國立政治大學資訊科學系研究所

碩士學位論文

社群媒體新詞偵測系統

以PTT八卦版為例

Chinese new words detection from social media

A Thesis submitted to Department of Computer Science

指導教授:陳 恭 博士

研究生:王力弘 撰

中華民國一0四年六月

June 2015

摘要

近年來網路社群非常活躍,非常多的網民都以社群媒體來分享與討論時事。不僅於此，網路上的群聚力量已經漸漸從虛擬走向現實，社群媒體的傳播力已經可以與大眾傳媒比擬。像台大 PTT 的八卦版就是一個這樣具指標性的社群媒體，許多新聞或是事件都從此版開始討論，然後擴散至主流媒體。透過觀察，網路鄉民常常會以略帶灰諧的方式，發明新的詞彙去討論時事與人物，例如：割闌尾、祭止兀、婉君、貫老闆…等。這些新詞的出現，很可能代表一個新的熱門話題的正在醞釀中。但若以傳統的關鍵詞搜索，未必能找到這些含有此類新詞的討論文章。因此，本研究提出一個基於「滑動視窗(Sliding window)」的技巧來輔助中文斷詞，以利找出這些新詞，並進而透過這些新詞對來探詢社群媒體中的新興話題。我們以此技巧修改知名的 Jieba 斷詞工具，加上新詞偵測的機制，並以 PTT 的八卦版為監測對象，經過長期的的監測後，結果顯示我們的系統可以正確的找出絕大多數的新詞。此外，經過與主流媒體交叉比對，本系統發現的新詞與新話題的確有極高的相關性。

關鍵字：中文斷詞、新詞偵測、社群媒體分析

Abstract

In recent years, a very large portion of Internet users are used to share and discuss current events over social media. Indeed, as more and more people actively participate in the various virtual communities over the Internet, it is fair to say that the spread power of social media can be compared with that of mass media. The popular PTT gossip board is one such indicator of social media. In Taiwan, many news or events originated in this board would spread to the mainstream media and then become hot topics in the society.

We notice that many Internet users often invent new vocabulary to discuss current events and characters. The emergence of these new words may later grow into a new hot topic in the society. However, if we apply the traditional keyword search, we may not be able to find these articles with such new words. Therefore, this thesis present a "sliding window" technique to assist Chinese segmentation tool for facilitating the identification of these new words. Besides, these new words often represent a key indicator for new discussion topics. We use this technique to extend the famous Chinese segmentation tool, Jieba, with a new word detection mechanism, and apply it to the PTT gossip board. After a long-term monitoring, we obtain the results showing that our system can correctly identify the vast amount of new words in the board. In addition, after a cross comparison with the mainstream media, the new words identified by this system are indeed related to the popular social topics in a very high manner.

Keywords: Chinese Words Segmentation、New Words Detection、Social Media Data Analysis

謝辭

兩年的時間一轉眼就要過去了，很榮幸可以在政大資科和各位同學和老師一同成長，之前的求學路唸的都是商科相關，對於電腦的涉略比較少，在就業之後開始從事軟體相關工作，漸漸對於完成電腦相關碩士的學業產生了想法，很感謝我的女朋友，一路支持我報考，當我想放棄的時候也不斷得鼓勵我，還有努力支持我的家人，希望可以將完成學業的喜悅獻與在天上的父親。

非常感謝指導教授陳恭，每學期的上課內容非常的充實讓我獲得不少新的技術及知識，讓我可以學藝致用在我的工作領域上，初期在論文的題目尋找上並不順利的，非常感謝老師不斷的教導提醒提供了我不少資源，另外也感謝傳播學院宇君老師和百齡老師，對於我的研究提出關於傳播方面的指導跟建議使得論文方向可以越來越明朗。

還有我的研究所同學及戰友們珂齊、瑞程、宗佐，上課的時候我們總是互相加油幫助勉勵，我才可以順利的完成課程，祝大家也可以儘快地一起畢業。

最後再次感謝陳恭老師，您讓我從一個只會不明究理寫Code自學程式人到慢慢瞭解原理，使我可以更喜歡寫程式更有興趣去面對技術，這對於將來非常的有幫助，感謝老師的教導。

政治大學資訊科學研究所 王力弘 104年6月8日

目次

表次

圖次

第一章 緒論

1.1 研究背景與動機

現今社群媒體的使用者數量越來越多，根據 Facebook 2015 公佈的每月 Active user 數量已達 1393 billion，從美國總統歐巴馬使用 Facebook 的社群宣傳運動贏得勝選到阿拉伯之春革命事件的成功，社群網路媒體的影響力已經從虛擬開始慢慢擴展到了現實。近幾年在台灣也發生不少事件由網路傳播擴展至整體社會響應，從洪仲丘案到後來的太陽花學運及黑箱課綱，許多都是經由社群媒體發起而演變成實際作為，對這個社會造成改變。

智慧型隨身裝置的出現後，隨時都可以連上網路，社群媒體因此快速擴展至每一個人的生活中成為數位世界的新住民，這些新住民喜歡使用社群媒體來分享心情和時事，網路的普及達成了高度的去『中間化』，許多現實中原本微小不被人注意的事情，因為網路的力量讓它更容易受人關注、討論，甚至演變成實際作為，數位世界不再只是虛擬，它可以改變社會，改善現實環境。從以上我們可以發現，社群媒體對普羅大眾的影響力越來越重要，越來越多學者開始投入社群媒體方面的相關研究，因為中文詞的歧異性還不容易解決的問題，目前大部份都是針對英文語系的社群媒體研究為主，因而本研究想針對台灣中文社群媒體進行研究，從中提出一個解決中文詞歧異性的一個方案。

本研究選定台大PTT八卦版作為研究對象，它是一個台灣獨有的社群媒體，由台灣大學架設的 BBS 站台，因為其開放自由中立的特性，使得它成為台灣年輕人的指標性網站之一，站台討論版中其中又以八卦版最具代表性，除了它是 PTT 上文章量討論量最大的討論版外，許多社會事件(如洪仲丘案[17])因為八

卦版的影響力而受到社會所重視 。在八卦版中存在一個現象，那就是不少網路鄉民會創作一些新詞去調侃社會現象或是政治人物，如：「慣老闆」(指的是被台灣勞工市場過度傾向企業所產生的名詞，指被寵壞的經營者)、「柯神」「賴神」（柯文哲及賴清德）…[附錄 1]，所以本研究想嘗試製作一個基於 PTT 八卦版的社群媒體新詞偵測系統，並希望藉此驗證從新詞觀測的角度找出輿論的風向及正在發酵話題的可行性。

1.2 研究目的

本研究想透過社群媒體新詞偵測系統去了解社群媒體上面的現況，實驗的素材我們選擇的是一台大 PPT 八卦版，它是一個台灣獨有具有指標性的社群媒體平台，有很多新聞事件都是從此版延伸出來的，本研究假設新出現的暴紅新詞的背後代表的可能是一個即將或正在發生的高價值新聞事件，透過觀察這些新詞的出現時機、熱門程度、共現詞關係到每日爆紅話題偵測，希望透過這樣的輔助系統可以提供後續研究者，一個研究中文社群的參考案例。

1.3 研究貢獻

目前許多社群媒體論文研究都是針對國外的英文語系的資料居多，因為中文的斷詞歧異性依舊是一個不容易解決的問題，本研究想嘗試針對這個問題提出一個解決方案， 然而在中文的文本探勘的分析領域中，中文斷詞的精準度是一個困難待解決的問題，中文斷字並不像英文斷字那麼簡單，英文可以從「空白」及「符號」就可以完成一篇文章的斷詞，而中文文章都是一整句接續詞語，句子中由許多的詞組成，目前主流的斷詞方法是採用辭典法來做斷詞，但鑑於社群媒體的輿論常常會出現一些創作詞或是網路用語，一般的常用詞辭典可能

無法精準的偵測出這些詞，本研究提出一種 Sliding Windows(SW)的概念嘗試去修正斷詞結果，經實驗後證實新詞偵測的正確率可達 96%，透過新詞偵測的結果再反饋產生屬於該領域的專屬詞典，透過這樣機器學習的方式可以得到更精準地的斷詞結果讓新詞偵測率及話題分析的成果同步的提升。

1.4 論文章節架構

本研究分為六個章節，第一章說明本研究動機及背景、第二章對於中文斷詞及文本探勘技術的相關研究及文獻做探討，內容主要針對中文斷詞的相關文獻及字庫斷字的原理方法及已知問題做介紹，第三章針對 Jieba 本身的斷詞原理進行說明，透過了解此斷詞工具的一些優缺點後，再對本研究提出的 Sliding Windows 做詳細介紹，第四章在解說本社群媒體新詞偵測系統實作架構，第五章為系統評估、實驗設計結果與新詞檢驗及實際案例介紹，最後在第六章做出結論並對於未來研究方向提出一些看法。

第二章 相關技術及研究背景

中文的文本探勘領域中，斷詞問題一直是一個不容易解決的問題，中文斷詞的已知的兩個主要的問題是歧異性及未知詞的擷取錯誤的修正方法，本章節將針對相關研究文獻去探討，期望從中學習來找出一個解決方案。

2.1 歧異性與未知詞

在中文斷詞中「歧異性」(Ambiguity)是一個常常被提出來討論的問題，例如：「我是政治大學的學生」，透過斷字處理可以變成「我/是/政治大學/的/學生」，而「政治大學」其實也有可能被視作為「我/是/政治/大學/的/學生」，斷詞結果正確與否有時候會因地制宜，例如說同樣是中文語系國家的居民，如果是土生土長的台灣人，它將會因為本身的詞彙及社會經驗的關係下，致使他可以分辨的出「政治大學」應該是一個完整的詞，但如果是國外長大的華人，對於他們來說或許「政治/大學」才是他們所預期的正確結果，這種譬喻對於電腦來說也是如此，斷詞的結果與系統本身的收錄詞彙語料有很大的關係，因此在斷詞系統中的詞典的語料豐富性，往往會導致它的斷詞結果有所不同。

本研究中在「未知詞」(Unknown Words)的定義指的是辭典中未收錄的詞，它有可能是人名、地名、組織名或是及縮寫名稱、事件延伸字、個人特定族群慣用語、甚或是專有名詞，由於社會及科技思想不停的再改變，新的詞語不斷的被產生出來，而社群媒體的文章詞的變化的速度較一般的新聞媒體來的更快，更大大提升了斷詞的難度。

在 (Chen & Bai 1998)[1] 裡中，該篇論文提到未知詞有以下幾種種類：

(a) 縮寫(abbreviation)：例.中油,台汽...

這種縮寫詞通常不會出現在辭典裡，縮寫詞的未知詞在傳統的方法不容易被辨認出來，因為它的組成往往都沒有任何規則。有的時候甚至在口說交談的文章中縮寫詞使用頻率比完整詞還要高。

(b) 特定名稱(proper name)：例.馬英九,台北市,台積電...

特定名稱可以是人名、地名、組織名稱，有一些特定名稱可以透過一些特別的指示符號可以辨認出人名（例如：百家姓），地名的部分可以透過“鄉”、“市”、“縣”的結尾字來擷取發現，而組織名稱就較沒有一些規律性可以辨認出來。

(c) 衍生詞(derived word)：例.電腦化...

(d) 混合字(compounds)：例.獲允,搜尋法,電腦桌…

未知詞有很大部份是自於混合字，它可以由數個個別單詞字義（字音）的合併組合去描述某一件事情或是物體，再透過網路及口語上的傳播流行，混合詞中詞與詞結合產生新詞時大致沒有特定的規則，所以這類型的字詞屬於最難偵測的類型。

(e) 數值型混合詞(numeric type compounds)：例. 2015年,19巷,三千...

這類型的詞它會混合著數字單位，像是：物體量詞，測量單位，日期，電話號碼，地址..等，這類型的字較有規則性，屬於較容易辨認出來的類型。

2.1.1 未知詞的擷取

中文未知詞偵測是非常困難的一件事，因為它可以出現在文章裡的任何一個地方，也沒有任何分隔條件可以容易的辨認出來。在(Chen & Bai 1998)[1] 中作者提到，如果不透過“造句法”或是“語義”的規則前提下，很難去辨別一個詞它究竟是屬於某個未知詞的一部分還是它自己本身就是一個獨特的詞，再

實作上我們不可能將所有詞的混合統計結果都放置入辭典中，也無法透過使用一個簡單的規則去辨認出來。

在(Chen & Bai 1998)[1]中提到，該研究使用了中研院的文件集當作測試資料，其中有4632個新詞被找出來，而這些新詞的出現中約有4572個詞其實是因為發生斷詞錯誤的問題而產生，而這些被誤判的新詞有一些特徵：

一、 斷詞結果的詞較原本預期中的詞還要短。

二、 該詞之中包含著一個以上單音節的字。

Chen[1]假設，當一個詞典不存在的詞出現時在該詞只有單音節的狀況下，表示該詞很有可能是一個未知詞的元素詞。但這種預言式假設方法的偵測率並不高，接下來透過一些簡單的統計過程使用傳統的字典法的詞典法去做斷詞，Chen發現了69733個詞具有單音節詞，但只有9434個是未知詞的元素詞。

2.1.2 未知詞的偵測

在 (Chen , Ma)[2]的研究中，談論到上一篇研究[1]討論過的問題，中文詞的偵測基本上還是非常困難的。這篇論文採用的是基於詞出現的頻率的統計結果辭典再加上該文章出現的次數機率去做文章斷詞。

在該論文中[2]舉出了一個例子：

原句：張明正要殺人。

擷取結果：(1) 張明正要殺人 ，(2) 張明正要殺人

這兩種擷取結果都可以成立，但因為不同的斷詞結果，論述的對象名稱可以有兩種不一樣的名字，且所論述動作執行的時間也會大不相同，如果單純地由過去歷史詞統計的次數去進行斷詞，這是一種投機性的方式，而且這樣的方式有時候可能出現天差地別的下場，以下是一個例子：

原句：小明改變態度。

擷取結果：(1) 小明改變態度 ， (2)小明改變態度

如果該辭典中變態的過去統計數量大幅高於改變，就會導致斷詞結果錯誤，形成連鎖效應導致後續衍生的字詞出現更多錯誤。

[2]中提出一個非常值得參考的方法，就是除了比較候選詞本身的歷史統計頻率外還需要去觀察這些候選詞在這該篇文章中出現的次數，將出現的次數一律考慮進去，才有可能去計算預測出最佳擷取詞，本研究中就是啟蒙於此概念而發展出Sliding Windows的修正方法。

2.1.3 **N-gram 斷詞**

前面我們談到了許多詞典斷詞的方法，基本上詞典斷詞，詞的辨認即走尋的過程是採用N-gram的方式，N-gram會經由不同單元的組合及統計結果找尋出一個基於過去經驗得知的最佳解，著名的例子就是Google的搜尋引擎，它就是使用大量的N-gram演算法，去猜測使用者輸入的字彙以及其最有可能接續的詞語。N本身是一個變數，根據文獻參考當N-gram size = 1 稱之為「Unigram」，N-gram size = 2 時為「Bigram」，size = 3 時為「Trigram」。

Type	Unigram	Bigram	Trigram
Example	政治, 獻金, 案.. 阿帕契,貴婦,團..	政治-獻金.. 阿帕契-貴婦..	政治-獻金-案.. 阿帕契-貴婦-團..

