



龍華科技大學

資訊管理系碩士班

碩士學位論文

應用文字探勘技術於客訴資料之研究-以
台大 PPT 論壇為例

Applying Text Mining to Customer Complaints
on the NTU PPT Forum

研究生：林名彥

指導教授：馬芳資 博士

中華民國 104 年 07 月

龍華科技大學

碩士學位考試委員會審定書

本校 資訊管理系 碩士班 林名彥 君

所提論文 應用文字探勘技術於客訴資料之研究-以台大PPT論壇為例

經本委員會審定通過，特此證明。

學位考試委員會

委員：郭展盛

胡鳳義

馬芳資

指導教授：馬芳資

系主任（所長）：

中華民國 104 年 7 月 15 日



摘要

論文名稱：應用文字探勘技術於客訴資料之研究-以台大 PPT 論壇為例

頁數：59

校所別：龍華科技大學

研究所：資訊管理系碩士班

畢業時間：103 學年度第 2 學期

學位：碩士

研究生：林名彥

指導教授：馬芳資 博士

關鍵詞：文字探勘、客訴、LDA Gibbs

網際網路的盛行下許多消費者會透過網路論壇來發表意見，尤其是網購商品的抱怨；目前企業對於顧客抱怨(又稱客訴)的處理，大多是以客戶服務中心人員來取得顧客抱怨資訊而進行處理，對於網路論壇上的抱怨資訊常常是無法來處理。因此，本研究搜集網路論壇的客訴文章進行文字探勘，以尋找抱怨文中的關鍵字詞，並瞭解網友們經常抱怨的主題和關聯的字詞。

本研究使用 R 語言撰寫程式來自動抓取台大 PTT e-shopping 論壇上的文章，每一篇文章存成一個文字檔以形成文檔庫；接著以人工方式挑選出 868 篇抱怨文，利用文字探勘套件將文檔庫內文章進行中文斷詞；出現次數 40 個以上的字詞來繪製關鍵字雲；然後利用 LDA Gibbs 方法計算每一個關鍵字的 TF-IDF，並且設定 10 個主題來找出關鍵字詞的關聯性。

研究結果顯示這些抱怨文經由中文斷詞後有 4004 個關鍵字，出現次數達 40 次的有 278 個重要關鍵字；利用 LDA Gibbs 模式從 278 個字中找出 10 個主題，分別是訂單、匯款、衣服、地址、朋友、貨運、顏色、公司、包裹、瑕疵等。從這 10 個主題中我們可以發現消費者在網購商品時常會發生的抱怨課題，而此可提供給電子商務業者在商品管理和服務上的參考，期能強化商品和服務以減少顧客的抱怨。

ABSTRACT

Thesis Title : Applying Text Mining to Customer Complaints on the NTU PPT Forum

Pages : 59

University : Lunghwa University of Science and Technology

Graduate School : Department of Information Management

Date : July, 2015

Degree : Master

Graduate Student : Lin, Ming-Yen

Advisor : Dr. Ma, Fang-Tz

Keywords : Text Mining 、Customer Complaint 、LDA

With the popularity of internet, many consumers comment in the forum, especially complaining about online shopping goods. Currently enterprises deal with customer complaints mostly focus on the complaints obtained by the staffs in the customer service center. They are often unable to handle the complains on the web forum. Thus, this research collected customer complaints in internet forum, and used text mining techniques to obtain the keywords in these complains and to understand the topics often complained by consumers and the association of terms.

This research wrote the R language programs to automatically crawl articles on the NTU PTT e-shopping forum. Each article was saved as a text file to set a document library. Then we manually picked out 868 complains, because some articles were not complains. We used text mining suite in R language to segment Chinese into words and get the most frequent terms of the complains. And we use the words which frequency is greater than 40 to draw a word cloud. Then we use the LDA Gibbs method to calculate the TF-IDF of these terms and set 10 themes to find

the word association.

The results show that there are 4,004 keywords in complains after the Chinese tokenization. And there are 278 terms which frequency is greater than 40. Use LDA Gibbs method to find 10 topics from 278 words, which are orders, money transfer, clothes, address, friend, freight, color, company, package and defect. From these 10 topics, we can understand what consumers often complained about the online shopping. These results can be a reference for e-commerce vendors on merchandise management and service delivery to strengthen the goods and services to reduce customer complaints.



誌謝

在龍華就讀資管研究所這段時間，有幸能加入龍華資管所這個大家庭，可以跟著大家學習、成長。在研究所這段時間，我從系上老師身上學到了許多新的知識，讓我不僅在研究上或是未來的工作上都是助益良多。感謝老師們耐心的教學，對學生總是傾囊相授，且聆聽學生的問題並適時指引方向，讓學生除了在學習專業領域外，更從研究的過程中體驗到認真學習態度的重要。

感謝口試委員 郭展盛教授、胡鳳義教授與馬芳資教授於口試時給予的指導以及建議，使學生獲益良多。同時，也感謝系上老師們這兩年來的教導，以及同班同學們，感謝你們在這段學習時間給予的鼓勵、扶持與協助。再一次感謝指導教授 馬芳資博士，承蒙老師的諄諄教誨，學生方得以順利完成研究。

最後，感謝我最親愛的家人，因為你們一路的支持與關懷，讓我能夠心無旁騖的專注於課業研究，得以順利完成碩士論文，也感謝所有曾經幫助過、鼓勵過我的人，此刻的喜悅願與大家分享，同時將這份榮耀，獻給我的老師、同學和家人們。

目錄

摘要	i
ABSTRACT.....	ii
誌謝	iv
目錄	v
圖目錄	vii
表目錄	ix
第一章 緒論	1
1.1 研究背景	1
1.2 研究動機	2
1.3 研究目的	3
1.4 論文架構	3
第二章 文獻探討	4
2.1 客訴	4
2.2 客訴相關研究	6
2.3 文字探勘	9
2.4 LDA.....	11
2.4.1 LDA 的貝氏網路結構	12
2.4.2 Gibbs 的抽樣方法	13
第三章 研究方法	15
3.1 研究架構	15
3.2 客訴資料收集方法	17
3.2.1 R 語言	17
3.2.2 RCurl	18
3.3 關鍵字詞搜尋	23
3.3.1 資料提取	24
3.3.2 中文斷詞	25
3.3.3 關鍵字詞雲	29
3.4 LDA.....	30
3.4.1 詞頻-反轉文件頻率 (TF-IDF)	30
3.4.2 LDA Gibbs 模式	31
第四章 研究結果	37
4.1 網路客訴分析流程	37
4.2 客訴分析結果	38
4.2.1 中文斷詞與關鍵詞字雲分析結果	38

4.2.2 關鍵字數量與 TF-IDF 分析結果	40
4.2.3 LDA Gibbs 模式分析結果	42
4.3 分析結果討論	49
第五章 結論與未來研究	50
5.1 結論	50
5.2 未來研究	51
參考文獻	52
附錄	56



圖 目 錄

圖 2.1 文字探勘技術金字塔	10
圖 2.2 文字探勘的處理流程	11
圖 2.3 LDA 主題及包含的字彙以顏色區分	12
圖 2.4 LDA 的貝氏網路結構	12
圖 2.5 LDA 的 Gibbs 抽樣方法流程	13
圖 3.1 研究架構	16
圖 3.2 抓取文章網址程式流程圖	19
圖 3.3 抓取文章網址結果	20
圖 3.4 撈取文章內容程式流程圖	21
圖 3.5 撈取文章結果(部分)	23
圖 3.6 text mining 套件	24
圖 3.7 檔案提取進 R 軟體裡的結果	25
圖 3.8 最初的中文斷詞結果(部份)	27
圖 3.9 矩陣統計字詞出現的數量出現結果(部分)	28
圖 3.10 最終中文斷詞結果	28
圖 3.11 關鍵詞雲	29
圖 3.12 LDA 程式流程圖	32
圖 3.13 分割矩陣結果	34
圖 3.14 矩陣合併結果	35
圖 3.15 LDA Gibbs 結果	36
圖 4.1 客訴分析流程圖	38
圖 4.2 868 篇抱怨文中文斷詞結果	39
圖 4.3 關鍵字詞雲分析結果	40

圖 4.4 278 個關鍵字	40
圖 4.5 關鍵字次數統計結果	41
圖 4.6 TF-IDF 結果	42
圖 4.7 278 個關鍵字 LDA 模式結果	43
圖 4.8 10 個 topic 關聯結果	43
圖 4.9 topic 2 與 topic 8 關聯圖	44



表目錄

表 2.1 客訴相關研究	8
表 3.1 抓取文章網址程式碼	20
表 3.2 撈取文章內容程式碼	22
表 3.3 檔案提取進 R 的程式碼	24
表 3.4 清除標點符號、數字、英文單字程式碼	26
表 3.5 中文斷詞程式碼	26
表 3.6 清除停用字符程式碼	27
表 3.7 矩陣統計字詞數量程式碼	27
表 3.8 關鍵字詞雲程式碼	29
表 3.9 計算 TF-IDF 的程式碼	30
表 3.10 LDA Gibbs 分析程式碼	33
表 4.1 LDA Gibbs topic 1	44
表 4.2 LDA Gibbs topic 2	45
表 4.3 LDA Gibbs topic 8	45
表 4.4 LDA Gibbs topic 3	46
表 4.5 LDA Gibbs topic 4	46
表 4.6 LDA Gibbs topic 5	47
表 4.7 LDA Gibbs topic 6	47
表 4.8 LDA Gibbs topic 7	48
表 4.9 LDA Gibbs topic 9	48
表 4.10 LDA Gibbs topic 10	49

第一章 緒論

1.1 研究背景

根據資策會在 2013 年的統計，台灣家庭至少擁有電腦普及率為 88.4%，家戶連網普及率為 84.8%，個人上網普及率 75%；加上平板電腦擁有率已經高於三成，並呈現持續成長趨勢。再從人口特性來看，「10-44 歲」（九成以上）民眾有每天上網的比例明顯高於其他年齡的民眾。從這樣的調查結果，隨著台灣電腦及網路的普及率逐年上升，網路已經成為人們不可或缺的生活重心。

網際網路的盛行，許多使用者已習慣透過網際網路交換他們所需要的資訊，這可從資策會的 2013 年調查顯示民眾所從事過的網路活動以「管理使用個人網誌、微網誌或社交網絡」63.2% 的比例最高，瀏覽「網路論壇」也有 27%。這份調查顯示民眾已習慣在網路上做資訊分享和交換訊息，其中包括在網路上購買各類商品的抱怨文。

根據 PC HOME (2014) 針對全台消費者進行的「奧客大調查」，從調查分析中發現，現在二十四歲以下的年輕族群們，嘴巴長在滑鼠上，願意當面向客訴中心反應僅四成。這群年輕的安靜不語型奧客，有 81% 會使用部落格、MSN、BBS、論壇討論區等管道，透過網路力量來抒發不滿的情緒。即使現今企業雖在自己的網站設有客戶意見服務區與客服專線，但對於消費者所提供的意見或是投訴產品問題的處理態度並不是很積極，導致消費者將不滿的情緒發洩在論壇的抱怨版面上。論壇上的網友看到了抱怨文之後，對於企業產品的購買慾望會下降，這結果將會使企業慢慢的流失客源。

所以現在越來越多的民眾會直接將自己買的產品有不好的經驗或是滿腹的抱怨公開在網路上，這即是在網路上的一個客訴行為，而過了一段時間之後，其他人也會針對自己所感興趣或是有相同痛苦經驗的來做討論，這樣的討論程度可視為大家對於這問題的重視程度，另外，網路上的客訴具備著重複的特性，也許有

眾多的客訴，但在性質上卻是相同的。

因此，本研究利用 R 軟體的文字探勘技術，針對台大 PTT 論壇上的網路拍賣的抱怨文進行資料收集、整理、中文斷詞。再進一步找尋抱怨文的關鍵字雲，並且使用 LDA Gibbs 模式找出多篇抱怨文章的關注課題，以及它們之間的關聯。

1.2 研究動機

現今產業競爭逐漸愈加劇烈的狀況，除了在成本、品質、產品上積極改善，也需引進以顧客為主的思維來改善服務。以顧客為出發點，瞭解真正的需求為何，讓顧客的滿意度提高。隨著資訊科技的發展，讓這樣的想法得以具體化。於是，顧客關係管理系統興起，可以有效率地得知顧客的需求。但是，企業目前對於客訴的處理方式，大多數還是處於被動的態度，仰賴著客服中心採用人工方式記錄客戶所反映的問題。但現今民眾喜歡按照自己的方式發表自己的意見，最多特別喜歡在 PTT、BBS、各大論壇、部落格上發表評論，所以漸漸地網路上這樣的客訴行為企業也不得不注意。

分析客訴內容，萃取出顧客所重視的問題，提供給網站經營者參考。顧客意見內容，可以作為企業內溝通的用途，以藉此發現工作人員的態度、網路服務上的問題，並謀求改善。客訴處理得當，顧客成為忠誠客戶的機率大增，可為企業帶來正向口碑廣告減少企業處理客訴時間、金錢等成本，對企業的成長與發展有益。

根據上述所提及的論述，本研究的研究問題如下：

- (一) 根據台大 PTT e-shopping 論壇上的抱怨文，瞭解拍賣網站會面臨到什麼樣的客訴？
- (二) 分析客訴資料可以給拍賣網站管理者什麼樣的資訊？
- (三) 分析客訴資料對拍賣網站管理者的經營管理有何益處？

1.3 研究目的

現今企業除了自己累積許多各類的資料外，再加上論壇上的抱怨（客訴）資料在這些龐大的資料中，存在著許多有用的資訊，可以針對顧客的行為加以分析。隨著消費者的習慣以及想法逐漸改變，顧客們不再像過去的消費者一樣，只願意透過企業所提供的方式來表達他們的意見，而是透過其他的方式（PTT、BBS、論壇、部落格）來抒發自己的意見或是搜尋相關的討論文章。但企業多半還是依賴客服中心相關技術，如：電話語音中心、自動語音服務等等。因此本研究，透過 PTT 網友們提供的客訴資料，然後經過過濾後，透過這些資料來做為一個分析的樣本，不僅可以作為現階段產品的參考資料，以及產品後續階段改進的依據；另一方面，可以有助於企業瞭解當下消費者的想法與需求。

根據以上分析，本研究目的有下列兩點：

- (一) 利用 R 語言對台大 PTT e-shopping 論壇上的 868 篇網拍抱怨文進行文字探勘，以關鍵字雲方法，找出各抱怨文的關鍵詞彙作為探勘的基礎。針對文字探勘所得到的相關詞彙，進一步探勘出可能的關聯。
- (二) 使用 LDA Gibbs 模式找出它們之間的關聯，進一步得到相關的知識，提供給拍賣網站決策參考，以防止它們的客源漸漸的流失掉。

1.4 論文架構

本研究首先確認研究動機與目的，接著進行相關文獻的探討，再利用 R 軟體的文字探勘技術及方法。針對相關方法進行分析後，根據研究結論的依據，並依此對拍賣網站提出具體的管理建議。

第二章 文獻探討

2.1 客訴

客訴，即是顧客抱怨，當顧客在進行交易性的購買行為中，對於商品或是服務感到不滿意時，對提供商品或是服務的企業所產生的抱怨行為便是客訴了。那客訴的定義是甚麼呢？Jacoby & Jaccard (1981)認為「顧客抱怨」是「由於顧客認知到由購買行為所產生的不滿意情感或情緒而引發的反應，沒有不滿意感覺之顧客的反應，就不能稱為顧客抱怨行為」。中村卯一郎(1992)認為「當顧客對於他們一向信賴而又抱著高度期待的商店（或企業）產生精神或物質上的不滿意或憤怒時，就會很容易將之表面化，也就是直接了當地產生抱怨。」。Singh (1988)的實證研究指出，大部份的顧客抱怨均來自於某些購買或產品使用上的不滿意經驗與問題。客訴是很主觀的意識行為，當有些人會對於商品或是服務不滿的時候，便會積極地向企業做申訴的行為，這就是所謂的客訴。

顧客抱怨對企業而言有什麼樣的價值？中村卯一郎(1992)認為，「抱怨並非囉唆、煩人的事或者是顧客存心找碴，而是由顧客內心發出的重要訊息，一種既難得而又重要的訊息」。陳耀茂(1997)根據研究指出，「對於申訴或抱怨的顧客，如果能夠善加處理，且能獲得顧客對於企業後續處理的滿意，則這些顧客對該公司的忠誠度將可能更高於未曾抱怨過的顧客。如果感到不滿意的顧客不提出抱怨，則表示他們已掉頭他顧或是正準備這樣做。顧客會提出抱怨，表示他們仍讓你先有機會保住他們的生意。」

- Fornell & Wernerfelt (1987) 認為，顧客抱怨對於企業的價值在於下列兩點：
- (一) 作為企業內溝通的用途，以藉此發現產品在設計、製造、品管和銷售上的問題，並謀求改善。
 - (二) 廠商可以藉由良好的抱怨處理，將不滿意的顧客轉變為滿意且更具忠誠的顧客。

佐藤知恭(1988)認為將會引起顧客不滿意與抱怨的原因有如下三點：

- (一) 起因於企業方面的問題：係由於企業越來越趨巨大，組織更加複雜，與顧客在人性上的接觸變成極為淡薄的關係，才會引起許多問題的發生，亦即買方與賣方之間的距離變得非常疏遠。
- (二) 起因於顧客方面的問題：即顧客價值觀已經發生了變化，但是，企業卻未能即時體認到此一變化，也未能及時提出與執行因應的對策，而致使顧客認為企業並未能真正跟得上他們改變後的需要。
- (三) 一開始就要從事正確的工作：儘管如何研究消費者的意識及其行為，同時也具備了完善的消費者部門，假如企業本身一開始即不好好地從事正確的工作，就會引起顧客的抱怨，那是理所當然的事。

另外，Kolter (1997)認為依照觀點的不同，顧客抱怨原因大致上可分成以下三大導向：過程時點、人員責任和服務機能。

以過程時點導向的觀點來看，Renoux (1973)將消費者對零售業者的抱怨依發生時點的先後劃分為：售前的銷售系統、售中的購買系統、售後的消費系統。Westbrook (1981)更依此架構針對大型傳統百貨公司在購買系統中的顧客滿意起因細分成：店員、店鋪環境、商品政策、服務定位、商品/服務、常客、價值/價格關連性、特別折扣等 8 大項。

以人員責任導向的觀點來看，Bitner et al. (1990)針對航空、旅館及餐飲業，探討消費者的不滿意主要是來自於服務傳送系統之員工失誤、消費者需求反應導致員工之失誤及員工自發性行為三大原因；此法在探討服務失誤類型上常被引用，如 Kelley et al.(1993)研究零售業失誤及挽回的類型、Hoffman et al. (1995)研究餐飲業的服務失誤時皆曾被採用。

以服務機能導向的觀點來按，另有一些研究者以賣場管理的項目進行分類，如 Jacoby & Jaccard (1981)將消費者抱怨分成四類：產品、零售環境、促銷活動及製造商方面；陳淑娟（1995）研究百貨公司賣場設計與現場消費行為關係時，以

