南華大學科技學院資訊管理學系

碩士論文

Department of Information Management
College of Science and Technology
Nanhua University

Master Thesis

運用文字探勘探討美容議題－以 PTT 美容版為例

Applying Text Mining Techniques to Salon Issues on PTT Beauty Salon

楊嬿庭

Yen-Ting Yang

指導教授：洪銘建博士

Advisor: Ming-Chien Hung, Ph.D.

中華民國 110 年 12 月

December 2021

南 華 大 學

科技學院資訊管理學系

碩士學位論文

運用文字探勘探討美容議題-以 PTT 美容版為例

Applying Text Mining Techniques to Salon Issues on PTT BeautySalon

研究生：楊嬿庭

經考試合格特此證明

口試委員：翁富美

謝定助

洪銘建

指導教授：洪銘建

系主任(所長)：陳信良

口試日期：中華民國 110 年 12 月 25 日

南華大學資訊管理學系碩士論文著作財產權同意書

立書人：楊嫿庭 之碩士畢業論文

中文題目：

運用文字探勘探討美容議題-以 PTT 美容版為例

英文題目：

Applying Text Mining Techniques to Explore Salon Issues on PTT Beauty Salon

指導教授：洪銘建 博士

學生與指導老師就本篇論文內容及資料其著作財產權歸屬如下：

☑ 共同享有著作權

☐ 共同享有著作權，學生願「拋棄」著作財產權

☐ 學生獨自享有著作財產權

學　　生：楊嫿庭 （請親自簽名）

指導老師：洪銘建 （請親自簽名）

中華民國 111 年 6 月 5 月

致　　謝

首先感謝身邊很多師長和同學，陳信良和謝定助教授的引薦下，讓我進入了資管所的大家庭，碩士期間在洪銘建教授的指導中，使我對於研究有更深一步的了解，平日也督促我們要向前邁進，藉此引發我自學的能力，學習如何撰寫程式和培養邏輯訓練，再來也要感謝碩士班的班導邱宏彬及陳宗義導師，謝謝您們對於碩士班學生的關心，每學期的一次聚餐活動讓大家在繁忙中利用課餘聯繫感情，最後謝謝系助理琪琪姐協助我們論文口試等前置作業。

也要謝謝 Pyhton 家族的大家，寒假期間我們互相勉勵互相學習，在大家的支持和陪伴下讓我能更加勇敢去面對自己不擅長的事情。此外也要非常感謝家人的支持和鼓勵，碩士就讀期間體諒我的情緒不給予我過多的壓力，疫情嚴峻時還願意花時間撥空來回接送返校，也要謝謝很多身邊的朋友的鼓勵和支持，最後我想謝謝自己，願意投入和學習新事物，雖然過程中較為波折，但終究還是完成了這份任務，使我從這過程中成長。

運用文字探勘探討美容議題－以 PTT 美容版為例

學生：楊嬿庭

南　華　大　學　資訊管理學系碩士班

摘　　要

科技網路的發達下，人與人之間的交流不再受時間和空間的限制，可以立即的透過網路分享資訊，讓資訊的傳遞更為快速，在新冠疫情影響之下群眾更加依賴網路帶來的生活便利，為了配合防疫措施在無法外出的情形下，民眾開始大量使用網路購買所需用品，但因為無法看到實體物品，消費者為了降低購買前的不確定感，因此形成消費者在購買前會先看網路產品評論的習慣。從 PTT 各大版中有數不清的產品評論，大量來自於消費者使用心得，潛在消費者會因為評論而影響原先的購買意願，因此本研究針對 PTT 美容版文章，利用文字探勘和爬蟲技術針對使用者對美妝產品的購後心得進行網路口碑分析。研究結果發現在 PTT 美容版中最常被提及的品牌為雅漾、蘭蔻、克蘭詩以及碧兒泉，並從各品牌之中以主題模型分析出關於品牌產品的「保濕」和「吸收」成效是消費者較為著重，以及在行銷上消費者會以紙本式「DM」當作購買前的參考指標，因此本研究建議品牌可以在產品目錄設計上更符合消費者的設計需求。

關鍵字:網路爬蟲、文字探勘、主題模型

Applying Text Mining Techniques to Explore Salon Issues on PTT Beauty Salon

Student: Yen-Ting Yang Advisor: Ming-Chien Hung, Ph.D.

Department of Information Management
Nanhua University
Master Thesis

ABSTRACT

With the intelligence of the technology network, the communication between people is no longer limited by time and space. Immediately share information through the Internet, and under the influence of the Covid-19, the people rely more on the convenience in life brought by the Internet. Cooperating with epidemic prevention measures, people began to use the Internet to buy supplies in large quantities under activities that were unable to go out. Due to being unable to see the products physically, consumers will read online product reviews before buying in order to reduce their sense of uncertainty. There are countless comments in the major editions of PTT, and a large number of them come from consumer experience. Potential consumers' initial purchase intentions will be affected due to the influence of the reviews on the product. because of product reviews. In order to extract the most frequently mentioned keywords from a large amount of text, this research aims at PTT beauty articles, using text mining and Web crawler to explore users' post-purchase experience of beauty products and conduct online word-of-mouth analysis. The results of the study found that the most frequently mentioned brands in the PTT beauty edition were "Avene", "Lancome", " Clarins" and "Biotherm". By analyzing the products from each brand using the LDA model, the results found

out that consumers pay more attention to the effects of "moisturizing" and "absorption", and consumers will use the paper-based "DM" as a reference index before purchase in marketing. Therefore, this study suggests the brands product catalog designs can made based on the needs of the consumers.

Keyword: Web Crawler, Text Mining, LDA

目錄

圖目錄

表目錄

第一章 緒論

第一節 研究背景與動機

「顏值經濟」的興起近年來越來越多人重視個人的整體外形和樣貌，過去除了女為悅己者容之外，男性也開始注重保養，讓美妝行業受消費者的重視，化妝品品牌也如雨後春筍般越來越多元化，消費者擁有眾多的選擇反而不力於品牌發展，但如何從大量的競爭者裡面做出口碑和特色，變成每一個企業所要重要的發展方向。品牌通路的選擇也會影響到品牌的定位和價值，以往化妝品的通路大多以百貨公司或實體店面為主，隨著科技蓬勃發展，訊息交通的發達使得人們已跟過去有所不同，最大的差異即是購物方式的轉變(黃嚴弘 & 黃瓊儀，2018)。

根據財團法人台灣網路資訊中心的 2019 年台灣網路報告，12 歲以上的民眾有近九成有上網經驗，而當中有六成五的民眾利用上網時間購物，以購物類別來看則以 PChome、momo 與蝦皮購物為主要的購物平台(財團法人台灣網路資訊中，2019)。疫情的驅使下，許多零售業都面臨了轉型的需求消費模式因此也產生了變化，根據經濟部統計處的數據指出，2020 年第二季，電商占整體零售的銷售 8.8%，與美國雖還是較低(11.5%)，卻已是台灣歷年來創新高(台灣經濟部統計處，2021)。凱度數據指出網購消費者的忠誠度開始逐漸增加，這代表網購已經是後疫情時期的新常態，在享受了網購的便利性之後，之後就很難回到過去了(KANTAR,2021)。

過去在購物的過程中，消費者會考慮到產品特色、價位、效能以及包裝，這些都是消費者在購買時會考量到產品的能提供的附加價值，需要準確抓住消費者的想法，品牌需要透過使用者的反饋才能得知消費者所想，了解消費者對於品牌

的期望以及想法。一直以來人除了透過口耳相傳的方式進行資訊傳遞外，網路社群論壇的興起消費者會在網頁上分享自己得知道相關資訊，社群論壇可以增進彼此的溝通，強化團體內的信仰，當社群論壇發生在網路上，人與人之間不但可以更迅速交換更多訊息，且所能觸及的範圍也更加廣泛(程婉婷，2017)。因此本研究想利用社群論壇中消費者所留下來的資訊和文字，做進一步的探討和研究，統整社群論壇中最常被提及的四大品牌，以及四大品牌中最常出現的關鍵詞，並分析關鍵詞所產生的主題模型，進一步挖掘出消費者對於品牌的使用建議和心得，最後提供給品牌作為一個參考。

第二節 研究重要性

Kolter（1997）曾調查歐洲七個國家的7000 位消費者，有60%的人承認是受親友影響而採用一項新品牌。Dichter（1966）表示，消費者對產品或使用經驗的涉入會導致後續的口碑溝通，過去因為網路並不普及，口碑傳遞僅限於人與人之間有限的範圍，在傳遞的過程中會到受距離和時間的阻礙，但由於近年來網路的興起，讓資訊分享變得更加及時和便利，因此消費者能透過網路不限時間和地點的方式，盡情的在網路社群中分享自己所擁有的知識和資訊。過去想瞭解消費者的想法可能是透過電話訪問以及問卷收集等方式，但現在網路社群的便利下，消費者可以到任何網路社群論壇進行產品的分享和討論，大量的網路資訊也影響到品牌的評價，無形中形成該品牌的網路口碑。

Godes(2004)認為由消費者在網路上提出關於產品資訊的評論，可被視為特殊類型的口碑，眾多的研究結果發現口碑比其他溝通媒介更具潛在影響力(Arndt, 1967)，因為口碑訊息是在非商業意圖的動機下，由有使用經驗的一方傳播(程婉婷,2017)，當消費者認為訊息來源的可信度越高，受到該口碑的影響越大，購買意願也隨之越高(Laffertyet al., 1999)。網路口碑的產生也導致潛在的消費者會觀看相關評價，間接地影響到潛在消費者是否想購買的想法，因此可以發現網路口碑對於品牌是有相當的重要性，可以很快速的瞭解消費者使用心得，讓品牌能提供更貼近消費者的服務和產品。

第三節 研究目的

網際網路和社群論壇的興起，使用者可以不受任何限制下傳遞各種訊息，而網路上資訊傳遞的方式包含了以文字型態、圖片展示以及影片內容等方法，在這些傳遞方式過程當中，純文字的結構更是發展歷史悠久，這些文字的產生方式都是利用平日裡人類所產生的自然語言，其中隱藏了具有影響力和價值的資訊，直接的表達出每個人的想法。因此文字探勘的目的，是希望利用資訊技術開採出重要的訊息，除此之外，許多研究都指出，消費者在購物前會先收集產品相關資訊和評論，社群論壇除了方便讓人進行虛擬社交之外，也變成資訊傳遞的網路平台。因此本研究目的以 PTT 美容版作為資料來源，已非結構化的文字來發掘和取得文章中潛藏的重要訊息，最後以主題模型進行分析和探討，從文章中找出被消費者最常提及的前四大熱門品牌，從熱門品牌文章中找出各品牌獨特的主題模型，挖掘重要訊息使品牌瞭解消費者最常發布甚麼型態的文章內容，和常出現的關鍵詞並提出各大品牌的獨特點和共通點，提供給品牌作為一個參考依據，協助企業瞭解消費者對於品牌的想法以及意見。

第二章 文獻探討

第一節 化妝保養品產業

一、概述

過去化妝品(Cosmetics)一詞為希臘文的 Kosmetikcs，意為裝飾技巧。洪偉章、李金枝、陳榮秀(1997)提出基礎化妝品也就是所謂的保養品(Skin Care Products)可分為有清潔和潤膚成分，隨著消費者的不同需求，化粧保養產業每年會產出許多不同種類的產品，為了更加明確定義和區分化妝保養品，台灣於 107 年修正化妝品衛生安全管理法，定義化妝品指「施於人體外部、牙齒或口腔黏膜，用以潤澤髮膚、刺激嗅覺、改善體味、修飾容貌或清潔身體之製劑」。

二、未來發展

近年來化妝和保養備受重視，根據我國統計部指出藥品、醫療及化粧品零售業於民國 108 年度之營業額將近新臺幣 2000 億元，其中以化粧保養品為大宗，佔近 5 成的銷售量。根據報導顯示美妝產業因電商和社交平台快速發展，以及關鍵領袖(Key Opinion Leader ,KOL)的影響力之下，推動了「顏值經濟」的新市場。此外，新冠疫此情下加速化妝保養品產業在網路上的發展，美妝市場十大趨勢的議題中指出未來美妝產業需朝向「線上線下，全通路購買。」市場調查機構益普索 Ipsos 調查下指出 60%的消費者會在線上購買之前在網上調查好想要購買的品牌與產品、27%的消費者會在實體店面完成購買。由此可知產品的網路資訊越來越為重要，DCB 產業分析師陳玲玉表示，全球的化粧保養品產業也逐漸進入新智慧科技，因為疫情爆發後更加速化粧保養品產業朝向智慧化發展，包含虛擬實境(Virtual Reality,VR)、擴增實境 (Augmented Reality,AR)、人工智慧(Artificial Intelligence,AI)和大數據(Big Data)等技術，化粧品結合物聯網和 AI 分析提供及時客製化保養品，透過數據即時了解消費者的需求，以及加以改善消費者的皮膚狀況，由此可知化粧保養品的網路市場以及新智慧科技的加入，會使台灣的化妝品產也有新的突破，讓美妝市場有未來蓄勢待發，開展出全新商機。