表 2-1 N-gram 範例

在（陳鍾誠、許聞廉 1998)[8]中提到，未知詞的處理方法目前大致上可分做為兩種方法：構詞律及詞雙連(bigram)的統計方法。

構詞的斷詞方法在如果有明確詞首或是詞尾的詞彙上表現較好,詞雙連統計方法在強健性 (Robustness)上的表現的較好,但是對於不具明顯詞首或詞尾的較長詞彙而言,則這兩種方法都難以正確辨認。

而本研究中使用的斷詞工具－Jieba，它是透過詞雙連統計方法再搭配語料庫的統計結果來做辨識，因為在社群媒體中所產生的新詞彙可能不會有很明確的詞首或是詞尾，這裡採用Sliding Windows後可以做即時偵測未知詞再將該詞反饋到原本的系統。

2.2 文章特徵詞擷取

中文文本探勘領域中除了斷詞的主要問題外，另一個問題是如何取樣斷詞結果。本研究在研究初期有收集了一些新聞資料作為試驗，當對每篇文章去做斷詞，一篇文章大致可以分做數十到數百個字，如果只是單純的對所有文章產生的詞，去做統計會遇到三個主要的問題：

問題一：大部份找出來的字詞都是一些特定的主詞、連接詞…等，如：台灣、你、我、他、的、今天、昨天…等。

問題二：斷詞很難做到完全的精準，許多斷詞的結果會有side effect的問題，只要一個詞的分割錯誤它會導致更多的錯誤詞產生。

問題三：所有的文章斷完詞後如果產生的詞數過於龐大，對於分析上面會有效率的影響，且造成大量雜訊的問題。

根據(陳聰宜. 2012) [9]中建議對於每篇文章取特徵的方式是採取關鍵字擷取法，因為對於整篇文章做斷詞，產生的詞若是出現頻率較少的詞，它可能對於整體文章或是該社群的整體趨勢分析上並沒有太大的幫助，所以本研究打算只對每篇文章找出具特定數量的代表詞，從這些擷取結果去去偵測新詞或者是熱門詞來進行分析。

第三章 斷詞系統設計與實作

社群媒體新詞偵測系統最困難的部分是斷詞的問題，斷詞的精準度會影響到新詞偵測的結果，本章節將針對斷詞工具的選擇、斷詞的基本原理及遭遇到的問題以及本研究提出的 Sliding Window 的運作應用方式做一個全面性的介紹。

3.1 斷詞工具的選擇

目前在中文斷詞方面主流選擇的斷詞工具有兩種:一為中研院製作的「中文斷詞系統」(以下稱 CKIP)，另一個是由中國基於 Python 實作的開源斷詞程式庫－結巴(Jieba)，因為 CKIP 較於封閉在使用上較無完整 API 文件且線上斷詞回應速度上較不友善於開發，所以本研究選擇較於開放的 Jieba 作為斷詞系統核心程式。

3.2 Jieba 的斷詞模式

Jieba 是一個 Python 的 open source library,目前有不少論文也選擇採用 Jieba 作為斷詞的工具[5][6][7]，Jieba 支援三種斷詞的模式［引用自[14]]：

精確模式(cut)：試圖將句子找出最精確的斷詞結果，適合使用文本分析。

全模式(cut_all)：把句子中所有的可能構成詞的詞語都掃瞄出來，速度尚可但無法解決歧異詞的問題。

搜索引擎模式(cut_for_search)：在精準模式的基準下對長詞再次切分，提高召回率適合對於搜索引擎斷詞。

本研究採用的是精準模式(cut)，因為在文本分析中，關鍵詞的精準度還是

最主要的考量，因為關鍵詞的品質的重要性遠高於可以斷出的詞數量多寡，在使用全模式(cut_all)的方式下會較原本的精準模式還要多出 29%個詞，而過多的詞不一定可以提升文本分析的效果，很多的時候反而會導致效率以及雜訊的產生導致分析的困難。而搜尋引擎模式(cut_for_search)大致上與全模式相同，但搜尋模式為了提升搜尋關鍵詞的廣度，不只會將所有可能性詞都斷詞出來，還會對於所有分出來的詞在做詞的元素分割，這會導致斷出來的詞都是一些長詞組成的元素詞，會使文本分析時產生更多的歧異性，所以綜合以上的評估之後本研究選擇使用精準模式作為斷詞的模式。

【全模式】: 我/ 來到/ 北京/ 清華/ 清華大學/ 華大/ 大學
【精確模式】: 我/ 來到/ 北京/ 清華大學
【搜索引擎模式】: 小明/ 碩士/ 畢業/ 於/ 中國/ 科學/ 學院/ 科學院/ 中國科學院/ 計算/ 計算所/ 後/ 在/ 日本/ 京都/ 大學/ 日本京都大学/ 深造

表 3-1 Jieba 各種模式的斷詞結果演示

[引用自[14]]

Jieba 是運用基於詞典實現高效的詞圖掃描，將生成句子中的所有可能構成詞的情況構成的有向無環圖（ＤＧＡ）採用動態規劃查找最大概率路徑，找出基於詞頻的最有可能組合。

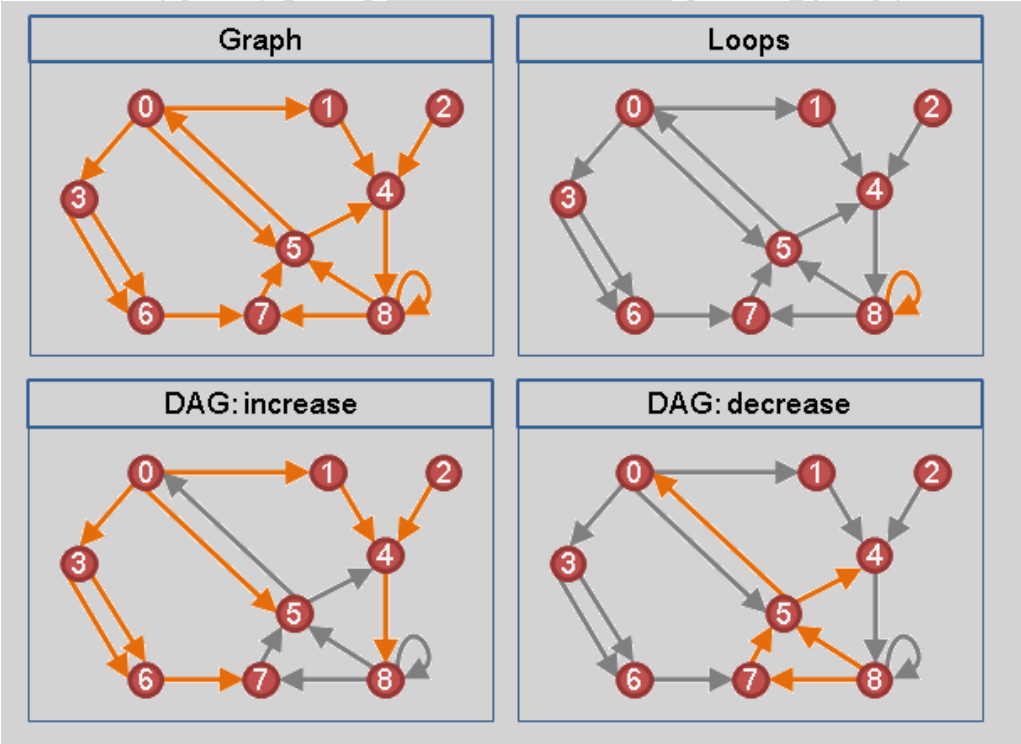
對於詞典未收錄詞，採用基於漢字成詞能力的隱性馬可夫模型（Hidden Markov Model），再加上維特比算法（Viterbi algorithm）找出最佳的斷詞結

果。

以下將會對於上述提到得一些名詞做一些介紹，一、有向無環圖（Directed Acyclic Graph，以下稱做 DGA）在介紹到這個圖形之前，必須要先提到資料結構常會看到的樹(Tree)，Tree 是一個無向無環的資料結構，而 DGA 以延伸擴展的觀念來看觀察，它就是一個具有大致上跟 Tree 相像的一個圖形，不同點在於 DGA 是一個有向的圖形，它跟 Tree 一樣沒有環，意思是路徑在走巡的過程中永遠不會回頭，只會不斷的向前進，它可以不斷的重新繪製，每個點有著先後次序的關係，透過不同的演算方式，像是索引由大至小，由小至大甚至是自動調整索引依據不同的行進準則，可以產生出許多不同的有向無環圖，每一個圖代表的是一種可能性，再透過類似像是維特比算法的方式，找尋出最佳的 DGA 圖作為最佳解的選擇。

圖 3-1 各種不同的 DGA 圖形

[引用自[10]]



二、隱性馬可夫模型（Hidden Markov Model）是一種統計模型，它用於描述一個隱含未知參數的一個模型，透過觀察可確定的參數與未知參數之間的模型關係，再利用這些參數去推測未知參數的結果。

以下是一個A案例，假設你有一個網友，他每天會在網路貼出他當天做的活動，這個網友每天對只會做三種活動：去公園散步、出門購物、清理房間。假設當天的天氣對於他選擇做什麼事情有很大的影響力，可是你無法直接得知他所在的天氣，但是你知道該地的前幾天的天氣趨勢，透過每天觀察他所做的活動基礎上，猜測該網友所在地當天的天氣概況。把天氣的運行想成是一個馬爾可夫鏈(Markov Train)，這個鏈裡有兩種狀態："雨"和"晴"，因為你無法直接觀察天氣，所以它們對於你來說是未知的參數，而你的朋友每天有一定的概率進行這三類活動:"散步"、"購物"、 或 "清理"。 而你可以透過朋友的回報得知他今天的活動，所以這些活動就是你的觀察數據，透過過去統計的數據概率關係，去推導今天他所在的地方的天氣，這就是一個隱馬爾可夫模型。

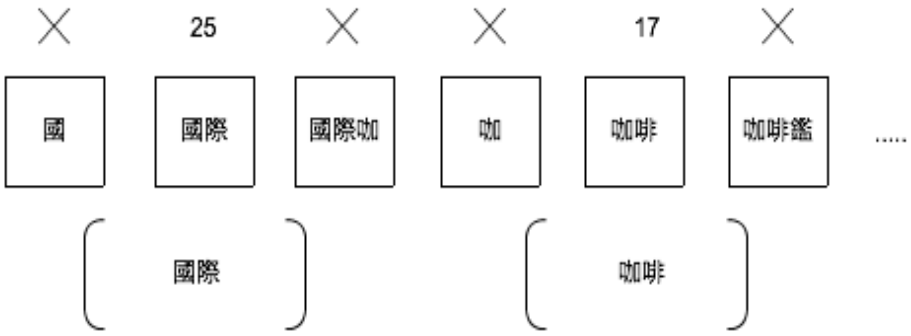
使用隱性馬可夫模型的情況是，當你想要預測一些事情可能發生的結果，但是你無法直接得知真正的結果，但你可以透過一些已知的相關參數再加上一些統計機率模型去推理猜測可能發展的發展路徑，而斷詞也是類似的一種情境問題，我們可以透過現有的詞典做基本的斷詞，對於未登錄詞的部分則可以使用維特比算法（Viterbi algorithm）來找出最佳的斷詞解答。

三、維特比算法由安德魯・維特比(Andrew Viterbi)於 1967 年提出，起初用於在數字通信線路中消除噪音之用，現今廣泛被應用語音辨識、關鍵字辨識、計算語言學和生物信息學中。維特比算法（Viterbi algorithm）是一種動態規劃演算法，常用於隱馬爾可夫模型中尋找出最有可能產生觀測事件序列的維特比路徑（隱含狀態序列）。Jieba 採用 N-gram 找尋前綴詞的方式再配合 HMM 模