人員、商品、空間為賣場管理的三大基本要素；此外，蔣麗君(1999)在探討國內百貨公司顧客抱怨原因的研究中，建立了一個新的分類模式，將探討的切入觀點歸納成：系統、人員及商品等三類失誤，其系統機能係指：店舖賣場內及賣場外的硬體設備及企業服務政策等所造成的顧客抱怨，包括賣場設施不良、停車問題、服務政策失誤、促銷活動失誤、延遲服務等。人員機能則是：指人員在服務過程中的服務失誤所造成的顧客抱怨，包括店員專業度不良、記帳或收款問題、店員欺騙、店員態度不佳、店員缺乏耐心及退換貨失誤等。商品機能：針對商品本身的瑕疵所造成的顧客抱怨，包括售價問題、品質問題等。Burnett et al. (1981)針對百貨公司中對服務人員的抱怨再細分成 6 項：一般態度、禮貌性、習性與外表、銷售技巧、關連性商品的知識及店員的耐性。

2.2 客訴相關研究

近年的客訴分析方法中，大多數是運用關聯規則來做為客訴文件的分析工具、建構 Web Service 服務的整合網站以代理人導向的方式處理客訴問題或是利用統計軟體、AHP 層級分析法問卷方式進行分析與驗證。運用關聯規則的分析工具，依據產業的性質，將客訴資料透過資料探勘等方法來發掘出客訴問題，然後再進一步關聯發生問題的原因，建構出「因應對策資料庫」來做為日後企業的客服單位，在面對客訴時能夠迅速的提供解決問題的相關參考知識。

蔡厚灼（2003）的「客訴文件探勘系統」研究中，便是運用關聯規則以及概念階層的方式分析顧客抱怨及解決方案來得到各種概念之間的關聯性與其中的相關程度，他的研究對客訴的文件進行分析，試著將這些客訴資料中，挖掘出可能的知識，進而能夠瞭解到客訴的原因，除了可以做為企業的決策參考外，也可以提供解決方法讓第一線的客服人員作為處理方式的參考依據。

陳俊達(2005)的「影響客服中心電話服務量之因素探討-以資訊服務業為例」的研究中，主要是以資訊服務業 A 公司的客服中心進行個案研究，在現有的客戶

服務資料庫挑選每日電話服務需求量進行分析。藉由 A 公司三整年度的每日電話服務量的資料，探索電話服務量起伏是否有週期性的表現，特定的外部事件與內部事件對客服中心電話服務量是否有明顯的影響，並以 SPSS 統計軟體進行資料分析與驗證。提出的管理建議，能對 A 公司在客服中心的管理與作業流程上，產生改善的效果。

陳瑞陽（2007）的「智慧型代理人 Web Services 整合平台—客訴下游問題回饋為例」研究中，他提出一個以代理人導向之模糊網路服務，來整合客訴下游問題和回饋流程。他先以代理人導向分析出客訴下游問題回饋流程系統，再建構 Web Service 服務的整合網站，將回饋資料在不同角色之間做整合，例如：問題徵狀，以模糊歸屬函數形式來呈現模糊狀況，並轉換儲存成 XML 格式的 Web service。最後再將 Web service 上的資訊提供給企業參考。

鄭麗珍和賴美惠（2011）的「結合知識地圖之公部門陳訴文件自動化分案系統」研究中，針對政府部門網站內提供「首長信箱」功能進行分析，他的研究先透過訪談方式，發現分案專家的腦中似乎有張各部門工作職掌的知識地圖，可以快速且準確的做好分案工作。因為過去文件分案的研究，都忽略這一塊，而直接用文字探勘的技術來做分類。所以他提出二階段分案處理模式，運用訓練資料及文字探勘技術來建立知識地圖，接著透過新進文件與知識地圖的比對進行文件分案的預測。提出一個智慧型知識地圖建立的機制，來協助政府部門以自動化的方式來進行分案處理民眾的意見，使民眾意見能快速分送至正確的處理單位，並迅速予民眾回應，以提升政府部門的處理績效及節省處理分送案件的人力。

根據以上研究得知，使用隱含狄利克雷分布（LDA）模型理論，再針對客訴來做分析的這方面研究是較少的。

表 2.1 客訴相關研究

作者(年份)	研究問題	研究方法	研究成果
蔡厚灼(2003)	客訴文件探勘系統	文件探勘 關聯規則	提出一個架構挖掘顧客抱怨文件中可能的關聯，以做為企業建立產品知識庫時的參考
陳俊達(2005)	影響客服中心電話服務量之因素探討。	以 SPSS 統計軟體進行資料分析與驗證	對企業在客服中心的管理與作業流程上，產生改善的效益
陳瑞陽 (2007)	智慧型代理人 Web Services 整合平台—客訴下游問題回饋為例	代理人導向、模糊歸屬函數	以代理人導向來分析具有複雜人類行為的產品客訴下游問題回饋流程、將本身具有模糊性的回饋資訊，以模糊歸屬函數形式來呈現其衡量模糊值的內容，並將此內容轉換成 XML 格式的 Web service。
鄭麗珍、賴美惠 (2011)	結合知識地圖之公部門陳訴文件自動化分案系統。	文件探勘 關聯規則	提出一個智慧型知識地圖建立的機制，來協助政府部門以自動化的方式來進行分案處理民眾的意見

資料來源：本研究整理

2.3 文字探勘

文字知識發掘(Knowledge Discovery from Text, KDT)，亦可稱為文字探勘(text mining)或文件探勘(document mining)，其應用了資訊檢索(Information Retrieval, IR)、資訊萃取(Information Extraction, IE)、計算語言學(computational linguistics)、自然語言處理(Natural Language Processing, NLP)、資料探勘技術、知識表示等技術，其中每門技術都有專門的領域，都有相當成熟的發展。但文字探勘特別著重於利用這些技術，自非結構或半結構的文字中，發掘出先前未知、隱含而有用的資訊。Sullivan (2001) 定義文字探勘為「一種編輯、組織及分析大量文件的過程，為了要提供特定使用者特定的資訊，以及發現某些特徵及其間的關聯」。相較於傳統資料探勘，文字探勘需要加上一些額外的資料選擇處理程序，以及較為複雜的特徵萃取步驟。

文字探勘的技術中，最基礎的為語言學分析及自然語言處理，其主要目的是為了辨別出文件的關鍵資訊，如：描述 who、what、when、where 等資料的關鍵字詞，或是文件的概念階層(concept hierarchy)。有了這些能代表文件的關鍵特徵資訊之後，才能進行進一步的分析，以找出隱含而有用的資訊，並有效率表示。圖 2.1 文字探勘技術金字塔顯示了目前常見的文字探勘技術及各技術間的層級關係。

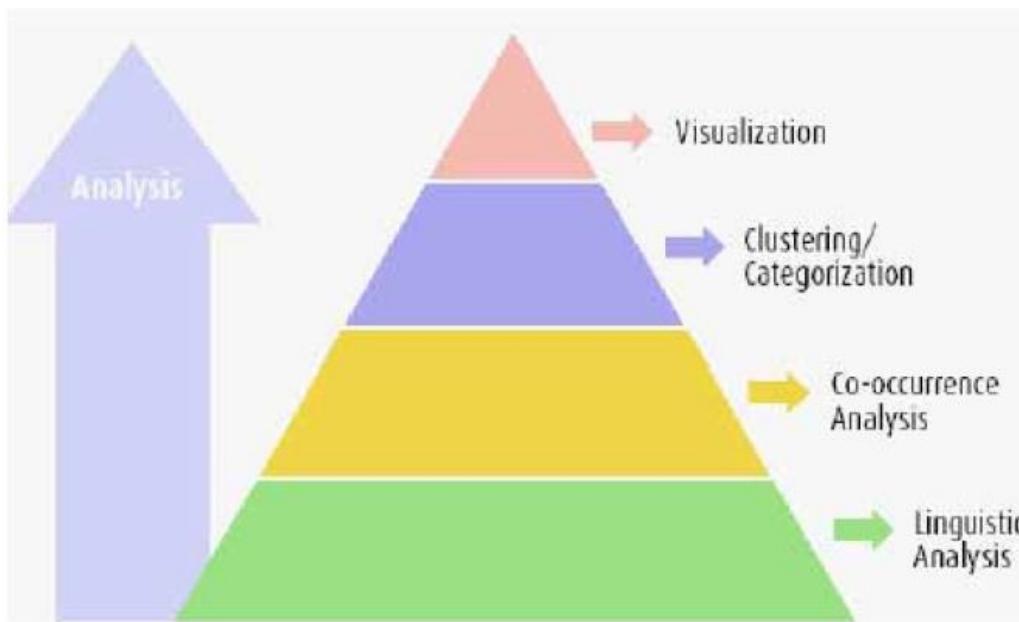


圖 2.1 文字探勘技術金字塔

資料來源：Knowledge Management Systems – A Text Mining Perspective

由於網際網路時代的來臨，網路資訊量的暴增，而這些資訊通常都是以文字的形式存在著，而文字探勘技術就是來處理非結構或是半結構化的資料，以挖掘出文字資料中所隱藏的規則與結構。

除了傳統的資料採礦技術外，文字探勘技術上還需要結合許多不同領域上的技術來配合，如：文字搜尋與分析、知識萃取，機器學習、人工智慧…等，因此文字探勘技術較傳統資料採礦技術複雜且費工。而文字探勘成功與否的關鍵在於語意分析，主要用來辨別文件的關鍵字詞，若能確實找到文件中關鍵字之資訊，則可得到有價值的資訊。

對於文字處理過程首先要擁有分析的文字語言資料庫(text corpus)，比如報告、信函等。而後根據這些語言資料建立半結構化的全文資料庫(text database)。而後生成包含詞性的結構化的資料術語矩陣(term-document matrix)。這個一般性資料結構會被用於後續的分析，比如：文字分類、語法分析、資訊提取和修復、文件資訊匯總。如圖 2.2 所示。

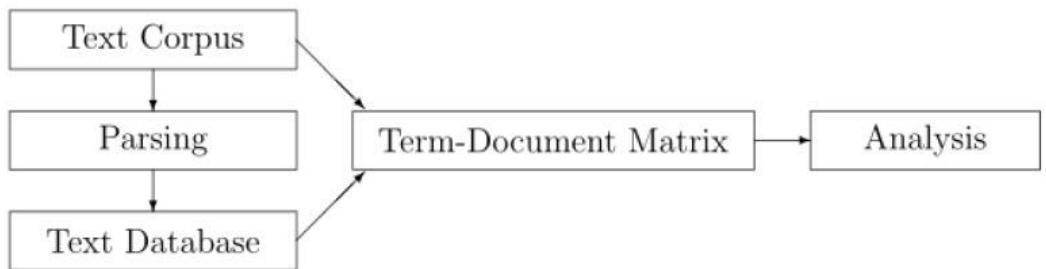


圖 2.2 文字探勘的處理流程 資料來源：Text Mining in R (2012)

2.4 LDA

隱含狄利克雷分布簡稱 LDA (Latent Dirichlet allocation)，為一種統計的主題模型，它可以將文件集中按照機率分布的形式給出每篇文件的主題。並且它是無監督學習演算法，在訓練時不需要標註的訓練集的類別，僅需要準備好文檔集及設定主題的數量 k 即可。對於每一個主題均可找出一些詞語來描述它是 LDA 的另一個優點(Wikipedia : LDA , 2015)。

在 LDA 模型當中假設文件是由一堆的主題按某種機率分佈隨機混合所產生，每一個主題是一個多項式分佈的組合，主題被所有的文件所共享，每一份文件包含各主題的分佈。Blei, et al. (2003)指出 LDA 模型一個主題當中包含有許多的字彙，同時一個文件是由主題的分佈所組合，範例中分別以不同的顏色來做區別，如圖 2.3 所示。

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

圖 2.3 LDA 主題及包含的字彙以顏色區分

資料來源：Latent dirichlet allocation. (Blei, Ng, & Jordan, 2003)

2.4.1 LDA 的貝氏網路結構

蕭文峰(2014)提到，在LDA中每個主題都能找到用以對其描述的字詞。LDA使用了「bag of word」方法，其方法是忽略文件中字詞出現的順序，只考慮文件中字詞出現的頻率，LDA生成模型如圖 2.4 所示。

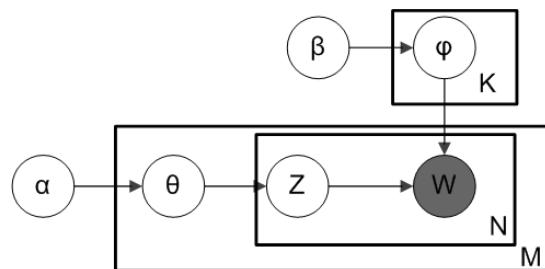


圖 2.4 LDA 的貝氏網路結構

資料來源：Wikipedia - Latent Dirichlet Allocation

以圖 2.4 來說明， α 為每篇文件的 Dirichlet 參數， β 為每個主題的字詞分布

Dirichlet 參數， θ_i 為第 i 篇文件的主題分佈， φ_k 為第 k 個主題的字詞分佈， Z_{ij} 為第 i 篇文件第 j 個字詞所屬的主題， W_{ij} 為最終生成的字詞，M 為所有文件的總數，N 為所有字詞的總數。建立 LDA 主題模型的流程有三：

1. 從該文件的主題分佈 θ 中隨機抽取一個主題 Z。
2. 從主題 Z 所對應的字詞分佈 φ 中抽取一個字詞 W。
3. 重複上述過程至抽取到文件中的每一個字詞為止。

2.4.2 Gibbs 的抽樣方法

Gibbs 抽樣方法其概念是在聯合分佈的未知而單一變數條件機率已知的情況下進行大量抽樣和迭帶計算，樣本會逐漸收斂並接近聯合分佈 (Gibbs Sampling)。Griffths 與 Steyvers (2004) 提出用 Gibbs 抽樣方法執行 LDA 的理論，由於容易執行，已經廣泛運用於許多研究中。LDA 的 Gibbs 抽樣方法流程如圖 2.5 所示。

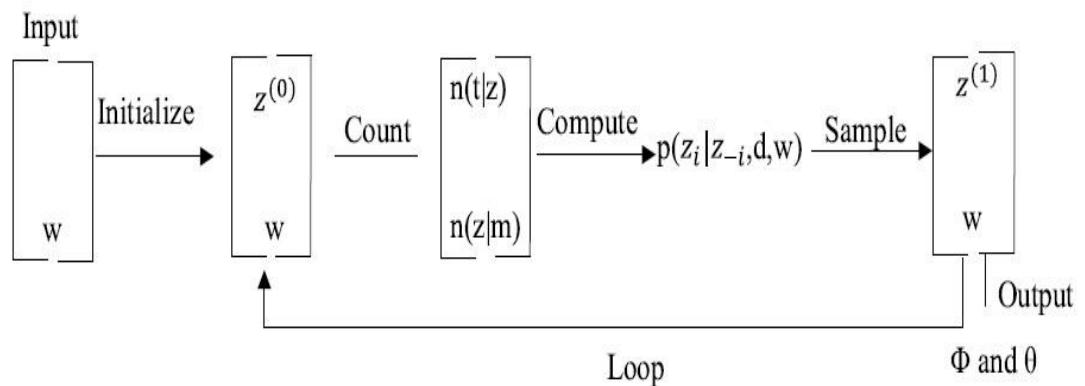


圖 2.5 LDA 的 Gibbs 抽樣方法流程

資料來源：Distributed Gibbs Sampling of Latent Topic Models : The Gritty Details
Technical report. (Wang,2008)

LDA 的 Gibbs 抽樣方法流程有五(概率語言模型及變形系列 LDA 及 Gibbs Sampling)：

1. 將文件中所有字詞隨機分配至主題 $z^{(0)}$ 。
2. 統計每個主題 z 的字詞 t 的數量與每個文件 m 中出現主題 z 中的字詞數

量。

3. 計算排除當前字詞後的主題分配，再藉由其他字詞的主題分配估計當前字詞分配至各主題機率。
4. 根據其機率分佈為該字詞抽樣出新的主題 $z^{(1)}$ 。
5. 重複以上步驟直到每個文件下的主題分佈 θ 和主題一字詞分佈 Φ 收斂為止。

Tang (2008) 指出用 Gibbs 抽樣方式執行 LDA 更加容易且易於擴展，同時能更快獲得極佳的近似值。因此本研究使用 Gibbs 抽樣方法執行 LDA 替中文斷詞過後的文章建立主題模型，以便於找出關鍵詞中的關聯性。



第三章 研究方法

3.1 研究架構

消費者在網路上提出對產品使用的不滿或是抱怨時，無非是想要引起企業重視該層面的問題。因此本研究以客訴資料的觀點出發，以顧客相關的議題為核心，使用 R 語言的技術來探勘顧客抱怨文件。首先，利用 R 軟體對台大 PTT 網站進行爬蟲，抓取所需資料，再進行文件的前處理動作，然後將文件資料提取進 R 裡面，再將文件中所有的詞彙進行中文斷詞，之後將中文斷詞結果利用關鍵詞字雲分析找出關鍵詞，再從 LDA Gibbs 模型找出關聯性，形成一個新的客訴分析模式，最後從分析的結果進行討論。研究架構如圖 3.1。



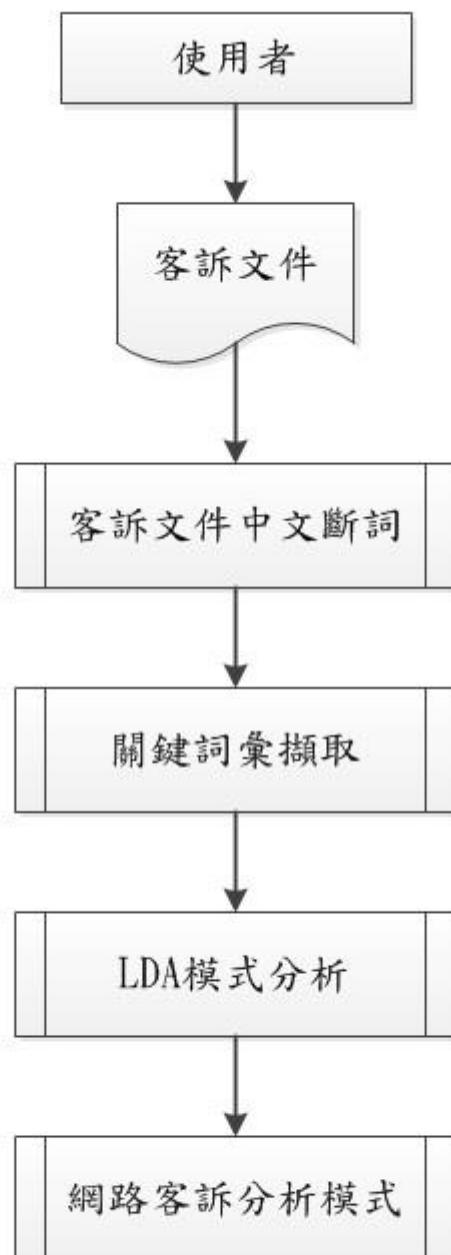


圖 3.1 研究架構

3.2 客訴資料收集方法

3.2.1 R 語言

R 語言，是一種自由軟體程式語言與操作環境，主要用於統計分析、繪圖、資料探勘。R 本來是來自紐西蘭奧克蘭大學的 Ross Ihaka 和 Robert Gentleman 開發，現在由「R 開發核心團隊」。R 是基於 S 語言的一個 GNU 計畫專案，所以也可以當作 S 語言的一種實作，通常用 S 語言編寫的代碼都可以不作修改的在 R 環境下執行。