第二節 網路口碑

口碑深深地影響每個人做選擇的意願，透過非正式交談的方式，人與人之間交流資訊的模式，口碑也是消費者搜尋外部資訊的工具，屬於人際之間的資料來源（Podoshen,2008）是消費者制定決策的重要依據(Cheung et al. ,2008) 尤其是進行高風險產品的選擇時，消費者更會傾向搜尋口碑資訊，以降低決策前的知覺風險（Perceived Risk）（Harrison-Walker, 2001），對企業而言，是成本低廉，傳遞快速的行銷策略（Trusov et al. , 2009），最終將口碑價值會轉換為購買意圖，完成整個口碑行銷的過程(江義平、蔡坤宏、黃耀德，2015)。

近年來網際網路的發達下，讓口碑的傳遞方式變得更加多元，因此形成了所謂的網路口碑(Electronic Word-of-mouth；eWOM)。影響層面十分廣大，因此被企業所重視，學術研究中以網路口碑為議題探討不同行業的網路口碑，利用不同的角度分析網路口碑和銷售之間的關係，劉思妤(2019)以高科技產品為研究對象，發現網路口碑與在線評論對於產品銷售量都有正向之影響。除此之外，網路口碑本身又能區分為正負向口碑，在呂珊妤(2018)研究發現，消費者之心中期望與實際經驗比較後，實際經驗高於期望將產生正面口碑，反之則產生負面口碑。李國榮、顏暄祐(2019)發現正面網路口碑可以顯著提高企業股價報酬率；負向網路口碑將不利於上市貿易百貨類之企業營收。從不同的行業都可證實，網路口碑確實有實質的影響性，因此需要更加運用在不同行業中進行研究。

在網路口碑的影響下，購物前的資料收集也越來越為普遍，網路帶給人們更加便利的生活，讓消費者擴大範圍對商品蒐集相關資訊的能力，也利用網路搜尋、回饋及分享訊息。就如網路口碑，例如社群媒體、聊天室、佈告欄、資訊網、電子郵件、LINE 群組、Messenger 及論壇。「網路社群口碑需求」調查發現，高達81%的消費者在購物前，有使用網路搜尋口碑訊息的習慣，前三名分別為社交網站（45.8%）、討論區（44.7%）、部落格（33.1%）(資策會，2014)。由 Etienne et.al.(2018)等人提出社群媒體與論壇更是搜索商品、服務及企業的主要來源，研究指出對企業而言，比較郵寄問卷或市場調查等方式，企業從網路討論區、論壇及網路社群中，更能發掘出顧客對產品的真實想法與實際需求，因此本研究以社群論壇的討論區為主要研究對象，探討化妝品牌的網路口碑，並給予適時的行銷建議。

第三節 文字探勘

近年來在社群平台的擁簇下，為了從錯綜複雜的文字訊息中擷取和挖掘有用的資訊，因此開始大量使用文字探勘技術從中找到文字之間的關聯性以及重要訊息，做為企業未來發展的參考依據。在可觀之商業利益的促使下，對電子化訊息解讀的需求速度增加，近幾年社群媒體越來越多元，像是 Facebook、PPT、Instagram 部落格等社群平台由人類的自然文字語言所形成，並且主要以非結構化的方式呈現，因此可利用文字探勘從大量且非結構性的文字訊息中獲取有價值意義的技術(潘彩君，2018)。

網際網路的盛行下，網購已經變成人類生活中不可或缺的一環，消費者會在購買完產品後分享至社群中，而林名彥(2015)以 PTT 社群論壇的網購版為研究對象，針對網購商品的客戶抱怨來做探討，以文字探勘技術尋找文中的關鍵字，並了解網友們時常關注的主題和關聯的字詞。除此之外，部落格盛行許久許多旅遊達人很樂於分享相關訊息，林孟龍、張浩為、張宗正(2019)以部落格的文章進行分析，利用文字探勘的形式來探討並了解遊客對日本環球影城的真實想法。近年來 Dacard 論壇也成為年輕人的最愛，吳俊緯(2021)以現在大學生常用的 Decard 的旅遊版為研究來源，探討顧客在網路上較為關注的需求識相，協助旅遊業者主動積極的了解顧客需求並提出相關意見，創造旅遊業者的競爭優勢。從中可以得知，文字探勘運用非常多元，可做為文字採礦、智慧型文字分析、文本資料庫中的知識挖掘，是一種對大量文件資訊進行編輯、整理及分析的過程，用來發掘文件資訊的關鍵資訊和關鍵資訊之間的關聯性(Sullivan,2001)。

Archak(2011)認為利用文字探勘來分析評論會是一個新的研究領域，每個人使用的方法不盡然相同，但是最終目標都是將雜亂的文字訊息轉換為有意義的文字訊息(Tanger al., 2015)。使得資料探勘熱潮延伸出文字探勘處理巨量資料的過程，仰賴電腦軟體的計算(陳世榮，2015)， Abdous & He(2011)將文字探勘其分為四步驟：

一、資料收集：

找出可用文字資料集，例如利用資料庫尋找或是撰寫爬蟲來擷取可用的資料。

二、資料預處理：

以字詞來說就是進行斷詞作業，將文件中的各個詞斷開來後整理，並可以過濾掉不適合分析的詞性及冗詞。

三、資料分析：

以關鍵字詞分析為主，擷取方法多以 TF-IDF 為主 TF (Term Frequency)，是指詞頻依詞彙出現的次數，以字詞在某一文件中的出現頻率代表該字詞的重要性，如果出現頻率越高則表示越重要，而 IDF(Inverse Document Fequency)是出現在文件數中的比例，並依一定公式計算關鍵詞的權重，屏除單一文件同一字詞出現過多次造成的偏差。

四、視覺化可用資訊：

將分析完成的資料以視覺化的圖表來呈現，例如文字雲可以看出哪些字詞出現次數多，關鍵詞網路圖可以判斷詞彙與詞彙之間的關係，顯示核心字詞出現時其它可能會隨之出現的字詞。

第四節 自然語言處理

自然語言(Nature Language)是人類溝通時所發產生的語言，而發展自然語言處理(Natural Language Processing, NLP)的目的就是讓電腦去理解人類的自然語言所做的事情。Chowdhury (2003)指出自然語言處理是一個研究和應用領域，探索如何使用計算機來理解和操縱自然語言文本或語音以做有用的事情。NLP 的發展有利於在網路蓬勃發展的世代，現今多數的資料型式是以非結構化所呈現，包含影音、圖片、音訊、文章等，尤其是文字佔據了非常大的空間，為了能分析和利用這些非結構化的文字，因此需要利用 NLP 技術，讓機器理解這些文字並從中取得更多有用的訊息。

NLP 是現今極具潛力的一種應用，像是進行內容分類、文本分析、情緒分析、語音到文本和文本到語音的轉換、文檔摘要以及機器翻譯等多元的應用。因此許多學者也朝向這方面來進行鑽研，像在中文文字偵錯的研究，張杰暄(2020)利用最先進之自然語言模型 BERT 進行文字的探勘，提高偵測中文文字錯別字的準確度。此外，NLP 也可以加入深度學習預測未來趨勢，夏鶴芸(2020)利用 BiLSTM 和 Transformer 預測股價走勢，並且以股價的漲跌和情緒辭典對新聞進行 BERT 情緒標籤的分類，結果發現 Transformer 模型搭配股價與相關指數資料集效果較佳。NLP 可以用來分析歷史言情小說，林紋羽(2020)以《如懿傳》和《延禧攻略》的小說進行實驗，利用文本斷詞和詞頻以人工方式辨識人物，以說話方、聽話方和對聽話方的稱呼三個對話元素來識別小說中角色人物的多元關係，由此可知 NLP 的技術已經成功進入到我們的生活，透過 NLP 理解自然語言，也可以挖掘語言背後所蘊含的深層含義。

第五節 主題模型

一、介紹隱含狄利克雷分佈

隱含狄利克雷分佈(Latent Dirichlet Allocation,LDA)於 2003 年由 Blet et al.(2003)等人提出，利用狄利克雷(Dirichlet)作為多項式的共軛先驗，是一種以非監督式學習法找文件集合或語料庫中潛藏主題，文檔由許多不同主題所構成，主題則和詞彙有共同關係所共同組成。以下是 LDA 的概念，針對不同主題下的詞彙以顏色作為標示，表達了 LDA 的思想：

"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 1 LDA 文檔案混合主題及詞彙

(圖片來源:Blei et al. 2003)

Blei et al.(2003)定義 LDA 生產過程:

(一)、對每篇文章抽取其中的一個主題。

(二)、從上述取出的主題中分析出對應的單詞進行主題單詞抽取。

(三)、反覆以上的操作過程，直到抽完文件的每一個單詞。

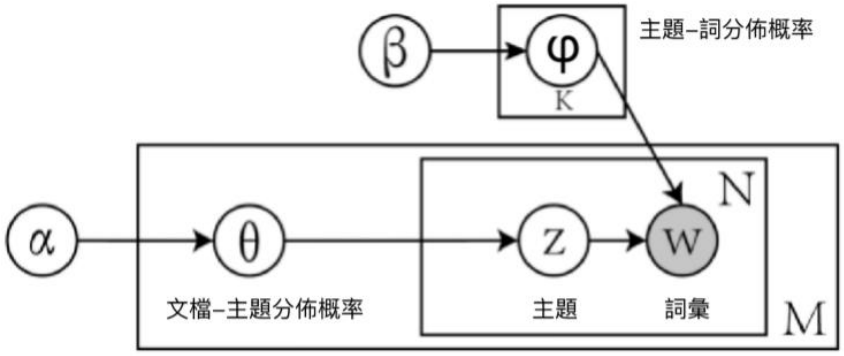


圖 2 LDA 概念表示圖

(圖片來源 : 沈嘉誠)

表 1 LDA 概念表示圖符號說明

符號	含意
α	主題事前狄利克雷先驗參數
β	詞彙事前的狄利克雷先驗參數
θ	主題分布概率
φ	分詞分布概率
k	主題數
z	文檔中的主題 z
w	文檔的詞彙
N	文檔中的詞彙總數
M	文本集合中的文檔總數

二、LDA 應用

現今文本挖掘領域應用於文本主題識別、文本分類以及文本相似度計算。文字一直以來是人們溝通的模式之一，因為網路的發達，讓人們可以不受時間和地點的限制，自由的使用社群、論壇以及部落格等發表個人言論及觀。因此讓許多學者開始研究社群論壇中的評論和留言，並利用主題模型來深入挖掘文字背後的重要訊息，張宇鋒(2020)以 TripAdvisor 旅遊資訊平台作為研究來源，來探討旅客在各星級飯店評論主題使用的字詞差異，並用主題模型進行比對發覺相互映的地方。

除了研究評論，也可以加入深度學習的方式，加深文字挖掘的應用，趙彥淳(2019)利用臉書粉絲頁貼文中的特徵關鍵詞彙，透過深度學習方法結合貼文屬性資料建構收視率預測模型，預設收視趨勢，並以 LDA 進行分析提供業者經營粉絲專業時的參考依據。也可以以 LDA 來對論壇的文章分類出主題群，郭泓志(2018)以 PTT 論壇作為研究資料來源，並以 Word2Vec 模型分類出文章的留言，觀察文章和留言之間的相關性，進一步瞭解論壇的狀況。由以上的研究可以發現 LDA 可以加入不同的研究方式來探討不同領域的文字，並且擷取出在社群論壇中的重要資訊，提供給經營者做為參考，利用文字來發現背後的重要資訊變得極為重要。

第三章 研究方法與工具

第一節 研究架構

研究架構如下圖 3 所示，本研究以社群論壇 PTT 美容版為研究對象，使用文字探勘技術和 Pyhton 的 Jupyter Notebook 作為編輯環境。整體的研究架構中，首先，以網頁爬蟲技術進行資料收集，其次，搭配中央研究院所研發的 Ckiptagger 工具結合文字雲擷取出主要的研究主題文本，再使用近期最熱門的 Jieba 工作進行中文斷詞，針對斷詞後的結果以隱含狄利克雷分布來進行文本挖掘之研究分析，加以深入探討並提供相關研究建議。

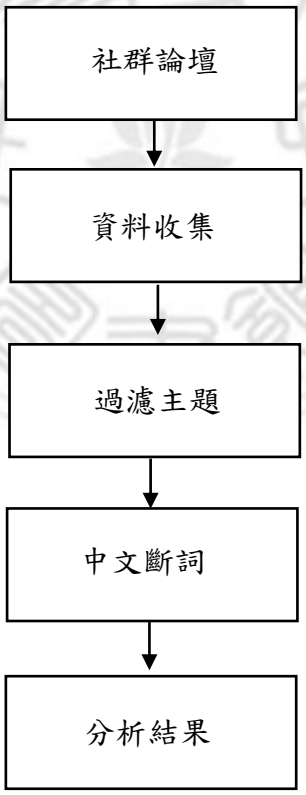


圖 3 研究架構

第二節 研究工具

一、資料擷取

Python 中可使用 Request 模組模擬瀏覽器向所要擷取的網站伺服器提出 HTTP 請求，並能取得回應內容。常見的方式有二，POST 與 GET，伺服器接受到請求後可以回傳各種檔案，像是 HTML、JSON 和 CSS 型態呈現回應，下表 2 和 3 將區分 POST 和 GET 的使用說明及優缺點比較。

表 2 POST 和 GET 資料請求方式說明

請求方法	說明	語法
POST	不會改變網址內容，POST 請求方式多半應用在擷取網頁表單上。	requests.post (網址, data = 字典)
GET	GET 請求方式會將傳遞的資料直接放置網址後方。	requests.get(網址)