型找出所有可能的結果再搭配詞典的詞頻分數作斷詞的演算，以下是一個案例來講述 Jieba 斷詞的過程：

原始句子：國際咖啡鑑定師打造專屬烘焙法

圖 3-2 Jieba N-gram 找尋斷詞過程



Jieba在演算斷詞的過程中會將詞作初步分段（Segment），利用文章中的換行符號、標點符號、中止詞(Stop Word)，將文章分成一個一個的片段，再依序對每一個片段由左至右使用N+1的方式去合併字元，由(圖3-2)示意圖，這是一個斷詞的演示過程，從文章的首字開始逐一由左至右的逐字移動去找尋字詞，將當前的字組查詢辭典（表3-2）是否存在符合的字詞以及該詞的統計分數，將所有可能性都全部列舉出來，再經由計算找出最符合的斷詞結果。

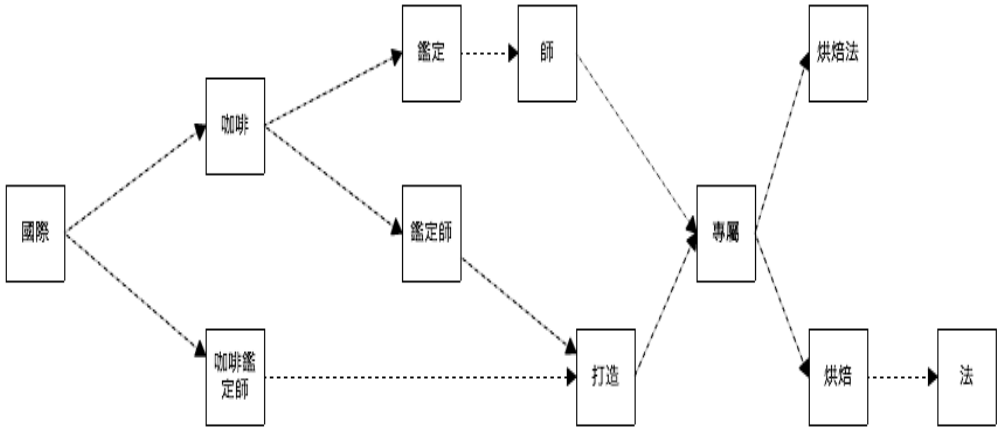
字詞	分數
國際	25
國際咖啡	8
咖啡	17
鑑定	15
鑑定師	5
咖啡鑑定師	3
打造	13
專屬	11
烘焙	7
烘焙法	9

表 3-2 範例辭典

斷詞組合
國際/咖啡/鑑定/師/打造/專屬/烘焙法
國際咖啡/鑑定/師/打造/專屬/烘焙/法
國際/咖啡鑑定師/打造/專屬/烘焙法
國際/咖啡/鑑定師/打造/專屬/烘焙法

表 3-3 斷詞候選詞組

圖 3-3 基於斷詞結果產生的 DAG 圖



（圖 3-3）是透過以上四組的候選斷詞詞組去產生出有向無環圖（DAG），由(表 3-2)的字典分數使用下述的計算式的計算結果，如果使用詞頻分數計算法斷詞我們將會得到「國際/咖啡/鑑定師/打造/專屬/烘焙法」為最佳斷詞結果。

從這些的計算過程，我們可以了解到，詞典收錄詞是否豐富的重要性，如果該詞典對於對象文章的詞大多都未收錄，它的斷詞結果會產生有許多超長詞，無法精確的將詞的元素分割出來，所以採用的詞典是否具有收錄該文章的特殊領域詞將會對於斷詞的結果有很大的影響，另一個問題是如果使用其他領域產生出來的辭典，它可能因為這些領域的專屬詞且這些詞的分數過於偏袒的話，甚至會導致斷詞的結果跟實際期望的結果有很大的歧異性，所以我們可以得到一個結論，不同領域的文章想要有精準的斷詞結果，就必須具有專屬的領域詞典。

3.3 斷詞的問題

使用 Jieba 的預設精準模式斷詞模式去斷詞，文章字數不同時一篇文章會產生數十至數百個詞，經過計算的高頻率詞大致上都是一些平常撰寫文章常見的主詞或是連結詞，例如：台灣、你、我、他、的、表示…等，這些詞在文本分析上面並不具有太大的意義，它並不是我們所關注的詞，所以如何選擇一個有效的詞統計模型，會是個重要的課題。

在資料分析中，資料特徵（Feature）選定是往往會面臨到的問題，(陳聰宜, 2012][9]) 中提到，並不是文件中的所有字詞都能直接代表該文章內容的主題，且每篇文章關鍵詞不一定相同，要選出真正具有代表的詞，必須對每個詞相對於該篇文章內的出現頻率作為權重計算，藉此找出對該篇文章的代表詞，所以本研究將採用 TF-IDF(Term Frequency Inverse Document Frequency)的統計模型，對每篇文章做 TF(Term Frequency)關鍵詞的擷取，再使用 IDF（Inverse Document）對該詞出現文章做統計，這樣的統計模型將會更有利於我們去找尋新詞。

在上一節我們討論到詞典的領域性問題可能會導致斷詞結果造成歧異性，而本研究使用的 Jieba 是經由中國撰寫出來的開源軟體，雖然它有提供內建的統計字庫，但它的統計的素材的來源來自於：一、1998 年中國人民日報的語料庫、二、MSR 機構提供的語料庫和開發者手邊的一些小說資料，所以許多的字詞還是以中國地區的用語為主，例如：「A 輪」、「QQ 號」、「D 盤」…等，由於字典的語料蒐集大致上是以中國地區為主，對於我們台灣的常用語、組織名、知名人事物這些是在它的詞典裡沒有收錄的，這會導致斷詞的結果會與預期不符，例如（表 3-4）的案例，因為「大寶」在中國算是一個菜市場名，因此在詞庫

的詞頻分數高達 92，而導致以下的文章在斷詞結果得到的詞是“大寶”而非我們所認知的知名證券公司 —“元大寶來”。

ETF 新兵登場 元大寶來推 ETF 傘型證券投資信託基金
IB 開放國外期貨 元大寶來證券、期貨協助全台推廣
OSU 進一步開放 元大寶來證：好事一樁
《台北股市》元大寶來投信：台股中多不變，長線首選 6 題材

表 3-4 錯誤詞偵測範例

從這個現象我們可以得知，當使用的素材的地區領域不同，產生出來的詞典特性可能也不盡相同，當某些領域的語料素材過多時，它會因此導致該詞的統計分數特別的高，致使包含這個元素詞的所有句子都斷出錯誤的詞，就像是（表 3-4）的例子，所有的文章都斷詞出了“大寶”而非“元大寶大”，雖然在事後觀察關鍵詞分群結果文章還是會被正確的分布在同一群，但如果某些文章如果真的在談論“大寶”的話，這對於後面研究分析會產生雜訊，若能夠自動修正錯誤的斷詞結果的話，便可以大大降低這類型的錯誤。

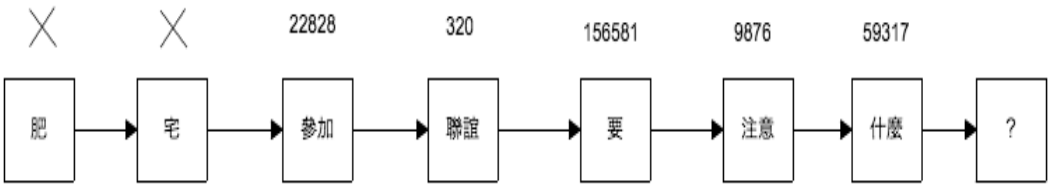
3.4 維特比算法新詞偵測模式與其缺點

從上一節，談到了關於斷詞錯誤會導致出來的問題，而本研究想基於社群媒體的文章中找出一些網友的創作詞，那這樣的詞自然不會存在我們的詞典中，所以如何去發掘未知詞這也是本研究的研究目標。

Jieba 使用的是維特比算法（Viterbi algorithm）來偵測找尋新詞，以下是一個例子「肥宅參加聯誼要注意什麼」，透過這個例子來解釋維特比算法是如何找尋新詞。首先 Jieba 會以 N-gram（N+1）的方式移動，以組成的詞去查詢

詞典，若預期正確的斷詞是「肥宅/參加/聯誼/要/注意/什麼」，但如果 Jieba 的預設詞典裡面並沒有肥宅這個詞，那麼 Jieba 在做斷詞時，它會先將所有已知的詞先切割出來，如：「參加」、「聯誼」、「要」、「注意」、「什麼」這些詞在預設字典裡都可以查詢到它們，而肥宅這個詞並不存在在詞典中，所以 Jieba 會將這個詞當作是一個新詞，這裏還有另外一個現象，因為 Jieba 預設是採用 N+1 的方式往前找尋前綴詞，所以在 Jieba 未知詞會是二個字元以上，但是這有一些例外。

圖 3-4 在 Jieba 預設的詞典中，基於維特比算法找尋新詞



在（圖 3-4）的例子中，我們可以發現“要“這個單一字元詞，這是因為當前接詞跟後接續詞都被 Jieba 分別辨識斷詞元素，所以”要“這個詞就被前後詞推擠出來變成一個詞，這也是維特比算法（Viterbi algorithm）發現新詞的一種方法，但是這種找尋新詞的方式僅限於一般的正規文章，它並不適合像是社群媒體的文章，例如以下兩個例子：案例A「請問有宇宙大覺者的八卦嗎？」，從理解社群的評論內容後人類可以很清楚的發現宇宙大覺者是一個新詞，但如果以 Jieba 預設詞典做斷詞，我們會得到以下斷詞結果「請問 / 有 / 宇宙 / 大 / 覺者 / 的 / 八卦 / 嗎 / ?」。

案例B：「鬼島紀實 - 全世界只有台灣可以生產高素質的輪班星人」，在使用 Jieba 預設詞典進行斷詞，我們會得到以下這個斷詞結果「鬼島 / 紀實 /

－ / 全世界 / 只有 / 台灣 / 可以 / 生產 / 高素質 / 輪班 / 星人」，在網友的創作詞中「輪班星人」指的是在科學園區輪班的技術員，這邊就產生了詞的歧異性，但是如果就了解文章的使用者來說「輪班星人」才是一個完整的詞，這個問題可以透過我們提出的方法來獲得改善，接下來的章節我將對於 Sliding Windows 來進行說明。

3.5 錯誤詞的修正及新詞偵測

從前面所提及的本研究中所遭遇到兩個問題「錯誤詞修正」、「新詞在社群媒體的偵測率」，本研究提出一種 Sliding Windows(以下稱 SW)來解決錯誤詞修正及新詞偵測率提升的問題。之前本研究提到的一個案例(表 3-4)，因為 Jieba 預設詞典的詞庫並沒有收錄「元大寶來」這個組織名稱，而在中國地區「大寶」是一個高出現頻率的詞，導致我們的斷詞結果取出了「大寶」而不是「元大寶來」，在事後的分析，我們發現我們可以針對文章內容的該詞所出現的樣式(Pattern)去偵測及修正錯誤詞，從下（圖 3-5）我們可以發現元大寶來的一些樣式特性。

圖 3-5 樣式方法說明範例文章

元大寶來　搶搭滬港通　獲利破百億

相較於去年全台券商整體獲利成長五三%，元大寶來創下成長一九○ %的佳績。深耕香港、今年在韓國更提早獲利，元大寶來要憑藉滬港通商機，打出自己的國際佈局戰。

二○ 一四年，證券業獲利龍頭元大寶來大豐收。營收大幅增加近六成，稅後純益一○ 三．九億，成長一九○ %，首度突破百億。

獲利成長近兩倍，到底有何祕訣？

元大寶來證券總經理賴宗武坦白地說，沒什麼祕密，元大寶來經紀業務市佔率一四%，只因台股好，成交量上來。

不過相較於去年全台券商整體獲利成長五三%，元大寶來的成本控制，與新市場爆發力不可小覷。

掌握客戶動態　控管成本

賴宗武指出，台股本質已發生劇變。過去台股上萬點時代湧入的投資人如今都老了，或將資金移往海外。年紀愈長，投資愈趨保守，殺進殺出買賣頻率降低。

從（圖 3-5）我們可以發現在目標詞(Target)的前詞（字）或是後詞（字）都是不同的樣式，例如：第一句「元大寶來創下」，接著的「元大寶來要憑藉」、「龍頭元大寶來大豐收」我們可以發現其中它接續或是前綴詞都是不同的接續字詞，這是文章作者在寫作文章的時候的主題內容標示特性，當一個詞如果它

在文章中扮演要角的時候，那麼它在一篇文章中出現的次數應該會有數次以上，透過這樣的前綴詞及接續詞的樣式比對，發展出一個修正的演算方法，接下來將會詳細地加以說明。

3.6 SW 修正法

這個一章節是本研究的主軸介紹，Sliding Windows 修正法（以下我們將以 SW 稱之），大致的運作過程為：首先是詞的樣式偵測及比對，接著如果樣式比對發現有共有的樣式就會將這個詞當作是一個新詞偵測出來，新詞的偵測出來之後會將這個詞放置入即時線上及批次線下詞典，使得這個修正方案可以擴散，讓無法使用 SW 的文章也有機會可以修正為正確的詞。

在詳細介紹修正過程的時候，我們想先說明 SW 的限制及使用情境：

本修正法建議使用在關鍵詞擷取的使用情境下（例如 TF-IDF 字詞擷取），因為如果在整篇斷詞的時候使用 SW 修正法可能會導致一些副作用（Side Effects），因為 SW 再找到類似的樣式的時候會將詞合併，某些時候可能會造成維特比演算法錯置的問題。