R 是一套完整的資料處理、計算和製圖軟體系統。其功能包括：資料存儲和處理系統；陣列運算工具（其向量、矩陣運算方面功能尤其強大）；完整連貫的統計分析工具；優秀的統計製圖功能；簡便而強大的程式設計語言；可操縱資料的輸入和輸出，可實現分支、迴圈，使用者可自訂功能。與其說 R 是一種統計軟體，還不如說它是一種數學計算的環境，因為 R 並不是僅僅提供若干統計程式、使用者只需指定資料庫和若干參數便可進行一個統計分析。R 的思想是：它可以提供一些集成的統計工具，但更大量的是它提供各種數學計算、統計計算的函數，從而使使用者能靈活機動的進行資料分析，甚至創造出符合需要的新的統計計算方法。該語言的語法表面上類似 C，但在語義上是函數設計語言（functional programming language）的變種並且和 Lisp 以及 APL 有很強的相容性。特別的是，它允許在“語言上計算”（computing on the language）。這使得它可以把運算式作為函數的輸入參數，而這種做法對統計模擬和繪圖非常有用。

R 是一個免費的自由軟體，它有 UNIX、LINUX、MacOS 和 WINDOWS 版本，都是可以免費下載和使用的。在那兒可以下載到 R 的安裝程式、各種外掛程式和文檔。在 R 的安裝程式中只包含了 8 個基礎模組，其他外在模組可以通過 CRAN 獲得。R 的原代碼可自由下載使用，亦有已編譯的執行檔版本可以下載，可在多種平臺下運行，包括 UNIX(也包括 FreeBSD 和 Linux)、Windows 和 MacOS。R 主要

是以命令列操作，同時有人開發了幾種圖形化使用者介面。R 內建多種統計學及數位分析功能。因為 S 的血緣，R 比其他統計學或數學專用的程式設計語言有更強的物件導向(物件導向程式設計)功能。

R 的另一強項是繪圖功能，製圖具有印刷的素質，也可加入數學符號。雖然 R 主要用於統計分析或者開發統計相關的軟體，但也有人用作矩陣計算。其分析速度可媲美 GNU Octave 甚至商務軟體 MATLAB。R 的功能能夠通過由用戶撰寫的套件增強。增加的功能有特殊的統計技術、繪圖功能，以及程式設計介面和資料輸出/輸入功能。這些套裝軟體是由 R 語言、LaTeX、Java 及最常用 C 語言和 Fortran 撰寫。下載的執行檔版本會連同一批核心功能的套裝軟體，而根據 CRAN 紀錄有過千種不同的套裝軟體。其中有幾款較為常用，例如用於經濟計量、財經分析、人文科學研究以及人工智能。

3.2.2 RCurl

Curl (2009) 是一種被設計來編寫網路程式的程式語言。它並不像 HTML 一樣是屬於一種文本標記語言，但 Curl 語言卻可以用於普通的文本顯示，又可以用於實現大規模的客戶端商業軟體系統。

RCurl 是 R 語言用來抓取網路資料的爬蟲工具，它是由 UC Davis 分校的 Duncan Temple Lang 副教授開發出來，他是根據 Curl 語言來撰寫，功能包含了：獲取網路頁面、網路認證、上傳下載、資訊搜索等。但在使用前需先安裝 Java 與 R Studio 開發工具才可使用。RCurl 廣泛用於在網際網路上抓取各種數據資源，例如金融資料，行情資料，價格資料，體育資料等，用於進行後續建模分析。資訊時代，網際網路已經毫無置疑地滲透到我們生活工作的各個細節，因此它本身亦成為日益重要的大資料來源。

本研究使用上述所提的 RCurl 套件，將台大 PTT 網站的 e-shopping 上的 2011 年 1 月到 2015 年 6 月的 50444 篇文章擷取下來。經過人力整理過後，將文章前面

主題有「抱怨」的挑選出來，共有 868 篇文章。程式流程如圖 3.2、圖 3.4。

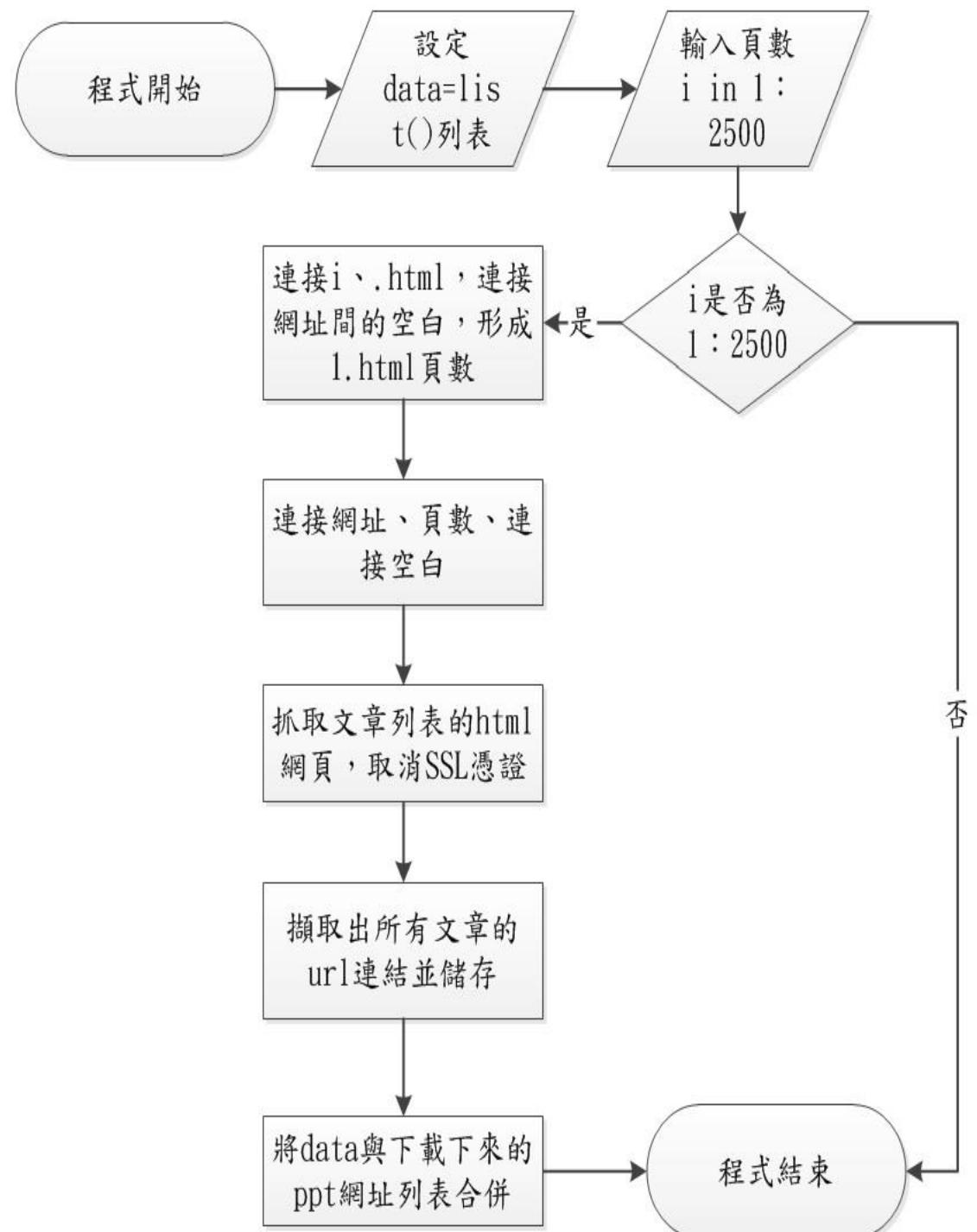


圖 3.2 抓取文章網址程式流程圖

抓取文章網址程式碼如下表所示：

表 3.1 抓取文章網址程式碼

```
1. for( i in 1:2500){  
2.   tmp <- paste(i, '.html', sep="")  
3.   url<- paste('https://www.ptt.cc/bbs/e-shopping/index',tmp,sep="")  
4.   html <- htmlParse(getURL(url,ssl.verifyPeer = FALSE))  
5.   url.list<-xpathSApply(html,"//div[@class='title']/a[@href]",xmlAttrs)  
6.   data <- rbind(data, paste('https://www.ptt.cc',url.list,sep=""))  
7.   data <- unlist(data)
```

首先使用 for 迴圈設定 i 要抓取的網頁頁數為多少。接著利用 paste 連接 i(頁數)、'.html'、然後使用 sep="將空白處連起來。再來我們需要利用 paste 連接 PTT 網址與 tmp，才能抓取到我們要的網址頁數，然後利用 htmlParse 抓取文章列表的 html 網頁與取消 SSL 憑證，接著使用 xpathSApply 檢取出所有文章的 url 連結並儲存。最後使用 rbind 將 data 與列表合併，再使用 paste 將'https://www.ptt.cc'與 url.list 連接起來，接著將 data 變為字串，即可抓取到文章列表。存取到的網址如圖 3.3。

```
Console C:/Users/test/Desktop/test/  
> data  
[1] "https://www.ptt.cc/bbs/e-shopping/M.1433262870.A.B5B.html"  
[2] "https://www.ptt.cc/bbs/e-shopping/M.1433330893.A.40F.html"  
[3] "https://www.ptt.cc/bbs/e-shopping/M.1433263518.A.3B4.html"  
[4] "https://www.ptt.cc/bbs/e-shopping/M.1433330958.A.A12.html"  
[5] "https://www.ptt.cc/bbs/e-shopping/M.1433264665.A.63C.html"  
[6] "https://www.ptt.cc/bbs/e-shopping/M.1433331186.A.36B.html"  
[7] "https://www.ptt.cc/bbs/e-shopping/M.1433264912.A.57E.html"  
[8] "https://www.ptt.cc/bbs/e-shopping/M.1433331832.A.FE6.html"  
[9] "https://www.ptt.cc/bbs/e-shopping/M.1433266268.A.7A1.html"  
[10] "https://www.ptt.cc/bbs/e-shopping/M.1433332006.A.6A7.html"  
[11] "https://www.ptt.cc/bbs/e-shopping/M.1433292254.A.1F3.html"  
[12] "https://www.ptt.cc/bbs/e-shopping/M.1433336819.A.C0E.html"  
[13] "https://www.ptt.cc/bbs/e-shopping/M.1433294182.A.D84.html"  
[14] "https://www.ptt.cc/bbs/e-shopping/M.1433338138.A.68A.html"  
[15] "https://www.ptt.cc/bbs/e-shopping/M.1433297097.A.BDD.html"  
[16] "https://www.ptt.cc/bbs/e-shopping/M.1433341017.A.961.html"  
[17] "https://www.ptt.cc/bbs/e-shopping/M.1433301568.A.404.html"  
[18] "https://www.ptt.cc/bbs/e-shopping/M.1433342323.A.C29.html"  
[19] "https://www.ptt.cc/bbs/e-shopping/M.1433301672.A.76B.html"  
[20] "https://www.ptt.cc/bbs/e-shopping/M.1433342508.A.5BF.html"
```

圖 3.3 抓取文章網址結果

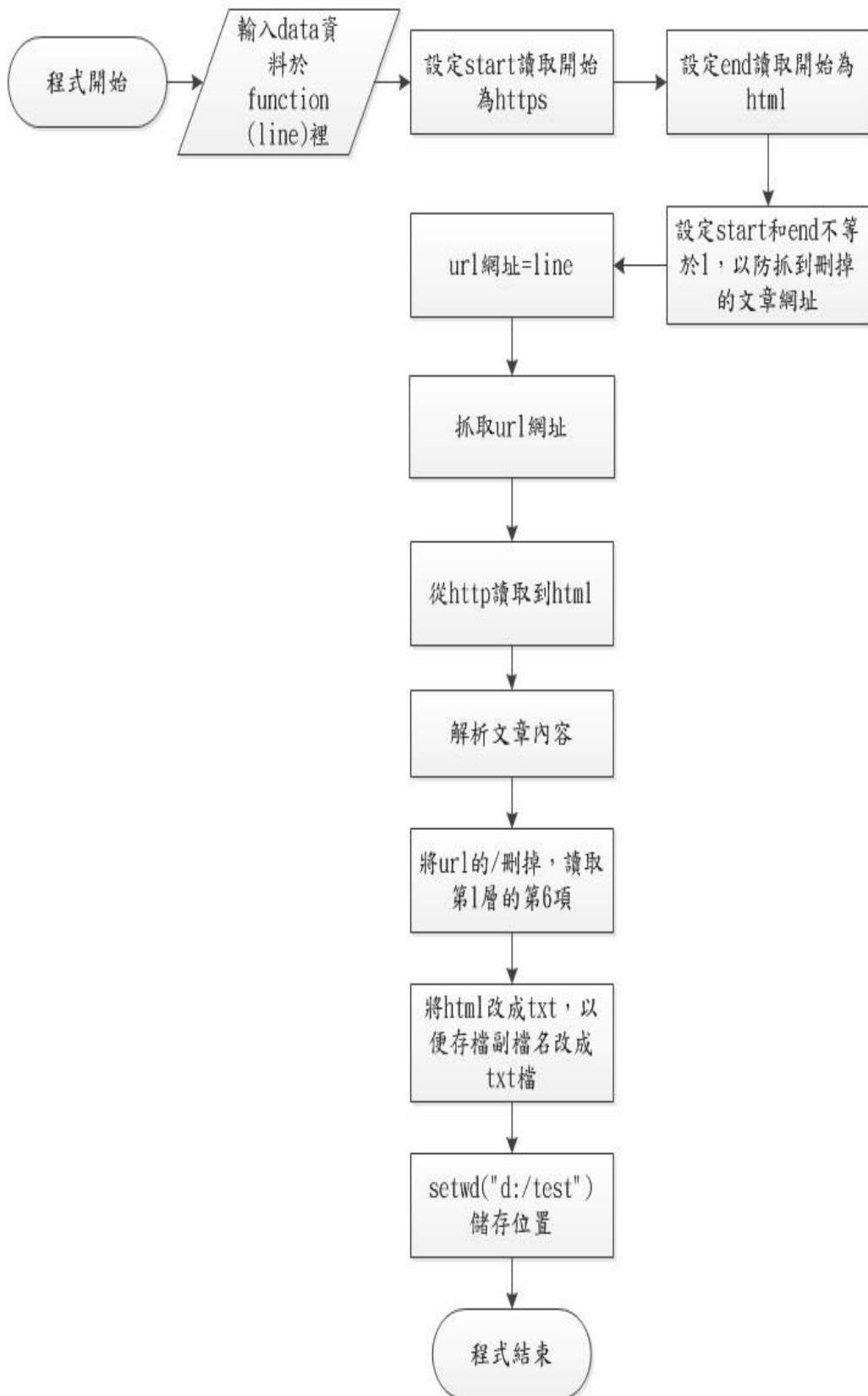


圖 3.4 撈取文章內容程式流程圖

擷取文章內容程式碼如下表：

表 3.2 擷取文章內容程式碼

```
1. getdoc <- function(line){  
2.   start <- regexpr('https', line)[1]  
3.   end <- regexpr('html', line)[1]  
4.   if(start != -1 & end != -1 ){  
5.     url <- line  
6.     html <- htmlParse(getURL(url, ssl.verifyPeer = FALSE), encoding='UTF-8')  
7.     url <- substr(line, start, end+3)  
8.     doc <- xpathSApply(html, "//div[@id='main-content']", xmlValue)  
9.     name <- strsplit(url, '/')[[1]][6]  
10.    write(doc, gsub('html', 'txt', name))  
11.    setwd("d:/test")
```

首先我們先設定 line 函數內容，接著利用 regexpr 設定 start 與 end 讀取開始為 https 和 html，後面的[1]表示從第一張列表讀取。再來將 start 與 end 設定不等於-1，以防抓到刪掉的文章網址。我們利用 htmlParse 抓取 url 網址，encoding 設定為 'UTF-8'，接著設定 url 網址為 line，再來利用 substr 從 https 開始讀取到 html，end 如果不+3 只會讀取到 h 就停止。接著利用 xpathSApply 解析 html 裡的文章內容，然後使用 strsplit 將 url 分割，以'/'做分界，[[1]][6]表示讀取第 1 層的第 6 項。例如："https:", " ", "www.ptt.cc", "bbs", "e-shopping","M.1432993278.A.063.txt"。最後用 write 寫入 doc 文章內容，將副檔名 html 替換成 txt，設定儲存位置之後，即可擷取文章內容。擷取好文章如圖 3.5，經整理後有 868 篇可用抱怨文。

The screenshot shows a Windows desktop environment. At the top, there is a taskbar with several icons. Below the taskbar, a file list is displayed in a table format. The table has three columns: file name, modification date, and file size. The files listed are all text files (txt) from a folder named 'M.1323848779.A.C9B'. The modification date for all files is '2015/6/18 下午 0...'. The file sizes range from 1 KB to 8 KB. Below this table, a 'Notepad' window is open with the title 'M.1323848779.A.C9B - 記事本'. The window contains text in Chinese. The text discusses the user's experience with Taobao sellers, mentioning issues like bad attitude and requests for offline transactions, which the user considers fraudulent. It also notes the safety of Taobao transactions due to the involvement of Alipay.