表 3 POST 和 GET 優缺點比較

請求方法	優點	缺點
POST	資料的隱密性和安全性較高。	資料打包需使用較長時間。
GET	傳遞資料的速度較快且方便。	資料較容易被看見，隱私性較低。

二、網路爬蟲

網路爬蟲（Web Crawler），也叫網路蜘蛛（Spider），是一個自動抓取網站資料的過程，在訊息爆炸的時代中，人工方式已經無法負荷大量的網路資料，因此利用網路爬蟲的技術來挖重要資訊已變成現今重要的資料收集方式。網路爬蟲包含了兩類，分別是網路爬蟲(Web Crawler)和網頁抓取(Web Scraper)其中運作模式並不相同。網路爬蟲，主要是透過搜尋引擎像是 Google、Yahoo 透過網路爬蟲將抓到的資料存取下來；網頁抓取，則是從特定的網頁中抓取頁面裡的內容，更著重在非結構化的資料上，常見的非結構化有 HTML 格式，抓取後再存到資料庫或者是電子試算表中，並近一步進行資料的分析，而在習慣上統稱這兩種方式為網路爬蟲。

本研究主要採取的方式為網頁抓取，抓取 PTT 美容版的網頁內容，並將爬取下來的資料各別儲存在 Excel 試算表中後續再進行收集、整理以及分析訊息，以利進行文字探勘之研究，網頁抓取的最典型常見方式為使用 Python 中的 Beautiful Soup 模組，以下將針對 Beautiful Soup 模組作初步介紹。

Beautiful Soup 名字取自劉易斯·卡羅爾在《愛麗絲漫遊仙境》裡的同名詩歌。是一個用來解析 HTML 解構的套件(Package)，要使用 Beautiful Soup 來解析網頁 HTML 結構前需要先使用 Request 模組把將要擷取的網頁 HTML 程式碼取回來，瞭解完網頁的架構後，就可使用 find 來進行節點的搜尋，傳入要搜尋的標籤名稱。

三、中文斷詞

文字探勘(Text Mining)是一種透過資訊技術分析文本中意見、情緒及感受的技術(Big Data Finance,2019)。廣泛運用在社群媒體上，探討民眾所發布的社群文章內容，從中擷取重要資訊做為未來的研究方向。近年來中文斷詞技術越來越為成熟，尤其是在這文字傳遞快速的時代裡，文字已經變成人跟人之間的溝通語言之一，因此若要理解一篇文章首先需要經過斷詞處理，也就是說要讓電腦擁有理解自然語言的能力，事先進行自然語言處理(NLP) 中研院資訊科學研究所的馬偉雲助研究員說明：「以中文來說，最基本的，要先教電腦學會「斷詞」和「理解詞的意思。」中文斷詞最為流行的就是由 Python 所發展的 Jieba，除此之外，台灣近年來也著重在中文斷詞的研究，中央研究院也組成了詞庫小組致力於研發中文斷詞系統。

本研究利用中研院所研發的 Ckiptagger 和 Jieba 進行斷詞，首先，在進行初次的 PTT 美容版爬蟲時先以 Ckiptagger 進行中文斷詞，雖然結果效果顯著但是因為電腦 CPU 的因素導致數度上較為緩慢，因此在進行關鍵字爬蟲時以 Jieba 進行二次斷詞處理，其中在 Jieba 斷詞中除了內建的詞典外，也自行加入了自建詞典和停用詞典，讓中文斷詞可以更為準確，以下將針對 Ckiptagger 和 Jieba 做詳細的介紹以及說明套件安裝方式。

(一) 、中研院斷系統 Ckiptagger

Ckiptagger 的特色加入了結合實體辨識(Named Entity Recognition,NER)，並且以深度學習模型為基礎的 NLP(自然語言處理) 應用。其用處在文本資料當中，可以辨識出 11 類一般領域的專有名詞以及 7 類數量詞，包含了「人名」、「團體」、「設施」、「組織」、「地理」、「地點」、「商品」、「事件」、「藝術品」、「法律」、「語言」、「日期」、「時間」、「比例」、「錢」、「數量」、「序數」和「數詞」共 18 類。下表為中央研究院詞庫小組 CkipTagger 和 Jieba-zh_TW 比較表，經過 ASBC 4.0 測試集可以觀察到 Ckiptagger 的斷詞以及詞性標註的效果優於結巴系統，在功能上相較於舊 CKIP 系統增加了中文處理工具包含了，斷詞(WS)、詞性標註 (POS)、實體辨識 (NER)。

此外也增加了以下幾種功能:

1. 加強斷詞表現
2. 無法自動增加或刪除
3. 修改文字、可支援不限制長度的句子
4. 可支援使用者自行加入的參考或強制辭典

表 4 CkipTagger 和 Jieba-zh_TW 比較

Tool	(WS) prec	(WS) rec	(WS) f1	(POS) acc
CkipTagger	97.49%	97.17%	97.33%	94.59%
Jieba-zh_TW	90.51%	89.10%	89.80%	--

(資料來源：中央研究院詞庫小組)

（二）、Ckiptagger 安裝步驟如下所示

步驟	方法
安裝套件	套件可分為三種方式安裝，安裝之前先安裝 tensorflow。 標準式 pip install -U ckiptagger[tf,gdown] 簡易式 pip install -U ckiptagger 完整式 pip install -U ckiptagger[tfgpu,gdown]
下載模型檔案	模型檔可在幾個不同地方下載，並存放在電腦路徑。 iis-ckip gdrive-ckip gdrive-jacobvsdanniel
載入模型	GPU 和 CPU 擇一安裝即可 使用 GPU 安裝 Tensorflow-Gpu 設定 Cuda_Visible_Devices 環境變數， 例如：os.environ["Cuda_Visible_Devices"] = “0” 設定 disable_cuda=False， 例如：ws = WS("./data", disable_cuda=False) 使用 CPU ws = WS("D:\data") pos = POS("D:\data") ner = NER("D:\data ")
釋放記憶體	del ws del pos del ner

（三）、Jieba 介紹

Jieba 所使用的演算法是基於 Trie Tree 結構去產生中文字構成的可能情況，並進行動態規劃（Dynamic Programming）來計算出最大機率的路徑。此外 Jieba 提供了三種分詞模式:

1. 全模式：掃描出可以成語或詞語的句子。

2. 精確模式：適合文本分析，試圖將句子精確地做分割。

3. 搜尋引擎模式：在精確模式之下，針對較長的詞做切割，適用在搜尋引擎。

(四)、Jieba 安裝說明

安裝模式	方法
建立全自動安裝	easy_install jieba 命列輸入 pip install jieba / pip3 install jieba
半自動安裝	下載檔案 http://pypi.python.org/pypi/jieba/ 解壓縮後執行 python setup.py install
手動安裝	將 Jieba 目錄放置在使用的目錄或者是 site-packages 目錄，輸入 import jieba

(五)、程式碼範例如下所示

```
import jieba
sent = '今天天氣不錯適合去花蓮玩 '
seg_list = jieba.cut(sent, cut_all=True)
print('全模式：', '/ '.join(seg_list))
全模式： 今天/ 天天/ 氣/ 不/ 錯/ 適/ 合/ 去/ 花/ 蓮/ 玩/ /  /
seg_list = jieba.cut(sent, cut_all=False)
print('精確模式：', '/ '.join(seg_list))
```

精確模式： 今天/ 天氣/ 不錯/ 適合/ 去/ 花蓮/ 玩/

```
seg_list = jieba.cut_for_search(sent)
print('搜尋引擎模式', '/ '.join(seg_list))
```

搜尋引擎模式：今天/ 天氣/ 不錯/ 適合/ 去/ 花蓮/ 玩/

(六)、TF-IDF 介紹

經過上述的中文斷詞作業後再來要取得文章中的主要詞彙，在此之前需要先去除掉不必要的無用詞像是「的」、「是」、「了」、「吧」、「喔」等，以及標點符號，因這些將會干擾關鍵字的產生效果，透過高詞頻(Term Frequency-Inverse Document Frequency,TF-IDF)計算，高詞平頻也就是所謂的關鍵詞，利用自然語言中 TF(Term Frequency)取得最常被提及到的主要關鍵字，之後再進行逆向詞頻(IDF, Inverse Document Frequency)過濾掉常見的單詞，保留重要的單詞，兩者透過加權技術可以評估單詞對文件的集合之重要程度。

詞頻(TF)，假設 j 是指某一文件，i 是其中一種該文件中所使用的單詞或單字，n(i , j)表示 i 在 j 當中出現的次數，也就是說 TF(i,j)的算法為 n(i,j)/(n(1, j)+n(2,j)+n(3,j)+…+n(i,j))。假設一篇文章中被篩選出兩個重要單詞「台灣」、「新冠」，「台灣」在文章中共出現 70 次，「新冠」出現 30 次，「台灣」TF=70/100＝0.7，而「新冠」的 TF=30/100=0.3。在第二篇文件裡，再次選出兩個名詞，為「台灣」、「新冠」，「台灣」共在該篇文件中出現 40 次，「新冠」出現 60 次，那「台灣」的 TF=40/100=0.4，「新冠」的 TF=60/100=0.6，TF 的數值愈高，表示該單詞愈重要。兩篇文章都有「台灣」、「新冠」單詞，透過計算後可以得知第一篇文章中單詞「台灣」0.7 比第二篇「台灣」0.4 文章來的重要；同樣在第二篇的「新冠」0.6 比文章一「新冠」0.3 的中更為重要。

逆向詞頻(IDF)，假設 D 是「所有的文件總數」，i 是網頁中所使用的單詞，

t(i)是該單詞在所有文件總數中出現的「文件數」，那麼 IDF(i)的算法為 log(D/t(i))=log D–logt(i)。假設有 100 個網頁，「台灣」出現在 10 個網頁當中，而「新冠」出現在 100 個網頁當中，那麼「台灣」的 IDF=log(100/10)=2–1=1，而「新冠」的 IDF=log(100/100)=2–2=0。所以，「台灣」出現的機會小，與出現機會很大的「新冠」比較起來，便顯得非常重要。最後，將 TF(i,j)* IDF (i)（例如：i=「台灣」一詞）來進行計算，以某一特定文件內的高單詞頻率，乘上該單詞在文件總數中的低文件頻率，便可以產生 TF-IDF 權重值，且 TF-IDF 傾向於過濾掉常見的單詞，保留重要的單詞，如此一來，「新冠」便不重要了，最終可以得知在 TF-IDF 的算法中字詞的重要性會隨著它在文件中出現的增加次數成正比，但會隨著它在文本中降低出現的頻率成反比。

(七)、TF-IDF 關鍵字權重範例

```
from jieba import analyse
s = """王國材今天再指揮中心記者會公布放寬大眾運輸場站飲食及 11 月 8 日起恢復高鐵自由座措施也可開放站票避免太多人潮擬開放更多自由座車廂今天疫情呈現三零王國材說交通部也有開放措施第一個是陸海空場站飲食開放從明天開始有三項開放包括客運轉運站包括國道市區公車客運明起候車區可以開放飲食再者是海空運場站明起國內場站已經開放飲食國際航空站海港部分出境也開放飲食但國際場站入境部分目前仍是嚴格分流管制高鐵自由座部分 11 月 8 日起開放自由座且有放寬站票王國材說 11 月 8 日會把所有班次從現在 899 班調整回疫情前 1016 班朝恢復正常進行大家如果擔心自由座太擠交通部將協調高鐵公司在尖峰時段增加 4 至 8 節自由座並且滾動檢討調整車廂數如果太擁擠尖峰時段可以增加其他車廂當作自由座椅分散人流也會視需求增加班次"""
data = analyse.extract_tags(s,topK=3,withWeight=True)
for key, weight in data:
    print('%s %s' % (key, weight))
```

關鍵字權重結果：

開放 0.7356780001784615

飲食 0.36783900008923076

自由 0.3177099776657692

程式碼介紹

1、sentence 為擷取的文本。

2、topK 為返回權重 3 以內的最大關鍵字。

3、withWeight 為是否一併算出關鍵詞權重值。

4、allowPOS 指定性的詞默認為空值不及於篩選。

5、jieba.analyse.TFIDF（idf_path = None）算對應的權重值

第四章 實作結果與分析

第一節 資料擷取

本研究以 PTT 美容版(https://www.ptt.cc/bbs/BeautySalon/index.html)作為研究之資料來源，利用 Python 工具進行資料收集，時間為 2020 年 7 月 10 日至 2021 年 6 月 17 日擷取共 4013 篇文章，本研究提供爬蟲作業程式碼可參閱附件(一)。

一、網路爬蟲

首先使用 Requests 模組透過 HTTP 請求網頁伺服器下載指定的 URL，利用 Beautiful Soup 解析伺服器回傳的資料，並且以 find_all 找到所屬物件的標籤進行部分資料的擷取，以標籤功能取得有關文章發文作者、時間、標題和內容，如圖 4 所示。

