民黨	125.1	台北	102.29	表示	71.53	阿扁	60.617
立委	203.48	總統	116.17	事實	89.388	市府	69.546	胡志強	57.641
~~國民黨~~	184.63	高雄	109.23	高雄市	89.388	中央	68.554	主席	56.649
選區	155.86	民主	107.25	民國	81.451	英文	68.554	執政	56.649

下圖 18 為模型 A 中，分群結果為主題 6 的關鍵字。

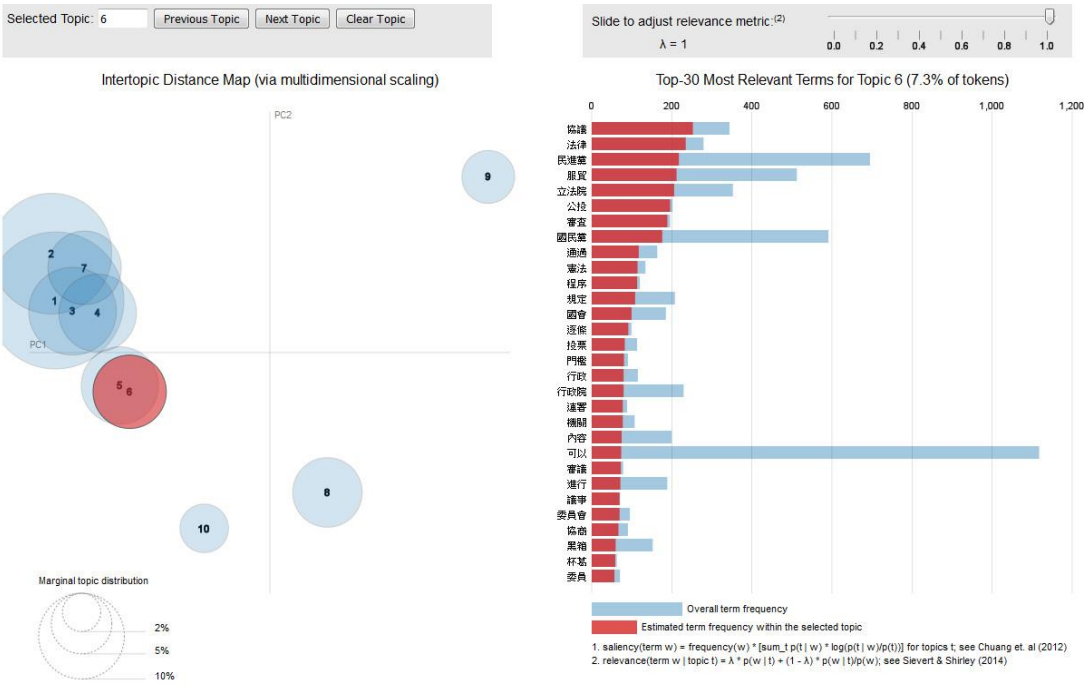

圖 18 模型 A 主題 6 之 LDA 圖像化及前 30 個關鍵字

表格 20 是在模型 A 的主題 6 中，較為重要的前 30 個關鍵字，觀察本群組與 Default 群組的差異，並扣除相同辭彙後，其餘辭彙為張慶忠、杯葛、審查、協商、條例等立法相關辭彙，猜測應為討論服貿立法過程，亦或者是黑箱通過服貿事件，本研究中將此主題命名為「立法院」。

表格 20 模型 A 主題 6 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
協議	252.61	審查	188.99	國會	99.513	連署	77.642	委員會	69.689
法律	234.72	國民黨	176.06	逐條	91.56	機關	77.642	議事	69.689
民進黨	217.82	通過	117.41	投票	82.613	內容	74.66	協商	66.707
服貿	211.85	憲法	114.43	門檻	80.625	可以	73.666	黑箱	59.748
立法院	205.89	程序	113.43	行政	79.631	審議	72.672	杯葛	58.754
公投	195.95	規定	108.46	行政院	79.631	進行	71.678	委員	56.765

下圖 19 為模型 A 中，分群結果為主題 7 的關鍵字。

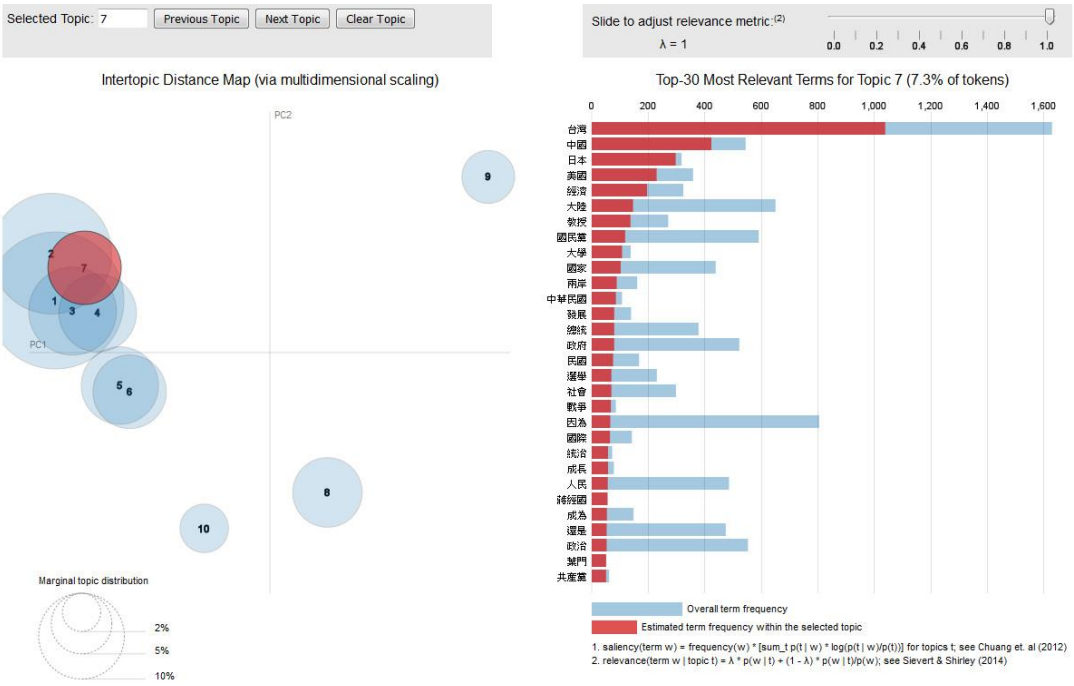

圖 19 模型 A 主題 7 之 LDA 圖像化及前 30 個關鍵字

表格 21 是在模型 A 的主題 7 中，較為重要的前 30 個關鍵字，觀察本群組與 Default 群組的差異，並扣除相同辭彙後，其餘辭彙為兩岸、政府、共產黨、國家、中華民國等，多為討論兩岸關係相關辭彙，本研究中將此主題命名為「兩岸」。

表格 21 模型 A 主題 7 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~台灣~~	1038.5	教授	137.22	政府	79.592	戰爭	67.668	蔣經國	55.744
~~中國~~	423.4	~~國民黨~~	118.34	發展	79.592	因為	65.68	成為	53.757
~~日本~~	297.2	大學	107.41	總統	79.592	國際	64.687	政治	52.763
~~美國~~	229.63	國家	102.45	民國	75.617	成長	57.731	還是	52.763
~~經濟~~	195.85	兩岸	88.534	社會	69.655	統治	57.731	共產黨	50.776
~~大陸~~	146.17	中華民國	85.553	選舉	69.655	人民	56.738	葉門	50.776