M.1323848779.A.C9B	2015/6/18 下午 0...	文字文件	5 KB
M.1337266649.A.9BC	2015/6/18 下午 0...	文字文件	7 KB
M.1352740096.A.6A8	2015/6/18 下午 0...	文字文件	3 KB
M.1358245840.A.15F	2015/6/18 下午 0...	文字文件	2 KB
M.1358251801.A.588	2015/6/18 下午 0...	文字文件	8 KB
M.1358256603.A.BDB	2015/6/18 下午 0...	文字文件	4 KB
M.1358266723.A.A4F	2015/6/18 下午 0...	文字文件	2 KB
M.1358353803.A.2D6	2015/6/18 下午 0...	文字文件	4 KB
M.1358356442.A.288	2015/6/18 下午 0...	文字文件	3 KB
M.1358357880.A.AD7	2015/6/18 下午 0...	文字文件	1 KB
M.1358393919.A.85A	2015/6/18 下午 0...	文字文件	2 KB
M.1358406146.A.F75	2015/6/18 下午 0...	文字文件	3 KB
M.1358409699.A.00D	2015/6/18 下午 0...	文字文件	2 KB
M.1358417974.A.F74	2015/6/18 下午 0...	文字文件	2 KB
M.1358497898.A.125	2015/6/18 下午 0...	文字文件	4 KB
M.1358503886.A.EA5	2015/6/18 下午 0...	文字文件	3 KB
M.1358513346.A.85A	2015/6/18 下午 0...	文字文件	9 KB
M.1358527933.A.03F	2015/6/18 下午 0...	文字文件	5 KB
M.1358754461.A.659	2015/6/18 下午 0...	文字文件	2 KB
M.1358768107.A.9C4	2015/6/18 下午 0...	文字文件	5 KB

圖 3.5 撷取文章結果(部分)

3.3 關鍵字詞搜尋

本研究使用 R 語言裡面的 tm 套件對 868 篇客訴進行文字探勘，對於中文環境

還需要使用一些套件來處理中文字符。在進行文字探勘前，得先使用 install.package 指令來安裝套件。如圖 3.6 讀取 text mining 套件。

```
library(tm)
library(tmcn)

## # tmcn Version: 0.1-2

library(Rwordseg)

## Loading required package: rJava
## # Version: 0.2-1
```

圖 3.6 text mining 套件 資料來源：Text Mining in R (2012)

3.3.1 資料提取

在文字探勘 (Tm) 中主要的管理檔的結構被稱為語言資料庫 (Corpus)，代表了一系列的文檔集合。語料庫是一個概要性的概念，在語料庫構成中，必須有一個說明資料來源 (input location) 的來源對象 (Source Object)。對於這些資料來源，Tm 套件提供了一些相關的函數，例如：

- (一) DirSource：處理目錄。
- (二) VectorSource：由文檔構成的向量。

完整信息的提取需要使用 inspect()，如以下程式碼所示。

表 3.3 檔案提取進 R 的程式碼

1. test<-Corpus(DirSource('c:/test'),readerControl=list(language = "CN"))
2. inspect(test)

先從Corpus中使用DirSource指令提取放置c:/test的資料，再利用readerControl的中文語系讀取資料，最後使用inspect()呈現結果。如圖3.7。

```

Console ~/test/ ↵
$M.1434268606.A.498.txt
作者Talomu (Talomu)看板e-shopping標題[抱怨] 露天拍賣帳號被盜，請幫我檢舉我自己時間sun Jun 14 15:56:43 2015

*商品/店家名稱：LV包包 香奈兒包包GUCCI包 斜跨包 牛皮 齊格 老花 咖啡格 側包
後背包 男包 女包 水桶包

*購買連結：http://goods.ruten.com.tw/item/show?21524798559760

*相關佐證：http://i.imgur.com/l1yb3wb.jpg
http://i.imgur.com/2cjozyc.jpg

*事發經過：下午兩點五十分左右發現一直有商品遭下標，但我根本不會賣東西
露天對我來說只會用來買東西，開上去便發現被盜了，請大家幫忙檢舉我
讓我帳號被停權吧，麻煩大家了。

-- 
※ 發信站：批踢踢實業坊(ptt.cc)，來自：123.194.61.69
※ 文章網址：https://www.ptt.cc/bbs/e-shopping/M.1434268606.A.498.html
推 error123：每一頁都有露天被盜是怎樣，也太頻繁 59.115.235.169 06/14 16:57
→ error123：露天是不是流資料出去= = 59.115.235.169 06/14 16:58
推 comomo：這幾個月我也被盜了2、3次 每次帳號都被111.241.155.201 06/14 17:50
→ comomo：鎖111.241.155.201 06/14 17:50
推 windlan：前幾天被盜+1203.217.101.127 06/14 18:07
推 vicky9884：前幾天被盜+1 寫信給客服鎖帳號了123.192.128.163 06/14 19:18
推 rabbitlock：前天也被盜+1 101.11.24.156 06/14 20:22
推 crystal601：露天也太誇張了～111.240.179.158 06/14 22:17
推 yeah0206：前個禮拜也被盜，超無言 59.115.97.204 06/15 19:32

$M.1434376007.A.DFA.txt
作者feeling13 (KOBE)看板e-shopping標題[抱怨] 小禁電器等了半個月的經驗時間Mon Jun 15 21:46:42 2015

*商品/店家名稱：Yahoo小禁電器

*購買連結：http://Orz.tw/xMn1J

*相關佐證：

*事發經過：

6/1 在該賣場下標除濕機並當天匯款
6/2 收到該賣場出貨通知

這段時間因遇到下雨很想快點買除濕機來使用，等到6/7尚未收到

詢問是否出貨，該賣家回覆如下

我：您好，週二收到出貨通知，但到今日尚未收到商品，可否幫忙查詢一下貨物運送狀況
呢

賣家：您好~6/4以拉貨 預估周一至二拉出喔!!小缺貨
(原文附上 沒修錯字)

```

圖3.7 檔案提取進R軟體裡的結果

3.3.2 中文斷詞

在英文結構中，每個單字都是獨立的，但中文的字元之間沒有空格連接。比如對於“花為什麼這樣紅”，包含了7個字元，雖然對於任意一個台灣人都可以分清這句話應該怎樣斷句，但對於電腦卻有些困難。因此，利用R語言裡的Rwordseg套件的segmentCN來進行斷詞。利用rJava去連結java分詞工具ansj來進行斷詞。另外，斷詞後的詞彙有詞性，例如動詞、名詞、形容詞、介係詞等，本研究只挑出名詞來進行分析。但在進行中文斷詞前，得先清除標點符號、數字、英文單字。

如下表程式碼所示。

表3.4 清除標點符號、數字、英文單字程式碼

1. test <- tm_map(test, removePunctuation)
2. test <- tm_map(test, removeNumbers)
3. test <- tm_map(test, function(word) {gsub("[A-Za-z0-9]", "", word)})

tm_map的作用是將參數function套用到語料庫中的每一篇語料。這個tm_map的好處是讓我們可以直接對一個語料庫的所有文章做操作，而不必使用迴圈，而在R語言裡，迴圈正是我們要避免的事，因為R對迴圈運算是很慢的。數據清理完畢之後，即可進行中文斷詞的動作。如下表程式碼所示。

表3.5 中文斷詞程式碼

1. test <- tm_map(test[1 : 868], segmentCN, nature = TRUE)
2. test <- tm_map(test, function(sentence) {
3. noun <- lapply(sentence, function(w) {
4. w[names(w) == "n"]
5. })
6. unlist(noun)
7. })
8. test <- Corpus(VectorSource(test))
9. inspect(test)

首先先讀取1到868篇文章，再使用segmentCN中文斷詞指令，接著從function(sentence)裡進行斷詞斷句。接著我們設定字詞只挑出名詞來進行分析，將名詞變為字串。然後利用向量提取方式呈現中文斷詞，最後再以inspect()指令將最初的中文斷詞呈現。經本研究自行整理如圖3.8所示。

```

"一側", "一連串", "一場空", "一審", "一鍋粥", "一體", "九州", "二話", "二線",
"人人", "人力", "人士", "人工", "人才", "人心", "人手", "人文", "人生", "人名", "人形", "人身", "人事", "人性", "人物", "人
人格", "人氣", "人馬", "人情味", "人造毛", "人間", "人群", "人像",
"人稱", "人數", "人頭", "人臉", "人證", "人類學", "人權", "人體", "八字",
"八卦", "刀子", "刀片", "刀把", "刀痕", "力度", "力氣", "力量", "十字",
"十字架", "廣家", "三合院", "三角", "三角褲", "三明治", "三星", "三部曲", "三節",
"三葉草", "三聯單", "上文", "上司", "上身", "上帝", "上限", "上班族",
"上級", "上將", "上集", "上圖", "上層", "下手", "下文", "下官", "下雨天",
"下限", "下風", "下馬威", "下場", "下期", "下款", "下策", "下集", "下落",
"下圖", "下篇", "下線", "下擺", "凡事", "千古", "口吻",
"口角", "口味", "口紅", "口音", "口感", "口碑", "口置",
"口頭", "口頭禪", "土匪", "大人", "大力", "大公", "大戶", "大手筆",
"大火", "大片", "大王", "大件", "大任", "大名", "大安", "大米",
"大衣呢", "大作", "大忌", "大我", "大事", "大使", "大叔", "大姐", "大幸",
"大拇指", "大門", "大雨", "大便", "大城市", "大洋", "大要", "大哥", "大哥大",
"大帥", "大氣", "大事", "大理石", "大略", "大眾", "大麻", "大喜",
"大筆", "大衛", "大隊", "大雅", "大集", "大媽", "大意", "大會", "大業",
"大爺", "大義", "大腦", "大腳", "大號", "大過", "大道", "大綱", "大腿",
"大餅", "大寫", "大樓", "大獎", "大器", "大學", "大學生", "大錢", "大頭",
"大頭針", "大餐", "大嬸", "女人", "女子", "女友", "女王", "女生", "女巫",
"女兒", "女性", "女朋友", "女娃", "女神", "女童", "女裝", "女聲",
"子公司", "子彈", "小人", "小心眼", "小戶", "小毛", "小生", "小件", "小企業",
"小字", "小米", "小妞", "小弟", "小事", "小妹", "小姑", "小朋友",
"小雨", "小便", "小便宜", "小孩", "小孩子", "小屋", "小巷", "小個子", "小家子氣",
"小家電", "小時候", "小鬼", "小偷", "小動作", "小區", "小康", "小組",
"小鳥", "小麥", "小幅", "小費", "小腳", "小葉", "小農", "小腿", "小說",

```

圖3.8 最初的中文斷詞結果(部份)

斷詞結束之後，接著進行清除停用字符，停用字符指的是一些文章中常見的單字，但卻無法提供我們資訊的冗字。如下表程式碼。

表3.6 清除停用字符程式碼

- myStopWords <- c(stopwords(), "編輯", "時間", "標題", "發信", "實業", "作者",
",商品", "店家名稱", "購買連結", "相關佐證", "事發經過")
- test <- tm_map(test, removeWords, myStopWords)

stopwords() 指令為設定停用字符，設定完之後使用 removeWords 移除 myStopWords 裡的字符。清除停用字符完之後，再將中文斷詞結果，利用矩陣的方式統計字詞出現的數量，之後再呈現出來，如下列程式碼所示。

表3.7 矩陣統計字詞數量程式碼

- tdm <- TermDocumentMatrix(test, control = list(wordLengths = c(2, Inf)))

TermDocumentMatrix 為矩陣的指令，wordLengths 為設定字詞數為兩個字以上方可呈現。經本研究使用 excel 整理如圖3.9所示，圖中的0代表著字詞沒出現在該篇文章裡。

上衣	心得	文章	外套	白色	衣服	物品	感覺
1	4	1	10	3	0	1	3
1	2	1	5	0	0	1	1
1	2	1	1	0	0	1	2
1	0	3	0	0	2	2	2
1	2	1	2	2	0	1	3
2	2	1	0	4	2	1	4
1	0	1	0	4	0	0	1
1	0	2	0	0	3	0	1
1	3	1	0	0	3	1	1
2	2	2	5	2	7	1	0
7	3	1	0	3	4	1	4
1	2	1	4	0	0	1	1
2	2	1	0	8	4	1	0
3	3	1	3	0	2	1	0
2	0	2	2	5	2	0	1
8	2	1	0	3	0	1	2
16	2	1	0	0	1	1	0
4	0	2	2	0	0	0	0

圖3.9 矩陣統計字詞出現的數量出現結果(部分)

最後利用指令findFreqTerms(tdm, 1)，來查看最終中文斷詞後的結果。如圖3.10。

> findFreqTerms(tdm, 1)	"一家"	"一帶"	"一連串"	"一體"	"人力"	"人工"	"人才"	"人手"	"人民幣"
[1]	"人生"	"人名"	"人身"	"人品"	"人員"	"人家"	"人格"	"人頭"	"人權"
[10]	"刀片"	"刀把"	"力氣"	"力量"	"十字"	"上衣"	"上限"	"下馬威"	"下集"
[19]	"下落"	"下標"	"下擺"	"千里馬"	"口子"	"口氣"	"口袋"	"大小"	"大火"
[28]	"大件"	"大任"	"大名"	"大衣"	"大我"	"大拇指"	"大哥"	"大略"	"大陸"
[37]	"大雅"	"大意"	"大腿"	"大學"	"女巫"	"女兒"	"女孩"	"女神"	"女裝"
[46]	"小人"	"小戶"	"小毛"	"小件"	"小弟"	"小事"	"小姐"	"小便宜"	"小孩"
[55]	"小時"	"小腳"	"小道"	"小劇場"	"小錢"	"小雞"	"工作室"	"工作量"	"工程"
[64]	"工廠"	"工讀生"	"不幸"	"中心"	"中央"	"中盤"	"內心"	"內衣"	"內容"
[73]	"內褲"	"公主"	"公司"	"公告"	"公函"	"分行"	"分機"	"化學"	"午餐"
[82]	"友人" "鏡"	"天才"	"天使"	"天空"	"天氣"	"天涯"	"天理"	"天數"	"太陽眼
[91]	"尺寸"	"尺碼"	"幻想"	"心力"	"心得"	"心情"	"心理"	"心態"	"戶名"
[100]	"戶頭"	"手工"	"手法"	"手段"	"手記"	"手袋"	"手腕"	"手機"	"手臂"
[109]	"手續"	"手續費"	"文件"	"文字"	"文章"	"文說"	"斗篷"	"方向"	"方式"
[118]	"方法"	"方面"	"日期"	"月亮"	"木條"	"木頭"	"比例"	"比價"	"毛巾"
[127]	"毛孔"	"毛衣"	"毛料"	"毛襯"	"水桶"	"火焰"	"牙齒"	"牛仔"	"牛仔褲"
[136]	"牛皮"	"牛角"	"牛馬"	"世界"	"主人"	"主角"	"主意"	"主觀"	"代表"
[145]	"代號"	"代價"	"代謗"	"出錯率"	"功能"	"包裝"	"包裝袋"	"包裝箱"	"包裹"
[154]	"半身"	"半圓"	"卡片"	"句號"	"台幣"	"右手"	"司機"	"四季"	"外套"
[163]	"外觀"	"奶油"	"左右手"	"左腳"	"巧克力"	"平台"	"平底鞋"	"平版"	"平面"
[172]	"本事"	"本事"	"正品"	"正負"	"正題"	"民宅"	"生日"	"生平"	"生氣"
[181]	"生理"	"生意"	"用字"	"用品"	"用意"	"白色"	"白花"	"白眼"	"白痴"
[190]	"白髮"	"皮衣"	"皮夾"	"皮革"	"皮膚"	"皮膚科"	"皮質"	"目的"	"立場"

圖3.10 最終中文斷詞結果

3.3.3 關鍵字詞雲

利用wordcloud套件，畫出在所有文件中出現次數最多的名詞。出現次數越多，字體就會呈現越大。如下表程式碼所示。

表3.8 關鍵字詞雲程式碼

1. m1 <- as.matrix(tdm)
2. v <- sort(rowSums(m1), decreasing = TRUE)
3. d <- data.frame(word = names(v), freq = v)
4. wordcloud(d\$word, d\$freq, min.freq = 10, random.order = F, ordered.colors = F,
5. colors = rainbow(length(row.names(m1))))

再畫關鍵字詞雲之前，我們須將tdm矩陣代入m1裡，然後利用sort指令將文字出現次數由小排到大，接著將文字詞頻代入d。最後設定min.freq最小值多少以下不呈現，顏色採用rainbow，之後即可畫出關鍵字詞雲。圖3.11為關鍵詞雲。



圖3.11 關鍵詞雲

3.4 LDA

3.4.1 詞頻-反轉文件頻率 (TF-IDF)

為防止出現高詞頻被高估和低詞頻被低估的問題，所以在進行 LDA 模式之前，需先計算 TF-IDF。TF-IDF (1988) 是一種用於資訊檢索與文字探勘的常用加權技術。它是一種統計方法，用以評估一個字詞對於一篇文件集或一個語料庫中的其中一份文件的重要程度。字詞的重要性隨著它在文件中出現的次數成正比增加，但同時會隨著它在語料庫中出現的頻率成反比下降。以下為 R 語言裡計算 TF-IDF 的方式，如下表程式碼所示。

表 3.9 計算 TF-IDF 的程式碼

```
1. dtm2 <- DocumentTermMatrix(test, control = list(wordLengths = c(2,Inf)))  
2. dtm = dtm2  
3. term_tfidf <- tapply(dtm$v/row_sums(dtm)[dtm$i], dtm$j, mean) *  
   log2(nDocs(dtm)/col_sums(dtm > 0))  
4. t = term_tfidf >= quantile(term_tfidf, 0.5)  
5. dtm <- dtm[,t]  
6. dtm = dtm[row_sums(dtm)>0, ]  
7. term_tfidf
```

首先先將 test 轉成 DocumentTermMatrix 矩陣，然後利用 tapply 指令計算 TF-IDF 公式。TF-IDF 公式為 (3.1) 式。

$$D_{ij} = \text{TF} \times \text{IDF} \quad (3.1)$$

- $D_{i,j}$ 表示某一詞彙 j 在文件 i 中的重要性。TF 為詞頻、IDF 為反轉文件頻率。

計算完後使用 quantile 指令取 50% 的數值，避免太高或太低，因為會有高估與低估詞頻的情況發生。接著將 t 放入 dtm 矩陣裡，然後 row_sums(dtm)>0 避免有負數發生。之後即可呈現出 TF-IDF 為多少。請參閱附錄 A 關鍵字出現次數與 TF-IDF

計算結果。

3.4.2 LDA Gibbs 模式

TF-IDF 計算完之後，再來利用 R 語言裡的 topicmodels 套件，這個套件是隸屬於 tm 套件裡的一部分，它主要是用來繪製 LDA 的三種模型 VEM、Gibbs、CTM。VEM 模型主要採用最大期望算法進行分析，使用 VEM 的話，它會直接將出現次數最多的當 topic 進行關聯，這樣會出現某些出現次數較多的字詞，重要性程度出現高估的情況發生。CTM 模型主要是將一篇文章當成一個主題來找出各個字詞的關聯性，但分析的文章一多，則將會有主題相同而衝突到的情況發生。Gibbs 則是經過 TF-IDF 計算而不會出現有字詞被高估的情況，它也不會侷限在一篇文章為一個主題的限制中。因此本論文採用 LDA Gibbs 的方式，來進行研究。程式流程圖如圖 3.13。