本研究 SW 的設計希望可以嚴謹一點不要因為 SW 的合併產生更多的歧異詞，所以本研究設定 SW 的合併條件須為同一個詞至少要有兩種以上的樣式且偵測之所有樣式都要相同才會執行合併，若其中有一個樣式與其他樣式不同 SW 新詞擷取合併即會停止。（例如：有一篇文章中偵測出「魏應」這個詞，該篇文章也發現兩次以上的樣式，但其中穿插出現「魏應充」、「魏應行」，儘管以上兩個詞出現的次數也超過兩次以上，但因為本研究的樣式判斷設定為一定要全部一樣，也就是所有樣式都必須含有同樣的組成字，才會執行 SW 合併，若有一個詞不符即會視為無新詞發現而停止合併，但這類似型的問題可以透過 SW 反饋線上詞典

有機會得到改善)。

每次 SW 移動順序為一次為限，若需要多次移動才能合併完全詞，現有機制會將這個動作交由下一個文章再次執行合併工作時實行，本研究的 SW 具有線上及線下詞典功能，下一篇文章可以直接接續上次的偵測結果前提下向下偵測。（例如：我們發現一篇文章的預設詞擷取出「宇宙」，在該次 SW 我們發現有「大」這個相同樣式，我們將會將「宇宙大」認為是一個新詞，下一次我們在遇到類似話題的文章我們會直接偵測出「宇宙大」而本次我們在 SW 的時候發現有「覺者」這個樣式可以合併，在這本次我們將會把「宇宙大覺者」這一個時事新詞偵測出來。

SW 修正法為新詞偵測，但本研究依然建議保留原本的斷詞結果，也就是原本預設詞典偵測出來的詞（例如：大寶），如果 SW 修正條件符合的情況下另外增加一組詞（例如：元大寶來），所以一篇文章將會有 N+1 個詞成為它的文章代表詞組。以上的限制跟使用情境主要是根據本研究的研究素材的原因而設計的 SW 修正法，若以後的研究後進有相關需求可以自我調整以上的條件。

3.6.1 **Sliding Windows 的運作過程**

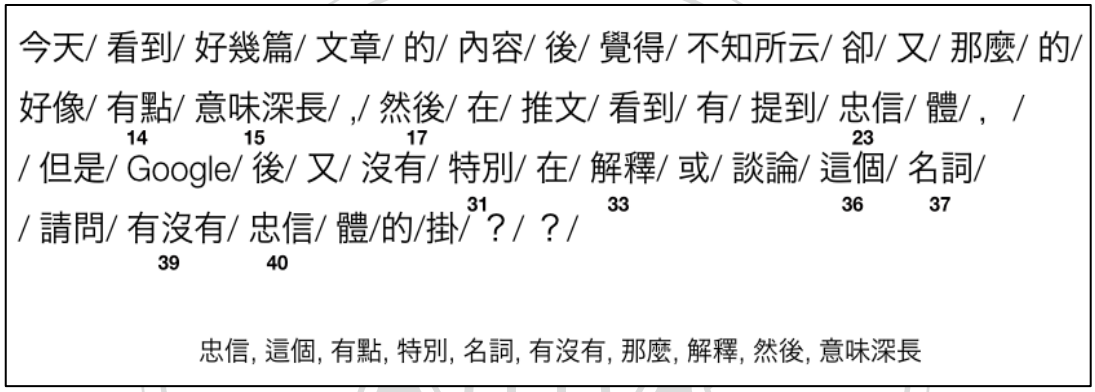
圖 3-6 SW 說明範例文

> [問卦] 有沒有忠信體的八卦？
>
> 今天看到好幾篇文章的內容後覺得不知所云卻又那麼的好像有點意味深長,
>
> 然後在推文看到有提到忠信體，
>
> 但是Google後又沒有特別在解釋或談論這個名詞
>
> 請問有沒有忠信體的掛？？

接下來我將使用一個實際案例解說SW,(圖 3-6)是一篇PTT八卦版的文章，本篇文章從人類閱讀的習慣下，人的角度會學習到「忠信體」是一個詞，且知道這是一個新詞，究竟人類是如何辨認出某個區塊它屬於一個詞?首先我們會先判斷一些已知詞，例如本文標題「有沒有忠信體的八卦」,「有沒有……的八卦」這樣一個句子是 PTT 八卦版常會出現的一個標題型態，就算閱讀者對於該討論版並不熟悉，但是一般來說「有沒有」、「的」、「八卦」是我們一開始很容易辨識出來的詞,如果我們嘗試將一些字元從這個句子之中拿掉讓這個標題變成「有忠信體的八卦？」,接下來辨識可能就沒有這麼順利，但是閱讀者是如何在閱讀完之後還是可以知道「忠信體」其實是一個詞？一般人會透過文章的樣式去評斷，例如「提到忠信體，」、「有沒有忠信體的」,我們可以發現相同的樣式出現以及其對應應該要辨認的詞,也就是「忠信體」,在一般來說我們直接使用 Jieba 的精準模式對本文分析去找出 TOP10 的詞，我們會得到:「忠信」、「這個」、「有點」、「特別」、「名詞」、「有沒有」、「那麼」、「解釋」、「然後」、「意味深長」。這

十個詞除了「忠信」之外的詞都只有出現一次，在 Jieba 在建立 TF Rank 的時候當有一個以上同頻率的詞，它的排序是程式內部在實作 HashDict 的底層排序自動決定的，這裏除了「忠信」之外其他的詞其實是隨機取出的，意義不大，而這裡得結果沒有出現「體」這個結果，是因為 Jieba 預設抓取的詞最小單位以一般的狀況下會以兩個字組成的詞為主，如果該詞只有一個字，就會被過濾掉。

圖 3-7 斷詞索引示意圖 I



當 Jieba 在算詞頻之前會先做一次全文精準斷詞，會將所有的詞邊上索引（Index），針對 TOP10 的詞，當我們加入 SW 的方法，我們會逐詞去計算每一個詞的 Index 所在位置（如圖 3-7），SW 會對所有的 TOP10 詞逐一檢查，像是「有點」「意味深長」…詞都只有一次的出現次數，這並不符合 SW 至少兩次樣式的規則，所以 SW 會直接跳過以上這些詞，而「忠信」這個詞具有兩個以上的樣式前提，所以 SW 機制會嘗試偵測比對是否有修正合併詞的必要。

圖 3-8 斷詞索引示意圖 II

今天/ 看到/ 好幾篇/ 文章/ 的/ 內容/ 後/ 覺得/ 不知所云/ 卻/ 又/ 那麼/ 的/
好像/ 有點/ 意味深長/ ,/ 然後/ 在/ 推文/ 看到/ 有/ 提到/ 忠信/ 體/ ， /
/ 但是/ Google/ 後/ 又/ 沒有/ 特別/ 在/ 解釋/ 或/ 談論/ 這個/ 名詞/
/ 請問/ 有沒有/ 忠信/ 體/的/掛/ ？/ ？/

-1 23 +1

-1 40 +1

SW 會針對「忠信」所在的 Index-23 及 Index-40 去向前向後找尋前綴詞以及後續詞，在範例(圖 3-8)我們可以發現前綴詞的樣式是「提到」及「有沒有」，後續詞地樣式皆是「體」，可以 SW 會將「忠信」合併「體」而成為「忠信體」。

圖 3-9 SW 修正新詞演算法：

$\langle words.article_{count} > 2 \rangle$

$$A := \sum_{w_1}^{w_{10}} [j..k] := words_{n-1} + words_n + words_{n+1}$$

$$B := \sum_{w_1}^{w_{10}} [j..k] := words_{n-1} + words_n$$

$$C := \sum_{w_1}^{w_{10}} [j..k] := words_n + words_{n+1}$$

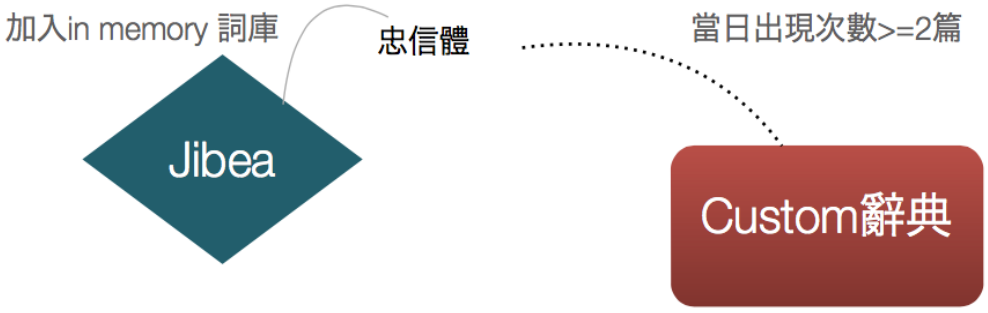
$$\begin{cases} A = true & words_{n-1} + words_n + words_{n+1} \\ B = true & words_{n-1} + words_n \\ C = true & words_n + words_{n+1} \\ else & words_n \end{cases}$$

透過維特比演算法以及 SW 的修正結果的幫助下，斷詞系統可以從中發現新詞，但是在某一些狀況下 SW 是無法使用的，例如如果「忠信體」這個詞出現在文章中，而這個詞只出現一次，我們就無法透過樣式去把這個新詞辨認出來，所以本系統又實作了反饋的方式，將先前文章所辨認出來的詞加入詞典，使得下一篇文章即使沒有特殊的樣式依然可以依靠詞典去辨識這個未知詞。

3.6.2 新詞的反饋模式

在前一節最後提到為了將某些文章透過 SW 的修正找到的未知詞的能力延續下去，本研究會將這些被偵測出來的未知詞加入詞典中，接下來本研究將會針對這個反饋模式進行說明。

圖 3-10 新詞反饋說明



這裏我們以上一節的例子「忠信體」為例，當 SW 辨識出來這個新詞之後，我們會將這個詞加入加入線上詞典，Jieba 在每次啟動的時候會先將所有的詞典檔讀取至記憶體中，在 Jieba 的 API 中它提供了 jieba.add_word()這個方法可以讓使用者編寫程序時將一些詞自動加入記憶體的詞典中，對於下一篇文章斷詞時可以加入這個新偵測的詞作為斷詞的依據，本系統除了實作線上反饋之外，也實作了線下反饋，也就是加入使用者自訂的詞典檔，這裏我們設定的規則是這一個新偵測的未知詞，必須至少出現在某兩篇文章的 TOP-N 中(如 圖 3-10)，

因為有時候偵測出來的未知詞可能只是某一些特定人士的口頭禪，它不一定是一般人所認識的詞，這裏可能會產生詞的歧異性，所以我們採取至少要有兩篇以上提到這個詞，才會將該詞加入使用者自訂辭典檔中，本系統也會每日蒐集結果去更新每個詞的詞頻分數，藉由以上這些做法讓電腦自己去學習去產生符合該系列文章的獨有辭典，以達到不斷提升修正斷詞的精準度的目的。

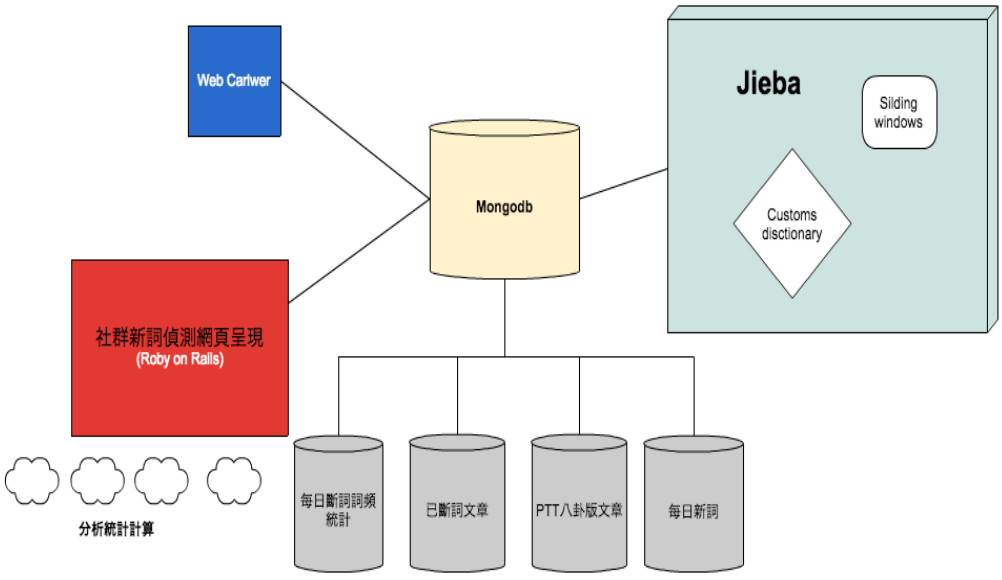
第四章 新詞偵測-系統分析與實作

PTT 八卦版的社群媒體文章為本研究的分析對象，希望透過線上討論分享的內容去找尋新詞、熱門詞，再進一步彙整結果嘗試去偵測新話題以及熱門話題。為達成此目的，本系統需要實作資料蒐集程式去定期擷取 PTT 八卦版的文章，強化版實作 SW 及詞反饋機制的 Jieba 分析工具，後端資料庫儲存分析資料以及前端網頁呈現系統。

4.1 系統設計架構

本研究使用 Ruby 作為主要框架實作語言去撰寫 Web Crawler 和前端網頁結果呈現（Ruby on Rails），透過 Python 在 Jieba 斷詞工具中實作 SW 及詞回饋機制，後端資料庫使用的是 Mongodb。