下圖 20 為模型 A 中，分群結果為主題 8 的關鍵字。

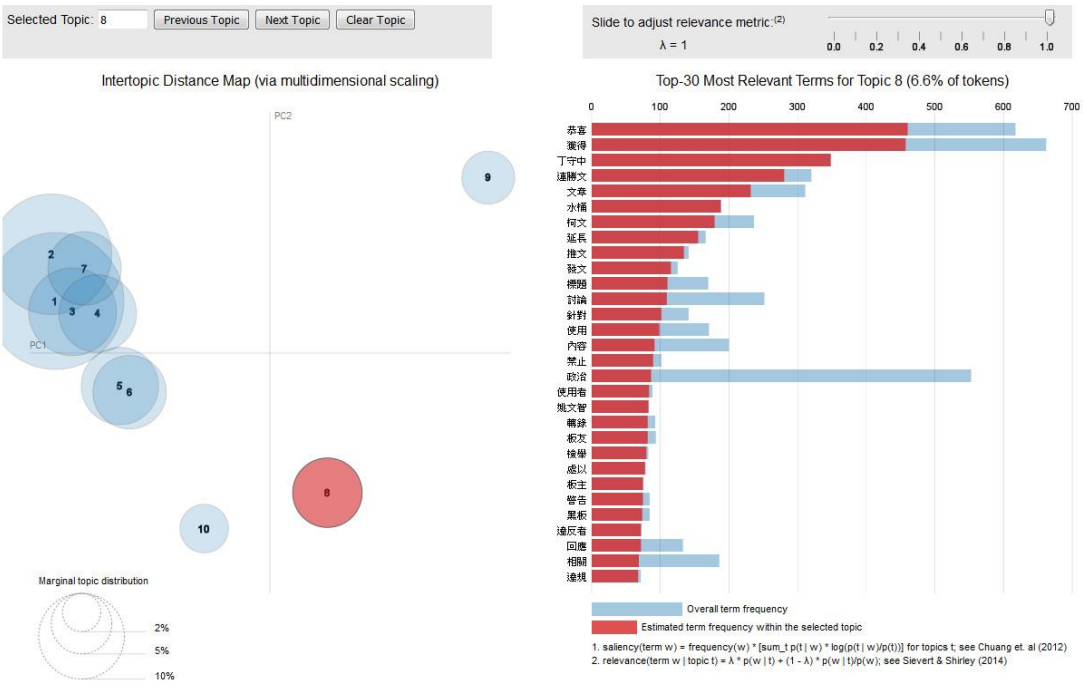

圖 20 模型 A 主題 8 之 LDA 圖像化及前 30 個關鍵字

表格 22 是在模型 A 的主題 8 中，較為重要的前 30 個關鍵字，觀察本群組與 Default 群組的差異，並扣除相同辭彙後，其餘辭彙多為 PTT 相關辭彙，如警告、回應、水桶、版主、違規等，本研究中將此主題命名為「PTT」。

表格 22 模型 A 主題 8 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~恭喜~~	460.34	柯文	179.03	針對	101.49	姚文智	82.606	警告	74.653
獲得	457.36	延長	155.17	使用	98.51	板友	81.612	黑板	73.659
~~于守中~~	348.02	推文	134.3	內容	91.552	轉錄	81.612	回應	71.671
連勝文	280.42	發文	115.41	禁止	89.564	檢舉	79.623	違反者	71.671
~~文章~~	231.71	標題	110.44	政治	86.582	處以	77.635	相關	68.689
水桶	187.97	討論	109.44	使用者	83.6	板主	74.653	違規	67.695

下圖 21 為模型 A 中，分群結果為主題 9 的關鍵字。

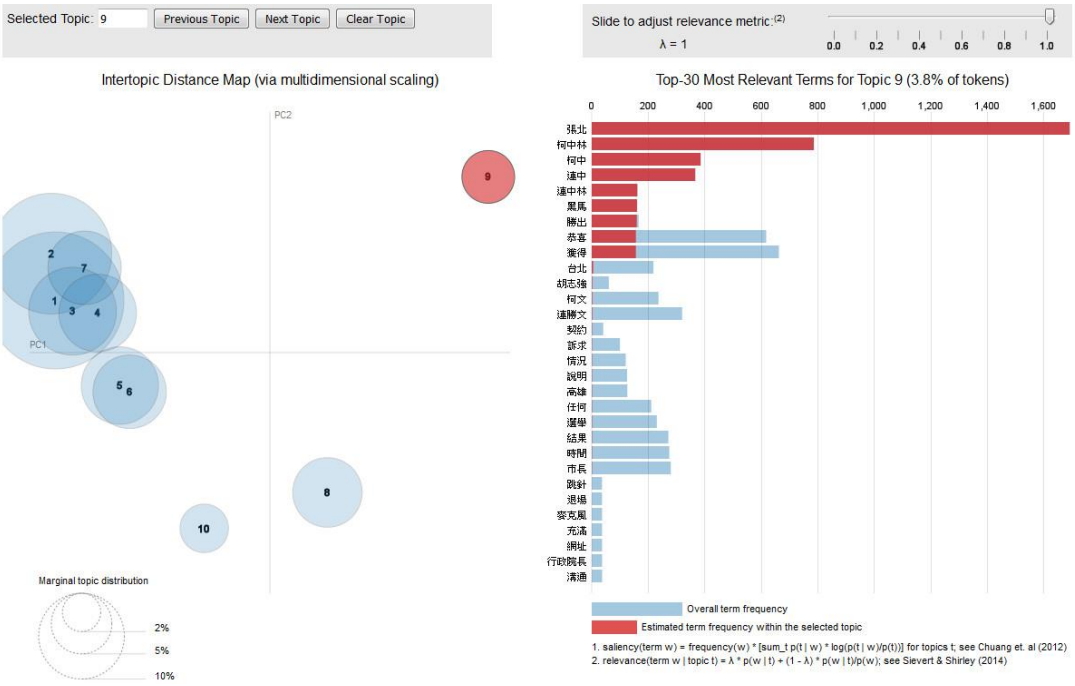

圖 21 模型 A 主題 9 之 LDA 圖像化及前 30 個關鍵字

表格 23 是在模型 A 的主題 9 中，較為重要的前 30 個關鍵字，觀察本群組與 Default 群組的差異，並扣除相同辭彙後，其餘辭彙為黑馬、市長、柯文、連中林、選舉等，本主題應是在討論柯文哲於台北勝出，林佳龍於台中勝出等議題，本研究中將此主題命名為「選舉」。

表格 23 模型 A 主題 9 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~張北~~	1690.9	勝出	159.23	~~連勝文~~	2.0886	情況	1.094	什麼	0.09945
~~柯中林~~	785.82	~~恭喜~~	156.25	~~市長~~	1.094	結果	1.094	他們	0.09945
~~柯中~~	385	~~獲得~~	156.25	任何	1.094	訴求	1.094	充滿	0.09945
~~連中~~	366.11	台北	5.0724	契約	1.094	說明	1.094	~~台灣~~	0.09945
連中林	161.22	柯文	2.0886	時間	1.094	選舉	1.094	因為	0.09945
黑馬	160.23	胡志強	2.0886	高雄	1.094	一個	0.09945	行政院長	0.09945