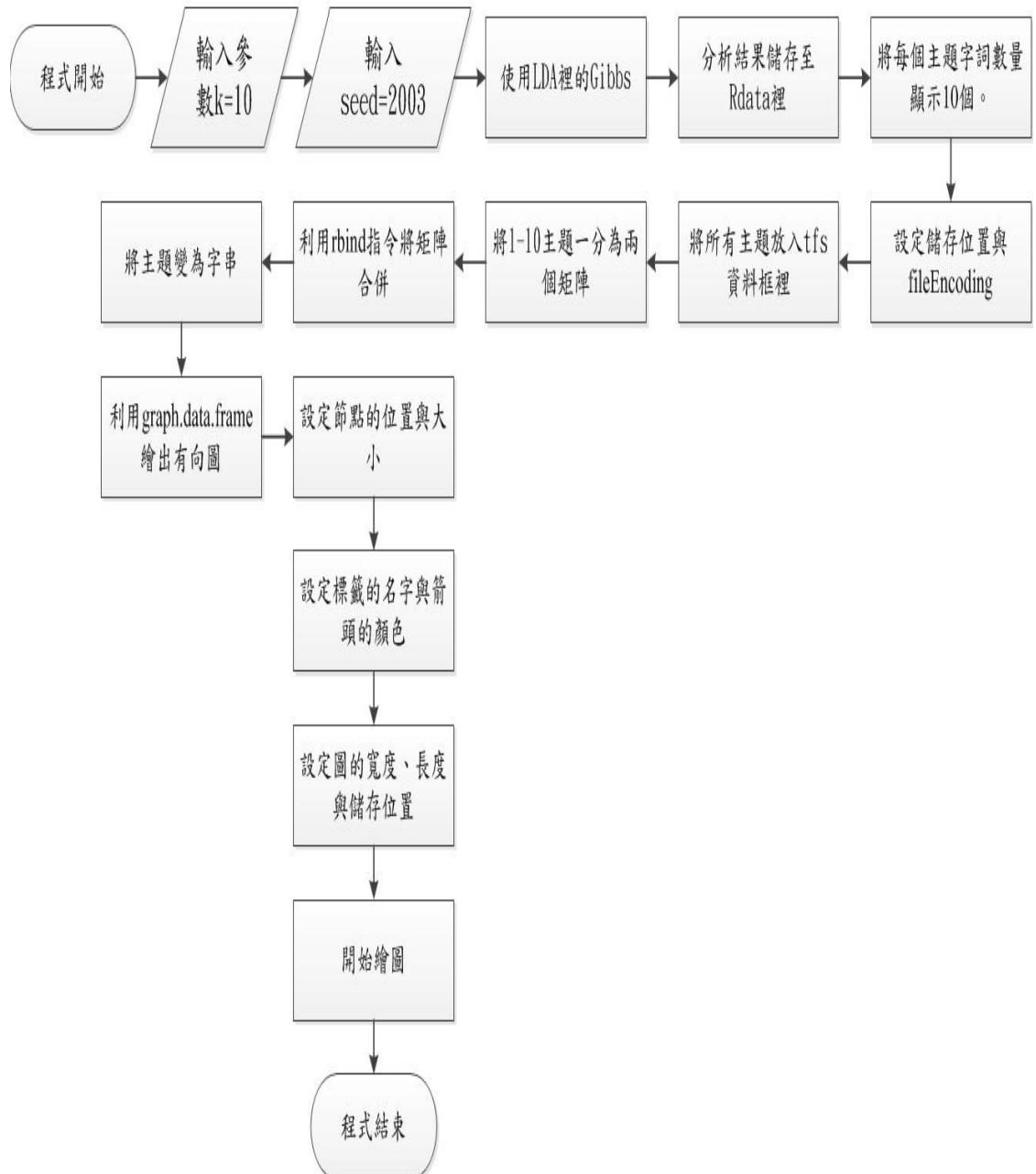


圖3.12 LDA程式流程圖

LDA 程式流程圖如下列程式碼所示。

表 3.10 LDA Gibbs 分析程式碼

```
1. k = 10
2. SEED <- 2003
3. LDA_TM <- list( Gibbs = LDA(dtm, k = k, method = "Gibbs", control = list(seed
4. = SEED, burnin = 1000, thin = 100, iter = 1000)),)
5. save(LDA_TM, file = paste(getwd(), "/LDA_TM.Rdata", sep = ""))
6. termsForSave1<- terms(LDA_TM[["Gibbs"]], 10)
7. write.csv(as.data.frame(t(termsForSave1)), paste(getwd(), "/topic-document_",
8. "_Gibbs_", k, "_2.csv", sep=""), fileEncoding = "UTF-8")
9. tfs = as.data.frame(termsForSave1, stringsAsFactors = F)
10. adjacent_list = lapply(1:10, function(i) embed(tfs[,i], 2)[, 2:1])
11. edgelist = as.data.frame(do.call(rbind, adjacent_list), stringsAsFactors =F)
12. topic = unlist(lapply(1:10, function(i) rep(i, 9)))
13. edgelist$topic = topic
14. g <- graph.data.frame (edgelist,directed=T )
15. l<-layout.fruchterman.reingold(g)
16. nodesize = centralization.degree(g)$res
17. V(g)$size = log( centralization.degree(g)$res )
18. nodeLabel = V(g)$name
19. E(g)$color = unlist(lapply(sample(colors()[26:137], 10), function(i) rep(i, 9)));
20. unique(E(g)$color)
21. png(  paste(getwd(), "/cl_graph_gibbs.png", sep=""),width=5, height=5,
i.      units="in", res=700)
22. plot(g, vertex.label= nodeLabel,  edge.curved=TRUE, vertex.label.cex =0.5,
甲、edge.arrow.size=0.2, layout=l )    dev.off()
```

首先我們將主題參數設為 10，thin=100 代表每 100 個字詞進行分析。種子參數設為 2003，接著將 dtm 矩陣使用 LDA 模型裡的 Gibbs 方法來進行分析，然後將分析結果儲存至 LDA_TM 至 Rdata 裡。我們再將等會繪出的每個主題字詞數量顯示 10 個，接著設置儲存 termsForSave1 的檔名以及 fileEncoding 設為 "UTF-8"，然後將 10 個主題放入資料框 tfs 裡之後，再將 1-10 個主題一分為兩個矩陣。如圖 3.14。

```
> adjacent_list
[[1]] [,1]     [,2]
[1,] "訂單"   "露天"
[2,] "露天"   "日期"
[3,] "日期"   "新品"
[4,] "新品"   "平台"
[5,] "平台"   "贈品"
[6,] "贈品"   "信用卡"
[7,] "信用卡" "灰塵"
[8,] "灰塵"   "作業"
[9,] "作業"   "罐頭"

[[2]] [,1]     [,2]
[1,] "匯款"   "現貨"
[2,] "現貨"   "郵資"
[3,] "郵資"   "郵局"
[4,] "郵局"   "幫手"
[5,] "幫手"   "折扣"
[6,] "折扣"   "名單"
[7,] "名單"   "手續費"
[8,] "手續費" "規則"
[9,] "規則"   "效率"
```

圖3.13 分割矩陣結果

分割矩陣之後，我們利用 rbind 指令將 adjacent_list 矩陣合併。確定主題關聯方向。如圖 3.15。

	> edgelist		
	v1	v2	topic
1	訂單	露天	1
2	露天	日期	1
3	日期	新品	1
4	新品	平台	1
5	平台	贈品	1
6	贈品	信用卡	1
7	信用卡	灰塵	1
8	灰塵	作業	1
9	作業	罐頭	1
10	匯款	現貨	2
11	現貨	郵資	2
12	郵資	郵局	2
13	郵局	幫手	2
14	幫手	折扣	2
15	折扣	名單	2
16	名單	手續費	2
17	手續費	規則	2
18	規則	效率	2

圖3.14 矩陣合併結果

我們確定主題字詞關聯方向後，將 topic 變為字串，接著利用 graph.data.frame 繪圖，directed 設 T 是讓他成為有向圖。將圖呈現之前，我們先設定節點的位置與大小之後，再設定標籤的名字與箭頭的顏色，然後設定圖的寬度與長度與儲存位置。最後我們再設定標籤為 vertex.label= nodeLabel、標籤大小為 0.5、箭頭大小為 0.2 之後開始繪圖。結果如圖 3.16。

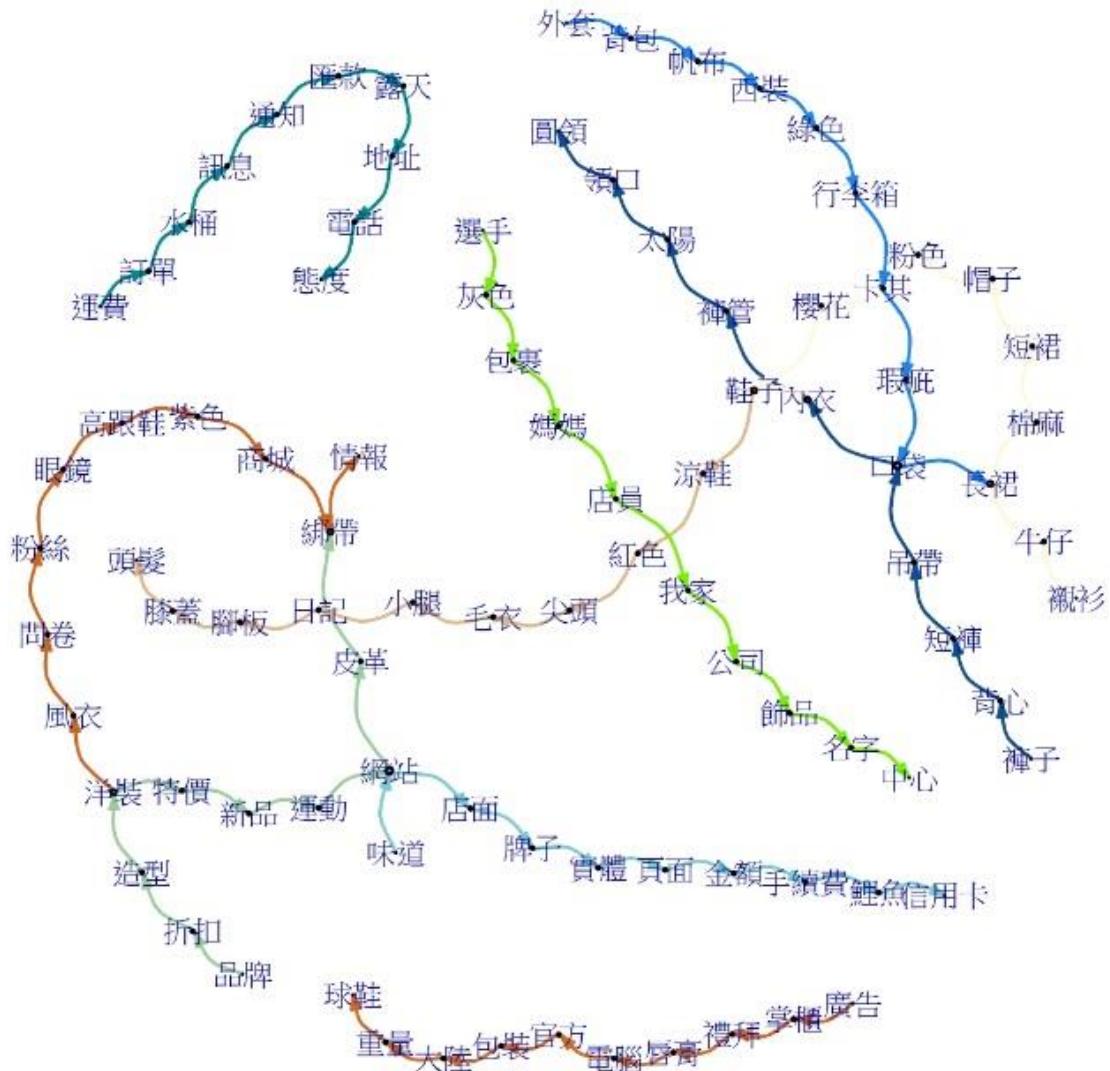


圖 3.15 LDA Gibbs 結果

第四章 研究結果

4.1 網路客訴分析流程

當某些產品的使用者在網際網路上的 PTT 討論區上留下了對於產品的批評或是抱怨文章時，我們可以透過 PPT Search 以及 R 語言裡的 RCurl 軟件幫助下，來收集這些資料，然後經過 Rwordseg 套件的 segmentCN 來進行中文斷詞，以及關鍵字擷取的處理流程之後，將這些關鍵詞彙儲存至 R 語言裡的關鍵詞彙庫，以方便供我們進行 LDA Gibbs 客訴分析，最後再對客訴分析結果進行討論。如圖 4.1。