圖 4-1 社群媒體新詞偵測系統架構一覽



4.1.1 資料蒐集程式

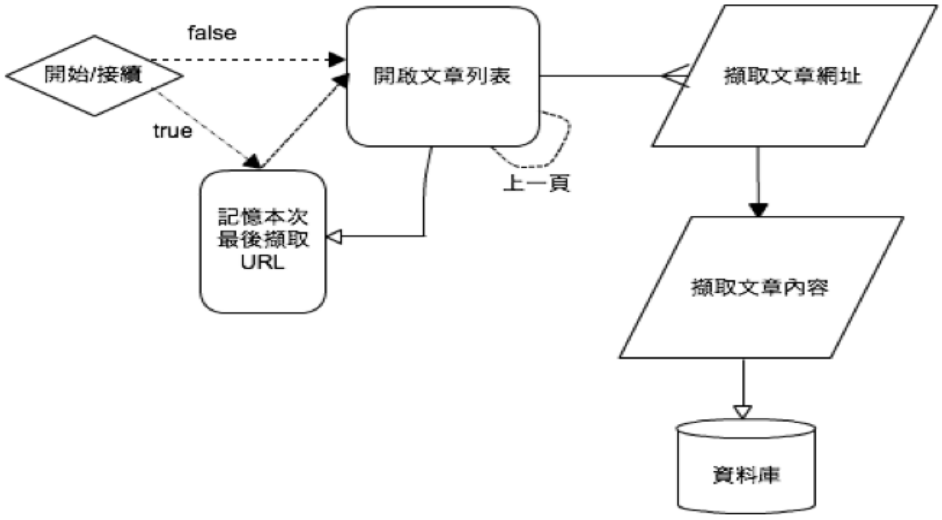
資料分析的過程中首要的任務是搜集資料，最常使用的手段就是撰寫 Web Crawler去搜集資料.本研究使用Ruby Nokogiri來擷取PTT八卦版的討論文章。Nokogiri 是一個 Ruby 上的一個 HTML、 XML、 SAX 的 parser library，藉由 XPath 或是 CSS3 selectors 經由 Tag/Class/ID 來尋找 XML/HTML 內元素(element)再透過正規表達式（Regular Expression）過濾擷取出網頁內容。

圖 4-2PTT 八卦版畫面

PTT 原本是一個 BBS(Bulletin Board System)站台，後來因為很多人有透過 Web 及手機 App 瀏覽的需求而發展了網頁版的平台讓使用者可以不用透過 Telnet 協定就可以瀏覽文章,本資料擷取城市主要透過網頁版去擷取文章資料，但是 PTT 網頁版平台介面還是依循原本 BBS 站的設計,每次開啟的首頁就是最新的文章列表，文章的瀏覽主要透過上一頁及下一頁來切換移動，這不同於一般的網頁設計，社群媒體的特性是隨時隨地都會有人發表文章，這裡本系統設計每隔一小時會執行 Crawler 去擷取新的文章然後存放至資料庫中，但是 PTT 網頁版的界面是不會顯示當前頁數及文章編號，每次執行的時候程式需要記憶

上次執行的位置便於下次可以接續上次頁面繼續擷取文章。

圖 4-3 Web crawler 執行過程



如(圖 4-3)所示，每次 Web crawler 執行的時候會先開啟(最新/最後訪問)的文章列表 URL，從文章列表中取得每一篇文章的連結之後逐篇訪問文章 URL 去擷取文章內容，並將擷取的內容存放至資料庫中，每次執行後終止點（文章列表的最後一頁)，將此最後一頁的網址記錄下來，在下次啟動的時候可以接續上次執行的位置繼續擷取文章資料。

4.1.2 後端資料庫

本研究使用 MongoDB 作為資料儲存的資料庫，它是目前最流行的 NoSQL 資料庫之一。NoSQL 一詞最早出現在 1998 年，訴求是開發一個輕量、開源、不提供 SQL 功能的資料庫。後來到了 2009 年，當時發起了一次關於分散式開源資料庫的討論，此時再次提出了 NoSQL 的概念，這時的 NoSQL 主要指的是非關係型、分布式、不提供 ACID 的資料庫設計模式，而至今的 NoSQL 資料庫大部分還是依舊具有 ACID 的設計模式概念 [10]。

ACID [引用自 10]:

1. Atomicity（原子性/不可分割性）：

一個事務（Transaction）要完成所有的動作，只要中間一個環節失敗了就立即還原（Rollback）到該事務開始之前。

2. Consistency（一致性）：

寫入資料必須符合所有原先設定的預設原則。

3. Isolation（隔離性）：

當數個事務同時被查詢或是修改，資料同一數據表示出的相互關係。

4. Durability（持久性）：

當事務完成之後，該事務修改後的結果會持久且完全地保留在資料庫中。

NoSQL 具有 Free-Schema 以及在分散式叢集上良好的執行特性，許多資料分析的平台都開始採用 NoSQL 作為儲存媒體，而 MongoDB 是一個 Documents Base Storage 的資料庫，Documents 指的是一種 JSON (JavaScript Object Notation)-Like 的 Key-Value pairs 的儲存格式，讓使用者可以就像直接透過 Key 查詢資料，就像查詢程式內部的結構化資料一樣，因為社群文章大多都是

Documents base 的文章，所以非常適合採用 MongoDB 作為儲存媒介,所以本系統採用 MongoDB 作為資料儲存的媒介。

在 MongoDB 中一個 Collections 代表的是一組相關的 Documents 並可能 Shared 共通的索引鍵(Indexed Keys)，Collections 的概念就類似於 RMDB 裡的 Table。在 MongoDB 存入 Document Records 時，使用者必須先選定要存入的 DATABASE，之後指定 Data Collection 使用 Insert Operation 將資料寫入。

圖 4-4 Insert Records into MongoDB

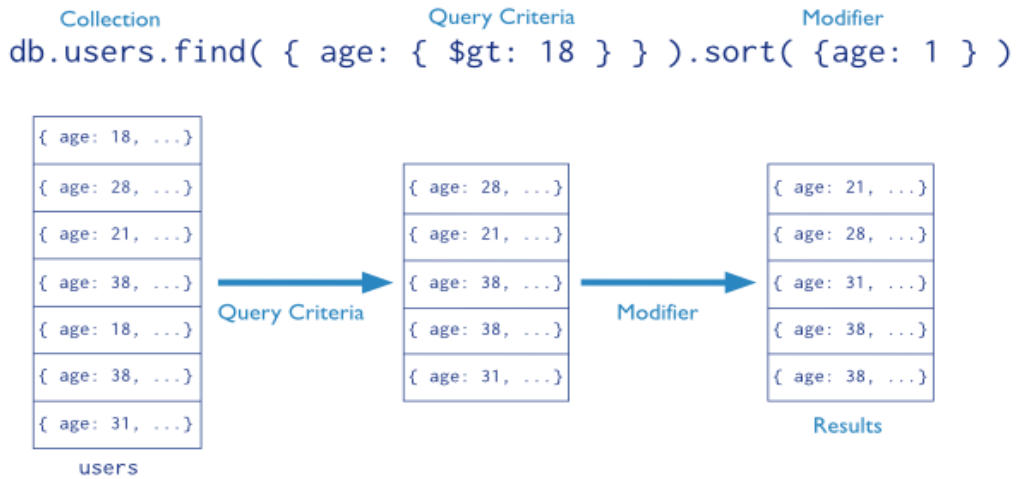
```
#set up database
use pttdb

#insert document into gossips collecation
db.gossips.insert(
  { "body" : [ "日本", "合作", "..", "東南亞", "越南", "支那", "南海", "台灣", "侵略", "軍事合作" ],
    "src" : "119.14.24.8", "author" : "bigbear2007",
    "url" : "https://www.ptt.cc/bbs/Gossiping/M.1419868249.A.EDA.html",
    "title" : "Re: [問卦] 為何亞洲不成立大日國協",
    "tdate" : 1419868245,
    "author_nick_name" : "哇最愛速完"
  }
)

```

使用者可以透過 Find Operation，指定 Query Criteria 查詢 Records(如下圖 4-5)，這樣的資料存放及查詢方式非常適合作為資料分析系統作為後台資料庫使用。

圖 4-5 在 MongoDB 查詢資料



4.2 分析平台查詢及排程運算

本研究實作的網頁框架使用的是 Ruby on Rails 的方式實作，它是一個 Ruby 程式語言的 Web Framework，使用者可以透過 MVC 的方式快速建構一個網站，使用者可以透過網頁的操作快速瞭解本平台分析的結果。

在上一節提到在資料儲存的部分提到資料庫使用的是 MongoDB，我們這裡會使用 Mongoid 與 Ruby on Rails 的後台做整合。透過 Mongoid 我們可以直接對 Rails Application 設定 Ruby 在撰寫 Controller 資料處理及查詢可以使用讀取使用的欄位以及 Query Methods，使得撰寫查詢分析條件的時候更加容易方便，因為 MongoDB 本身是可動態調整欄位(Schemaless)的特性，所以此設定檔可以隨時更改增加或是修改欄位或是 Customs query methods(如圖 4-6 所示)。

圖 4-6 透過 id 定義使用者自訂查詢語法

```
1  class DailyWord
2    include Mongoid::Document
3    include Mongoid::Timestamps
4    field :tdate, type: Integer
5    field :words, type: Array
6
7    scope :finddategt, ->(ctime){ where(:tdate.gt => ctime ) }
8    scope :finddatelt, ->(ctime){ where(:tdate.lt => (ctime + 86400) ) }
9
10 end
```

資料蒐集程式會每小時會將文章蒐集至 MongoDB 的「文章原始資料」Collection 中(圖 4-7)，於每天 00:30 的時候使用 Jieba 加強版斷詞工具進行斷詞計算每一篇文章的 TF(term frequency) TOP10 特徵詞存入「已斷詞文章」Collection 中，再將本日所有文章特徵詞比對現有詞典，如果詞典中找尋不到的特徵詞而且該詞符合反饋標準（當日該詞出現在兩篇文章以上且為特徵詞），就會將加入 MongoDB「每日新詞列表」Collection 及 Jieba 的使用者自訂詞典中，

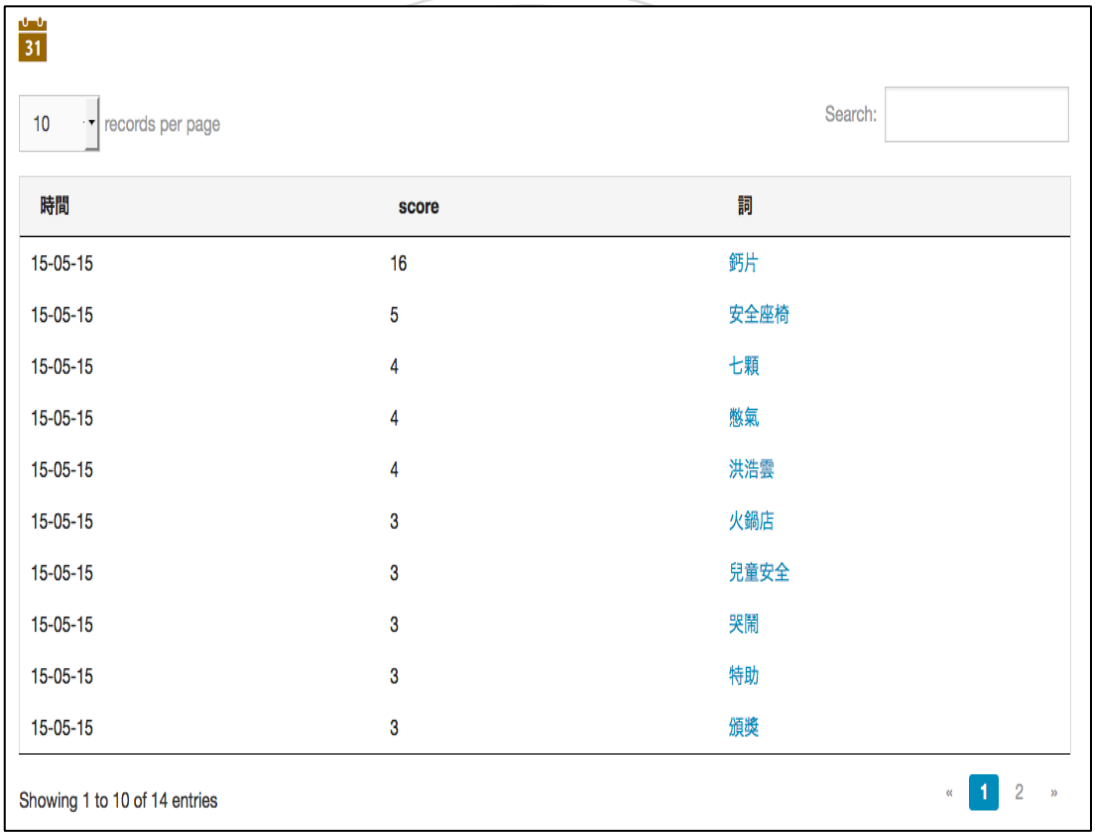
最後將本日所有的特徵詞的 IDF (inverse document frequency)統計結果存入「每日斷詞詞頻統計」Collection 中。

圖 4-7 社媒新詞分析系統 Mongo collections

4.3 社群媒體新詞分析系統頁面

根據上面的資料蒐集、後端資料庫、文章斷詞、每日排程計算結果實作之後，社群媒體新詞分析系統實作了幾種詞的分析頁面讓使用者去觀察詞與社群話題文章之間的關聯性，藉由透過新詞去找尋新的話題或是透過特徵詞的共現關係去瞭解文章的議題框架。