(一)、鎖定目標網址的，了解網頁結構。

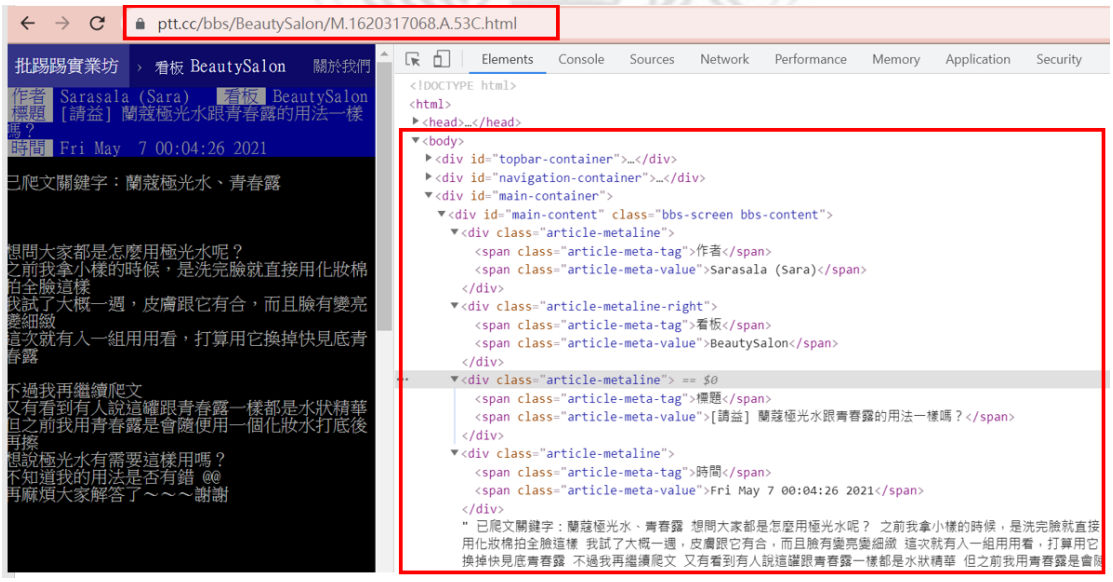


圖 4 PTT 網頁

(二)、利用 Beautiful Soup 模組向目標網址發送 HTTP 請求封包，取得網頁的 HTML 原始碼，如圖 5 所示。

```
In [4]: #將原始碼做整理
soup = BeautifulSoup(response.text, 'lxml')

#使用find_all()找尋特定目標
#results = soup.select("div.title")
#articles = soup.find_all('div', 'push')
print(soup)
```

```
<!DOCTYPE html>
<html>
<head>
<meta charset="utf-8"/>
<meta content="width=device-width, initial-scale=1" name="viewport"/>
<title>[請益] 蘭蔻極光水跟青春露的用法一樣嗎？ - 看板 BeautySalon - 批踢踢實業坊</title>
<meta content="all" name="robots"/>
<meta content="Ptt BBS 批踢踢" name="keywords"/>
<meta content="已爬文關鍵字：蘭蔻極光水、青春露
想問大家都是怎麼用極光水呢？
之前我拿小樣的時候，是洗完臉就直接用化妝棉拍全臉這樣
我試了大概一週，皮膚跟它有合，而且臉有變亮變細緻
這次就有入一組用用看，打算用它換掉快見底青春露
" name="description"/>
<meta content="Ptt 批踢踢實業坊" property="og:site_name"/>
<meta content="[請益] 蘭蔻極光水跟青春露的用法一樣嗎？" property="og:title"/>
<meta content="已爬文關鍵字：蘭蔻極光水、青春露
想問大家都是怎麼用極光水呢？
之前我拿小樣的時候，是洗完臉就直接用化妝棉拍全臉這樣
```

圖 5 網頁原始碼

(三)、分析 HTML 網頁，使用瀏覽器開發人員工作，利用 Beautiful Soup 模組中的 find 功能找尋特定標籤的資料位置，如圖 6 所示。

```
with open('BeautySalon.txt','w',encoding='UTF-8') as f:
    for article in articles:
        #去除掉冒號和左右的空白
        messages = article.find('span','f3 push-content').getText().replace(':','').strip()
        print(messages)
        f.write(messages + "\n")
```

圖 6 程式標籤

（四）、解析 HTML 網頁，利用剖析器建立樹狀結構標籤集合，如圖 7 所示。

```
#將原始碼做整理
soup = BeautifulSoup(response.text, 'lxml')

#使用find_all()找尋特定目標
#results = soup.select("div.title")
articles = soup.find_all('div', 'push')
print(articles)
```

```
[<div class="push"><span class="f1 hl push-tag">→ </span><span class="f3 hl push-userid">ianwu777</span><span class="f3 push-co
ntent">: 極光水洗完臉用化妝棉擦拭 不用再打底</span><span class="push-ipdatetime">49.158.135.251 05/07 01:47
</span></div>, <div class="push"><span class="hl push-tag">推 </span><span class="f3 hl push-userid">Aoife</span><span class="f3
push-content">: 借地方問一下，可以先用極光水再用青春露嗎</span><span class="push-ipdatetime">118.169.5.92 05/07 07:27
</span></div>, <div class="push"><span class="f1 hl push-tag">→ </span><span class="f3 hl push-userid">Aoife</span><span class
="f3 push-content">: 我的印象也是極光水可直接用 但青春露最好</span><span class="push-ipdatetime">118.169.5.92 05/07 07:28
</span></div>, <div class="push"><span class="f1 hl push-tag">→ </span><span class="f3 hl push-userid">Aoife</span><span class
="f3 push-content">: 前面先用化妝水之類的東西會比較好吸收</span><span class="push-ipdatetime">  118.169.5.92 05/07 07:28
</span></div>, <div class="push"><span class="hl push-tag">推 </span><span class="f3 hl push-userid">boya0518</span><span class
="f3 push-content">: 極光水我當化妝水使用，前面不打底</span><span class="push-ipdatetime">101.12.36.37 05/07 10:51
</span></div>, <div class="push"><span class="f1 hl push-tag">→ </span><span class="f3 hl push-userid">boya0518</span><span cla
ss="f3 push-content">: 回A大，我不是敏感肌，但是我兩個一起</span><span class="push-ipdatetime">101.12.36.37 05/07 10:52
</span></div>, <div class="push"><span class="f1 hl push-tag">→ </span><span class="f3 hl push-userid">boya0518</span><span cla
ss="f3 push-content">: 用會有一點太刺激，建議先測試看看比較</span><span class="push-ipdatetime">101.12.36.37 05/07 10:52
</span></div>, <div class="push"><span class="f1 hl push-tag">→ </span><span class="f3 hl push-userid">boya0518</span><span cla
ss="f3 push-content">: 好</span><span class="push-ipdatetime">101.12.36.37 05/07 10:52
</span></div>, <div class="push"><span class="f1 hl push-tag">→ </span><span class="f3 hl push-userid">EIngXuan</span><span cla
ss="f3 push-content">: 直接用就好了</span><span class="push-ipdatetime">202.70.179.109 05/07 12:04
</span></div>, <div class="push"><span class="hl push-tag">推 </span><span class="f3 hl push-userid">ryofang</span><span class
="f3 push-content">: 極光水洗完臉直接擦就可以 不用像青春露</span><span class="push-ipdatetime">49.216.29.185 05/07 14:33
</span></div>, <div class="push"><span class="f1 hl push-tag">→ </span><span class="f3 hl push-userid">ryofang</span><span clas
s="f3 push-content">: 那麼麻煩 而且比青春露更保濕</span><span class="push-ipdatetime">49.216.29.185 05/07 14:33
</span></div>, <div class="push"><span class="hl push-tag">推 </span><span class="f3 hl push-userid">funnyrain</span><span class
="f3 push-content">: 我都把極光水當化妝水使用</span><span class="push-ipdatetime">114.47.231.231 05/07 18:25
</span></div>, <div class="push"><span class="hl push-tag">推 </span><span class="f3 hl push-userid">connielu</span><span class
="f3 push-content">: 覺得用拍的比用化妝棉效果好</span><span class="push-ipdatetime">112.78.90.110 05/07 22:46
</span></div>, <div class="push"><span class="f1 hl push-tag">→ </span><span class="f3 hl push-userid">j05a</span><span class
="f3 push-content">: 用化妝棉擦拭～～</span><span class="push-ipdatetime">1.165.147.234 05/08 00:13
</span></div>, <div class="push"><span class="hl push-tag">推 </span><span class="f3 hl push-userid">Maurine</span><span class
="f3 push-content">: 極光水不能濕敷</span><span class="push-ipdatetime"> 223.136.130.55 05/08 13:14
```

圖 7 樹狀結構標籤集合

（五）、擷取資料，把資料儲存成特定格式，如圖 8 所示。

```
with open('BeautySalon.txt','w',encoding='UTF-8') as f:
    for article in articles:
        #去除掉冒號和左右的空白
        messages = article.find('span','f3 push-content').getText().replace(':','').strip()
        print(messages)
        f.write(messages + "\n")
```

```
極光水洗完臉用化妝棉擦拭 不用再打底
借地方問一下，可以先用極光水再用青春露嗎
我的印象也是極光水可直接用 但青春露最好
前面先用化妝水之類的東西會比較好吸收
極光水我當化妝水使用，前面不打底
回A大，我不是敏感肌，但是我兩個一起
用會有一點太刺激，建議先測試看看比較
好
直接用就好了
極光水洗完臉直接擦就可以 不用像青春露
那麼麻煩 而且比青春露更保濕
我都把極光水當化妝水使用
覺得用拍的比用化妝棉效果好
用化妝棉擦拭～～
極光水不能濕敷
```

圖 8 擷取資料

第二節 中文斷詞

一、 Ckiptagger

在網路社群論壇文章中擷取大量地的文字訊息，首先需經過斷詞作業才能萃取出文本中的重要訊息，因此本研究利用中研院提供的 Ckiptagger 進行資料預處理，過程中會透過斷詞(WS)、詞性標注(POS）和實體辨識（NER）等功能來進行中文斷詞作業。如圖 6 中研院針對 Ckiptagger 公開相關程式碼和使用操作方式存放 github 給使用者運用，Ckiptagger 的程式碼可參閱附件(二)。

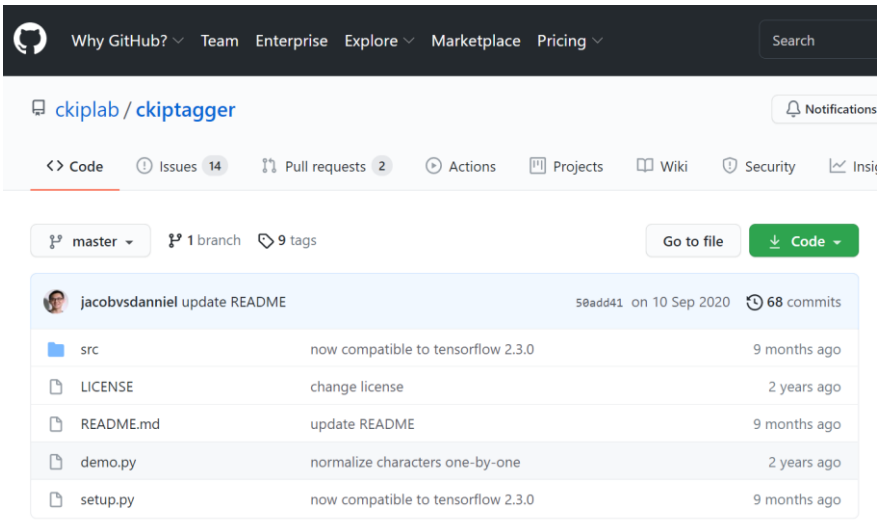


圖 9 斷詞公開程式碼

(一)、斷詞(WS)

語言單位中詞擁有最小意義，在進行任何語言處理中都必須先分辨文本中的詞才能更進一步的處理，過程會以詞典中收入的詞自行和文本進行比對，找出可能包含的詞。

```
'['Re', ':', ' [', '請益', '] ', '雅漾', '輕透', '防曬液', '無', '香版', ' -
坊', ':', '看到', 'momo', '有', '特價', '+', '回饋', '100', ' momo幣', '就',
'買', '的', '輕透', '防曬液', '，', '剛好', '可以', '分享', '兩', '條', '之間',
'都', '沒有', '賣', '了', '，', '感覺', '是', '輕透', '防曬液', '都', '改', '無
':', '\nAvene ', '雅漾', '全效', '極', '護', ' ', '輕透', '防曬液', ' ', '無'
善', '防曬', ')', '\n']'
```

圖 10 斷詞(WS)

(二)、詞性標注(POS）

可以利用詞性來判別字詞在語法和語言結構中所扮演的角色,經過詞性的分類賦予句子中每個字詞適當的標記和詞性的符號,可謂為自然語言處理研究領域的根本。

```
Re(FW) :(COLONCATEGORY) [(FW) 請益(VB) ](FW) 雅漾(Nb) 輕透(VJ) 防曬液(Na) 無(VJ) 香版(Na) -(FW) 看板(Na) Beauty
(FW) Salon -(FW) 批踢踢(Na) 實業坊(Nc) :(COLONCATEGORY) 看到(VE) momo(FW) 有(V_2) 特價(Na) +(FW) 回饋(VC) 100(Neu)
momo幣(Na) 就(D) 入手(VA) 了(Di)
(WHITESPACE) 手(Na) 邆(Ncd) 也(D) 有(V_2) 去年(Nd) 買(VC) 的(DE) 輕透(VH) 防曬液(Na) ,(COMMACATEGORY) 剛好(Da) 可以
(D) 分享(VJ) 兩(Neu) 條(Nf) 之間(Ng) 的(DE) 差異(Na)
(WHITESPACE) 不過(Cbb) 舊款(Na) 的(DE) 好像(D) 都(D) 沒有(D) 賣(VD) 了(Di) ,(COMMACATEGORY) 感覺(VK) 是(SHI) 輕透(VH)
防曬液(Na) 都(D) 改(VC) 無(VJ) 香(Na) 新上(Nc) 市
(Na) 新款(A) 的(DE) 全名(Na) :(COLONCATEGORY)
Avene (FW) 雅漾(Nb) 全效(A) 極(Dfa) 護(VC) (WHITESPACE) 輕透(VH) 防曬液(Na) (WHITESPACE) 無(VJ) 香(Na) SPF50+ (FW)
50(Neu) ml(Nf) ((PARENTHESISCATEGORY) 海洋(Na) 友善(VH) 防曬(VH) )(PARENTHESISCATEGORY)
```