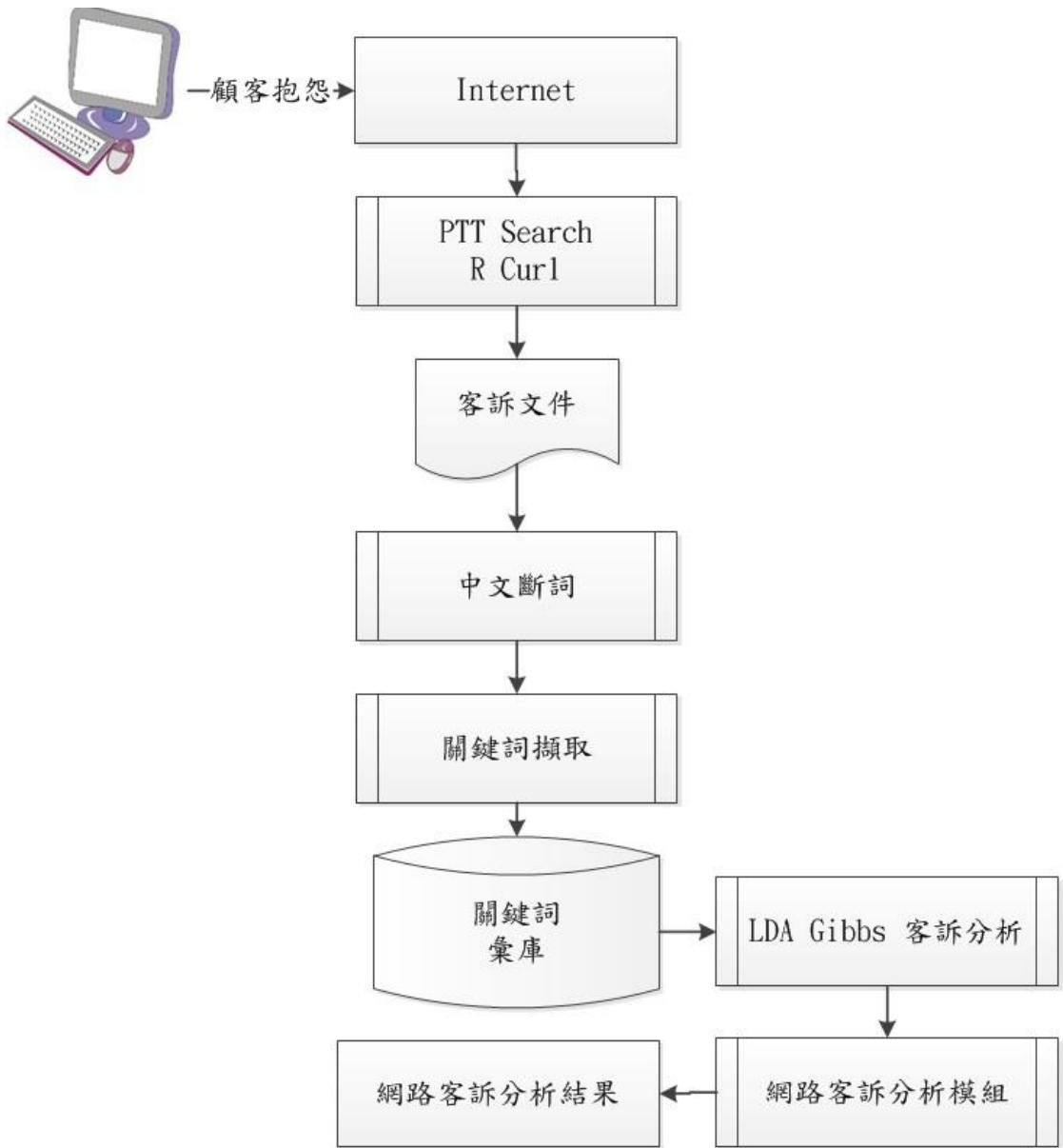


圖 4.1 客訴分析流程圖

4.2 客訴分析結果

本章節探討 R 分析客訴資料的結果，將各項資料整理並分析。

4.2.1 中文斷詞與關鍵詞字雲分析結果

本論文將台大 PTT e-shopping 版面的 868 篇抱怨文，經過 Rwordseg 套件的

segmentCN 中文斷詞後，有 4004 個字詞，如圖 4.2。

[3441]	"膠質"	"膠囊"	"蓮花"	"蔬菜"	"蠶米"	"蝴蝶"	"蝴蝶結"	"蝸牛"
[3449]	"衛生局"	"衛生紙"	"衛生棉"	"衛星"	"課稅"	"謀業"	"調調"	"調調"
[3457]	"請求"	"論文"	"論壇"	"論點"	"豬皮"	"豬肉"	"賠款"	"賠償金"
[3465]	"賣方"	"賣家"	"賣賣"	"質料"	"質量"	"質感"	"質疑"	"質戶"
[3473]	"賬號"	"輩子"	"輪廓"	"遭遇"	"遮隱傘"	"鄰居"	"銅量"	"銅合金"
[3481]	"靠山"	"靠枕"	"鞋子"	"鞋底"	"鞋底"	"鞋帶"	"鞋跟"	"鞋跟"
[3489]	"鞋墊"	"駕色"	"骷髏"	"髮夾"	"髮乳"	"髮型"	"髮廊"	"髮網"
[3497]	"鬧劇"	"鬧鐘"	"魅力"	"鮋魚"	"器具"	"噩夢"	"噱頭"	"壓板"
[3505]	"堅紙"	"學生"	"學長"	"學徒"	"學校"	"學校門"	"學院"	"學期"
[3513]	"學費"	"學會"	"學歷"	"學歷"	"導火線"	"導向"	"導航儀"	"憑單"
[3521]	"憑據"	"憑證"	"戰士"	"戰利品"	"戰爭"	"戰術"	"戰場"	"擔船"
[3529]	"據點"	"整版"	"整體"	"歷史"	"曆史"	"曆年"	"橘子"	"橙汁"
[3537]	"機子"	"機身"	"機車"	"機制"	"機油"	"機型"	"機密"	"機械"
[3545]	"機票"	"機場"	"機殼"	"機會"	"機構"	"機種"	"機器"	"機器人"
[3553]	"機關"	"橡皮"	"橡皮筋"	"橫批"	"歷史"	"歷程"	"星光"	"燈會"
[3561]	"獨舞"	"瘡肉"	"痘紙"	"獨立國家"	"獨角獸"	"磚頭"	"磨耗"	"積分"
[3569]	"積蓄"	"糖果"	"縣政府"	"與趣"	"蕃茄"	"希麥"	"螃蟹"	"蝴蝶"
[3577]	"褲子"	"褲腳"	"褲管"	"親人"	"親友"	"親戚"	"貓膩"	"貓膩"
[3585]	"輸家"	"辦公室"	"辦公桌"	"辦法"	"庭到者"	"選集"	"選項"	"選擇性"
[3593]	"選擇題"	"鋼化玻璃"	"鋼圈"	"鋼絲"	"鋼管"	"鎗音"	"錢包"	"錯字"
[3601]	"錯別字"	"錯事"	"錯金"	"錯話"	"錯誤"	"錯覺"	"錶帶"	"難花"
[3609]	"龍鱗色"	"靜電"	"頭人"	"頭皮"	"頭頂"	"頭腦"	"頭髮"	"頭頭"
[3617]	"頻率"	"餐券"	"餐椅"	"餐轎"	"駕駛"	"優惠券"	"優惠價"	"優勢"
[3625]	"優點"	"壓力"	"嬰兒"	"幫手"	"徽章"	"戲劇性"	"戲碼"	"樺木"
[3633]	"檔案"	"檢察官"	"溫氣"	"溫疹"	"營業所"	"營業稅"	"營業額"	"環島"
[3641]	"環節"	"環境"	"療效"	"療程"	"鐵衣針"	"踏卻"	"鐵達"	"總公司"
[3649]	"總成"	"總值"	"總裁"	"總隊"	"總會"	"總價"	"總機"	"總額"
[3657]	"繁體"	"繁體字"	"聯合報"	"聯邦"	"聯絡官"	"聯繫處"	"聯盟"	"聰明人"
[3665]	"聲明"	"聲音"	"膽小鬼"	"膽部"	"膽囊"	"臉孔"	"臉皮"	"臉色"
[3673]	"舉動"	"薄片"	"薄荷"	"薄胎"	"語言"	"嫌水"	"螺紋"	"螺絲"
[3681]	"諱言"	"謠話"	"講法"	"講話"	"嫌頭"	"購買者"	"購買證"	"隱私"
[3689]	"趨勢"	"避孕藥"	"鍋子"	"鍋底"	"鍋蓋"	"隱形眼鏡"	"隱私"	"隱性"
[3697]	"隱私權"	"霜花"	"韓元"	"韓幣"	"顆粒"	"鷄子"	"鷹鹿"	"鷹頭"
[3705]	"黏度"	"黏膠"	"點數"	"點點"	"櫻桃"	"櫃子"	"櫃台"	"櫃檯"
[3713]	"濾波器"	"湯樂"	"獵人"	"禮服"	"禮物"	"禮品"	"禮拜"	"簡介"
[3721]	"簡訊"	"簡章"	"簡體字"	"鐵女"	"鐵口令"	"翻版"	"翻領"	"翻領"
[3729]	"難工"	"職務"	"職責"	"職業"	"職業裝"	"職業道德"	"畜貨"	"駁色"
[3737]	"認識"	"豐功"	"蹤影"	"轉角"	"轉速"	"轉運站"	"轉輪"	"醫生"
[3745]	"醫院"	"醫油"	"鎖骨"	"鎖頭"	"鎖子"	"雙人床"	"雙子座"	"雙手"
[3753]	"雙方"	"雙胞胎"	"雙喜"	"雙追"	"雜念"	"雜物"	"雜面"	"雜貨店"
[3761]	"雜費"	"雜誌"	"雞內"	"雞胸"	"雞蛋"	"雞蛋黃"	"雞子"	"雞子"
[3769]	"題字"	"韻度"	"額頭"	"顏色"	"騎料"	"騎士"	"鬃毛"	"緊緊"
[3777]	"鬆緊帶"	"雀魚"	"娛人"	"娛人"	"寵物"	"懲罰性"	"懷抱"	"簽名"
[3785]	"簽呈"	"簿子"	"繩子"	"藝術"	"葉水"	"葉品"	"薌劑"	"證人"
[3793]	"證件"	"證物"	"證據"	"贈品"	"轎車"	"辭典"	"過卡"	"過角"
[3801]	"過角料"	"過長"	"過界"	"過綠"	"鏡子"	"鏡頭"	"鏡局"	"難處"
[3809]	"關卡"	"關係"	"關說"	"關建"	"關鍵字"	"關鍵詞"	"難度"	"寶石"
[3817]	"類別"	"類型"	"騙子"	"騙局"	"騙術"	"鬍子"	"嚴重性"	"解石"
[3825]	"寶貝"	"寶號"	"寶貴"	"瘤結"	"礦泉水"	"競爭力"	"壽碼"	"解打"
[3833]	"琥果"	"警方"	"警察"	"警語"	"警衛"	"警衛室"	"議價"	"嫌獨"
[3841]	"餽菜"	"麵包"	"屬性"	"攝影嘲"	"攝影機"	"櫻花"	"櫻桃"	"露天"
[3849]	"襪子"	"護照"	"護衛"	"鐵片"	"鐵盒"	"鐵盒"	"鐵絲"	"權力"
[3857]	"霸氣"	"薩摩"	"顧客"	"顧慮"	"魔爪"	"魔法"	"顧峰"	"驕氣"
[3865]	"權利"	"權限"	"權益"	"權責"	"墨字"	"墨衫"	"讀者"	"邏輯性"
[3873]	"辭話"	"編號"	"編組"	"蘿蔔"	"葵慈"	"葵壓器"	"邏輯"	"邏輯性"
[3881]	"顯示器"	"驚嘆號"	"驚魂"	"體系"	"體育"	"體制"	"體重"	"體態"
[3889]	"體質"	"體積"	"確子"	"趣頭"	"鹽水"	"鹽進"	"觀世音"	"觀念"
[3897]	"觀眾"	"觀感"	"觀點"	"鑄匙"	"鑽石"	"鑑魚"	"鴉鴉"	"鬱金香"

圖 4.2 868 篇抱怨中文斷詞結果

因此本論文為了減少出現次數較少的字詞，而使用 wordcloud 裡的 min.freq = 40 將出現 40 次以上的字詞呈現出來，來確定關鍵字有哪些，之所以參數設 40，是因為 LDA 的模式分析是以每 100 個字詞進行分析，如果設 50 以上會低於 100，分析時 LDA 會有錯誤的情況發生，所以設為 40 是最為妥當的。確定之後使用 stopwords()指令將 40 次以下的字詞給刪除掉。結果如圖 4.3。

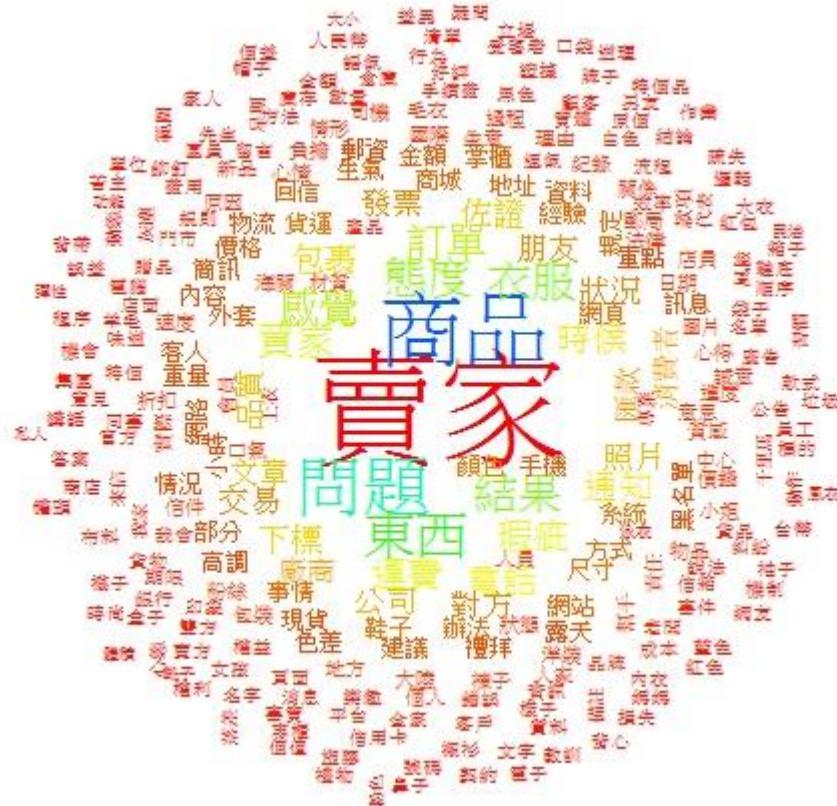


圖 4.3 關鍵字詞雲分析結果

關鍵字詞雲結果出來後，我們需知道出現在關鍵字詞雲的字詞有哪些，這裡可以使用 `findFreqTerms(tdm,40)` 指令，來得到想要的答案，分析過後從 4004 個字詞刪減掉只剩 278 個字詞，然後將 278 個字詞作進一步的分析。

```
> findFreqTerms(tdm, 40)
[1] "人民"   "人員"   "人家"   "上衣"   "下標"   "千里馬" "口氣"   "口袋"   "大小"   "大衣"   "大陸"   "女孩"   "小姐"   "小時"
[15] "中心"  "內衣"   "內容"   "公司"   "公告"   "尺寸"   "心得"   "心情"   "心態"   "手機"   "手續費" "文字"   "文章"   "方式"
[29] "方法"  "日期"  "毛衣"   "代表"   "功能"   "包裝"   "包裹"   "台幣"   "司機"   "外婆"   "布料"   "平台"   "民法"   "生氣"
[43] "生意"  "白色"   "立場"   "交易"   "先生"   "全家"   "全額"   "印象"   "同事"   "名字"   "名單"   "回信"   "地方"   "地址"
[57] "奸評"  "成本"   "反塵"   "羊毛"   "老闆"   "色差"   "行為"   "衣服"   "味道"   "垃圾"   "官方"   "店面"   "折扣"   "材質"
[71] "男友"  "私人"   "系統"   "事件"   "事情"   "事實"   "來信"   "受著者" "味道"   "金額"   "門市"   "信用卡" "信件"   "朋友"
[85] "東西"  "法律"   "冰衣"   "物品"   "物流"   "狀況"   "狀態"   "紛糾"   "紀錄"   "紅包"   "紅色"   "背心"   "背帶"   "品牌"
[99] "品質"  "契約"   "客人"   "客戶"   "建議"   "星期"   "洋裝"   "流程"   "紀錄"   "紅包"   "紅色"   "背心"   "吉主"   "差異"
[113] "訂單"  "負擔"   "重量"   "重點"   "頁面"   "風衣"   "風險"   "個人"   "原因"   "原價"   "員工"   "家人"   "粉色"   "絲絨"
[127] "庫存"  "效率"   "時尚"   "時候"   "海關"   "消息"   "消費者" "特價"   "高價"   "高價"   "清留"   "現貨"   "高調"
[141] "動作"  "動態"   "商店"   "商品"   "商城"   "問題"   "國際"   "專業"   "專業"   "高價"   "清留"   "現貨"   "高調"
[155] "理由"  "產品"   "流失"   "盒子"   "習慣"   "袋子"   "袖子"   "規定"   "規則"   "貨物"   "貨品"   "貨運"   "責任"
[169] "速度"  "部分"   "備註"   "單位"   "帽子"   "掌櫃"   "期限"   "款式"   "款項"   "貨品"   "發票"   "程序"   "通知"
[183] "結論"  "費用"   "買家"   "進度"   "郵局"   "郵資"   "集團"   "順序"   "黑名單" "黑色"   "匯款"   "塑膠"   "結果"
[197] "意思"  "感覺"   "損失"   "新品"   "會員"   "照片"   "瑕疵"   "經驗"   "落差"   "號碼"   "誠意"   "資料"   "資訊"
[211] "逼黃"  "過程"   "道理"   "鉤針"   "電子"   "電腦"   "電話"   "靴子"   "圖片"   "團員"   "實際"   "實體"   "對方"
[225] "詮問"  "網友"   "網頁"   "網站"   "網路"   "語氣"   "誤差"   "說法"   "說法"   "銀行"   "桌子"   "價值"   "價差"
[239] "廠商"  "廣告"   "彈性"   "數量"   "樂趣"   "標的"   "樣子"   "箱子"   "範圍"   "賣方"   "賣家"   "質料"   "質感"
[253] "鞋底"  "機制"   "機會"   "牌子"   "辦法"   "錯譯"   "幫手"   "講話"   "禮物"   "禮拜"   "簡訊"   "藍色"   "鞋子"
[267] "證據"  "贈品"   "關係"   "襪子"   "露天"   "顧客"   "權利"   "權益"   "潤滑"   "邏輯"   "體積"   "難頭"   "顏色"
```

圖 4.4 278 個關鍵字

4.2.2 關鍵字數量與 TF-IDF 分析結果

經過前面一小節的中文斷詞與刪除字詞後，首先要先知道 278 個關鍵字的在

868 篇文章裡出現的次數與 TF-IDF 為多少，才能進行 LDA 模式分析，因此本研究

先使用 row_sums(tdm)這項指令，把關鍵字出現次數統計出來，如圖 4.5。

> row_sums(tdm)																							
	人民幣	人員	人家	上衣	下標	千里馬	口氣	口袋	大小	大衣	大陸	女孩	小姐	小時	中心	內衣	內容	公司					
47	193	152	79	647	98	188	46	41	56	180	76	149	224	163	52	247							
549	公告	尺寸	心得	心情	心態	手機	手續費	文字	文章	方式	方法	日期	毛衣	代表	功能	包裝	包裹	台幣					
80	351	158	145	43	273	95	66	589	326	65	153	136	91	41	186	682							
49	司機	外套	布料	平台	民法	生氣	生意	白色	立場	交易	先生	全家	全額	印象	同事	名字	名單	回信					
135	263	52	120	41	295	149	95	45	400	65	118	64	101	82	75	107							
288	地方	地址	好評	成本	灰塵	羊毛	老闆	色差	行為	衣服	佐證	作業	我家	我會	折扣	材質	男友	私人					
183	306	93	85	82	54	116	206	70	1180	638	52	86	123	138	185	51							
41	系統	事件	事情	事實	來信	受害者	味道	垃圾	官方	店面	店員	朋友	東西	法律	泳衣	物品	物流	狀況					
357	104	328	74	89	55	101	50	105	101	151	438	1447	141	40	155	227							
403	狀態	糾紛	金額	門市	信用卡	信件	信箱	品牌	品質	契約	客人	客戶	建議	星期	洋裝	流程	紀錄	紅包					
140	75	274	122	93	184	115	118	584	65	247	117	274	84	165	115	151							
57	紅色	背心	背帶	苦主	訂單	負擔	重量	重點	頁面	風衣	風險	倉庫	個人	原因	原價	員工	家人	差異					
64	52	42	43	965	128	247	209	122	41	97	90	153	127	57	55	46							
44	庫存	效率	時尚	時候	海關	消息	消費者	特價	特價品	留言	粉色	粉絲	訊息	高調	動作	動態	商店	商品					
87	89	43	709	159	121	506	116	52	87	40	195	235	207	58	46	50							
2803	商城	問題	國際	專業	專櫃	情形	情況	教訓	清單	現貨	理由	產品	疏失	盒子	習慣	袋子	袖子	規定					
300	2043	146	81	72	132	224	65	52	240	146	99	47	57	58	69	61							
326	規則	貨物	貨品	貨運	責任	通知	速度	部分	備註	單位	帽子	掌櫃	期限	款式	款項	牌子	發票	程序					
88	84	95	354	174	709	178	238	85	43	47	298	86	62	127	51	427							
46	答案	結果	結論	買家	費用	進度	郵局	郵資	集團	順序	黑名單	黑色	匯款	塑膠	媽媽	意見	意思	感覺					
53	1187	57	898	84	156	154	207	54	51	243	99	479	71	66	77	181							
979	損失	新品	會員	照片	瑕疵	經驗	落差	號碼	誠意	資料	資訊	運氣	運費	過程	道理	鈎針	電子	電腦					
65	128	127	563	845	304	46	68	161	273	120	57	834	146	46	64	50							
87	電話	靴子	圖片	團員	實際	實體	對方	態度	疑問	網友	網頁	網站	網路	語氣	誤差	說法	銀行	鼻子					
746	41	158	68	135	103	389	1119	42	62	332	312	247	57	44	85	80							
43	價值	價差	價格	價錢	廠商	廣告	彈性	數量	樂趣	標的	樣子	箱子	範圍	賣方	賣家	質料	質感	鞋子					
45	44	256	164	482	75	41	80	97	63	112	55	44	77	4601	59	195							
349	鞋底	機制	機會	褲子	辦法	錯誤	幫手	講話	禮物	禮拜	藍色	雙方	顏色	證據	贈品	關係	襪子						
57	51	61	177	330	121	203	45	50	238	252	63	79	367	73	104	147							
50	露天	顧客	權利	權益	襯衫	邏輯	體積	罐頭															
229	81	50	123	70	57	40	44																