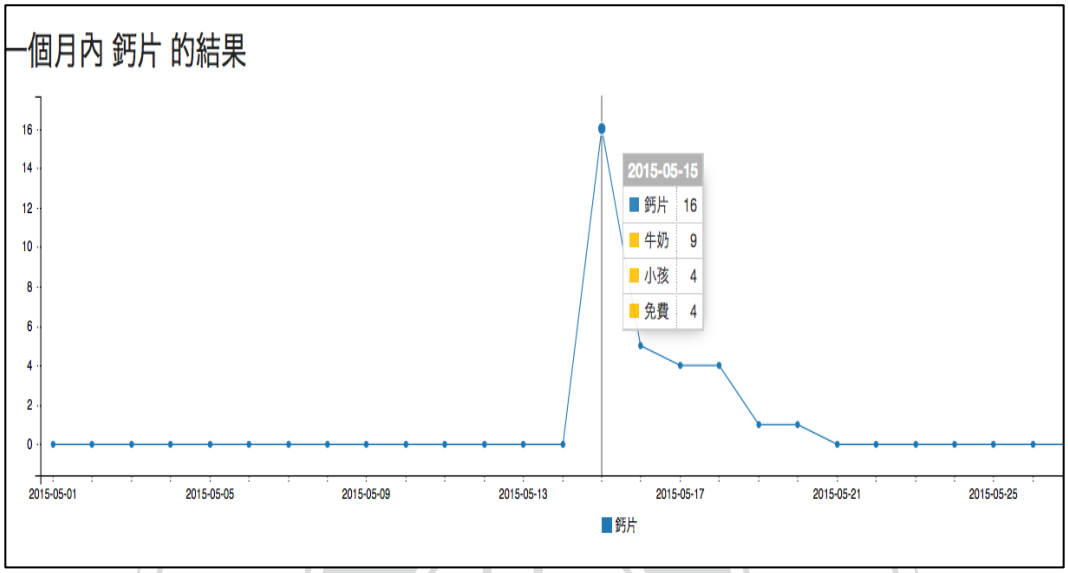
圖 4-8 每日新詞列表



根據每天的斷詞結果比對 Jieba 的當前詞典，如果該特徵詞在當日出現超過兩次以上就會被當作新偵測出來的新詞，這裡評估超過三篇以上的新詞才具有觀察的意義，如果該新詞出現的篇數超過十篇以上，代表這是一個新的熱門話題，值得使用者去關注，例如（圖 4-10）中的「鈣片」，這個詞的出現是在於台北市長提出鈣片替代牛奶提供給小學生作為鈣補充的來源，這樣的議題導

致網友的討論，透過本研究的新詞偵測的方式，可以快速掌握每天新出現的詞，藉由此去了解每天新產生的話題。

圖 4-9 特徵詞與共現詞



有時候只有單獨的一個詞，使用者無法快速地了解到這個詞背後所代表的涵意，當然使用者也可以直接查詢包含這個特徵詞的文章列表，但是透過共現詞的圖表，使用者不必直接去閱讀文章就可以更快速的從共現詞去瞭解本詞背後所代表的議題，如（圖 4-11）我們可以透過「鈣片」與「牛奶」、「小孩」、「免費」的詞語共現關係，由此知道鈣片這個議題在講述的跟小孩免費牛奶的議題有所相關，本圖表另外有提供雙擊功能，使用者就可以直接看到所有與鈣片相關的文章。

熱門話題偵測演算法:

FS: Post Frequency, each Post is represented by the Top-N words list.

$P^T = (w_c^T \ldots w_{10}^T)$

$PF(w^T)$ *For word w at day T.*

For each past P at day T, we will compute a ? score for it as follows.

$$Score(P^T) = \sum_{K=1}^{10} Y_K^T * PF(W_K^T)$$

$$Where\ Y_K^T = \begin{cases} 1 & if\ PF(w_K^T) \geq 1.5 * PF(W_K^{T-1}) \\ 0 & else \end{cases}$$

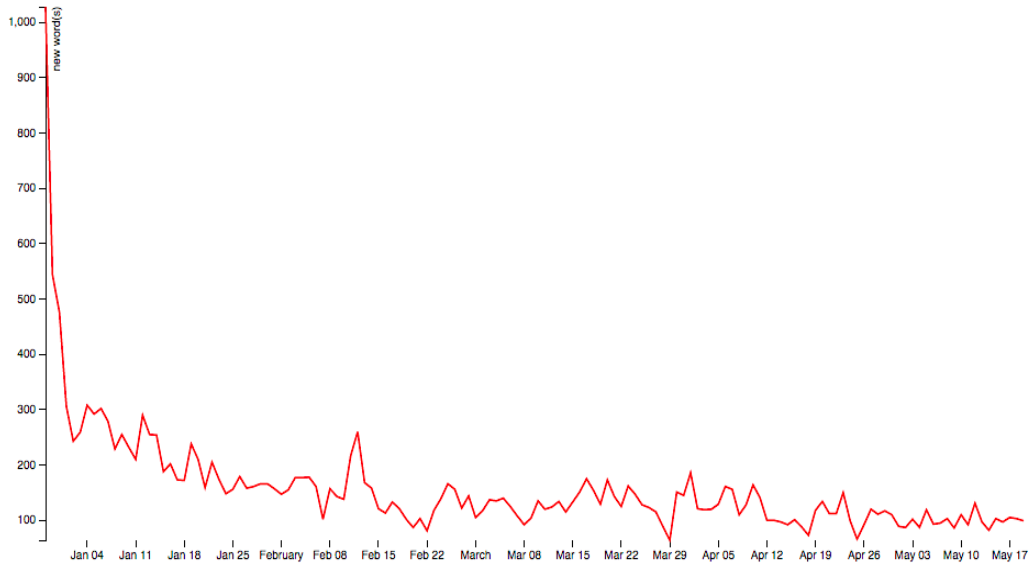
有時候我們發現只關注新詞有時候會錯過一些熱門議題，因為有一些詞可能在其他議題已經出現過，本研究針對這個問題提出文章 Ranking 的演算方法，針對每一篇文章十個特徵詞，透過每日詞頻統計資料庫的結果去比對昨日這些詞出現的文章次數作為一個比較基準，如果該特徵詞今天出現的文章篇數比昨天的次數高於 1.5 倍以上，就把該詞當作是熱門詞，如果一篇文章找出多個熱門詞這一些就是這篇文章的 Ranking 基準，將這些熱門詞今天出現的數量加總得出該篇文章的分數(如圖 4-13 的分數欄位)，由最高分數的第一篇文章為例，這篇文章內包含了五個熱門詞，分數基準: {寵物 => 15},{課稅 => 9},{農委會 => 4},{飼主 => 4},{棄養 => 2}，加總後得到 34 分，透過這樣的方式，可以去補足新詞所偵測不到的熱門議題的不足。

圖 4-10 每日新話題偵測列表

時間	標題	分數	代表詞	words	作者
15-05-25 06:21:55	[新聞] 反對達98% 課寵物稅好嗎	34	寵物	表示,政府,棄養,農委會,管理,反對,飼主,民眾,課稅,寵物	vvus
15-05-25 00:43:01	[新聞] 中華民國萬萬稅 「寵物稅」蠢蠢欲動	31	寵物	表示,政府,農委會,犬管理,流浪,民眾,課稅,寵物,台北市,完整新聞	qwesd611
15-05-25 11:04:12	Re: [新聞] 反對達98% 課寵物稅好嗎	30	寵物	課徵,政府,流浪動物,抑制,飼主,鼓勵,課稅,寵物,精準,只會	liquidbox
15-05-25 12:27:46	[新聞] 侯孝賢微博粉絲爆增 張震貼舒淇照祝賀	25	侯孝賢	聶隱娘,土人,坎城影展,一種,大陸,侯孝賢,臺灣,導演,文說,網友	ilyj2012
15-05-25 10:59:34	Re: [問卦] 有沒有政府很大方補助寵物的卦?	24	寵物	起來,課稅,一邊,了避免,都需要,現在,店家,寵物,負擔,難道	zcc921
15-05-25 21:46:43	Re: [新聞] 「殘障不易找工作」 柯P失言後道歉	23	殘障	歧視,要求,錄用,人士,名額,保障,工作,道歉,殘障,身障者	zu00405479
15-05-25 23:23:10	Re: [新聞] 「殘障不易找工作」 柯P失言後道歉	22	殘障	03,態度,22,而是,05,殘障人士難,人士難,工作,殘障,部分,殘障人士,真的很讓	IloveBlack2
15-05-25 00:37:16	[問卦] 有沒有鄉民們都看什麼台灣導演的八卦？	22	侯孝賢	媒體,侯導,聶隱娘,法國,坎城影展,曹晉,侯孝賢,表示,導演,公司	sedition
15-05-25 18:05:42	[新聞] 農地非法工廠 政府擬使合法引譁然	21	輔導	農地,農委會,非法,合法,輔導,農業區,特定,工廠,經濟部,審查	KahoJyumonji
15-05-25 11:36:50	[新聞] 議員要柯炒洪智坤 柯：殘障人士不好找工	20	殘障	找工作,要求,政風處,不是,洪智坤,資料,柯文哲,殘障,對於,殘障人士	NewPacers

第五章 斷詞驗證及系統成果

圖 5-1 八卦版每日新詞偵測數量曲線圖



（圖 5-1）是針對本研究的資料-PTT 八卦版的每日新詞偵測數量所繪製的曲線圖，本研究的資料區間約為 5 個月的 PTT 八卦版文章，從此線圖我們可以發現在第一天的新詞偵測的數量是最多的，因為 Jieba 預設的辭典主要還是以中國地區的用語詞為主，以下我們所說的新詞指的是原本文章裡面不具有的詞，在第一天本研究對於八卦版文章做斷詞後發現 1028 個新詞，經由反饋系統我們會將這些詞加入使用者自訂詞典中，大約在第三天之後新詞偵測的量就趨於穩定，一個月後每天詞的偵測量就只剩下 100 ~ 200 個詞（per day），而從 2015/1/1 至 2015/5/31 的八卦版文章中，總共偵測出 23314 個新詞。

5.1 Jieba 強化版的新詞偵測評估

強化版 Jieba(加入 SW 版本)的偵測結果做數據上評估：

1.以五天的 PTT 八卦版的文章最為斷詞素材。

2.就所有斷詞結果比對 Jieba 強化版的現有字典,如果符合成為一個新詞的原則,就將該詞認作是一個新詞。

3.人工比對新詞偵測結果，計算正確率，因為本系統主要著重於新詞偵測對於 Jieba 原生斷詞的結果會於予保留，所以本驗證方法只會對新詞的正確率做計算。

4.將所有偵測到的新詞，扣除偵測錯誤的結果，比較 CKIP 及原生版 Jieba 的新詞偵測涵蓋率（Coverage）。

我們採用五天的八卦版文章共 10218 篇,使用 Jieba 強化版對這些文章進行斷詞，斷詞結果約十萬字，比對當前的詞典之後發現 458 個不存在於目前詞典的新詞，人工判斷後判定正確率為 96%，扣除偵測錯誤的詞後有 445 個新詞被偵測出來(詳細結果如附錄 2),接下來將會使用這些詞與 CKIP 和原生版的 Jieba 做偵測涵蓋率的比較。

5.1.1 SW 新詞偵測成果及效能比較

從新詞的偵測結果發現 CKIP 對於一些人名具有較高的偵測率(例:鄧福如...),但是對於一些新的複合詞較弱(例:黑箱課綱、高鐵財改案、防磚條款...)而 Jieba 對於這些複合詞都可以偵測出來，但是相較於 CKIP 的名詞則偵測率較低,但透過SW和詞反饋的機制後可以使得原生版結巴提升32%的新詞偵測率，接下來將會對 Jieba 強化版及 Jieba 原生版進行效能的評估比較。

(表 5-1)是由 Mac air2014 年的機器，Intel 2 core 1.6G，4G memory，SSD Disk 上對 445 篇文章對斷詞的效能評比，我們可以發現加入 SW 的強化版 Jieba 會較原生版需要多花一點運算時間（16%），在程式中還有許多可以優化的部分，但本研究主要在於提升新詞的偵測率及創建專屬領域詞典，在效能優化部分著墨有限，以上數據供使用者參考評估效能與新詞偵測率的比較。

Command	Efficiency Rate
python jiebaSW.py	6.03s user 0.30s system 97% cpu 6.463 total
python jieba.py	5.20s user 0.27s system 99% cpu 5.513 total

表 5-1 效能評估

5.1.2 新詞偵測結果觀察

後續觀察半年的偵測結果字詞觀察到一些有趣的例子，在八卦板內，網路鄉民習慣以「魯蛇」自稱，這是一個奇怪的網路現象，而另一方面的解讀可以發現台灣網路鄉民普遍覺得自己處於壓抑對於未來不抱太多希望的現象，而這樣的現象中也衍生出一個有趣的新詞「慣老闆」，它指的是要求低薪勞力壓榨員工但卻因為台灣人刻苦耐勞而被寵壞的老闆。另外在政治議題可以發現鄉民對於政黨的特殊取向，產生不少的詞來調侃執政的政府(例如:黨證、割闌尾、騷...) 。造詞的規則大致上都是「會意」及「形聲」，其中又以形聲為主，(例如:八嘎冏、霉體、妓者、溫拿、魯蛇)，詳細的新詞整理表可以參照附件。