圖 11 詞性標注(POS)

(三)、實體辨識（NER）

可以判斷文字資料中詞語的意義，例如:地名、組織、人名、藥品名、分子式等專有名詞。NER 讓機器能自動找尋文本中提到我們感興趣的實體，例如公眾人物等，並加以分析，其產出亦作為人工智慧理解自然語言的重要資訊。

```
(31, 33, 'GPE', '台南')
(56, 58, 'ORG', '海洋')
(71, 73, 'ORG', '海洋')
(92, 96, 'DATE', '去年冬天')
(129, 131, 'PERSON', '雅漾')
(196, 198, 'CARDINAL', '一半')
(207, 209, 'PERSON', '雅漾')
(382, 384, 'CARDINAL', '10')
(492, 494, 'PERSON', '夏天')
(505, 507, 'PERSON', '海洋')
```

圖 12 實體辨識(NER)

二、 TF-IDF

從上述爬蟲擷取下來的 4013 篇發文中，利用 TF-IDF 詞頻演算法計算，取得每篇文章中最常出現的關鍵詞，從文字雲中得知 PTT 美容版文章中消費者最常提到的品牌關鍵詞為「雅漾」、「蘭蔻」、「克蘭詩」和「碧兒泉」，如圖 13。

圖 13 文字雲

三、 Jieba

因此本研究利用 Jieba 工具再次進行品牌關鍵詞爬蟲，擷取 PTT 文章有關於「雅漾」、「蘭蔻」、「克蘭詩」和「碧兒泉」這四個品牌的文章內容作為資料分析來源，利用程式碼擷取文章中有提到關鍵詞的網址儲存成 EXCEL 檔案，再以爬蟲方式擷取出發文者文章內容作為後續的分析，如圖 14 所示。

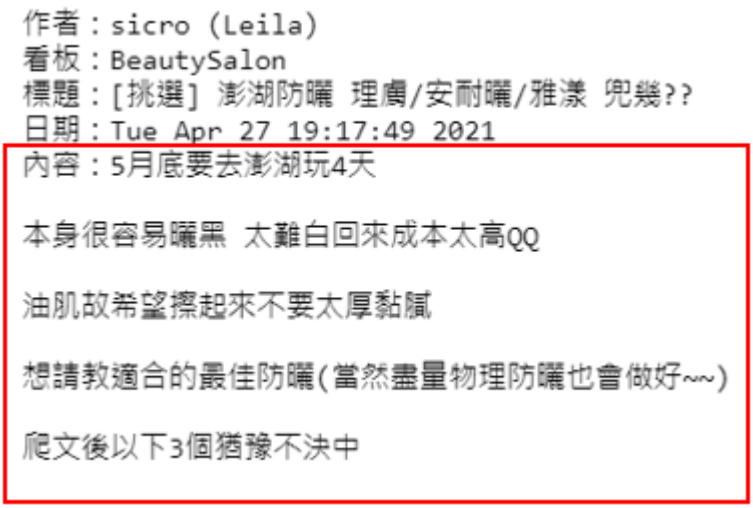


圖 14 擷取發文內容

Jieba 為開源式的特性，開發者可以依造自己的需求做調整，在使用上面以利新增修改或刪除詞彙，在美容版中因為許多詞彙為專有名詞，為了讓中文的斷詞效果更為顯著，因此加入了 Jieba 的內建詞庫、新增詞典以及停用詞典，希望藉此提高中文斷詞的精準度。

(一)、內建詞庫

使用 Jieba 提供的內建的下載繁體中文詞庫，下載完畢後在 Jupyter Notebook 匯入中文詞庫 jieba.set_dictionary('dict.txt.big')就可以利用詞庫做中文斷詞作業，進行初步中文斷詞後若發現專有名詞詞彙並無被判斷出來，因此放入自建詞典的新增詞彙讓文章的斷詞結果可以更加精準和豐富。

(二)、新增詞典

Jieba 套件內部提供詞庫可以進行斷詞，但會發現斷詞結果不如預期，像是品牌和產品名稱以及在美容產品中的專有名詞，原先的詞典中缺少很多相關詞彙，這時候可以利用自行定義詞彙新增詞典，本研究自建詞典有「雅漾」、「蘭蔻」、「克蘭詩」和「碧兒泉」等品牌名稱，建立詞典後存成(.txt)檔案以 jieba.load_userdict('userDict.txt')方式匯入，可以讓文本中的詞彙可以更加精確地被判斷出來。

(三)、停用詞典

停用詞典可以過濾掉文章中許多沒有意義的詞彙，可以使 Jieba 能呈現的更加準確的詞彙，利用 extract_tags 函數，要取出 TF-IDF 關鍵詞之前可以先去除掉無意義的單詞或詞彙，此外也可以用人工方式指定停用詞典程式碼如圖 15，本研究以 github 的 Stopword(**https://github.com/kdchang/python-jieba-chart**)所提供的停用詞以及在現有的停用詞典中自行加入 32 個停用詞，如表 5，可用來去除英文、中文單字、中文詞、標點符號等，如表 6 所示。

```
import jieba
import jieba.analyse
content= open(r'ETA2.txt', "r",encoding="utf-8").read()
jieba.analyse.set_stop_words('stopword.txt')
content = jieba.analyse.extract_tags(content,30)
print(content)
```

圖 15 停用詞程式碼

表 5 網路停用詞典

英文	中文單字				中文詞	標點符號
the	的	個	各	她	一個	，
of	了	其	給	哇	沒有	。
is	和	已	跟	嗡	我們	；
and	是	無	何	往	你們	、
to	就	小	還	哪	妳們	」
in	都	我	即	些	他們	「
that	而	們	幾	向	她們	！
we	及	起	既	沿	是否	！
for	與	最	看	喲		,
an	一	再	據	用		[
are	不	今	距	於		]
by	在	去	靠	咱		~
be	人	好	啦	則		‘
as	有	只	了	怎		)
on	為	又	另	曾		(
with	以	並	麼	至		)

can	於	或	每	致		/
if	上	很	們	著		？
from	他	亦	嘛	諸		？
which	後	某	拿	自		.
you	之	把	哪	得		...
it	來	那	那	打		
this	因	你	您	凡		
then	下	乃	憑	兒		
at	可	它	且	爾		
have	到	吧	卻	該		
all	由	被	讓	誰		
not	這	比	仍	雖		
one	也	別	啥	隨		
has	此	趁	如	同		
or	但	從	若	所		
that	當	到	使			

表 6 自建停用詞

1. 覺得	2. 感覺	3. 不會	4. 不錯	5. 不過
6. 因為	7. 比較	8. 雖然	9. 時候	10.一樣
11.有點	12.這次	13.這罐	14.還有	15.真的
16.還是	17.可以	18.現在	19.這樣	20.大家
21.其實	22.所以	23.一下	24.一點	25.一罐
26.什麼	27.應該	28.起來	29.發現	30.這個
31.而且	32.所以			

第三節 文字雲

Python 文字雲，在中文斷詞進行初步的處理後，經過 TF-IDF 計算文本中的出關鍵字，以圖形和顏色視覺化方式呈現，字體越大代表在文本中出現的次數越多，也可以解釋成文本中的重要關鍵詞。本研究利用文字雲顯示「雅漾(Avene) 」、「蘭蔻(Lancome) 」、「克蘭詩(Clarins)」以及「碧兒泉(Biotherm)」這四個品牌中的前 50 個關鍵詞彙，來分析品牌最被關注的議題。

(一)、雅漾

在雅漾的文字雲中可以看到「防曬」、「精華」、「乳液」和「保濕」等產品為消費者在文中主要分享到的產品類型，可以發現雅漾品牌中以「防曬」關鍵詞為最常出現之詞彙，可以知道發文者常分享有關防曬類的產品內容。除此之外，使用雅樣的產品中消費者會感受到產品「清爽」和「控油」的效果，但也有反應出產品會使肌膚有「過敏」、「敏感」、「黏膩」、「痘痘」和「粉刺」等問題，如圖 16 所示。

圖 16 雅漾文字雲

(二)、蘭蔻

在蘭蔻的文字雲中可以看到，消費者關注內容為「小黑瓶」、「玫瑰」系列的產品以及「極光水」和「精華」都是經常出現在文章中的關鍵詞，此外也關注百貨公司的「櫃姐」、「贈品」、「組合」等，從中可以了解蘭蔻的文章中，以產品介紹和優惠訊息為文章中最常被提到的關鍵詞，也可以得知是發文者較喜歡分享得議題，如圖 17 所示。

圖 17 蘭蔻文字雲

(三)、克蘭詩 Clarins

在克蘭詩的文字雲中可以看到，消費者關注內容為「精華」、「雙激萃」和「眼霜」等產品，其中也講到產品的「質地」、「保濕」、「味道」和「美白」，此外也關注百貨公司的「櫃姐」、「贈品」和「組合」等，從中可以了解克蘭詩的文章中，發文者會喜歡分享有關產品名稱、使用效果以及優惠方案等議題，如圖 18 所示。

圖 18 克蘭詩文字雲

(四)、碧兒泉 Biotherm

在碧泉兒的文字雲中可以看到，消費者關注內容為「保濕」、「特嫩」、「精華」、「活源」等關鍵詞，經查詢後發現這幾個關鍵詞出自於碧兒泉的奇蹟修護系列，因此可以發現奇蹟修護系列的產品為使用者較常分享的內容，也可以知道消費者非常種皮膚的保養效果，並且能透過碧兒泉文字雲了解到發文者會喜歡分享有關產品名稱、使用效果等議題，如圖 19。

圖 19 碧兒泉文字雲

第四節 主題模型

本研究利用隱含狄利克雷分配主題模型來深入挖掘文字背後的重要訊息，每個文檔都是由數個主題所構成的，且每個主題都由數個重要詞彙來描述，因此本研究以各品牌共有 5 組主題模型，且由 15 個詞彙所組成，最終會從主題模型的詞彙中取得主題名稱，最後統整四大品牌的主題模型，以利於分析各品牌的特色從中擷取品牌相似處以及重要資訊。

一、雅漾

(一)、產品介紹

根據表 7 發現主題(一) 詞彙「SPF30」、「SPF50」的 SPF 的全名為 Sun Protection Factor，為針對抗紫外線 UVB 的防曬系數標準，SPF 數字越高防護的時間越久，觀察後可以得知雅漾以 SPF30 和 50 為主要防曬商品，由於現今天氣炎熱人長期在太陽的曝曬下會導致皮膚受傷，因此大家開始也注重防曬產品上的系數選擇，此外雅漾的乳液和舒護活泉水為保養類的產品，根據內文的留言觀察舒護活泉水的使用上會以皮膚有泛紅和過敏時會進行使用有助於改善肌膚狀況，如圖 20。

表 7 產品介紹主題

划算	泛白	名單	**SPF30**	**舒護**
櫃姐	momo	**乳液**	特賣會	入手
安敏	**太陽**	保養	寶水	**SPF50**

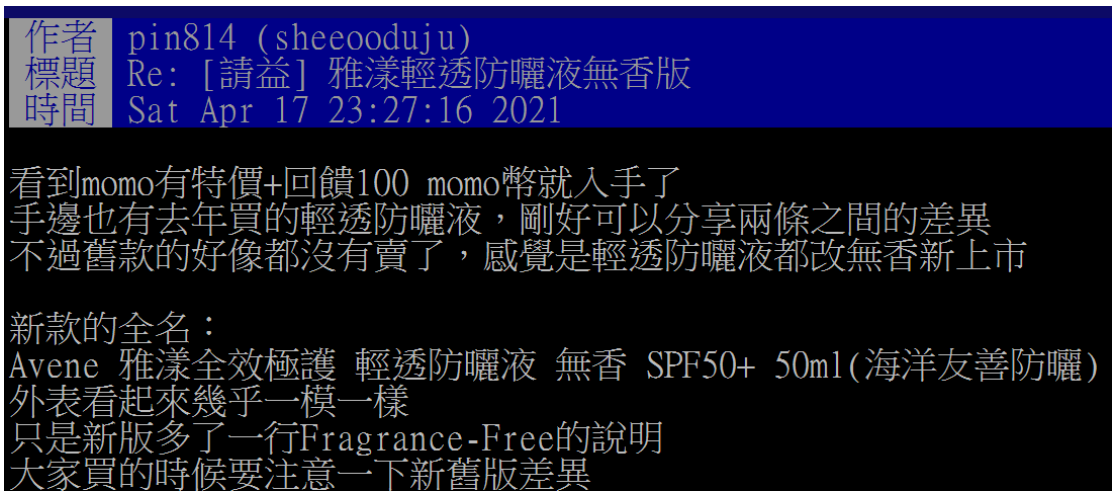

作者 pin814 (sheeooduju)
標題 Re: [請益] 雅漾輕透防曬液無香版
時間 Sat Apr 17 23:27:16 2021
看到momo有特價+回饋100 momo幣就入手了
手邊也有去年買的輕透防曬液，剛好可以分享兩條之間的差異
不過舊款的好像都沒有賣了，感覺是輕透防曬液都改無香新上市
新款的全名：
Avene 雅漾全效極護 輕透防曬液 無香 SPF50+ 50ml(海洋友善防曬)
外表看起來幾乎一模一樣
只是新版多了一行Fragrance-Free的說明
大家買的時候要注意一下新舊版差異