圖 4.5 關鍵字次數統計結果

接著使用章節 3.4.1 提供的程式碼，來算出各個關鍵字的 TF-IDF 為多少。計算結果如圖 4.6。

```

> term_tfidf
      1     2     3     4     5     6     7     8     9     10    11
0.14942649 0.09860736 0.05858483 0.10932127 0.06042444 0.27397986 0.05578914 0.18007290 0.08094636 0.17312341 0.11334568
      12    13    14    15    16    17    18    19    20    21    22
0.14973543 0.13136957 0.08604210 0.16196409 0.24118128 0.05298754 0.09479147 0.10052315 0.11707471 0.08940964 0.05386430
      23    24    25    26    27    28    29    30    31    32    33
0.06182203 0.08275146 0.25901734 0.06601507 0.02990683 0.04887818 0.06552848 0.11164528 0.24341613 0.05474424 0.09307151
      34    35    36    37    38    39    40    41    42    43    44
0.09679756 0.09613048 0.13252185 0.24430134 0.15227210 0.12961397 0.13778574 0.12093573 0.05349688 0.06037213 0.18678422
      45    46    47    48    49    50    51    52    53    54    55
0.06821210 0.06186706 0.16943798 0.17778684 0.09899454 0.07133166 0.13175671 0.12411833 0.09598169 0.08948582 0.05359570
      56    57    58    59    60    61    62    63    64    65    66
0.11402224 0.11630980 0.06837311 0.33381702 0.39438857 0.12286427 0.19644406 0.06538200 0.09827917 0.01449377 0.12472216
      67    68    69    70    71    72    73    74    75    76    77
0.09751606 0.05266033 0.09968454 0.12581701 0.14755103 0.12138617 0.07841873 0.05915662 0.04835136 0.05377888 0.08644586
      78    79    80    81    82    83    84    85    86    87    88
0.07503059 0.30955769 0.10661502 0.14888658 0.08968161 0.16644619 0.09362339 0.03761260 0.07312429 1.44962262 0.06942917
      89    90    91    92    93    94    95    96    97    98    99
0.13155791 0.04907703 0.08749618 0.06065707 0.07747391 0.20747125 0.17069061 0.06715964 0.08831857 0.13527866 0.08749060
      100   101   102   103   104   105   106   107   108   109   110
0.09134831 0.06311440 0.08238548 0.04634549 0.08747331 0.17936578 0.07624385 0.09182575 1.59452838 0.19785088 0.25532262
      111   112   113   114   115   116   117   118   119   120   121
0.67906912 0.09879046 0.10279159 0.09011495 0.27229202 0.04825855 0.06559140 0.16076078 0.06845564 0.12652945 0.05500338
      122   123   124   125   126   127   128   129   130   131   132
0.06663510 0.10796579 0.14651823 0.07832547 0.09135681 0.11326293 0.09443717 0.18813584 0.04121938 0.31431862 0.07008237
      133   134   135   136   137   138   139   140   141   142   143
0.07478951 0.12717582 0.14321936 0.08765123 0.18179572 0.10963848 0.08337030 0.14539470 0.06606084 0.10994561 0.09306774
      144   145   146   147   148   149   150   151   152   153   154
0.01117903 0.18437842 0.03120685 0.13910136 0.13260674 0.15819036 0.07262087 0.05375506 0.06282959 0.13379291 0.12805222
      155   156   157   158   159   160   161   162   163   164   165
0.06186101 0.10453842 0.07023946 0.15612562 0.04810001 0.15349071 0.18529614 0.08863536 0.11547619 0.10143218 0.07615325
      166   167   168   169   170   171   172   173   174   175   176
0.14660914 0.05005287 0.06068261 0.08082120 0.05589208 0.07958315 0.12079404 0.17550457 0.26206587 0.09511407 0.09300348
      177   178   179   180   181   182   183   184   185   186   187
0.08423962 0.15976212 0.25627071 0.10756808 0.08264428 0.02657098 0.05929605 0.05746372 0.09717718 0.07803133 0.09926673
      188   189   190   191   192   193   194   195   196   197   198
0.13041270 0.19358543 0.14709097 0.08800509 0.17376818 0.09356985 0.14076306 0.14142302 0.06540940 0.04688878 0.04017004
      199   200   201   202   203   204   205   206   207   208   209
0.06405551 0.12745085 0.22684437 0.07697487 0.11250632 0.04636889 0.09194427 0.11059741 0.08228562 0.08494899 0.07562476
      210   211   212   213   214   215   216   217   218   219   220
0.08501180 0.07944236 0.06033093 0.05233801 0.22405196 0.18191639 0.10722070 0.08083773 0.23374673 0.06524155 0.19549428
      221   222   223   224   225   226   227   228   229   230   231
0.06767127 0.07661571 0.08039440 0.05335383 0.05407127 0.08144545 0.06660441 0.08703774 0.06398451 0.07558081 0.11923805
      232   233   234   235   236   237   238   239   240   241   242
0.05533323 0.16008996 0.08446351 0.06535065 0.08786973 0.06816728 0.07406701 0.14497188 0.09831728 0.18448427 0.11439624
      243   244   245   246   247   248   249   250   251   252   253
0.63411344 0.07794006 0.05429602 0.11822059 0.07616633 0.08695667 0.07182463 0.09961291 0.11069210 0.19862560 0.19338131
      254   255   256   257   258   259   260   261   262   263   264
0.09251970 0.06595475 0.20407476 0.04959619 0.07805585 0.11454082 0.06797552 0.12132416 0.06892984 0.11696427 0.15513344
      265   266   267   268   269   270   271   272   273   274   275
0.06853733 0.14965844 0.07433161 0.33700488 0.04904478 0.40939975 0.24074834 0.08510399 0.08179473 0.05819882 0.11389954
      276   277   278
0.08303783 0.27805412 0.11489369

```

圖 4.6 TF-IDF 結果

本研究將 278 個關鍵字出現次數與 TF-IDF 整理過後表格再附錄 A 裡有收入。

4.2.3 LDA Gibbs 模式分析結果

將 278 個關鍵字詞出現次數與 TF-IDF 計算出來後，直接套用 3.4.2 章節提到的 LDA 模式的程式碼進行分析，分析結果 LDA 將 278 個字詞裡的其中 100 個重要關鍵字詞分為 10 個 topic，一條線即代表著一個 topic。如圖 4.7。

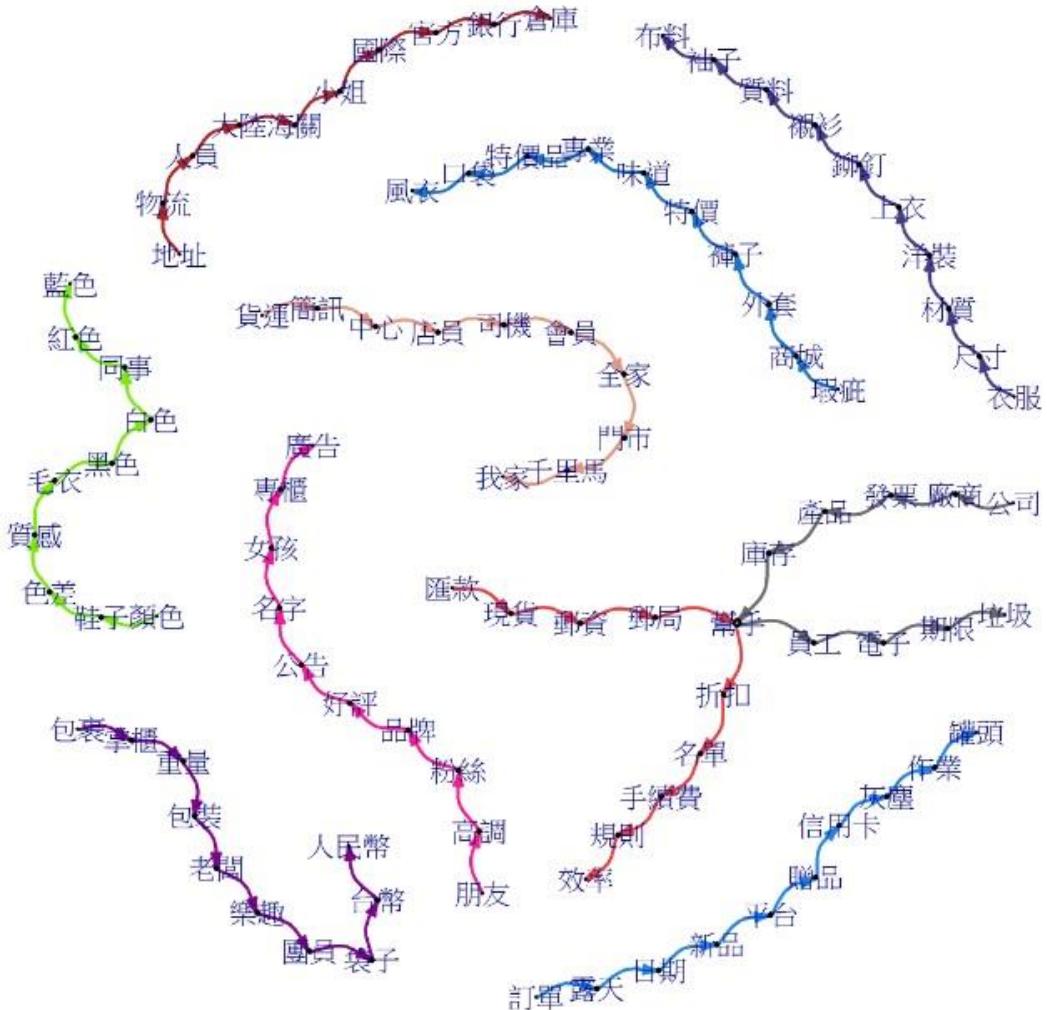


圖 4.7 278 個關鍵字 LDA 模式結果

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	訂單	匯款	衣服	地址	朋友	貨運	顏色	公司	包裹	瑕疵
2	露天	現貨	尺寸	物流	高調	簡訊	鞋子	廠商	掌櫃	商城
3	日期	郵資	材質	人員	粉絲	中心	色差	發票	重量	外套
4	新品	郵局	洋裝	大陸	品牌	店員	質感	產品	包裝	褲子
5	平台	幫手	上衣	海關	好評	司機	毛衣	庫存	老闆	特價
6	贈品	折扣	鈕釦	小姐	公告	會員	黑色	幫手	樂趣	味道
7	信用卡	名單	襯衫	名字	名字	全家	白色	員工	團員	專業
8	灰塵	手續費	質料	女孩	門市	千里馬	同事	電子	袋子	特價品
9	作業	規則	袖子	專櫃	我家	我家	紅色	日期	台幣	口袋
10	罐頭	效率	布料	廣告	廣告	藍色	藍色	期限	人民幣	風衣

圖 4.8 10 個 topic 關聯結果

LDA Gibbs 模式分析它是先以字詞出現次數分出 10 個 topic，如果出現字詞次數相近，再以 TF-IDF 計算找出 10 個 topic 的關聯性。首先我們從 topic 1 開始，topic 1 以〔訂單〕為主題，經過 LDA 分析過後，從〔訂單〕這個詞彙出現的文章裡，LDA 找出其他 9 個跟〔訂單〕有關聯的關鍵字。由此表 4.1 推斷在台大 PTT

e-shopping 論壇上，消費者常抱怨露天平台上的新品、贈品、罐頭常有灰塵，而這些產品在使用信用卡訂單作業上，常出現問題。

表 4.1 LDA Gibbs topic 1

字詞	訂單	露天	日期	新品	平台
次數	965	229	153	128	120
TF-IDF	0.1027	0.2407	0.1116	0.1274	0.1377
字詞	贈品	信用卡	灰塵	作業	罐頭
次數	104	93	82	52	44
TF-IDF	0.3370	0.1706	0.3338	0.1247	0.1148

接著看 topic 2，topic 2 以〔匯款〕為主題，經過 LDA 分析後，從〔匯款〕這個詞彙出現的文章裡，LDA 找出其他 9 個跟〔匯款〕有關聯的關鍵字，我們發現 topic 2 與 topic 8 因為有〔幫手〕這詞彙產生了關聯性，而 topic 8 是以〔公司〕為主題，如圖 4.9。

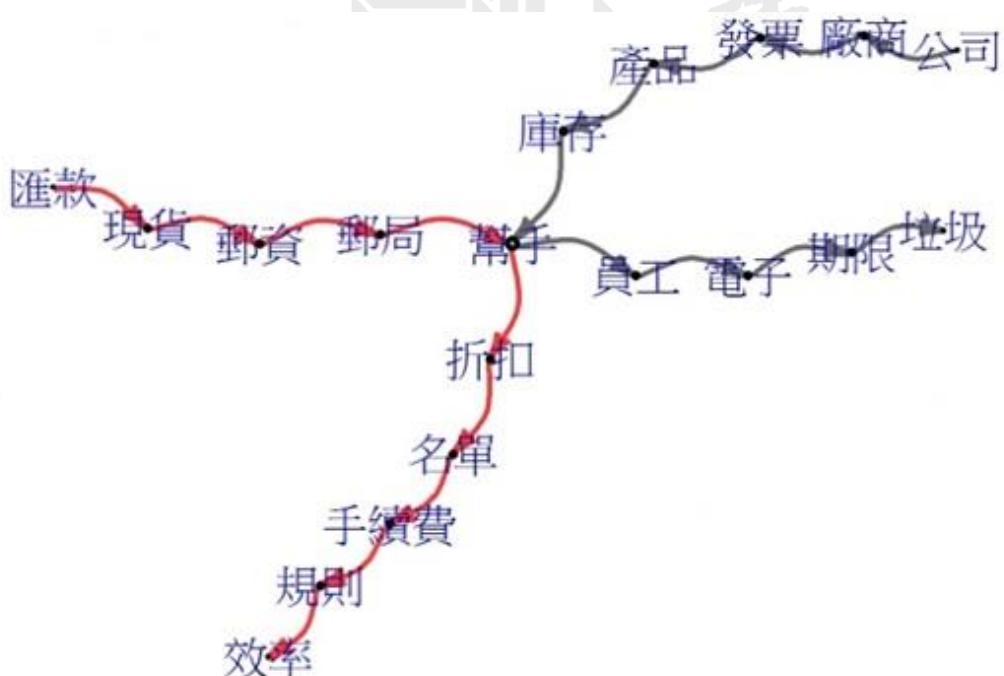


圖 4.9 topic 2 與 topic 8 關聯圖

這代表著兩個主題出現的關鍵字因為〔幫手〕這詞彙的緣故，可以推斷說它

們的關鍵字在同篇文章出現的頻率極高。因此我們從表 4.2 與 4.3 得知台大 PTT e-shopping 論壇裡，常有消費者購買現貨時，因為折扣的關係，匯款出現郵局郵資、手續費等問題，那這個員工幫手照著購買名單、規則進行，所以處理效率極差。另外消費者也抱怨公司廠商明明有庫存產品，員工幫手開的電子發票期限卻與購買時間不符合。關鍵字會有〔垃圾〕是消費者所表達的情緒化字眼。

表 4.2 LDA Gibbs topic 2

字詞	匯款	現貨	郵資	郵局	幫手
次數	479	240	207	154	203
TF-IDF	0.0935	0.1280	0.1304	0.0992	0.1145
字詞	折扣	名單	手續費	規則	效率
次數	138	107	95	88	89
TF-IDF	0.0996	0.0959	0.2590	0.1154	0.0944

表 4.3 LDA Gibbs topic 8

字詞	公司	廠商	發票	產品	庫存
次數	549	482	427	99	87
TF-IDF	0.0947	0.1449	0.2562	0.1045	0.1132
字詞	幫手	員工	電子	期限	垃圾
次數	203	55	50	86	50
TF-IDF	0.1145	0.1465	0.1819	0.0951	0.1066

從 topic 3 看，topic 3 以〔衣服〕為主題，經過 LDA 分析過後，從〔衣服〕這個詞彙出現的文章裡，LDA 找出其他 9 個跟〔衣服〕有關聯的關鍵字。由此表 4.4 推斷在台大 PTT e-shopping 論壇上，消費者常抱怨他們在拍賣網站購買衣服、洋裝、上衣、襯衫、袖子的布料與材質上有問題。除了衣服尺寸上有問題，消費者在拍賣網站上購買的鉤釘尺寸也常出現問題。

表 4.4 LDA Gibbs topic 3

字詞	衣服	尺寸	材質	洋裝	上衣
次數	1180	351	185	165	79
TF-IDF	0.0982	0.1170	0.1258	0.1793	0.1093
字詞	鉤釘	襯衫	質料	袖子	布料
次數	64	70	59	61	52
TF-IDF	0.2240	0.1138	0.0996	0.1852	0.1296

從 topic 4 看，topic 4 以〔地址〕為主題，經過 LDA 分析過後，從〔地址〕這個詞彙出現的文章裡，LDA 找出其他 9 個跟〔地址〕有關聯的關鍵字。由此表 4.5 推斷在台大 PTT e-shopping 論壇上，出現〔大陸〕兩個字，大部分都是與淘寶網有關，因此我們推斷消費者抱怨淘寶官方小姐在處理地址物流配送時，常會出現台灣海關人員把商品放入國際銀行裡的保稅倉庫裡，以至於要多繳一份費用。

表 4.5 LDA Gibbs topic 4

字詞	地址	物流	人員	大陸	海關
次數	306	227	193	180	159
TF-IDF	0.1140	0.1315	0.0986	0.1133	0.3143
字詞	小姐	國際	官方	銀行	倉庫
次數	149	146	105	80	90
TF-IDF	0.1313	0.1391	0.1488	0.1600	0.1265

從 topic 5 看，topic 5 以〔朋友〕為主題，經過 LDA 分析過後，從〔朋友〕這個詞彙出現的文章裡，LDA 找出其他 9 個跟〔朋友〕有關聯的關鍵字。由此表 4.6 推斷在台大 PTT e-shopping 論壇上，消費者與朋友在拍賣網站的女孩專櫃裡尋找有好評、粉絲多、廣告的知名品牌，但在女孩專櫃裡購買產品時，可能出現問題，而知名品牌處理方式，以高調公告名字的方式處理，引起消費者們不愉快。

表 4.6 LDA Gibbs topic 5

字詞	朋友	高調	粉絲	品牌	好評
次數	438	207	195	118	93
TF-IDF	0.0936	0.1453	0.1096	0.1352	0.1163
字詞	公告	名字	女孩	專櫃	廣告
次數	80	75	76	72	75
TF-IDF	0.1005	0.1241	0.1497	0.1581	0.0983

從 topic 6 看，topic 6 以〔貨運〕為主題，經過 LDA 分析過後，從〔貨運〕這個詞彙出現的文章裡，LDA 找出其他 9 個跟〔貨運〕有關聯的關鍵字。由此表 4.7 推斷在台大 PTT e-shopping 論壇上，消費者抱怨淘寶網的千里馬貨運中心在送貨至全家門市或是消費者家時，經常不以簡訊通知貨物已送達指定地點，以至消費者要常去門市詢問店員貨物是否已經送達。

表 4.7 LDA Gibbs topic 6

字詞	貨運	簡訊	中心	店員	司機
次數	354	252	163	151	135
TF-IDF	0.1466	0.1169	0.1619	0.1664	0.2443
字詞	會員	全家	門市	千里馬	我家
次數	127	118	122	98	86
TF-IDF	0.2268	0.1777	0.2074	0.2739	0.0975

從 topic 7 看，topic 7 以〔顏色〕為主題，經過 LDA 分析過後，從〔顏色〕這個詞彙出現的文章裡，LDA 找出其他 9 個跟〔顏色〕有關聯的關鍵字。由此表 4.8 推斷在台大 PTT e-shopping 論壇上，消費者抱怨與同事從拍賣網站購買的鞋子、毛衣經常會有嚴重的顏色色差問題，質感上也不是很好。

表 4.8 LDA Gibbs topic 7

字詞	顏色	鞋子	色差	質感	毛衣
次數	367	349	206	195	136
TF-IDF	0.1496	0.1986	0.1964	0.1106	0.2434
字詞	黑色	白色	同事	紅色	藍色
次數	99	95	82	64	63
TF-IDF	0.1737	0.1867	0.1317	0.1978	0.1551

從 topic 9 看，topic 9 以〔包裹〕為主題，經過 LDA 分析過後，從〔包裹〕這個詞彙出現的文章裡，找出其他 9 個跟〔包裹〕有關聯的關鍵字。由此表 4.9 推斷在台大 PTT e-shopping 論壇上，消費者抱怨與團員團購樂趣購物網裡的袋子，從老闆那包裝成包裹的人民幣費用因為秤重重量緣故與台幣費用不符合。

表 4.9 LDA Gibbs topic 9

字詞	包裹	掌櫃	重量	包裝	老闆
次數	682	298	247	186	116
TF-IDF	0.0961	0.2620	0.2722	0.0967	0.1228
字詞	樂趣	團員	袋子	台幣	人民幣
次數	97	68	69	49	47
TF-IDF	0.6341	0.1954	0.1534	0.1325	0.1494

從 topic 10 看，topic 10 以〔瑕疵〕為主題，經過 LDA 分析過後，從〔瑕疵〕這個詞彙出現的文章裡，找出其他 9 個跟〔瑕疵〕有關聯的關鍵字。由此表 4.10 推斷在台大 PTT e-shopping 論壇上，消費者抱怨在拍賣商城裡委託專業代購購買的特價品外套，褲子有瑕疵，口袋裡有怪味道。