5.2 社群媒體新詞偵測系統成果展示

有了良好的斷詞基礎以及尋找新詞的能力，以下是本系統對於每天的特徵詞擷取的結果實作了每日新詞偵測表以及基於熱門詞去計算找尋熱門話題兩個

資料觀測視圖，以下是本研究的實驗結果及介紹。

新詞偵測演算法: $\#W(D_1) - \#W(D_2) \geq 2$

時間	score	詞
15-02-04	69	復興
15-02-04	47	墜機
15-02-04	42	復興航空
15-02-04	14	ATR72
15-02-04	10	機身
15-02-04	9	機上
15-02-04	7	機艙
15-02-04	7	機體
15-02-04	6	遷建
15-02-04	6	飛行員

表 5-2 2015/2/04 新詞偵測表

這是在每日新詞偵測紀錄中分數最顯著的偵測結果，那一天發生了復興空難事件，許多人在轉貼討論復興空難的話題以及新聞，當天大部份的新詞都在是講述澎湖空難的相關詞。

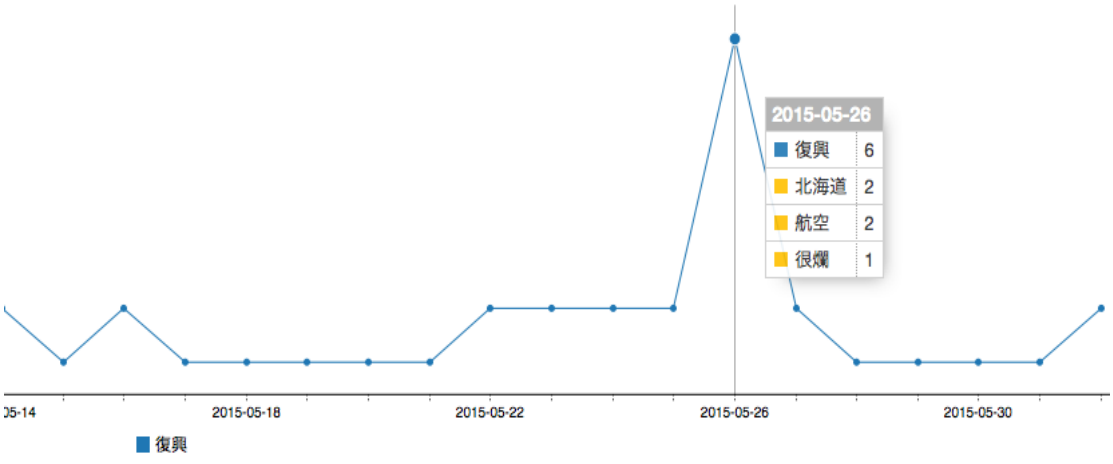
同樣的話題在熱門話題偵測分析，也是偵測出同樣的話題，當新詞及熱門詞出現一樣的話題的時候，代表著這個話題的討論熱度非常高，而且這是一個今天新發生的議題，值得使用者後續關注。

時間	標題	分數	代表詞	words	作者
15-02-04 13:59:25	[新聞] 復興同機型再失事 羅淑蕾籲追究	232	復興	表示,航空,羅淑蕾,基隆河,墜機,松山機場,去年,復興,機型,失事	OmegaWind
15-02-04 23:26:16	Re: [新聞] 柯文哲：請一起幫尚未獲救的乘客祈禱	208	復興	萬分,全力,獲救,04,搜救,乘客,新聞,空難,復興,航空	deitly
15-02-04 20:59:13	[新聞] 復興航空空難 佛光人慰問關心！！！	195	復興	台北,復興航空,航空GE235,家屬,復興,法師,班機,佛光山,空難,佛光	BlackRequiem
15-02-04 16:16:18	[新聞] 復航才隔半年多又墜機！機齡不到8個月	195	復興	表示,ATR72,航空,復興,乘客,搶救,班機,客機,起飛,這架	dearevan
15-02-04 11:17:01	[新聞] 輕航機失事！墜落基隆河　10多人待救援	186	復興	來源,10,救援,復興航空,基隆河,飛機,復興,完整新聞,失事,目前	drjc
15-02-04 16:49:46	[新聞] 陸機長：中國早就淘汰ATR72系列客機	185	復興	淘汰ATR72系列,航空,墜機,ATR72系列客機,淘汰ATR72,中國,復興,系列客機,ATR72系列,完整新聞,客機,淘汰,報導	striving

表 5-3 2015/2/4 熱門話題偵測

經過後續的觀察三個月後復興還是有一些負評存在，後續有觀察到復興有維修出包的事件，以及有旅客表示抱怨的問題透過本系統快已快速蒐集觀測話題的出現、趨勢流行、後續這個話題的延續話題狀態。

圖 5-2 與復興最近相關的共現詞



時間	score	詞
15-01-12	63	國稅局
15-01-12	18	借據
15-01-12	12	查稅
15-01-12	9	贈與
15-01-12	6	檢舉案件
15-01-12	6	涂醒
15-01-12	6	檢舉人
15-01-12	6	贈與稅
15-01-12	6	許立民
15-01-12	5	契約

表 5-4 2015/1/12 新詞偵測表

在 2015/1/12 偵測到的最熱門詞是國稅局，這是柯文哲被檢舉向父母借錢但是沒有借據，懷疑是贈與有逃漏稅的議題，這裏可以觀測到一個特性就是新詞的出現會有相關的聯帶性，像是「借據」、「查稅」、「贈與」、「契約」都是在講關於這一個議題事件的話題。

日期	偵測新詞	新詞的隱含話題
2015-01-02	陳彥衡	[新聞]林俊傑遭歌迷毆打事件
2015-01-04	導盲犬	[爆卦] 台北車站某家鐵板燒拒絕導盲犬入店
2015-01-05	寬宏	[新聞] 江蕙引爆搶票潮 寬宏網站癱瘓
2015-01-12	國稅局、查稅、借據	[新聞] 遭國稅局查稅　柯媽：借給孩子誰寫借據
2015-01-19	經濟艙	[新聞] 北市官員出差 商務艙砍為經濟艙
2015-01-20	殉職	[新聞] 20 幾歲消防員殉職 惡火無情
2015-01-21	街友	[新聞] 2 屁孩暴打街友濺血遭肉搜　1 人到案
2015-01-23	江肇國	[新聞] 違規停車不能開單？議員江肇國遭爆嗆警
2015-01-26	破銅爛鐵	[新聞] 外賓贈錶　柯 P 失言：可拿去破銅爛鐵賣
2015-01-28	交保、魏應充、楊蕙	[新聞] 頂新魏應充凌晨交保 5 分鐘就備妥 1 億
2015-02-04	復興航空、墜機、復興、ATR72、侯友宜…	[新聞] 復航才隔半年多又墜機！機齡不到8個月
2015-02-11	劫獄、鄭立德	[新聞] 典獄長遭 6 犯挾持　法務部：談判優先！

表 5-5 每日新詞及偵測話題表一

日期	偵測新詞	新詞的隱含話題
2015-02-23	釋昭慧、	[新聞] 慈濟內湖開發案 釋昭慧發文挺
2015-02-28	藍黑	全球趣聞／藍黑VS白金　一條裙子引發的顏色大戰！
2015-03-10	陳碧瑤	[新聞]聯開宅抗爭代表被爆房仲裝可憐陳碧瑤將現
2015-03-13	國宴	[爆卦] 義美FB：總經理受邀國宴，該桌只有一人
2015-03-20	撿拾、漂流木、亞杉、木材行	[新聞] 鋤錯了?! 漂流木爭議 暫緩「拔掉」內湖分局長
2015-03-24	刀塔	[新聞] 暴雪對刀塔傳奇提起刑事告訴
2015-04-01	F18	[新聞] 美軍F-18迫降台南機場　軍方協助維修
2015-04-02	李蒨蓉	[新聞] 飛官私帶李蒨蓉上阿帕契　陸軍：違法將嚴逞
2015-04-03	勞乃成、記過	[新聞] 李蒨蓉登阿帕契　軍官勞乃成大過一次移送
2015-04-10	趕工	[新聞] 捷運工程出人命 網友：林佳龍是殺人兇手
2015-04-23	網路霸凌、實名制	[新聞] 楊又穎兄哽咽：網路霸凌逼妹上絕路
2015-04-25	尼泊爾	[新聞] 尼泊爾強震聖母峰雪崩兩百年古塔倒了

表 5-6 每日新詞及偵測話題表二

（表 5-5）及（表 5-6）為一月到四月的新詞偵測結果及其偵測出來的話題一覽表，因為數量過於眾多，這裏本研究沒有全部都列舉上來，我們選擇的新詞的出現篇數必須大於 20 篇以上才會表列上面列出來，從上面的一覽表中再去對照當時的新聞焦點，證實新詞偵測的效果是精準的，後來本研究嘗試將新字偵測結果與 Google Trend 網站的熱門搜尋字做比較，新詞偵測的結果與 Google Trend 的搜尋結果常常是吻合的，由此我們可以推估 PTT 八卦版是一個社會的縮影，觀察八卦版可以了解台灣網路的流行風向，值得一提的是，有的時候新詞偵測的時間點會較 Google Trend 熱門字晚二至三天，這裏本研究的猜測是，當人們在電視新聞上或是其他媒介上得知一個新的消息，會嘗試先從 Google 搜尋它的關鍵字，了解到這個事情的來龍去脈之後才會慢慢的從網路上討論開來，所以被本新詞系統偵測到的時候往往是幾天之後。

本新詞偵測系統不只偵測到的是新聞事件，本系統也有偵測到一些網路熱門事件，像是過年期間許多人回家過年會遭到親戚的詢問近況，某網友在網路上提出說「ＯＯＯ要如何抵擋親戚的攻勢」，當天在板上這句話流行了起來，衍生出系列文章，例如：「[問卦] 心理系要如何抵擋親戚的攻勢」、「[問卦] 富二代該如何抵擋親戚的攻勢」、「[問卦] Gay 要怎麼抵擋親戚的攻勢」…等，然後在 2015-4-1 當天因為「美軍 F-18 迫降台南機場　軍方協助維修」這個新聞，造成板上開始有不少網友發起惡搞的系列文章像是：「[問卦]UFO 停在夜店門口會怎樣?」、「[問卦] 停一台阿斯拉在夜店門口會怎麼樣」、「[問卦] 有沒有停一台經國號在夜店門口會怎樣??」…等的系列文章，由以上的結果來看，從新詞去偵測話題是可以嘗試的一種文本分析的方案。

第六章 結論及未來研究

本研究選定八卦版作為社群媒體的分析資料，為了解決中文斷詞的歧異性問題，本研究在原本的 Jieba 斷詞工具，基於隱性馬可夫模型和維特比模型、最長詞優先的基礎上提出了 Sliding Windows 修正方法以及詞的反饋機制，經實驗後發現可以較原本的 Jieba 斷詞工具提升 32%的偵測率及 96%正確率，透過強化版 Jieba 斷詞工具的實作及評估後，我們可以把新詞偵測及特定領域的詞典產生問題交給強化版 Jieba 斷詞工具，針對所有文章的特徵詞加入一些演算法的方式去找出新詞及熱門詞。

經過半年的實驗結果，本研究發現新詞與熱門詞的偵測對於文本分析的效果比本研究當初預期的結果還要更好，經由一些其他支援的交叉比對，像是 Google Trend 每日熱門搜尋關鍵字，Yahoo 網路新聞…等主流媒體去觀察本新詞偵測系統所偵測出來的詞及話題，與社會議題的緊密性非常的高，甚至有時候可以反映一些潛在話題，像「支付命令」在該話題還在網路醞釀的時候，本新詞系統在新聞發表的當下就發現網路話題開始在燃燒，直到一個多禮拜後開始有一些立法行政的動作出現，新聞才後續反應報導這個話題，但本系統早在一個多禮拜之前就偵測出這個議題，所以透過每日新詞及熱門詞的觀察，加上傳播領域專家的本身知識，可以快速判讀這個議題的後續發展性，而一個熱門議題的主角（像是「復興航空」），也可以透過本分析系統去後續觀察關於這個詞的其他相關共現詞去掌握該議題主角在事件發生幾個月之後的網路評價及討論熱度，使得使用者可以從議題的開始、發展到後續追蹤都可以透過本社群媒體新詞分析系統完成。

展望未來，我希望從兩個方向出發：一、從 PTT 八卦版的實驗結果中本系

統可以有效的反應該該社群上所發生的一些事件而且這些事件可以緊密的與現實社會議題結合，但目前實作的研究素材只局限於 PTT 八卦版，對於其他社群媒體資源的適用性需要更多的實驗去證明，例如我們可以收集新聞以及社群的文章，對於雙方不同素材進行斷詞，當一個新的詞出現時候，我們可以經由觀察去發現該詞在新聞以及社群的趨勢及走向，對於探討一個新詞的起源，擴散以及發展應該是一個蠻有價值的研究方向。二、目前斷詞系統的主要斷詞依據還是在於詞典的領域相關性，雖然目前的實作方式可以達到針對斷詞領域的文章素材透過新詞偵測的方式及詞回饋的機制產生較吻合該領域的詞典，對於一些特殊的應用需求可以再增加一些條件去幫助偵測，例如新爆紅人名偵測我們可以依據 SW 再加上百家姓的規則去判斷找尋某日的爆紅新人物。本研究嘗試提出一個方法讓使用者可以透過一個簡單的方式去解決斷詞及關鍵詞擷取的問題，往後的研究可以專注於應用的探討及開發，希望可以對於文本分析及社群傳播研究可以有一些基礎貢獻。

參考文獻

[1] Chen. & Bai. (1998), Unknown word Detection for Chinese by Corpus-based Learning Method.