下圖 22 為模型 A 中，分群結果為主題 10 的關鍵字。

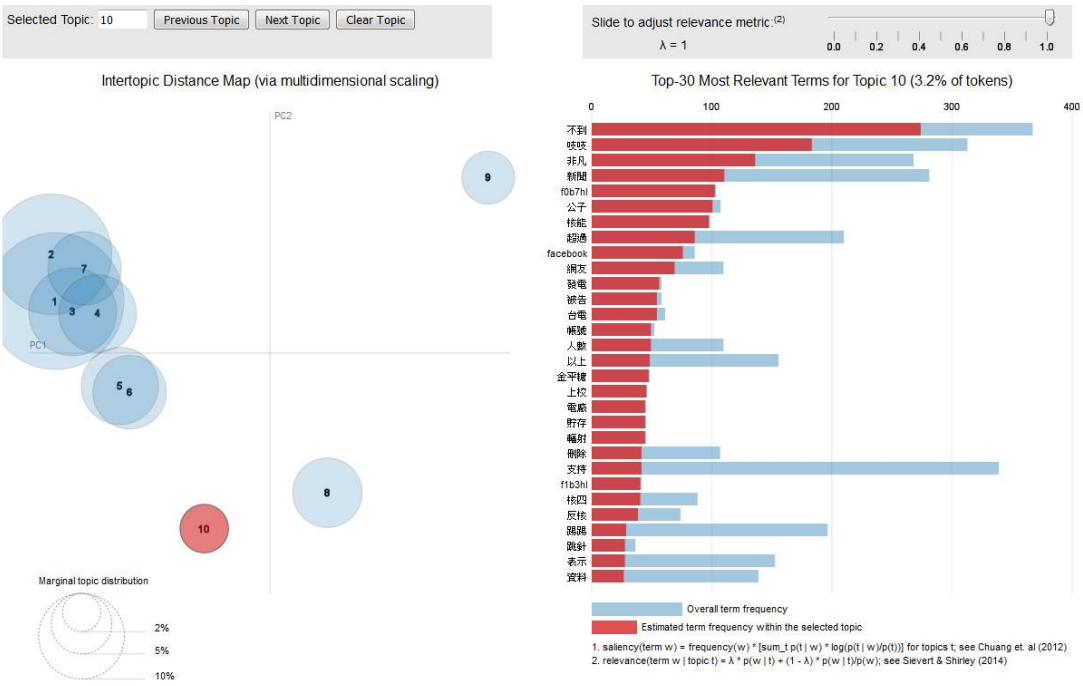

圖 22 模型 A 主題 10 之 LDA 圖像化及前 30 個關鍵字

表格 24 表 24 是在模型 A 的主題 10 中，較為重要的前 30 個關鍵字，觀察本群組與 Default 群組的差異，並扣除相同辭彙後，其餘辭彙為反核、台電、輻射、核四等辭彙，推測本主題應在討論廢核相關議題，本研究中將此主題命名為「核電」或者是「反核」。

表格 24 模型 A 主題 10 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~不到~~	273.88	核能	97.597	被告	54.264	貯存	44.416	核四	40.477
~~吱吱~~	183.28	超過	85.779	人數	49.34	電廠	44.416	反核	38.507
~~非凡~~	136.01	facebook	75.931	帳號	49.34	輻射	44.416	踢踢	28.659
~~新聞~~	110.4	網友	69.037	以上	48.355	支持	41.462	表示	27.674
f0b7h	102.52	發電	56.234	金平糖	47.371	刪除	41.462	跳針	27.674
公子	100.55	台電	54.264	上校	45.401	f1b3h	40.477	資料	26.689

2. 模型 B

以下因篇幅限制，僅列出關鍵字，並推測該群組討論何項議題，模型 B 中共有 5 個主題(含 Default 共 6 個)，表格 25 為模型 B 中 Default 群組的關鍵字。

表格 25 模型 B 所有資料集前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
台灣	1379	獲得	500	協議	339	日本	284	革命	244
我們	1187	恭喜	462	丁守中	338	學運	273	太陽	218
學生	639	民主	454	立法院	338	非凡	269	公投	200
大陸	591	中國	432	臺灣	311	法律	269	審查	194
民進黨	530	國民黨	413	吱吱	308	立院	254	延長	167
服貿	501	不到	352	連勝文	293	立委	254	柯文	161

表格 26 是在模型 B 的主題 1 中，較為重要的前 30 個關鍵字，觀察本群組與模型 B 中 Default 群組的差異，發現其餘辭彙多無特殊意義，如我們、這些、已經、其實...等，本研究中針對此主題不作命名。

表格 26 模型 B 主題 1 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~我們~~	838.01	不是	600.88	但是	412.58	還是	320.92	不要	256.16
什麼	718.45	因為	561.03	知道	398.63	大家	300.99	這些	248.18
一個	667.64	自己	504.24	現在	359.77	不會	290.03	已經	245.2
就是	650.7	可以	468.37	所以	346.82	你們	282.06	其實	244.2
沒有	637.75	~~台灣~~	441.47	如果	345.82	真的	274.09	覺得	232.24
他們	627.79	這個	424.53	這樣	333.87	怎麼	265.12	只是	232.24

表格 27 是在模型 B 的主題 2 中，較為重要的前 30 個關鍵字，觀察本群組與模型 B 中 Default 群組的差異，發現其餘辭彙如經濟、美國、企業、開放、發展...等，應為討論開放服貿後影響台灣經濟狀況等議題，本研究中將此主題命名為「經濟」或「服貿」。

表格 27 模型 B 主題 2 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~台灣~~	723.61	經濟	215.95	民國	165.99	所以	115.02	但是	95.036
~~大陸~~	572.71	開放	212.96	就是	144	政府	112.02	發展	92.038
~~中國~~	355.86	美國	187.97	企業	141	~~協議~~	105.03	核能	90.039
~~臺灣~~	310.89	教授	181.98	公司	141	問題	103.03	國家	88.04
可以	261.92	服務	178.98	投資	128.01	可能	102.03	~~服貿~~	88.04
~~日本~~	253.93	銀行	166.99	沒有	123.02	大學	101.03	如果	85.043

表格 28 是在模型 B 的主題 3 中，較為重要的前 30 個關鍵字，觀察本群組與模型 B 中 Default 群組的差異，發現其餘辭彙如警察、占領、抗議、自由...等，應為討論學運期間，學生占領行政院遭到警察驅逐等事件，本研究中將此主題命名為「學運」。

表格 28 模型 B 主題 3 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~學生~~	442.78	~~吱吱~~	227.39	~~服貿~~	159.9	馬英九	122.18	你們	95.384
~~我們~~	327.64	~~台灣~~	213.5	發信站	151.96	人民	114.24	黨工	92.406
~~非凡~~	258.16	~~太陽~~	211.51	踢踢	150.97	文章	111.26	佔領	90.421
~~立院~~	247.24	警察	201.59	~~立法院~~	139.06	行政院	111.26	可以	90.421
~~學運~~	237.32	新聞	169.82	報導	137.07	事實	108.29	抗議	87.443
~~革命~~	236.33	實業	159.9	~~民主~~	136.08	公子	106.3	自由	84.466

表格 29 是在模型 B 的主題 4 中，較為重要的前 30 個關鍵字，觀察本群組與模型 B 中 Default 群組的差異，其餘辭彙多為國會、通過、主題、程序、逐條...等，應為討論立法過程，本研究中將此主題命名為「立法院」。

表格 29 模型 B 主題 4 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~民進黨~~	474.86	~~公投~~	196	政府	118.04	國家	94.052	內容	79.06
~~國民黨~~	283.96	~~立法院~~	196	規定	114.04	行政院	86.056	機關	77.061
~~法律~~	236.98	~~審查~~	188	國會	113.04	可以	86.056	中國國民黨	73.063
~~協議~~	233.98	人民	184.01	通過	109.04	投票	84.057	委員會	73.063
~~服貿~~	224.99	~~民主~~	167.02	程序	105.05	政治	84.057	主席	73.063
~~立委~~	212.99	憲法	127.04	逐條	99.05	門檻	83.058	審議	71.064

表格 30 是在模型 B 的主題 5 中，較為重要的前 30 個關鍵字，觀察本群組與模型 B 中 Default 群組的差異，其餘辭彙分為兩類，一部分為 PTT 相關用詞，一部分為台北、市長、人數...等，猜測應為討論選舉議題，本研究中將此主題命名為「PTT 用詞」或「選舉」。