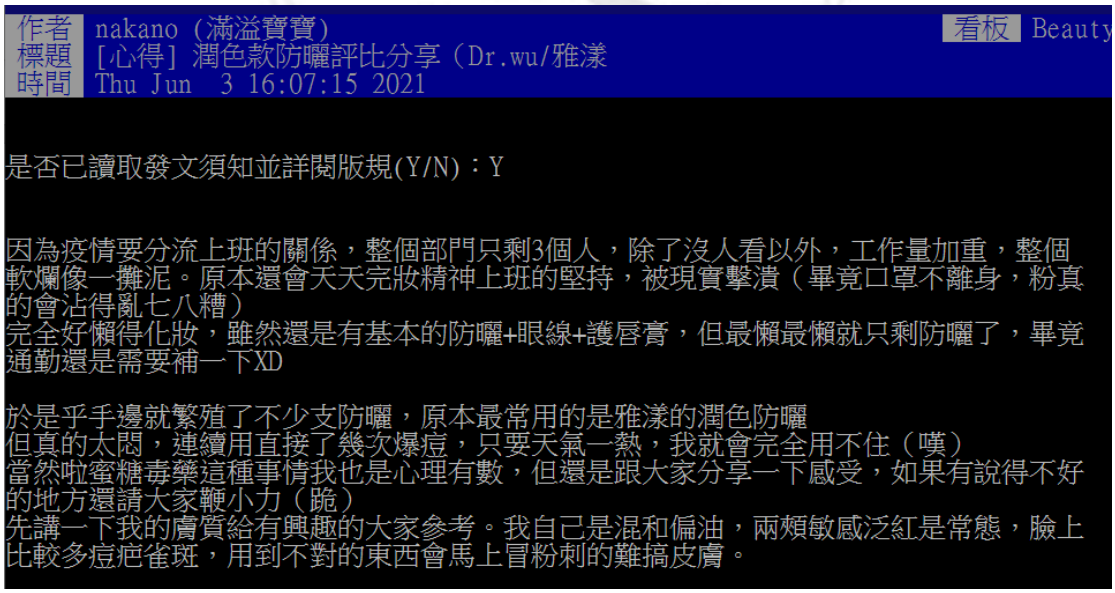
圖 20 PTT 產品介紹

(二)、防曬功效

根據表 8 發現主題(二)出現保養效果「皮膚」、「保濕」、「潤色」、「控油」和「品牌」等詞彙，可以觀察到消費者會在社群論壇中分享使用完產品後皮膚所產生的變化，包含保濕程度、皮膚的修護和控油效果，以及膚色上的變化，是消費者在使用雅漾的產品後會在發文中常出現的詞彙，除此之外，消費者還會以同類型產品但不同品牌來做實際使用比較，並且為不同品牌做一個產品效果的評比，如圖 21，提供給想購買某類產品的使用者，可以做品牌上的挑選，在社群論壇中可以獲得更充足的資訊，以減少購買到不適用的產品。

表 8 防曬功效主題

保濕	ml	**修護**	爆汗	**潤色**
Avene	使用	成分	**品牌**	有變
皮膚	味道	確定	建議	**控油**


作者 nakano (滿溢寶寶) 看板 Beauty
標題 [心得] 潤色款防曬評比分享 (Dr.wu/雅漾
時間 Thu Jun 3 16:07:15 2021

是否已讀取發文須知並詳閱版規(Y/N)：Y

因為疫情要分流上班的關係，整個部門只剩3個人，除了沒人看以外，工作量加重，整個軟爛像一攤泥。原本還會天天完妝精神上班的堅持，被現實擊潰（畢竟口罩不離身，粉真的會沾得亂七八糟）
完全好懶得化妝，雖然還是有基本的防曬+眼線+護唇膏，但最懶最懶就只剩防曬了，畢竟通勤還是需要補一下XD

於是乎手邊就繁殖了不少支防曬，原本最常用的是雅漾的潤色防曬
但真的太悶，連續用直接了幾次爆痘，只要天氣一熱，我就會完全用不住（嘆）
當然啦蜜糖毒藥這種事情我也是心理有數，但還是跟大家分享一下感受，如果有說得不好的地方還請大家鞭小力（跪）
先講一下我的膚質給有興趣的大家參考。我自己是混和偏油，兩頰敏感泛紅是常態，臉上比較多痘疤雀斑，用到不對的東西會馬上冒粉刺的難搞皮膚。

圖 21 PTT 防曬功效

(三)、心得分享

根據表 9 發現主題(三)保養效果「膚況」、「狀況」、「缺點」、「心得」、「美白」、「滋潤」和「吸收」等詞彙，可以發現消費者會針對產品使用後進行優缺點的評估，像是使用後的膚況、和保養品的吸收程度以及皮膚的美白、滋潤效果，都是消費者用來評估一個產品是否好用的關鍵詞，透過本身使用結果在社群論壇中分享給其他消費者進行購買前的資料搜尋和評估，可以讓其他不熟悉該類產品和品牌的消費者可以透過觀看心得分享而有初步的認知，如圖 22 所示。

表 9 心得分享主題

吸收	乳狀	面膜	個人	耐曬
滋潤	**膚況**	**心得**	活泉	**狀況**
適合	**缺點**	特別	Wu	**美白**

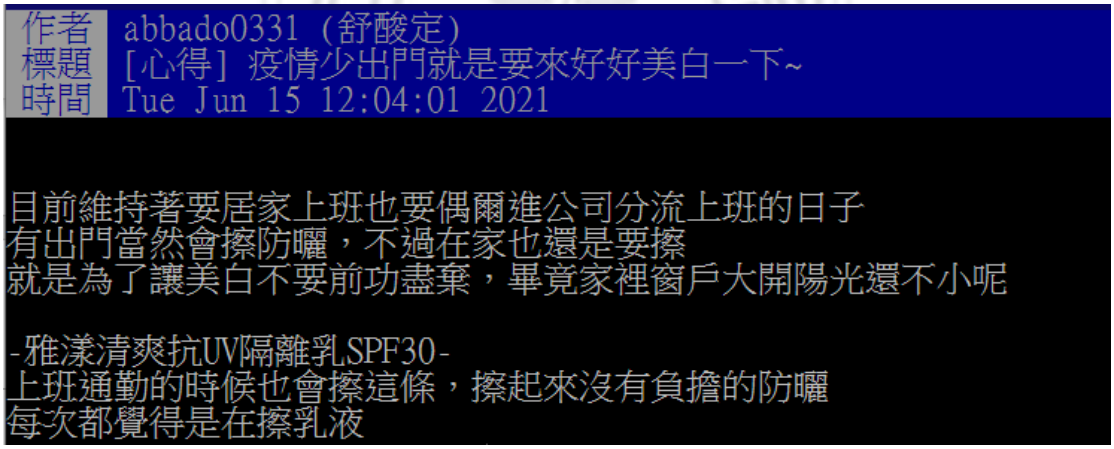


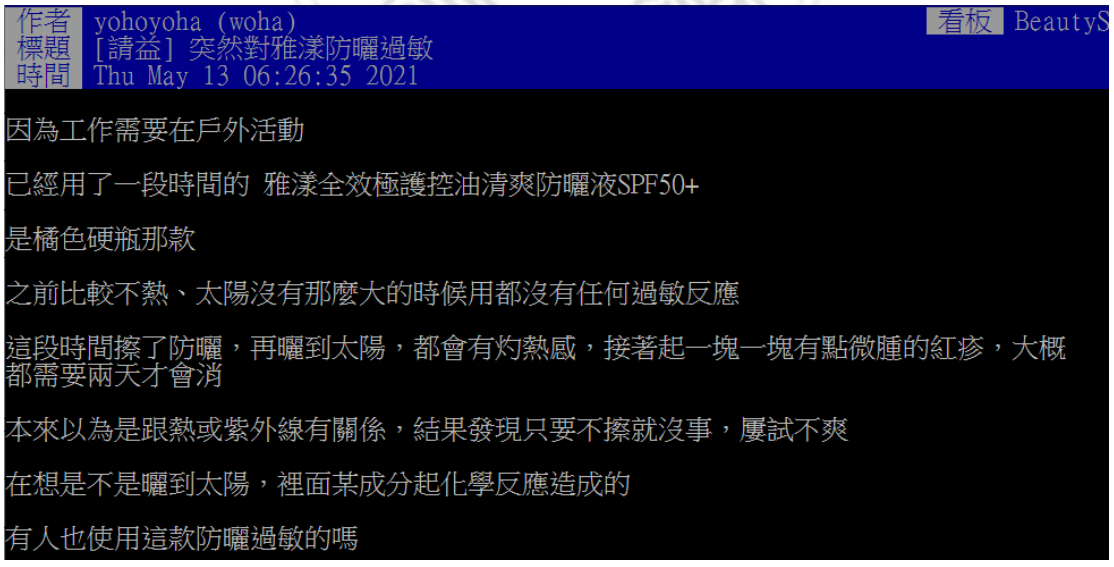
圖 22 PTT 心得分享

（四）、使用狀況

根據表 10 發現主題(四)保養效果「防曬」、「曬黑」、「過敏」和「討論」等詞彙，觀察後發現雅漾的產品中以防曬類產品最受矚目，消費者使用完產品後會注意自己的皮膚是否會再曬黑以及過敏等症狀。根據發文中發現，部分消費者使用該類產品時會有類似皮膚過敏的情況產生，由於不知道該如何處理，因此會透過社群論壇平台進行發問，瞭解是否也有其他消費者有相同狀況出現，並加以詢問是否有能提供改善的方法，希望可以解決皮膚過敏問題，由於產品使用上產生的負面效果，也會影響到該消費者往後購買時會不再考慮使用該品牌的產品，如圖 23 所示。

表 10 使用狀況主題

曬黑	用量	**討論**	修復	希望
化妝水	退紅	DR	上妝	興趣
質地	**防曬**	價格	冷氣	**過敏**


作者 yohoyoha (woha) 看板 BeautyS
標題 [請益] 突然對雅漾防曬過敏
時間 Thu May 13 06:26:35 2021

因為工作需要在戶外活動

已經用了一段時間的 雅漾全效極護控油清爽防曬液SPF50+

是橘色硬瓶那款

之前比較不熱、太陽沒有那麼大的時候用都沒有任何過敏反應

這段時間擦了防曬，再曬到太陽，都會有灼熱感，接著起一塊一塊有點微腫的紅疹，大概都需要兩天才會消

本來以為是跟熱或紫外線有關係，結果發現只要不擦就沒事，屢試不爽

在想是不是曬到太陽，裡面某成分起化學反應造成的

有人也使用這款防曬過敏的嗎

圖 23 PTT 使用狀況

(五)、購買組合

根據表 11 發現主題(五)保養效果「ml」、「組合」、「分享」和「購入」等詞彙，查詢後發現雅漾會利用 MOMO 網購平台並搭配節慶進行產品組合的促銷活動。除了網路通路外，在連鎖的藥妝店康透過舉辦醫美節期間進行價格優惠來吸引消費者進行產品購買，也會在百貨公司裡舉辦特賣會，可以瞭解到雅漾在行銷上面透過網路平台、連鎖藥妝店以及百貨商圈等方式推出一些產品的優惠價格，吸引許多消費者可以在這時間進行購物，並且在社群論壇平台中會有消費者進行資訊上的分享，也會引起其他消費者的共鳴，從主題中可以發現開架式品牌的銷售通路，如圖 24 所示。

表 11 購買組合主題

ml	毛孔	**組合**	尿酸	敏感話題
口罩	粉餅	**分享**	香味	感覺
隔離	推開	凡士林	極護	**購入**

圖 24 PTT 購買組合

二、蘭蔻

(一)、檔期優惠

根據表 12 發現主題(一)保養效果「回購」、「購物」、「回饋」、「贈品」和「百貨」等詞彙，觀察後發現蘭蔻會推出許多優惠方案來吸引消費者進行購買，文章中常提到會員贈品，達到一定的消費金額後也會贈送贈品給予消費者回饋，打造出消費者除了購買產品外也能有而外獲得免費產品的感受，使消費者很關注百貨周年慶或使官網的贈品回饋等。透過以上這些關鍵詞可以發現，除了價格優惠外贈送產品也是蘭蔻的行銷手法之，消費者會在社群論壇中不僅分享喜悅外，也會吸引到更多人去關注檔期的優惠方案，如圖 25 所示。