[2] Chen. & Ma. (2002), Unknown Word Extraction for Chinese Document.

[3] L. Jin. (2013), Number in Chinese: A Corpus-Based Computational Investigation.

[4] QX Lin. (2010), 結合長詞優先與序列標記之中文斷詞研究。

[5] Yi-Lun Wu. (2011), 多語語碼轉換之未知詞擷取。

[6] Zhihui. Wu, Hongwei. Liu, Li. Chen. (2014), 统计学与应用, 3, 30-35, 高效朴素贝叶斯 Web 新闻文本分类模型的简易实现,The Simply Implement of Effective Statistical and Application

[7] Z. Wu (2014), The Simply Implement of Effective Naive Bayes Web News Text Classification Model

[8] 陳鍾誠、許聞廉. (1998), 結合統計與規則的多層次中文斷詞系統

[9] 陳聰宜. (2012), 新聞事件偵測與追蹤結合時間區間之分群分類演算法評比

[10] DAG, http://www.csie.ntnu.edu.tw/~u91029/DirectedAcyclicGraph.html#1

[11] ACID, http://zh.wikipedia.org/wiki/ACID

[12] MongoDB, http://docs.mongodb.org/manual/core/crud-introduction/

[13] NoSQL, http://zh.wikipedia.org/wiki/NoSQL

[14] Jieba 斷詞工具, https://github.com/fxsjy/jieba

[15] 隱馬可夫模型, http://zh.wikipedia.org/wiki/隐马尔可夫模型

[16] 维特比算法, http://zh.wikipedia.org/wiki/维特比算法

[17] 洪仲丘案在 PTT,

http://zh.PTTpedia.wikia.com/wiki/%E6%B4%AA%E4%BB%B2%E4%B8%98%E4%BA%8B%E4%BB%B6%E5%9C%A8PTT

附錄 1:新詞偵測結果表

字詞	人為註釋
小魯我	Loser 稱謂
魯弟	Loser 稱謂
黨國	國民黨
嗡嗡嗡	柯 P
雜魚	日語
MG149	時事
粉粉	粉絲
婉君們	造詞
柯屁	柯 P
馬奶	馬英九
阿魯	Loser 稱謂
洪仲丘	時事(人名)
5F	ＰＴＴ用詞
靠爸	凡事倚靠父母
傲嬌	日語
魯肥宅我	Loser 稱謂
太陽花	時事
本嚕	Loser 稱謂
小魯妹	Loser 稱謂
阿共	中國
魯魯我	Loser 稱謂
魯蛇我	Loser 稱謂
魯宅我	Loser 稱謂
妮妮	歐陽妮妮
黨校	泛藍學術單位
腦補	日語

潮潮	新潮
優文	ＰＴＴ用詞
黑絲	黑絲襪
天龍人	台北
割闌尾	時事
洩洩	謝謝
八嘎囧	八家將
怒買	造詞
阿幹	國罵
台灣不會好	國民黨不倒，台灣不會好
祭止兀	蔡正元
腐女	日語
北七	白痴
八嘎冏	八家將
中國黨	國民黨
方濟各	羅馬教宗
本魯肥宅	魯蛇
旺中	時事（香港）
食安	時事（議題）
霉體	媒體
勝文	連勝文
歪國人	外國人
柯粉	柯文哲的粉絲
天龍國	台北（詞出自海賊王）
魯肥宅	Loser 稱謂
瘦宅	宅男
小魯哥	Loser 稱謂
肺紋	廢文
野雞大學	時事（洪秀柱的學歷）

酸民	造詞（說話很尖酸的網友）
苛屁	柯文哲（負面字眼）
非核家園	時事（議題）
好冰冰	郝龍斌
時代力量	組織（第三勢力政黨）
彎彎	人名（插畫家）
貧乳	日語
公督盟	組織
鄭捷	時事（捷運殺人案兇手）
小魯妹我	Loser 稱謂
鬼父	動漫遊戲
嘖嘖	狀聲詞
餿水油	時事（味全）
賣台	出賣台灣
螞蝗	馬英九
台 GG	台積電
魯叔	Loser 稱謂
支付命令	時事（詐騙事件）
黨證	國民黨證明文件
歐巴	韓語
廢文大賽	ＰＴＴ事件
水水	漂亮女生
肥宅魯蛇	Loser 稱謂
補刀	遊戲用語
法拉利姊	人名（新聞爆紅人物）
全民買單	全民幫政府的收爛攤子
廢宅	宅男稱謂
壁咚	日語
估狗	Google

魯肥	Loser 稱謂
太陽花學運	時事（服貿）
小七	7-11
人生勝利組	Winner（高薪成功人士們）
廢死團體	組織（廢除死刑）
廢死聯盟	組織（廢除死刑）
統神	知名遊戲玩家
藏頭	藏頭詩
蛇蛇妹	Loser 自稱
深藍	國民黨
覺醒公民	時事
島民	台灣人
柯批	柯文哲
本魯蛇	Loser 稱謂
肥魯	Loser 稱謂
小魯蛇	Loser 稱謂
補血	遊戲用語
夢境	用夢境說爆料
作夢	用夢境說爆料
妓者	記者
魯哥	Loser 稱謂
統二	7-11（統一）
台勞	Loser 稱謂
穩拿	Winner 稱謂
小當家	動漫角色
乳題	如題
大覺者	時事（慈濟）
宇宙大覺者	時事（慈濟）
柯神	柯文哲

強者我朋友	造詞（我厲害的友人表示）
課綱微調	時事（課綱事件）
奈米屌	造詞
護家盟	愛護家庭大聯盟
亞投行	時事
當當	麥當勞
多元成家	時事
民國黨	政黨名稱
張淑晶	時事（惡房東）
仇女	抱怨女生很現實
蔣光頭	蔣公
老魯	Loser 稱謂
萬鎂	美元
太魯	Loser
嘉瑜	高嘉瑜（民進黨市議員）
酸宗痛	選總統(諧音)
肥宅朋友	宅男
本肥宅	宅男
漂漂	漂亮
學匪	造詞（學運參與學生）
消波塊	黑社會黑話
李蒨蓉	人名（阿帕契事件）
勞乃成	人名（阿帕契事件）
乳提	如題
柱柱姐	洪秀柱
不願役	義務役
胖宅	宅男
國冥黨	國民黨
滑進	阿基師外遇梗

加薪四法	時事（政府政策）
涊魯弟	Loser 稱謂
超級後悔投給柯 p	特定人士反串簽名檔
聶隱娘	電影人物
島國前進	組織
小夫媽	洪秀柱
沃草	時事觀察組織
網路霸凌	時事（討論話題）
統媒	偏向一中的媒體
巴嘎囧	八家將
喇牙	蜘蛛
慣老闆	壓榨員工成性的老闆
理工宅	宅男
八嘎炯	八家將
阿陸	中國
八嘎窘	八家將
阿 Q	泡麵
紅牛	提神飲料
吊嘎	衣服款式
宅哥	宅男
起風	社論情勢轉向
長照法	長期照顧服務法
肥宅室友	宅男
哪尼	日語
胎嘎	台語（骯髒）
肛肛	同志癖好
夜衝	晚上跑去….(參與活動)
屌絲	中國網路用語
哥是	中國網路用語

大給	大家，台語諧音
K 黨	國民黨
支持死刑	時事（廢死）
反廢死	時事（廢死）
政問	問政治人物的八卦
戰文	挑釁文
歐冠	歐洲足球聯賽
鏡報	香港媒體
戰鬥民族	俄羅斯人
同文同種	同文化同種族
本小魯	Loser 稱謂
柱姐	洪秀柱
柱柱姊	洪秀柱
柱柱	洪秀柱
腦就	人一藍、腦就藍

附錄 2:詞比對素材

※灰色底色表格為誤判詞

19.5K	嚴重性	呵呵呵呵	海樂	燒雞
山難	防彈衣	陸戰 66	許崑源	顆星
徐永明	不虧	66 旅	兩倍	3000 公尺
日常業務	國學	和平倡議	永豐餘	4K
硬著	萬劍彈	來過	色情按摩	退休養老
陰廟	講些	燒鵝	沒寫	匯豐
黑箱課綱	什麼啦	28k	志氣	率高
油槍	咬斷	蹺課	電扇	購買力
豬豬	不像話	著重	守門員	坑殺
各部會	區時	專利授權	球門	強到
輿情	廠房	19K	遼寧隊	燈火通明
阿靠	K2	白皮書	降溫	超搭
轎班	聯準會	陳保仁	韓恩	指考英文
黃萬翔	所以希望	法網	靠自己	這種蟲
張博崴	豐滿	萬磁王	蠻多	愛是
放風箏	一顆蛋	泥鰍	青黃不接	塑膠袋
園長	布拉布拉	到頭	千鳥	搜救不力
大品牌	瓜地馬拉	小真	松葉	瑞典出生
緩緩	這餐	柏慎	茶臼	海軍陸戰隊
魏國	社經地位	也不敢	臨時人員	90 年代
太一	鹹魚雞粒豆腐煲	會留	研發替代役	廠藥
60 公升	開班	雨太大	既判力	每組
棉條	不允許	大給	厄本	退出時代力量
軟綿綿	燒酒雞	押韻	麻州	宣布退出
秀秀	防磚	好壞	波士頓	宣布退出時代力量
輸送	籌措資金	沒什麼好	庫賈	退休養老理想
夠味	聲寶	陸戰66旅	進口關稅	養老理想
噴街	貞子	魔幻力量	汽車產業	佳子公主
貼心	實質薪資	拆線	夜宿陸戰隊	張父

勾結	雪崩式	莊翁	第一盤	金車
音響設備	灣灣	歡迎光臨	如果你首先想到	摸魚
血本無歸	游藝	整顆	呵呵呵呵呵呵	疑似患者
油箱裡面	全哥	充滿希望	如果你首先	紙類
同儕	雨傘運動	鉅亨網	柏南克	打針
軍事戰略	常常有	巴著	柏南奇	先噓
爆炸聲	山友	蔡練生	碳酸氫鈉	常吃
剛跑	警分局	塵蟎	政見發表	底價
連長寢室	通靈	艾奎諾	腫大	胡克定
林中斌	舉旗	萬 8000 元	捐精	補習班人員
聯合國大會	高標	8000 元	自耕農	匪區
名將	供花	萬 8000	櫃子	賞罰
鄧福如	保守基督教	中永和	馬達加斯加	閃電俠
高鐵財改案	日行性	紫外線	一個法國人	圖形
葉紜萱	首次斷交	產假	攝護腺腫大	超廢
小蕊蕊	復交；	長照保險	網路連署	高中班
與果敢	搞內需	並不是	張國周強胃散	陳智雄
勇於	高雄廠	奶粉罐	周強胃散	有本事
晨操	你以為	陸戰隊 66	桃園市衛生局	190 萬
19.5k	就要針對	京東方	張國	烙鐵
不想用	台英	撈起來	張國周強	林燕祝
日女	彈道飛彈	身心障礙	龍虎山	莎莎亞
注射劑	湧出	週記	拉進去	你會養
學名	飄浮	專精	侯導	腐腐
想領	狀元	樂高	甄選	浮屍
博崴	獲獎	課輔	培訓	18 禁
韓文	叫聲	爭鮮	滯台中國人	諾貝爾經濟學獎
鍋蓋	洩漏公文	蛋殼	鏟子	紐澤西
鐘點	發錢	模擬考	除靈	澤西州

澳洲右翼政黨	六點半	倡議	還越	找工作
澳洲右翼	合太	藍皮書	罩門	核彈頭
右翼政黨	投擲	直選	車內	施雪蕙
劍士	過火	舊版	怎麼寫	殘障人士
集團結婚	張雅茹	育嬰	暗批司法逢迎	肄業
公開信	上戰場	邱姓	暗批司法	原民
網外	上戰場人數	美樂家	秋山澪	農業區
軍政	大學教授	研究員	熱浪	楊婦
副部長	債務人	數乙	力氣	必點
羽球	到演	菸草	一腳	武士刀
基努斯之槍	LV9	求職	繞行	假檢警
朗基努	自桶	鯡魚	睡午覺	鄭藝仁
山難救援	打起來	萬噸；	張鵬	現場表演
凋盡	洪旭東	眼紅	吉普森	燒到
一頓飯	蓮蓬頭	聖火	打用	時雨
戶籍謄本	棒棒堂	影視	愛德華茲	財神
裝來防	兩億人	宅女	職等	變通
葡萄園	婚姻無效	給我大	下個月	孫亞夫
成癮性	上天給我	關個	在坎城影展	玉婷
購併	陳麗珠	遲遲	小芬	丟的
長笛	億個	大仁哥	源頭	重摔到
多難相處	殲滅戰	過啦	濕氣	新崛江
得到規模化	最難打	這行	新聘	收傘
良種	動怒	以後會	不過堂堂一個	葛拉瑟
合作基地	一怒為	給力	不過堂堂	代書
播放清單	小額	遷廠	堂堂一個	參予
有天	蕭女	1080p	研究中心	棒協

落葉	物理學家	當選總統	陳福海	上戶
慣性	幾位	憑空	縣長陳	主動解散國會
塵土	膽識	五萬七	金門縣長	防磚條款
藍田國小	張靚穎	美人魚	金門縣長陳	包師傅
62.29	晉升	停班停課	施工項目	寬鬆
台海和平穩定	馬來	小三通	高個	
巴拿馬運河	虎尾交流道	精神領袖	漁港	