表格 30 模型 B 主題 5 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~恭喜~~	461.81	~~獲得~~	459.8	說明	3.1183	活動	1.1065	~~學生~~	2.1124
~~于守中~~	338.08	~~不到~~	261.63	原則	1.1065	市長	2.1124	責任	0.10059
~~連勝文~~	292.82	超過	76.549	時間	4.1242	完全	1.1065	國內	0.10059
~~柯文~~	161.04	人數	9.1537	踢踢	3.1183	台北	1.1065	崩潰	0.10059
姚文智	80.573	情況	6.136	實業	3.1183	報導	1.1065	承認	0.10059
~~延長~~	157.02	結果	9.1537	產生	1.1065	一下	1.1065	再來	0.10059

3. 模型 C

模型 C 中共有 10 個主題(含 Default 共 11 個)，表格 31 為模型 C 中 Default 群組的關鍵字。

表格 31 模型 C 所有資料集前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
台灣	23220	馬英九	10880	表示	7450	朱立倫	5073	討論	3620
國民黨	22709	中國	10771	但是	7111	新聞	4782	罷免	2617
柯文	13533	政府	9740	結果	6602	比較	4664	注意	2106
總統	12742	台北	9274	台北市	5667	市府	4240	警告	1743
市長	12209	選舉	8214	主席	5561	經濟	4110	張北	1730
民進黨	11789	立委	7795	美國	5306	大陸	3703	男生	1652

表格 32 是在模型 C 的主題 1 中，較為重要的前 30 個關鍵字，觀察本群組與模型 C 中 Default 群組的差異，其餘辭彙如就是、什麼、自己、真的、應該...等，較無其意義，故本研究針對此主題不做命名。

表格 32 模型 C 主題 1 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
就是	11075	沒有	7153.8	知道	5580.6	所以	4611.3	大家	4234.2
什麼	9517.1	現在	7152.8	還是	5373.6	我們	4606.4	這些	4116.8
可以	8323	真的	6823.4	因為	5311.9	這種	4547.7	問題	3828.2
不是	8065.3	引述	6127.9	如果	4969.6	出來	4516.8	只是	3584.4
自己	7883.2	之銘言	6047.3	他們	4916.8	這個	4413.3	~~但是~~	3551.6
一個	7483.2	這樣	5884.1	覺得	4702.9	不會	4331.7	應該	3525.7

表格 33 是在模型 C 的主題 2 中，較為重要的前 30 個關鍵字，觀察本群組與模型 C 中 Default 群組的差異，其餘辭彙如王金平、連勝文、支持、政治...等，應再討論王金平支持連勝文競選台北市長等事件，本研究中將此主題命名為「選舉」。

表格 33 模型 C 主題 2 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~國民黨~~	16246	現在	4538.9	連勝文	3401	可能	2758.8	王金平	2325.7
~~總統~~	9059.8	就是	4410.5	沒有	3280.5	還是	2715	應該	2277
~~馬英九~~	7227.9	英文	4376.6	~~柯文~~	3144.1	真的	2624.4	之銘言	2274
~~民進黨~~	6224.4	出來	3976.4	可以	2994.8	一個	2444.2	政治	2265
~~主席~~	4777.8	如果	3644.9	不會	2900.2	支持	2409.4	不是	2230.2
~~朱立倫~~	4686.2	~~選舉~~	3415.9	~~市長~~	2848.4	引述	2363.6	自己	2121.7

表格 34 是在模型 C 的主題 3 中，較為重要的前 30 個關鍵字，觀察本群組與模型 C 中 Default 群組的差異，其餘辭彙如國家、中華民國、中共、兩岸、政治...等兩岸關係議題，本研究中將此主題命名為「兩岸」。

表格 34 模型 C 主題 3 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~台灣~~	12763	兩岸	2687.4	因為	1839.9	問題	1484.2	這個	1220.3
~~中國~~	10363	一個	2544.5	中共	1749	~~經濟~~	1440.2	歷史	1205.3
~~美國~~	4415.3	就是	2412.6	臺灣	1736	政治	1407.2	如果	1186.4
~~大陸~~	3243.1	日本	2301.7	不是	1720	民主	1350.3	~~政府~~	1137.4
國家	3154.1	中華民國	2266.7	可以	1659.1	他們	1275.3	什麼	1108.4
我們	2777.4	沒有	2164.7	共識	1612.1	人民	1240.3	所以	1096.4

表格 35 是在模型 C 的主題 4 中，較為重要的前 30 個關鍵字，觀察本群組與模型 C 中 Default 群組的差異，其餘辭彙如巨蛋、專用道、市民、工程等，為討論選戰過後台北市長柯文哲各項政策，本研究中將此主題命名為「台北市長」或者「柯文哲」。

表格 35 模型 C 主題 4 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~柯文~~	6675.1	~~政府~~	2004.9	公車	1413.3	林佳龍	974.52	中央	784.63
~~市長~~	5807.6	~~表示~~	1889	市民	1371.3	如果	895.56	應該	775.64
~~市府~~	4077.7	問題	1739.1	工程	1280.3	交通	846.59	一個	768.64
~~台北~~	3648.9	遠雄	1598.1	局長	1236.4	時間	822.61	就是	764.64
~~台北市~~	3128.2	捷運	1529.2	可以	1165.4	專用道	813.61	議員	760.64
巨蛋	2492.6	沒有	1460.2	上任	1053.5	要求	810.61	郝龍斌	758.65

表格 36 是在模型 C 的主題 5 中，較為重要的前 30 個關鍵字，觀察本群組與模型 C 中 Default 群組的差異，其餘辭彙如調查、事件、陳水扁、馬英九等，應為討論政治人物案件調查等議題，本研究中將此主題命名為辦案、弊案、或總統等。

表格 36 模型 C 主題 5 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~表示~~	3327.5	~~總統~~	1932.2	自己	1098.6	時間	815.71	漂流	724.76
報導	2419	沒有	1588.3	指出	1002.6	調查	805.72	認為	719.76
~~柯文~~	2248	~~馬英九~~	1469.4	今天	992.63	事件	797.72	是否	708.76
~~台北~~	2130.1	記者	1405.4	臉書	913.67	當時	796.72	要求	689.77
媒體	2066.1	~~市長~~	1382.4	公文	835.7	強調	764.74	處理	688.77
~~新聞~~	1992.2	網友	1293.5	陳水扁	833.71	相關	729.75	知道	683.78

表格 37 是在模型 C 的主題 6 中，較為重要的前 30 個關鍵字，觀察本群組與模型 C 中 Default 群組的差異，其餘辭彙如民主、學運、蔡正元、為廷等，應為討論學運議題，本研究中將此主題命名為「學運」。

表格 37 模型 C 主題 6 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~立委~~	3275	政治	1561	公民	1219.6	太陽	1070.4	可以	833.09
~~罷免~~	2556.1	~~表示~~	1460.8	我們	1218.6	學運	1001.3	連署	830.09
~~民進黨~~	2242.8	~~選舉~~	1454.8	政黨	1213.5	蔡正元	934.21	~~政府~~	823.08
投票	2208.7	為廷	1447.8	門檻	1204.5	決定	872.14	通過	813.07
~~國民黨~~	2130.6	社會	1393.8	人民	1174.5	提出	867.13	反對	812.07
民主	1625	立法院	1292.6	支持	1143.5	民意	857.12	認為	810.06

表 38 是在模型 C 的主題 7 中，較為重要的前 30 個關鍵字，觀察本群組與模型 C 中 Default 群組的差異，其餘辭彙如政策、市場、年輕人、薪資、問題...等，應為討論服貿開放後之影響，本研究中將此主題命名為「服貿」。

表格 38 模型 C 主題 7 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~台灣~~	4279.8	政策	1183.4	薪資	1049.3	就是	803.99	現在	755.93
~~政府~~	3723.2	可以	1180.4	房價	1039.2	因為	800.98	財團	743.92
~~經濟~~	2374.7	國家	1169.4	發展	1022.2	增加	776.96	高鐵	740.92
問題	2106.4	產業	1147.4	年輕人	940.14	沒有	771.95	可能	723.9
投資	1461.7	工作	1067.3	社會	919.11	成長	759.94	應該	717.89
公司	1199.4	企業	1050.3	市場	824.01	人民	755.93	造成	690.86