表 12 檔期優惠主題

美白	心得	**回購**	**購物**	**百貨**
精露	**回饋**	花瓣	粉刺	肌膚
感受	**贈品**	嬌蘭	提亮	過敏

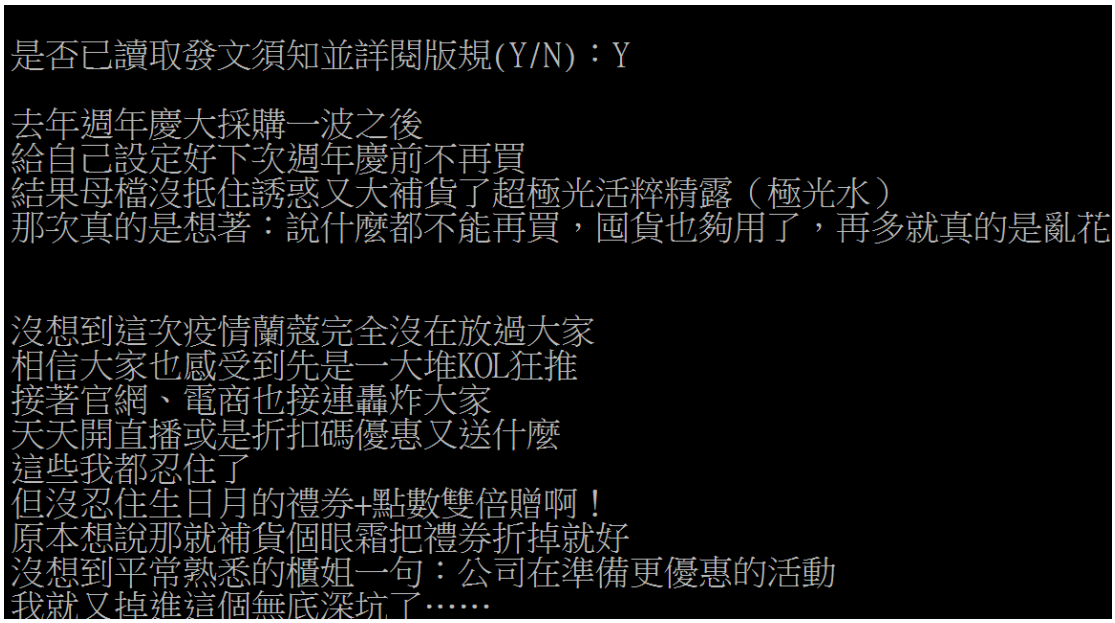


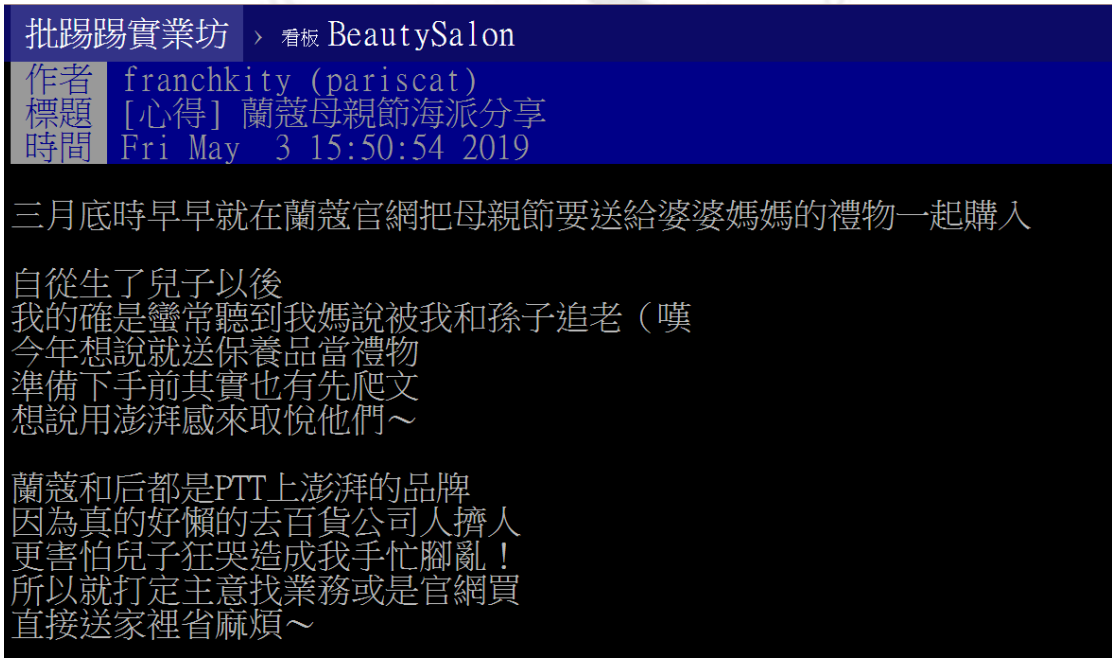
圖 25 PTT 檔期優惠

(二)、節日折扣

根據表 13 發現主題(二)保養效果「母親節」、「補貨」、「組合」、「購入」和「櫃姐」等詞彙， 櫃姐除了負責銷售外還會與顧客保持聯繫，像是在母親節期間就會寄送出相關優惠訊息，特別是消費族群多為女性因此母親節的優惠活動會使許多人利用優惠的價格進行保養品的補貨，或者購買產品當作母親節的禮物贈送給婆婆或媽媽，此外母親節會提供很多優惠的產品組合讓消費者利用母親節時期購入產品，消費者也會在社群論壇平台中分享自己得到的相關優惠訊息，如圖 26 所示。

表 13 節日折扣主題

ml	特別	**母親節**	保養	UV
安瓶	精華液	**補貨**	**櫃姐**	乳液
購入	痘痘	**組合**	面膜	保濕度


批踢踢實業坊 › 看板 BeautySalon
作者 franchkity (pariscat)
標題 [心得] 蘭蔻母親節海派分享
時間 Fri May 3 15:50:54 2019

三月底時早早就在蘭蔻官網把母親節要送給婆婆媽媽的禮物一起購入

自從生了兒子以後
我的確是蠻常聽到我媽說被我和孫子追老（嘆
今年想說就送保養品當禮物
準備下手前其實也有先爬文
想說用澎湃感來取悅他們～

蘭蔻和后都是PTT上澎湃的品牌
因為真的好懶的去百貨公司人擠人
更害怕兒子狂哭造成我手忙腳亂！
所以就打定主意找業務或是官網買
直接送家裡省麻煩～

圖 26 PTT 節日折扣

主題(三)、使用方式

根據表 14 發現主題(三)保養效果「步驟」、「參考」、「用量」和「產品」等詞彙，觀察後可以發現對於消費者會主動性的分享自己使用的產品，文章內容包含了介紹產品名稱、使用的用量和使用方法等，並且會附上圖片讓其他消費者可以觀看使用前後的肌膚改善效果，這樣的實驗性質的產品使用分享文章會吸引到讓有想購買或者皮膚有同樣狀況的消費者，以及買了卻不知道如何使用的使用者，可以透過網路上的資訊作為使用產品的參考，如圖 27 所示。

表 14 使用方式主題

步驟	**參考**	記得	泛紅	會員
導購	**用量**	滋潤	天氣	保養品
喜歡	**產品**	化妝水	改善	長期

圖 27 PPT 使用方式

主題(四)、皮膚狀況

根據表 15 發現主題(四)保養效果「吸收」、「味道」、「皮膚」和「膚質」等詞彙，觀察後可以發現消費者購買完產品後，最在意的是產品是否真的能改善自己的肌膚狀況，可以透過保養品的吸收程度來做為一個檢視外，也在產品的味道是否喜歡，文章中有提到因為疫情關係長期攜帶口罩，若味道刺鼻會讓使用者降低想繼續使用的慾望，並且不喜歡保養品太過於黏膩以及不意吸收，造成肌膚長痘痘或者有泛紅的負面效果，因此使用產品後的肌膚狀況使用者會利用社群論壇平台分享，讓其他消費者可做為參考資訊，如圖 28 所示。

表 15 皮膚狀況主題

開箱	習慣	白金	冰珠	品牌
吸收	**味道**	**皮膚**	專櫃	超導
膚質	出來	使用	台灣	整體

圖 28 PTT 皮膚狀況

主題(五)、官網購物

根據表 16 發現主題(五)保養效果「下單」、「DM」、「預購」和「小樣」等詞彙， 發現特殊節日像是周年慶和母親節會推出優惠方案和價格外，蘭蔻的官網內容也是消費者經常關注的，除了 DM 上的產品介紹，不少人喜歡利用蘭蔻的官網購買產品，一方面覺得購買上很方便不必外出，此外官網上會利用打折價和折扣碼來吸引喜歡網路購物的消費者，在官網購買產品外，也會贈送小樣產品是許多消費者在分享時會提到的，並且有購物經驗的消費者就會在社群論壇平台上分享折扣碼和優惠來吸引更多消費者前往購買，如圖 29 所示。

表 16 官網購物主題

ml	進去	包裝	**下單**	再來
小樣	**DM**	按摩	建議	脫皮
選手	**預購**	凝膠	修護露	評價

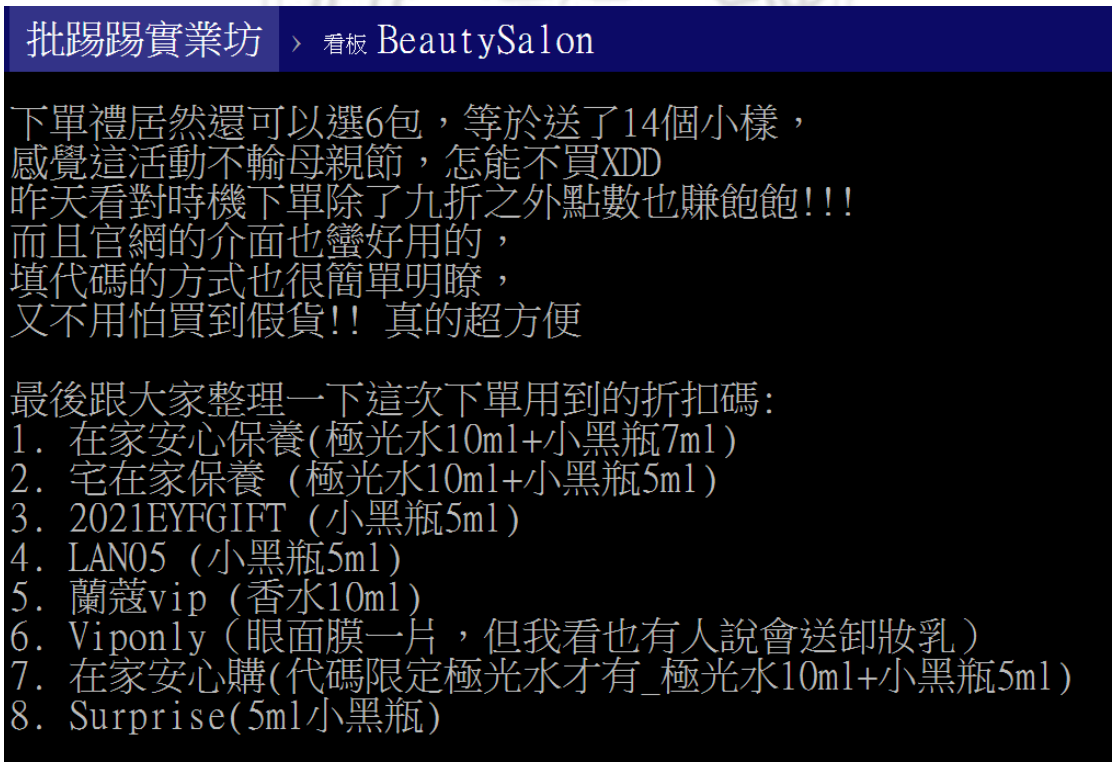


圖 29 PTT 官網購物

三、克蘭詩

(一)、產品效果

根據表 17 發現主題(一)保養效果「保濕」、「肌膚」、「吸收」和「護理」等詞彙，可以觀察到使用者會分享產品使用後的效果，像是保濕程度和吸收速度，會是使用者能直接透過肌膚感受到的，因此使用者會在 PTT 美容版社群論壇中實際分享自己的使用近況，以及透過產品的護理效果改善自己原本的皮膚狀況並且對產品進行一個評價，除此之外，發文者會先告知自己的膚質屬性，讓擁有相同膚質的消費者可以把該文章所介紹的產品做為一個參考依據，如圖 30 所示。