表 4.10 LDA Gibbs topic 10

字詞	瑕疵	商城	外套	褲子	特價
次數	845	300	263	177	116
TF-IDF	0.1125	0.1843	0.1522	0.2040	0.1271
字詞	味道	專業	特價品	口袋	風衣
次數	101	81	52	46	41
TF-IDF	0.3095	0.1326	0.1432	0.1800	0.1607

經過 1-10 個 topic 討論後，本論文發現有些詞彙出現次數較高，卻沒有排進 1-10 個 topic 裡，例如：〔賣家〕出現了 4601 次、〔態度〕出現了 1119 次。這是因為〔賣家〕這個詞的 TF-IDF 只有 0.0718，〔態度〕這個詞的 TF-IDF 也只有 0.0533，它們兩個詞彙在所有文件中所佔的比重並不是很高，所以 LDA 自動把它們給刪除掉。

4.3 分析結果討論

原本在 PTT 論壇上還不大清楚消費者想表達什麼，因為消費者有時在論壇上的抱怨，會參雜了情緒化謾罵字眼，而導致想表達出的意思與原本不同。因此本研究將台大 PTT e-shopping 論壇上的 868 篇抱怨文經過使用 R 語言裡的中文斷詞、關鍵字雲篩選關鍵字、LDA 模式分析後，大致了解在 PTT e-shopping 論壇上，消費者常抱怨的種類有哪些。

分析的結果，消費者經常抱怨的問題：信用卡訂單作業流程、匯款與手續費、衣服布料與材質、物流貨運配送、包裹秤重費用不一致、鞋子毛衣嚴重色差、特價品常有瑕疵。因此本研究認為，賣家應該針對上述所提到的問題反省、檢討、改善，買家才不會對賣家失去信心，而導致賣家自己流失一堆客源，失去信譽的問題。

第五章 結論與未來研究

5.1 結論

客訴不單單只有看它如何形成以及如何去應對，我們也應該從另外一個角度去看，因為現今網路的發展再加上時間上的變化，時間具有突顯問題的效果，這對於客訴的質量產生了影響性的變化。目前新世代的想法，他們習慣透過網路，以直接的態度和方式來表達他們對於產品的喜好，所以我們認為企業應該去瞭解他們的想法，然後針對他們提供的缺失以及建議去做產品後續的改善或是新產品的開發，客訴這個詞彙對我們來說，不單只是抱怨而已，也包含了對於企業的期待。

客訴從產品的各個階段來看，在於企業是否有積極且有誠意的來面對這一個客訴問題，而消費者所希望看到的企業精神，是在面對客訴這一塊，是否有努力去解決，而不是藉口逃避，視而不見。

目前的企業在面對客訴時，大多還是用被動的態度去思考如何快速的處理顧客的抱怨，仰賴著客服中心採用人工方式記錄客戶反映的問題，這樣往往忽略掉哪些客訴才是重要的，哪些客訴才是需要重視的，哪些客訴才是該列為優先處理的項目，這些都是值得企業來思考的。網路上的客訴行為已成為一種趨勢，時代的進步，企業對於網路客訴的思考模式也要跟得上這一波的競爭。所以企業應該好好思考如何運用這樣的資源。

所以透過我們所提出來的 R 語言 LDA Gibbs 的網路客訴分析模式，可以提供拍賣網站從論壇裡找出自家產品的問題，也多了另外一個層面的思考以及參考資訊，在經過我們提供出來的網路客訴分析模式，期許可以使客訴從另一個角度來觀察，能夠幫助企業在處理客訴這一塊上多了個分析的工具，同時讓使用者多了個參考資料。

5.2 未來研究

目前本論文中所做的 R 語言 LDA Gibbs 的網路客訴分析的研究方法，仍然未達到十全十美的分析，因此在未來還是有蠻多改善的空間：

1. 目前的分析方法僅針對 PTT 雜亂的抱怨文章進行分析，未來可以將 PTT 抱怨文章分門別類，分析的結果準確度應該會提高許多。
2. 目前的分析還未對 LDA 的參數做設定，未來如果分析的文章一多，繪圖時可能會有雜亂的情況發生，所以未來會增加對 LDA 分析的參數設定，來增加 LDA 分析的繪圖結果穩定性。
3. 文字探勘重點在於關鍵字擷取的精準度，未來 SQL Server 裡會有 R 語言的功能，相信 SQL Server 結合了 R 語言的功能可以提升關鍵字擷取的準確度。
4. 最後，在分析的過程，我們發現許多賣家用英文做 id 名稱，未來會增加中英文的擷取，來確定哪個拍賣網站的賣家問題最多，在提供資訊給拍賣網站。

參考文獻

1. 資策會（2013），**2013 年我國家庭寬頻現況與需求調查-個人**，
<http://www.find.org.tw/find/home.aspx?page=many&id=386>。
2. 資策會（2013），**2013 年我國家庭寬頻現況與需求調查-家戶**，
<http://www.find.org.tw/find/home.aspx?page=many&id=376>。
3. PC HOME 新聞（2014），**全台奧客大調查-宅世代上網抱怨殺傷力最強**，
<http://news.pchome.com.tw/magazine/print/po/cw/1195/124767360089295035001.htm>
4. 中村卯一郎（1992），**抱怨處理讀本—化「抱怨」為企業「利潤」的法則**，台北：遠流出版公司。
5. 陳耀茂（1997），**服務品質管理手冊**，台北：遠流出版公司。
6. 佐藤知恭（1988），**顧客抱怨處理實務**，台北：臺華工商圖書公司。
7. 陳淑娟（1995），**「零售賣場設計與現場消費行為關係之探索性研究--以台北市百貨公司為例」**，碩士論文，元智工學院管理研究所，桃園。
8. 蔣麗君（1999），**「國內百貨公司顧客抱怨原因之實證研究」**，碩士論文，靜宜大學企業管理研究所，台中。
9. 蔡厚灼（2003），**「客訴文件探勘系統」**，碩士論文，國立成功大學資訊管理研究所，台南。
10. 陳俊達（2005），**「影響客服中心電話服務量之因素探討-以資訊服務業為例」**，碩士論文，國立中央大學資訊管理研究所，桃園。
11. 陳瑞陽（2007），**「智慧型代理人 Web Services 整合平台—客訴下游問題回饋為例」**，**電子商務研究**，第五卷，第三期，第 361-386 頁。
12. 鄭麗珍、賴美惠(2011)，**「結合知識地圖之公部門陳訴文件自動化分案系統」**，**資訊管理學報**，第十八卷，第 4 期，第 1-20 頁。

13. Wikipedia, LDA (2015) ,<https://zh.wikipedia.org/wiki/隱含狄利克雷分布>。
14. 蕭文峰 (2014),「以 LDA 為基礎的英文論文重點擷取暨測驗系統」，資訊管理暨實務研討會，第 20 期，第 812-826 頁。
15. Wikipedia, Curl 語言 (2009) ,[http://zh.wikipedia.org/wiki/Curl_\(編程語言\)](http://zh.wikipedia.org/wiki/Curl_(編程語言))。
16. 晏文珍 (2005),「利用資料探勘技術於文件分類之研究」，碩士論文，南台科技大学，資訊管理研究所，台南。
17. 許文娜 (2008),「從資訊系統導入的觀點分析客服系統之評選架構-以某壽險公司資訊系統專案為例」，碩士論文，銘傳大學管理研究所在職專班，台北。
18. 劉凱銘 (2010),「以分類時間為基礎之網際網路客訴分析」，碩士論文，輔仁大學資訊工程研究所萬維運算研究室，台北。
19. 楊明青、何曼娟、呂萬吉 (2003),「自助餐廳顧客抱怨原因之實證研究」，觀光休閒暨餐旅產業永續經營學術研討會，第 3 期。
20. 蔡明達、陳怡謙 (2007),「運動俱樂部客訴處理」，大學體育，第 92 期，第 72-78 頁。
21. 顏財發 (2012),「餐飲顧客投機客訴行為的決定因素」，運動休閒餐旅研究，第七卷，第 2 期，第 29-44 頁。
22. Jacoby, J. and Jaccard, J. J. (1981) , "The Sources, Meaning & Validity of Consumer Complaining Behaviour : A Psychological Review," Journal of Retailing, 57,4-24.
23. Singh, J. (1988) , "Consumer Complaint Intentions & Behaviour : Definitional & Taxonomical Issues", Journal of Marketing, 52,93-107.
24. Fornell, C. and Wernerfelt, B.(1987), "Defensive Marketing Strategy by Consumer Complaint Management : A Theoretical Analysis" , Journal of Marketing Research, 24 , 337-346.
25. Kolter, P.(1997), Marketing Management : Analysis, Planning, Implementation, &

Control.

26. Renoux, Y. and F.C.Allvine,ed (1973) , "Consumer Dissatisfaction and Public Policy, in Public Policy and Marketing." ,Chicago : American Marketing Association, 53-65.
27. Westbrook, R. A. (1981) , "Sources of Consumer Satisfaction with Retail Outlets",Journal of Retailing, 57,4,68-85.
28. Bitner, M. J., Booms, B.G., and Tetreault, M. S. (1990), "The Service Encounter : Diagnosing Favorable and Unfavorable Incidents",Journal of Marketing, 54,71-84.
29. Kelley, S.W., Hoffman, K.D. and Davis, M.A. (1993) , "A Typology of Retail Failures and Recoveries",Journal of Retailing, 69,4,429-452.
30. Hoffman, K.D., Kelley, S.W. and Rotalsky, H.M. (1995) , "Tracking Service Failures and Employee Recovery Efforts",Journal of Services Marketing, 9,2,49-61.
31. Burnett, J. J., Amason, R.D. and Hunt, S.D. (1981) , "Feminism :Implications For Department Store Strategy and Salesclerk Behavior",Journal of Retailing, 57,4,71-85.
32. D. Sullivan, (2001) , Document Warehousing and Text Mining, Wiley Computer Publishing.
33. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003), "Latent dirichlet allocation. " ,the Journal of machine research, 3, 993-1022.
34. Griffiths, T. L, & Steyvers, M. (2004) , "Finding scientific topics, " Proceedings of the National academy of Sciences of the United States of Americ,101(Supp11),5228-5235.
35. Wang Y. (2008). "Distributed Gibbs Sampling of Latent Topic Models : The Gritty Details", <https://cxwangyi.files.wordpress.com/2012/01/llt.pdf>

36. Tang,L (2008) ."Gibbs Sampling for LDA", <http://leitang.net/presentation/LDA-Gibbs.pdf>
37. Salton, Gerard and Buckley, C. (1988) , "Term-weighting approaches in automatic text retrieval", Information Processing & Management, 24,5,513-523
38. Salton, Gerard. (1989) ,Automatic Text Processing,Addison-Wesley Publishing Company,.
39. Feldman, R. and Hirsh, H. (1997) , "Exploiting Background Information in Knowledge Discovery from Text",Journal of Information System, 83-97.
40. Dörre, J., Gerstl, P. and Seiffert, R. (1999) , "Text Mining: Finding Nuggets in Mountains of Textual Data,Proceedings of the 5's ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" , 398-401.
41. Singh et al., L. (1999) , "An Algorithm for Constrained Association Rule Mining in Semi-structured Data",PAKDD-99,148-158.
42. Klemettinen, M., Mannila, H., and Verkamo, A. I. (1999) , "Association Rule Selection in a Data Mining Environment",PKDD-99,372-377.
43. H. Chen, (2001) , Knowledge Management Systems — A Text Mining Perspective.

附錄

附錄 A 關鍵字出現次數與 TF-IDF 計算結果

人民幣	人員	人家	上衣	下標	千里馬	口氣	口袋	大小
47	193	152	79	647	98	188	46	41
0.1494	0.0986	0.0585	0.1093	0.0604	0.2739	0.0557	0.1800	0.0809
大衣	大陸	女孩	小姐	小時	中心	內衣	內容	公司
56	180	76	149	224	163	52	247	549
0.1731	0.1133	0.1497	0.1313	0.0860	0.1619	0.2411	0.0529	0.0947
公告	尺寸	心得	心情	心態	手機	手續費	文字	文章
80	351	158	145	43	273	95	66	589
0.1005	0.1170	0.0894	0.0538	0.0618	0.0827	0.2590	0.0660	0.0299
方式	方法	日期	毛衣	代表	功能	包裝	包裹	台幣
326	65	153	136	91	41	186	682	49
0.4887	0.0655	0.1116	0.2434	0.0547	0.0930	0.0967	0.0961	0.1325
司機	外套	布料	平台	民法	生氣	生意	白色	立場
135	263	52	120	41	295	149	95	45
0.2443	0.1522	0.1296	0.1377	0.1209	0.0534	0.0603	0.1867	0.0682
交易	先生	全家	全額	印象	同事	名字	名單	回信
400	65	118	64	101	82	75	107	288
0.0618	0.1694	0.1777	0.0989	0.0713	0.1317	0.1241	0.0959	0.0894
地方	地址	好評	成本	灰塵	羊毛	老闆	色差	行為
183	306	93	85	82	54	116	206	70
0.0535	0.1140	0.1163	0.0683	0.3338	0.3943	0.1228	0.1964	0.0653
衣服	佐證	作業	我家	我會	折扣	材質	男友	私人
1180	638	52	86	123	138	185	51	41
0.0982	0.0144	0.1247	0.0975	0.0526	0.0996	0.1258	0.1475	0.1213

附錄 A 關鍵字出現次數與 TF-IDF 計算結果(續)

系統	事件	事情	事實	來信	受害者	味道	垃圾	官方
357	104	328	74	89	55	101	50	105
0.0784	0.0591	0.0483	0.0537	0.0864	0.0750	0.3095	0.1066	0.1488
店面	店員	朋友	東西	法律	泳衣	物品	物流	狀況
101	151	438	1447	141	40	155	227	403
0.0896	0.1664	0.0936	0.0376	0.0731	1.4496	0.0694	0.1315	0.0490
狀態	糾紛	金額	門市	信用卡	信件	信箱	品牌	品質
140	75	274	122	93	184	115	118	584
0.0874	0.0606	0.0774	0.2074	0.1706	0.0671	0.0883	0.1352	0.0874
契約	客人	客戶	建議	星期	洋裝	流程	紀錄	紅包
65	247	117	274	84	165	115	151	57
0.0913	0.0631	0.0823	0.0463	0.0874	0.1793	0.0762	0.0918	1.5945
紅色	背心	背帶	苦主	訂單	負擔	重量	重點	頁面
64	52	42	43	965	128	247	209	122
0.1978	0.2553	0.6790	0.0987	0.1027	0.0901	0.2722	0.0482	0.0655
風衣	風險	倉庫	個人	原因	原價	員工	家人	差異
41	97	90	153	127	57	55	46	44
0.1607	0.0684	0.1265	0.0550	0.0666	0.1079	0.1465	0.0783	0.0913
庫存	效率	時尚	時候	海關	消息	消費者	特價	特價品
87	89	43	709	159	121	506	116	52
0.1132	0.0944	0.1881	0.0412	0.3143	0.0700	0.0747	0.1271	0.1432
留言	粉色	粉絲	訊息	高調	動作	動態	商店	商品
87	40	195	235	207	58	46	50	2803
0.0876	0.1817	0.1096	0.0833	0.1453	0.0660	0.1099	0.0930	0.0111

附錄 A 關鍵字出現次數與 TF-IDF 計算結果(續)

商城	問題	國際	專業	專櫃	情形	情況	教訓	清單
300	2043	146	81	72	132	224	65	52
0.1843	0.0312	0.1391	0.1326	0.1581	0.0726	0.0537	0.0628	0.1337
現貨	理由	產品	疏失	盒子	習慣	袋子	袖子	規定
240	146	99	47	57	58	69	61	326
0.1280	0.0618	0.1045	0.0702	0.1561	0.0481	0.1534	0.1852	0.0886
規則	貨物	貨品	貨運	責任	通知	速度	部分	備註
88	84	95	354	174	709	178	238	85
0.1154	0.1014	0.0761	0.1466	0.0500	0.0606	0.0808	0.0558	0.0795
單子	帽子	掌櫃	期限	款式	款項	牌子	發票	程序
43	47	298	86	62	127	51	427	46
0.1207	0.1755	0.2620	0.0951	0.0930	0.0842	0.1597	0.2562	0.1075
答案	結果	結論	買家	費用	進度	郵局	郵資	集團
42	1187	57	898	84	156	154	207	54
0.0826	0.0265	0.0592	0.0574	0.0971	0.0780	0.0992	0.1304	0.1935
順序	黑名單	黑色	匯款	塑膠	媽媽	意見	意思	感覺
51	243	99	479	71	66	77	181	979
0.1470	0.0880	0.1737	0.0935	0.147	0.1414	0.0654	0.0468	0.0401
損失	新品	會員	照片	瑕疵	經驗	落差	號碼	誠意
65	128	127	563	845	304	46	68	161
0.0640	0.1274	0.2268	0.0769	0.1125	0.0463	0.0919	0.1105	0.0822
資料	資訊	運氣	運費	過程	道理	鉤釘	電子	電腦
273	120	57	834	146	46	64	50	87
0.0849	0.0756	0.0850	0.0794	0.0603	0.0523	0.2240	0.1819	0.1072

附錄 A 關鍵字出現次數與 TF-IDF 計算結果(續)

電話	靴子	圖片	團員	實際	實體	對方	態度	疑問
746	41	158	68	135	103	389	1119	42
0.0808	0.2337	0.0652	0.1954	0.0676	0.0766	0.0803	0.0533	0.0540
網友	網頁	網站	網路	語氣	誤差	說法	銀行	鼻子
62	332	312	247	57	44	85	80	43
0.0814	0.0666	0.0870	0.0639	0.0755	0.1192	0.0553	0.1600	0.0844
價值	價差	價格	價錢	廠商	廣告	彈性	數量	樂趣
45	44	256	164	482	75	41	80	97
0.0653	0.0878	0.0681	0.0740	0.1449	0.0983	0.1844	0.1143	0.6341
標的	樣子	箱子	範圍	賣方	賣家	質料	質感	鞋子
69	112	55	44	77	4601	59	195	349
0.0779	0.0542	0.1182	0.0761	0.0869	0.0718	0.0996	0.1106	0.1986
鞋底	機制	機會	褲子	辦法	錯誤	幫手	講話	禮物
57	51	61	177	330	121	203	45	50
0.1933	0.0925	0.0659	0.2040	0.0495	0.0780	0.1145	0.0679	0.1213
禮拜	簡訊	藍色	雙方	顏色	證據	贈品	關係	襪子
238	252	63	79	367	73	104	147	50
0.0689	0.1169	0.1551	0.0685	0.1496	0.0743	0.3370	0.0490	0.4093
露天	顧客	權利	權益	襯衫	邏輯	體積	罐頭	
229	81	50	123	70	57	40	44	
0.2407	0.0851	0.0817	0.0581	0.1138	0.0830	0.2780	0.1148	