表格 39 是在模型 C 的主題 8 中，較為重要的前 30 個關鍵字，觀察本群組與模型 C 中 Default 群組的差異，其餘辭彙如議員、賴清德、李全教、議長、桃園、台南...等，應為討論選舉議題以及台南市議長賄選事件，本研究中將此主題命名為「選舉」或者「黑箱」。

表格 39 模型 C 主題 8 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~立委~~	2345	議員	1627.4	苗栗	1002.9	可能	822.75	參選	688.65
~~民進黨~~	2204.9	這次	1340.2	當選	980.88	賴清德	821.75	李全教	681.64
~~國民黨~~	1998.7	高雄	1331.2	~~台北市~~	945.85	縣市	774.72	選民	670.63
~~選舉~~	1910.6	台南	1244.1	縣長	944.85	~~結果~~	772.71	新北	663.63
選區	1834.6	桃園	1126	議長	918.83	應該	768.71	出來	627.6
~~市長~~	1708.5	地方	1106	~~台北~~	882.8	候選人	692.65	萬票	626.6

表格 40 是模型 C 的主題 9 中，較為重要的前 30 個關鍵字，觀察本群組與模型 C 中 Default 群組的差異，其餘辭彙如學校、同年級、高中女生、教育、高中...等，應為討論教育議題，本研究中將此主題命名為「教育」。

表格 40 模型 C 主題 9 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~男生~~	1651.4	不起眼	615.91	大學	512.26	手機	478.73	年紀	424.87
女生	894.35	高中女生	615.91	女性	509.21	那個	471.62	相當	420.81
學生	893.34	~~但是~~	608.8	一起	505.15	男人	468.57	不爽	406.58
平常	722.61	~~比較~~	571.2	開始	505.15	想要	454.34	外省人	405.56
學校	633.19	~~結果~~	541.73	教育	503.12	是不是	452.31	~~台灣~~	404.55
同年級	622.01	老師	519.38	喜歡	501.08	別人	446.21	高中	399.47

表格 41 是在模型 C 的主題 10 中，較為重要的前 30 個關鍵字，觀察本群組與模型 C 中 Default 群組的差異，其餘辭彙如國軍、國防部、阿帕契、柯中林、連中...等，分別討論議題為阿帕契事件與 2014 年選戰議題，本研究中將此主題命名為「選舉」或者「阿帕契事件」。

表格 41 模型 C 主題 10 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~警告~~	1735.5	版主	783.84	吱吱	469.33	禁止	408.26	連中	370.6
~~注意~~	1735.5	政黑	706.48	沒過	453.04	判決	397.06	~~新聞~~	360.42
~~張北~~	1729.4	~~討論~~	632.18	標題	434.72	柯中	396.04	板友	354.31
崩潰	806.23	國軍	524.29	刪除	432.68	國防部	376.7	板主	351.26
柯中林	805.21	阿帕契	488.66	恭喜	430.65	轉錄	375.68	水桶	351.26
文章	799.11	推文	486.63	獲得	429.63	黑板	373.65	發文	334.97

4. 模型 D

模型 D 中共有 5 個主題(含 Default 共 6 個)，表 42 為模型 D 中 Default 群組的關鍵字。

表格 42 模型 D 所有資料集前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
張北	1691	可以	212	連中林	161	選區	110	大陸	61
柯中林	785	國民黨	179	黑馬	160	市長	94	標題	57
柯中	385	民進黨	168	恭喜	160	討論	87	中國國民黨	52
連中	359	獲得	167	民主	129	水桶	70	發文	49
台灣	254	因為	165	文章	114	美國	70	進步黨	44
沒有	233	勝出	162	中國	113	使用	62	成長	43

表格 43 是在模型 D 的主題 1 中，較為重要的前 30 個關鍵字，觀察本群組與模型 D 中 Default 群組的差異，其餘辭彙如馬英九、陳水扁、總統、選舉...等，本研究中將此主題命名為「總統」。

表格 43 模型 D 主題 1 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
一個	197.97	我們	150.24	陳水扁	106.49	這個	85.613	但是	78.652
~~沒有~~	178.09	什麼	128.37	他們	98.539	問題	84.618	現在	73.68
~~國民黨~~	163.17	自己	124.39	不是	89.59	事情	83.624	這些	72.686
就是	161.18	~~台灣~~	116.44	馬英九	87.601	這樣	81.635	真的	70.697
~~民進黨~~	156.21	~~因為~~	111.47	開始	87.601	其實	79.646	當時	69.703
~~可以~~	152.23	政治	110.47	如果	87.601	總統	78.652	選舉	68.709

表格 44 是在模型 D 的主題 2 中，較為重要的前 30 個關鍵字，觀察本群組與模型 D 中 Default 群組的差異，其餘辭彙綠營、選舉、連勝文、胡志強、柯文...等，本研究中將此主題命名為「選舉」。

表格 44 模型 D 主題 2 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~張北~~	1691.4	~~勝出~~	159.38	柯文	2.0905	~~市長~~	1.095	所謂	0.099549
~~柯中林~~	785.54	~~恭喜~~	157.39	說明	1.095	帳號	0.099549	活動	0.099549
~~柯中~~	385.35	~~獲得~~	155.4	情況	1.095	綠營	0.099549	幾乎	0.099549
~~連中~~	359.47	台北	5.077	時間	1.095	動作	0.099549	越來越	0.099549
~~連中林~~	161.37	連勝文	3.086	結果	1.095	產業	0.099549	參與	0.099549
~~黑馬~~	160.37	胡志強	2.0905	選舉	1.095	兩岸	0.099549	關係	0.099549

表格 45 是在模型 D 的主題 3 中，較為重要的前 30 個關鍵字，觀察本群組與模型 D 中 Default 群組的差異，其餘辭彙如歷史、社會、國家、開明、專制、國家...等，猜測應為探討民主國家與專制國家等議題，本研究中將此主題命名為「民主」。

表格 45 模型 D 主題 3 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~台灣~~	131.74	~~因為~~	53.549	華人	37.712	社會	27.814	債務	25.834
~~中國~~	110.96	經濟	52.559	但是	30.783	還是	27.814	戰爭	25.834
~~美國~~	67.407	葉門	50.58	我們	30.783	什麼	27.814	發展	25.834
~~大陸~~	61.468	~~可以~~	47.61	這個	29.794	市場	26.824	專制	24.844
~~民主~~	57.508	國家	46.62	不是	29.794	歐洲	26.824	皇軍	23.855
~~沒有~~	55.529	現在	39.692	歷史	27.814	反對	26.824	開明	23.855

表格 46 是在模型 D 的主題 4 中，較為重要的前 30 個關鍵字，觀察本群組與模型 D 中 Default 群組的差異，其餘辭彙如引戰、推文、刪除、版友，使用者...等，多數辭彙為 PTT 相關用語，本研究中將此主題命名為「PTT」。

表格 46 模型 D 主題 4 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~文章~~	82.869	推文	46.411	板友	32.616	檢舉	27.689	違反者	24.733
~~水桶~~	70.059	禁止	41.484	黑板	31.63	轉錄	27.689	時間	24.733
~~討論~~	58.235	刪除	37.542	違規	30.645	點名	26.703	看板	23.747
~~標題~~	52.323	內容	37.542	政治	28.674	代稱	25.718	發表	23.747
~~使用~~	51.338	板主	35.572	使用者	27.689	相關	25.718	回應	22.762
~~發文~~	47.396	針對	34.586	處以	27.689	引戰	24.733	攻擊	22.762

表格 47 是在模型 D 的主題 5 中，較為重要的前 30 個關鍵字，觀察本群組與模型 D 中 Default 群組的差異，其餘辭彙如議會、候選人、參選、柯文、違建...等，本分群中關鍵字討論議題應分為 2014 年選戰，以及選戰後台北市長柯文哲拆除違建等事件，本研究中將此主題命名為「選舉」或者「柯文哲」。

表格 47 模型 D 主題 5 之前 30 個關鍵字

關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻	關鍵字	詞頻
~~選區~~	109.93	~~成長~~	37.361	網友	23.633	派系	20.691	胡志強	13.827
~~市長~~	54.032	違建	27.555	無黨	21.672	帳號	19.71	政黨	12.846
~~中國國民黨~~	52.07	議會	27.555	可能	21.672	柯文	19.71	選舉	12.846
~~民主~~	47.167	薪資	26.575	泛綠	20.691	競選	18.73	但是	12.846
~~進步黨~~	44.226	黑派	24.613	總部	20.691	~~國民黨~~	15.788	~~可以~~	12.846
議員	40.303	泛藍	23.633	物價	20.691	候選人	13.827	參選	11.865

4.2.4 主題分群比較

各模型主題討論議題如下表格 48

表格 48 各模型主題

模型 / 主題	A	B	C	D
Topic1	民主	無特殊意義	無特殊意義	總統
Topic2	無特殊意義	經濟	選舉	選舉
Topic3	學運	學運	兩岸	民主
Topic4	服貿	立法院	台北市長、柯文哲	PTT 用詞
Topic5	選舉	PTT 用詞、選舉	辦案、弊案、總統	選舉、柯文哲
Topic6	立法院		學運	
Topic7	兩岸		服貿(	
Topic8	PTT		選舉、黑箱	
Topic9	選舉		教育	
Topic10	反核		選舉、阿帕契事件	

模型 A 各主題應其資料範圍為學運起至學運後一年，期間除學運外，亦經歷了 2014 年選舉，討論之議題也應此主要分為學運相關與選舉相關議題，在模型 A 中主題 1、主題 3、主題 4、主題 6、主題 7 各主題與學運較為相關，主題 1、主題 5、主題 9、主題 10 則與選舉有相關。模型 B 的資料範圍落於學運起至同年 10 月，該資料區間之資料在取得時已經被 PTT 刪除過一次，故剩餘資料在眾多資料中可能會較具有其意義，而結果顯示，本模型中的主題除了 PTT 各項用語被分為一個主題以外，其餘主題多為討論學運相關議題，符合訴求結果。

在模型 C 和模型 D 中，資料區間為 2014 年 11 月起至學運滿周年(隔年 3 月)，其中模型 C 為刪文前之資料，模型 D 為刪文後之資料，以資料量來說，模型 D 之各項主題應會出現在模型 C 中，而實驗結果中，模型 D 除了 PTT 用詞以外，各項主題也都有在模型 C 中，且也符合本期間發生之事件、如選舉、阿帕契事件...等。

在模型 B 和模型 D 的部分，則因為兩模型資料區間的不同，可以很明顯的看出兩模型中各議題討論的趨向，在模型 B 主要討論服貿與學運相關議題，而在模型 D 中則在討論 2014 年選戰相關議題。

而在模型 A、模型 B 與模型 D 之間，就資料區間來說，模型 A 為模型 B 與模型 D 之聯集，而在主題分群的結果中，可以看到在模型 B 與模型 D 中討論之議題皆出現在模型 A 中。故接續章節之探討將以模型 A 為主。

4.2.5 相關研究比較

此外，我們亦將本研究的成果其他的研究進行比較，在洪綾君等 2016 年的研究中，針對本研究議題依專家法提出了四個構面，分為政治、經濟、學運、人物...等，而該四個構面下共有 52 個關鍵字，內容如下表格 49：

表格 49 四個構面與 52 個關鍵字列表

構面	政治	經濟	學運	人物
關鍵字	獨立 統一 兩岸 台灣 中華民國 大陸 中共** 中華人民共和國* 民進黨 DPP dpp 國民黨 KMT kmt 民主 民粹** 認同 黑箱 政府 親中** 綠營 藍營**	服貿 反服貿協定* ECFA** 出口** 經濟發展 邊緣化** 貿易 自由貿易協定* 市場 經濟效益* 失業* 財團** GDP** 服務貿易協定* 兩岸經濟合作架構協議*	太陽花學運* 社會運動* 大學生* 立法院 黑島青* 暴力 街頭** 網路 社會媒體* 黑色島國青年陣線* 社交媒體*	林飛帆* 張慶忠 蔡正元** 祭止兀**

將表格 49 語本研究中模型 A(表格 48)之各項主題進行比較，模型 A 中多數主題均可被收納在表格 49 的構面之中：如主題 1(民主)包含在政治構面、主題 3(學運)包含在學運構面、主題 4(服貿)包含在經濟構面、主題 6(立法院)包含在學運構面 、主題 7(兩岸)包含在政治構面等。但也有部分主題不在其中，如主題 5 和 9(選舉)不包含在內，但在重新檢視該主題下的關鍵字後，應可將其歸納在政治的構面中；而主題 10(反核、核電) 也不包含在內，但在重新檢視該主題下的關鍵字後，再次確認該議題為爭辯已久的社會關注之重大議題，故將其歸納在於政治、經濟構面下(表格 50)。由此可見，本研究所歸納之主題名稱，大致符合洪綾君等(2016)的研究。

表格 50 主題與構面之對照

模型 / 主題	模型 A 主題名	構面對照
Topic1	民主	政治
Topic2	無特殊意義	無特殊意義
Topic3	學運	學運
Topic4	服貿	經濟
Topic5	選舉	政治*
Topic6	立法院	學運
Topic7	兩岸	政治
Topic8	PTT	無
Topic9	選舉	政治*
Topic10	反核	政治*、經濟*

此外，我們亦將表格 49 中專家法所得出的關鍵字與本研究模型 A 之關鍵字做對照比較，結果呈現於表格中的(*)與(**)記號上。表格中沒有標記的關鍵字表示該詞與本研究所採用的關鍵字完全相符，總計 24 個。而表格中出現(*)者，表示該詞的部分子詞有出現在本研究的關鍵詞表之中(總計 15 個)；而最後被標記成(**)者，表示該詞完全沒出現在模型 A 的前 1000 個關鍵字當中(總計 13 個)。不過，「林飛帆」一詞，在我們的關鍵字中是以「非凡」一詞呈現，因其代表同一個人，故將其列為部分相符。此外，值得注意的是「蔡正元」一詞並未出現在本研究所擷取出之前 1000 重要的關鍵字。

而反向檢視本研究之關鍵字詞，發現在人物面尚包含幾位政治人物，依其計算出的重要權重依序為「柯文哲」、「丁守中」、「連勝文」、「陳水扁」、「蔡英文」、「姚文智」、「王金平」、「李登輝」、「胡志強」、「蔣經國」、「方仰寧」、「楊秋興」、「謝長廷」、「蔣中正」。

4.2.6 情緒語意分析

本研究中情緒語意分析採用台大語言實驗室陳信希教授的 NTUSD 詞庫以及中科院知網情緒辭典，將兩者結合後與主題分群之結果進行比較，對個別主題進行正負情緒的判斷。表格 51 為部分正向與負向情緒詞之舉例。

表格 51 擷取部分情緒辭典辭彙

正向				負向			
喜愛	適合	誠實	深愛	不過	撞擊	逃掉	落下
歡歡喜喜	真理	喜感	正直	殺死	違失	逃走	惡臭
歡喜	急迫	不錯	難怪	弄死	違反	逃犯	惡毒
喜歡	合適	善進	善心	死傷	落差	臭狗	殺掉
喜好	合理	愛情	正當	擊落	落空	臭屁	殺低
安適	天真	精力	急欲	壓壞	煩悶	冒犯	破碎
愛好	歡樂	真誠	健美	壓破	毀損	昏倒	破掉
精美	心喜	美妙	柔和	暴亂	搞亂	走狗	害慘
安好	真好	衝破	正道	傷害	損毀	扯破	殘害
精明	未違反	精心	正巧	雜亂	傷殘	扭傷	惡搞
美好	明智	奮進	確知	擊倒	傷亡	低沉	惡劣
正好	確實	有助和平	得力	壓碎	亂搞	低劣	兇惡
明理	真實	天才	快樂	壓倒	亂屁	危急	兇殺
愛好和平	安定和平	好神	快快樂樂	壓低	殘暴	危困	欠罵
敬愛	愛心	熱誠	高妙	壓下	殘殺	兇暴	假冒

後續依據辭典與主題分群後之關鍵字進行比對，可以得到在該主題內的關鍵字數，後續在依據該辭彙在主題中出現的頻率，進行加權，若為正數，即該

主題為正向，若為負數則反之，。以下就各模型討論其取得情緒詞數以及加權後分數。

1. 模型 A

表格 52 辭典與模型 A 主題關鍵字對應筆數

主題	正向	負向	中立
Default	7	0	1
Topic1(民主)	19	10	9
Topic2(無)	12	3	8
Topic3(學運)	4	4	7
Topic4(服貿)	9	3	1
Topic5(選舉)	5	3	1
Topic6(立法)	10	8	1
Topic7(兩岸)	8	3	1
Topic8(PTT)	9	2	5
Topic9(選舉)	9	0	2
Topic10(反核)	3	3	6

表格 53 模型 A 主題加權後各主題分數

主題	正向	負向	中立
Default	3083	0	0
Topic1(民主)	2016	-1640	0
Topic2(無)	1974	-1383	0
Topic3(學運)	262	-440	0
Topic4(服貿)	1036	-64	0
Topic5(選舉)	431	-44	0
Topic6(立法)	1266	-80	0
Topic7(兩岸)	640	-67	0
Topic8(PTT)	1420	-386	0
Topic9(選舉)	634	0	0
Topic10(反核)	225	-535	0

觀察兩表格，在關鍵字與辭典對應詞數與加權後的分數做比較，可以發現大多數主題皆屬於正向，僅有主題 10 為負向情緒，可能因為資料範圍較廣，無

法確切表示出學運期間群眾應該呈現負向情緒的情形。(表格 52、表格 53)

2. 模型 B

表格 54 辭典與模型 B 主題關鍵字對應筆數

主題	正向	負向	中立
Default	7	3	1
Topic1(無)	13	2	5
Topic2(經濟)	10	4	2
Topic3(學運)	5	4	4
Topic4(立法)	10	7	3
Topic5(PTT)	9	3	4

表格 55 模型 B 主題加權後各主題分數

主題	正向	負向	中立
Default	244	-352	0
Topic1(無)	282	-1838	0
Topic2(經濟)	141	-226	0
Topic3(學運)	608	-283	0
Topic4(立法)	133	-190	0
Topic5(PTT)	930	-338	0

觀察模型 B 以上兩表格,在關鍵字與辭典對應詞數與加權後的分數做比較,可以發現大多數主題皆為負向，較符合本資料區段之情緒，而在主題 3 中討論議題為學運,卻為正向情緒,可能是因分類到該主題的辭彙與辭典較無對應到。(表格 54、表格 55)

3. 模型 C

表格 56 辭典與模型 C 主題關鍵字對應筆數

主題	正向	負向	中立
Default	3	2	2
Topic1(無)	11	6	13
Topic2(選舉)	8	4	5
Topic3(兩岸)	7	5	6
Topic4(柯文哲)	4	9	4
Topic5(弊案)	11	6	6
Topic6(學運)	7	10	3
Topic7(服貿)	13	4	5
Topic8(黑箱)	6	2	0
Topic9(教育)	5	1	4
Topic10(選舉)	4	1	6

表格 57 模型 C 主題加權後各主題分數

主題	正向	負向	中立
Default	28925	-4360	0
Topic1(無)	45126	-2505	0
Topic2(選舉)	30989	-9527	0
Topic3(兩岸)	6985	-7535	0
Topic4(柯文哲)	3501	-4604	0
Topic5(弊案)	7050	-3339	0
Topic6(學運)	7142	-4039	0
Topic7(服貿)	10575	-4171	0
Topic8(黑箱)	4777	0	0
Topic9(教育)	2238	-1650	0
Topic10(選舉)	2774	-3958	0

觀察模型C以上兩表格,在關鍵字與辭典對應詞數與加權後的分數做比較,可以發現多數主題為正向情緒，本模型資料是在刪文前取得之資料，雖對應到

詞數並無比其他模型多，但因文章量大，故加權後分數也較為龐大。(表格 56、表格 57)

4. 模型 D

表格 58 辭典與模型 A 主題關鍵字對應筆數

主題	正向	負向	中立
Default	8	1	1
Topic1(總統)	11	2	8
Topic2(選舉)	10	0	6
Topic3(民主)	8	5	6
Topic4(PTT)	6	2	7
Topic5(柯文哲)	5	4	2

表格 59 模型 D 主題加權後各主題分數

主題	正向	負向	中立
Default	1145	-233	0
Topic1(總統)	715	-586	0
Topic2(選舉)	633	0	0
Topic3(民主)	231	-167	0
Topic4(PTT)	177	-202	0
Topic5(柯文哲)	86	-47	0

觀察模型 C 以上兩表格，在關鍵字與辭典對應詞數與加權後的分數做比較，可以發現除了主題 4 以外，其餘主題皆為正向情緒。(表格 58、表格 59)

實驗後發現，各模型分群所討論主題多數為正向情緒，但此結果與我們的認知有差距，初步檢討應該是與情緒辭典的詞量不足，或政治議題的討論與一般性情緒辭典對正負向詞的定義不盡然等同之故。故本研究再進一步依支持與不支持太陽花學運的二分法，而非情緒之正負分類。我們針對模型 A 中各主題中的各關鍵字重新進行分類，所得之結果如下表格 60 所示，可以發現，大部分主題為支持學運。

表格 60 模型 A 各主題關鍵字分類後字數

主題	不支持學運字數	支持學運字數
Topic1	2	10
Topic2	0	5
Topic3	5	23
Topic4	1	18
Topic5	2	14
Topic6	2	23
Topic7	3	12
Topic8	2	3
Topic9	2	3
Topic10	0	5

第五章　結論

近年來越來越多的使用者在社群網站上發表言論，因此在社群網站中可能包含了許多重要的訊息，而也由於網路的普及，在網路上發言的人越來越多，文章量也就越來越大，因此如何在大量的資料中找尋出隱藏的主題就成了很重要的議題。本研究以 318 學運為例，透過收集台大電子布告欄中政治黑特版的文章，並利用 LDA 試圖找出文章背後所討論的議題，並透過情緒辭典，針對每個主題給予支持學運，或者不支持學運的判斷，來了解人們對於相關議題討論的情形為何。

而從實驗結果中可以得知，在資料蒐集區間內，若有事件發生，LDA 所能得到的結果就會較為集中針對某項議題，而若區間內無事件發生，LDA 所分群出之結果就會較為發散。而本研究中分為四個模組做分群雖是因為 PTT 機制所致，卻也發現因刪文而使資料量變少的 A 模型討論議題較為集中，B 模型、D 模型因為區間較小，議題也較為集中，C 模型則因為資料量相當龐大，使得結果較為發散。在情緒語意分析的方面，各模型一開始得到的結果偏向反對學運，而後續重新檢查關鍵字與辭典的對應程度，並針對每一個詞重新判定後，結果卻完全相反，大部分主題都是以支持學運為主。

而在未來研究中也有些部分可以改進，在情緒語意分析中，本研究雖使用了大部分研究都使用的兩份辭典，但初始得到的結果卻不盡理想，原因則是在辭典中與本議題相關的辭彙較少，而部分對應到的辭彙在辭典中為正向，但應用在不同議題中，同一個詞可能會是負向的意涵，導致結果的不同、判斷錯誤。本研究後續僅能針對分群後的辭彙進行支持或是反對議題來進行正負向的判斷，而未來若能改善本問題，對於後續的研究會有相當大的幫助。

參考文獻

一、英文文獻：

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), 'Latent Dirichlet Allocation', Journal of Machine Learning Research, Vol. 3, No. 1, pp. 993-1022.

Cimiano, P. (2006). Ontology Learning and Population from Text: Algorithms, Evaluation and Applications (1st ed.). Berlin: Springer-Verlag.

Chen，Don-yun ,(2010), An Experimental Research on UserBehaviors on Web2.0 SocialNetworking Sites and E-governance

Ellison, Nicole B. et al. (2007). "The Benefits of Face book Friends: SocialCapital and College Students Use of Online Social Network Sites."Journal of Computer-Mediated Communication 12, pp.1143-1168

Eaton, C., Deroos, D., Deutsch, T., Lapis, G., Zikopoulos, P. (2012). Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming data. New York: McGraw-Hill Press.

Jones,K. (1972), "A Statistical Interpretation of Term Specificity and Its

Application in Retrieval," Documentation Journal, Vol. 28, No. 1, pp. 11-20.

Kushin, Matthew James and Masahiro Yamamoto. (2010). "Did Social Media Really Matter? College Students' Use of Online Media and Political Decision Making in the 2008 Election." Mass Communication and Society Vol.13,No 5: pp.608-630

Luhn, (1957), "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," IBM Journal of Research and Development, Vol. 1, No. 4,pp. 309-317.

Macnamara, Jim , Gail Kenning. (2011). "E-Electioneering 2010: Trends in Social Media Use in Australian Political Communication." Media International Australia 139, 7-22.

Salton, and C., Buckley, (1988), “Term-weighting Approaches in Automatic Text Retrieval,” Information Processing & Management, Vol. 24, Issues 6, pp.512–523.

Sullivan, D. (2001). Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing and Sales (1st ed.). New York: John Wiley & Son, Inc.

二、中文文獻：

王泰俐（2013）。「臉書選舉」？ 2012 年台灣總統大選社群媒體對政治參與行為的影響, 東吳政治學報/2013/第三十一卷第一期

邢光愷（2012）。創新選舉戰略研究-以歐巴馬競選 2008 年美國總統大選為例。

呂建億(2015)。民眾對政府輿情分析方法之信任研究-民意調查與網路輿情分析的比較。政治大學公共行政研究所論文。

林冠宇(2014)，可擴展式基於文字分析之股票趨勢預測系統，國立臺灣大學資訊工程學研究所論文

吳政毅(2015)，採用混合式特徵詞於 LDA 新聞主題萃取，國立雲林科技大學資訊管理系論文

邱怡菁(2015)，以 LDA 為基之英文課程文字稿摘要法，國立屏東大學資訊管理學系碩士班論文

洪綾君、郭迺鋒、謝雨豆，2015，318 學運期間政治人物網路發言探究—大數據爬文技術之應用，2015 年「臺灣公共治理研究中心」暨「臺灣公共行政與公共事務系所聯合會」年度研討會，台北：台灣大學政治學系。

洪綾君、劉育津、郭迺鋒、謝雨豆，2016，社群網站關鍵意見與網民情緒變化之探索—以太陽花學運為例，國科會計畫結案報告。2016 年 5 月 31 日。

夏林清、游慧卿（1984）「工作價值觀問卷介紹與初步修訂報告」，測驗與輔導，第 60 期，30-36。

孫瑛澤、陳建良、劉峻杰、劉昭麟、蘇豐文(2010)。中文短句之情緒分類。計算語言學學會，國立暨南國際大學主辦，自然語言:第二十二屆自然語言與語音處理研討會，南投市。

徐詠絮（1996）「從文化帝國主義到媒介國際化的再思考：世界文化互動的理論比較」，台北：輔仁大學大眾傳播研究所碩士論文，32。

徐筱雁(2014)，情感分析中屬性詞與情感詞的關係之探討-以牛肉麵食評為例，國立聯合大學資訊管理學系碩士班論文

陳言熙(2006)，運用文字探勘技術協助建構公司治理本體知識，國立政治大學會計研究所論文

陳昱年(2013)，電影評論中情緒辭彙之極性分析，國立臺灣師範大學資訊工程學系論文

陳冠瑜(2015)，利用語意分析模型分析谷歌部落格搜尋引擎效能，國立東華大學資訊管理碩士學位學程論文

黃承龍、陳穆臻、王界人(2004)。支援向量機於信用評等之應用計量管理期刊，1， 頁 155~172。

黃亦筠（2012）。Big Data: 政府、企業的下一場戰爭。天下雜誌，495， 頁 50-56。

黃祖菁(2012)，運用多模式情感運算技術設計智慧型家教系統之人機介面，國立臺南大學數　　位學習科技學系碩士班論文

黃信華(2013)，以 facebook 塗鴉牆文本分析情緒文字的關係，國立臺南大學數位學習科技學系碩士班論文

張日威(2014)，應用 LDA 進行 PLURK 主題分類及使用者情緒分析，國立雲林科技大學資訊管理系論文

彭桂香(2014)，從臉書中文使用者之動態貼文預測其人格特質 ，國立清華大學通訊工程研究所論文

賴東河（1999）「世界廣告花費排行榜」，廣告雜誌，第 7 期，11 頁。

楊穎鈞(2014)，YouTube 點閱率是否影響實體唱片銷售，國立中山大學企業管理學系研究所論文

廖洲棚，國發會電子治理研究中心（2014）。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。

謝家蘭(2006)。智勝文化出版，淺介資料探勘與 XBRL，會計研究月刊，第 245 期：56-63。

蕭昱維(2014)，基於多階 LDA 技術尋找 Twitter 文章的隱含主題之研究

網路資料

維基百科－太陽花學運。
網址：
https://zh.wikipedia.org/wiki/%E5%A4%AA%E9%99%BD%E8%8A%B1%E5%AD%B8%E9%81%8B#.E5.8F.8D.E6.87.89
取得日期：2015/12/28

維基百科－阿拉伯之春
網址：
https://zh.wikipedia.org/wiki/%E9%98%BF%E6%8B%89%E4%BC%AF%E4%B9%8B%E6%98%A5
取得日期：2015/12/28

智庫百科-新媒體
網址：http://wiki.mbalib.com/zh-tw/%E6%96%B0%E5%AA%92%E4%BD%93
取得日期：2015/12/28

李慶堂（2014）。Text Mining 技術淺談
網址：http://www.cc.ntu.edu.tw/chinese/epaper/0031/20141220_3101.html
取得日期：2015/12/28

為甚麼不能忽略迅速成長的數位廣告市場（圖）
網址：
http://www.inboundjournals.com/rapidly-growing-digital-advertizing-market-infographic/
取得日期：2015/12/28

Advice Interactive Group
網址：
http://www.adweek.com/socialtimes/files/2013/01/social-media-help-desk.jpg?red=at
取得日期：2015/12/28

如何使用結巴分詞程式
網址：
http://blog.fukuball.com/ru-he-shi-yong-jieba-jie-ba-zhong-wen-fen-ci-cheng-shi/
取得日期：2016/01/05

董振東(1988)。知網
網址：http://www.keenage.com/html/c_index.htmL
取得日期：2016/01/05

NIELSEN 廣告信任度
http://www.adweek.com/socialtimes/files/2013/01/social-media-help-desk.jpg?red=at
取得日期：2016/01/05

國立交通大學統計學研究所-巨量資料帶來的契機與挑戰
http://www.stat.nctu.edu.tw/data/super_pages.php?ID=data1
上網日期:2016/04/26
MBA 百科－輿情
http://wiki.mbalib.com/zh-tw/%E8%88%86%E6%83%85
取得日期：2015/12/28

Internet World Stats 網路人口統計
http://www.internetworldstats.com/stats.htm
取得日期：2015/12/28