表 17 產品效果主題

保濕	滿額	斑點	乳液	參考
檔期	**肌膚**	**吸收**	嘴唇	黃金
護膚	紅瓶	噴霧	彈力	精華液

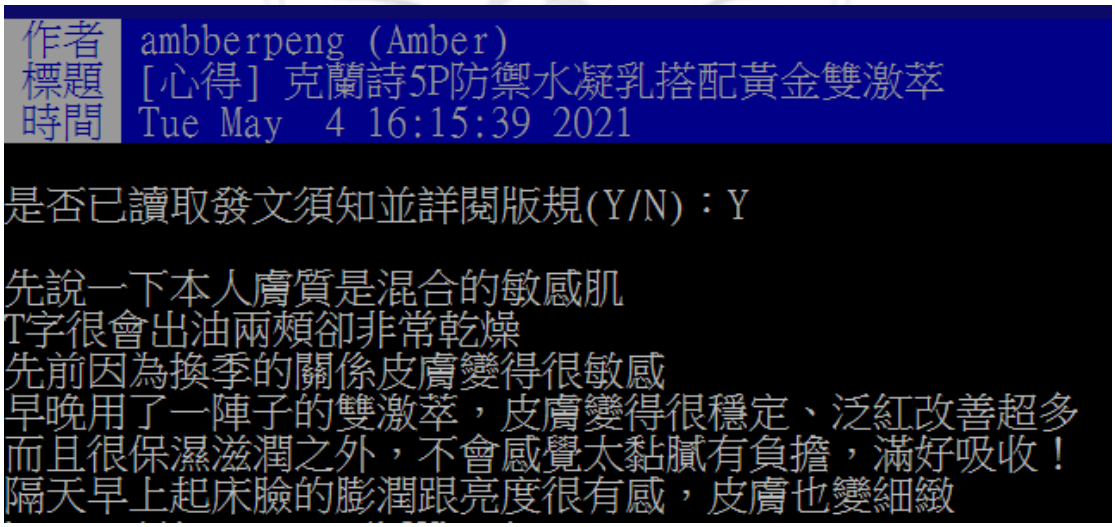


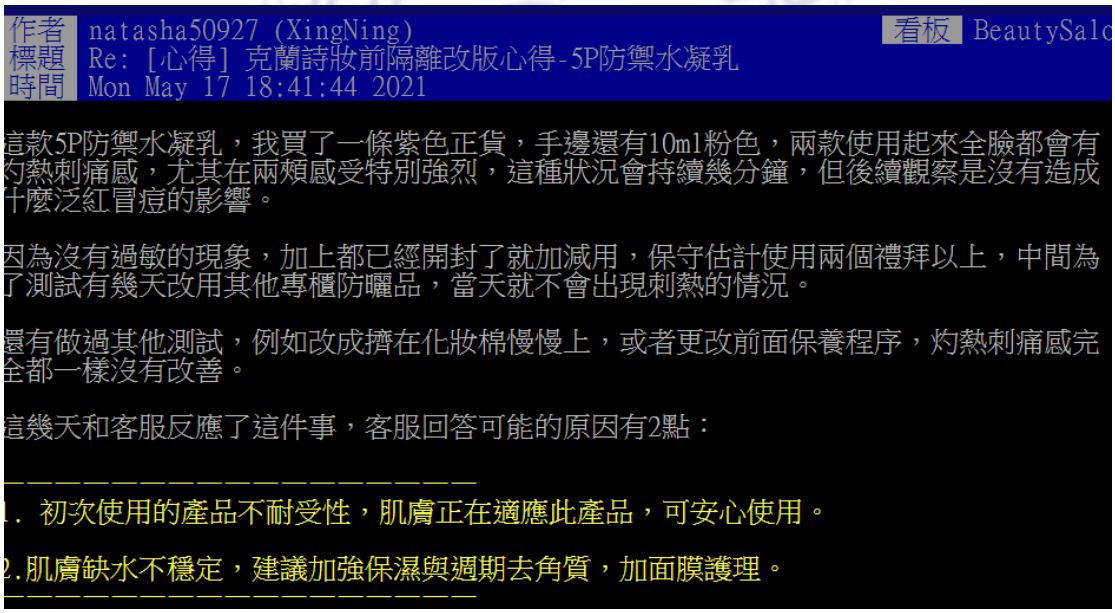
圖 30 PTT 產品效果

(二)、使用分享

根據表 18 發現主題(二)中出現「兩頰」、「臉部」、「過敏」和「詢問」等詞彙，可以觀察到克蘭詩的商品中消費者著重於臉部上的保養，而在臉部產品的使用過程中，可能會出現一些肌膚不適的情況，消費者會透過客服人員進行詢問，來了解自己的皮膚狀況和解決產品使用後的不適情形。售後服務的部分可以使消費者和品牌保持緊密的聯繫，並且品牌也可以透過消費者詢問的問題中加以了解消費者使用完後的感受，除此之外，使用者也會在 PTT 分享自己的臉部的狀況讓有相同困擾的消費者可以一同在社群論壇中進行討論如圖 31 所示。

表 18 使用分享主題

防禦	雪花	**兩頰**	包裝	**詢問**
味道	感覺	**臉部**	習慣	正貨
雅詩蘭黛	洗臉	小樣	修護	**過敏**


作者 natasha50927 (XingNing) 看板 BeautySalo
標題 Re: [心得] 克蘭詩妝前隔離改版心得-5P防禦水凝乳
時間 Mon May 17 18:41:44 2021

這款5P防禦水凝乳，我買了一條紫色正貨，手邊還有10ml粉色，兩款使用起來全臉都會有灼熱刺痛感，尤其在兩頰感受特別強烈，這種狀況會持續幾分鐘，但後續觀察是沒有造成什麼泛紅冒痘的影響。

因為沒有過敏的現象，加上都已經開封了就加減用，保守估計使用兩個禮拜以上，中間為了測試有幾天改用其他專櫃防曬品，當天就不會出現刺熱的情況。

還有做過其他測試，例如改成擠在化妝棉慢慢上，或者更改前面保養程序，灼熱刺痛感完全都一樣沒有改善。

這幾天和客服反應了這件事，客服回答可能的原因有2點：

1. 初次使用的產品不耐受性，肌膚正在適應此產品，可安心使用。

2.肌膚缺水不穩定，建議加強保濕與週期去角質，加面膜護理。

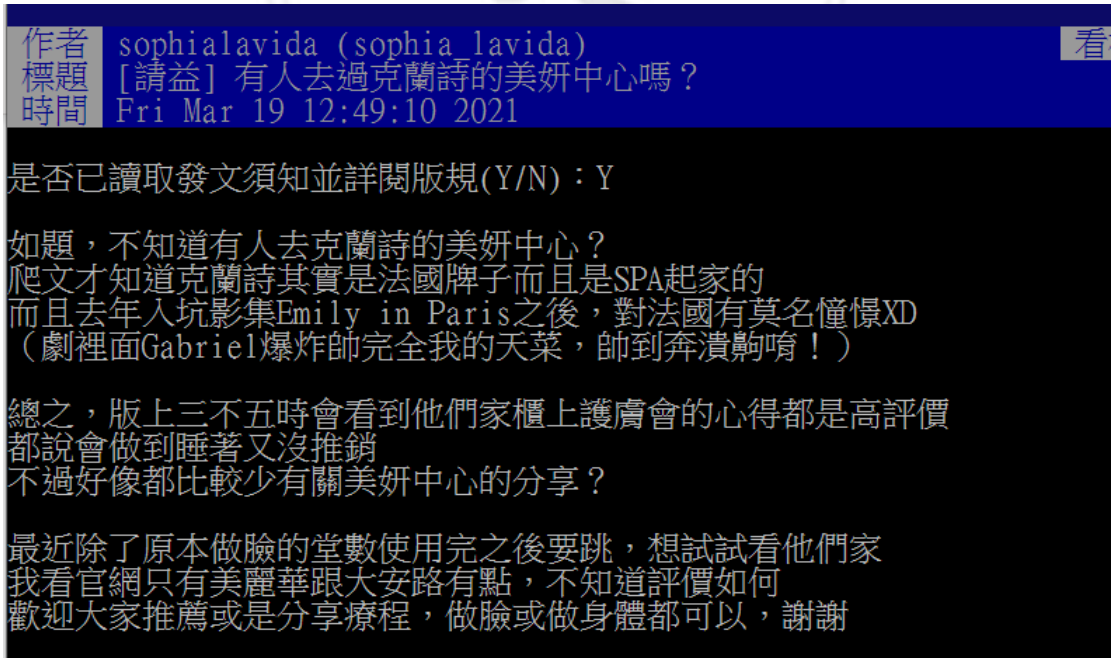
圖 31 PTT 使用分享

(三)、護理療程

根據表 19 發現主題(三)中出現「SPA」、「身體」和「預約」等詞彙，可以觀察到克蘭詩除了化妝保養品的銷售外還推出 SAP 美妍中心， 官網查詢後得知目前台灣有兩間美妍中心位於台北的大安店和美麗華百樂園，美妍中心主打客製化的專業療程、身體護理以及身心靈放鬆療程，共有 17 款肌膚護理療程可以讓消費者自行搭配和挑選，並且需要提前進行預約，才能讓店家有足夠的時間了解客戶的肌膚護理需求。如圖 32 表示有想嘗試 SPA 服務的消費者也會 PTT 美容版進行詢問。

表 19 護理療程主題

SPA	**身體**	簡訊	產品	眼霜
結帳	回購	逆轉	膚色	**預約**
膚況	申請	氧氣	提亮	媽媽


作者 sophialavida (sophia_lavida)
標題 [請益] 有人去過克蘭詩的美妍中心嗎？
時間 Fri Mar 19 12:49:10 2021

是否已讀取發文須知並詳閱版規(Y/N)：Y

如題，不知道有人去克蘭詩的美妍中心？
爬文才知道克蘭詩其實是法國牌子而且是SPA起家的
而且去年入坑影集Emily in Paris之後，對法國有莫名憧憬XD
（劇裡面Gabriel爆炸帥完全我的天菜，帥到奔潰齁唷！）

總之，版上三不五時會看到他們家櫃上護膚會的心得都是高評價
都說會做到睡著又沒推銷
不過好像都比較少有關美妍中心的分享？

最近除了原本做臉的堂數使用完之後要跳，想試試看他們家
我看官網只有美麗華跟大安路有點，不知道評價如何
歡迎大家推薦或是分享療程，做臉或做身體都可以，謝謝

圖 32 PTT 護理療程

(四)、產品介紹

根據表 20 發現主題(四)中出現「全效」、「蘭花」、「美白」、「乳霜」、「鼠尾草」和「精萃」等詞彙，這些在這個模組中有大量出現一些特殊名詞，經過查詢後得知，相關詞彙為克蘭詩的產品名稱，包含有「全效緊緻眼霜」、「花純萃賦活乳」、「美白精華液」、「乳霜系列」、「快樂鼠尾草舒敏系列」、「超級精華黃金雙激萃」等相關產品。如圖 33 所示，使用者會在發文時會打上產品名稱，讓搜尋產品文章的消費者可以立即找到想瞭解的品牌產品，並且在品名底下分享使用心得，消費者可以依據自己需求做更深入的資訊收集。

表 20 產品介紹主題

ml	**全效**	**蘭花**	喜歡	效果
美白	**乳霜**	特別	點數	顏色
鼠尾草	質地	客服	**精萃**	參加

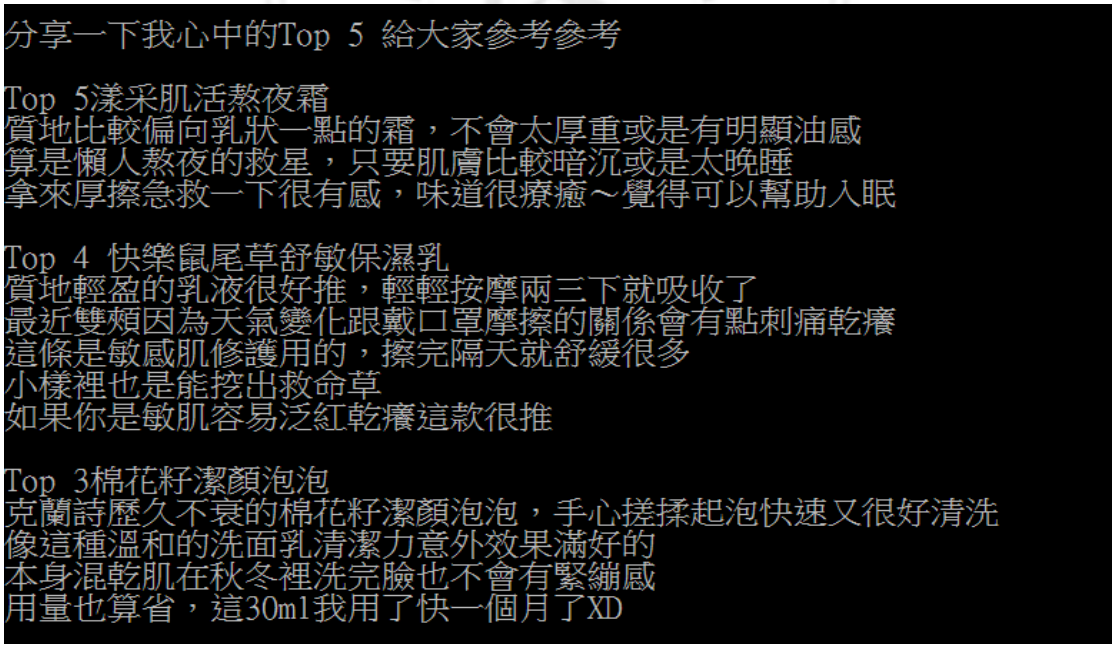


圖 33 PTT 產品介紹

(五)、資訊收集

根據表 21 發現主題(五)中出現「DM」、「官網」、「推薦」和「建議」等詞彙，可以得知消費者會在購物前會收集不同網站和專櫃的產品資訊，像是最新優惠方案的 DM、以及官網的產品介紹和 PTT 美容版的產品推薦，會瞭解到消費者希望以優惠的價格去購買自己想要的產品外，也會把相關資訊透過自身的收集後經過資訊的彙整作業並在社群論壇平台上分享給給其他消費者，從這裡就可以觀察到，消費者與消費者之間會在社群平台中產生資料的收集和資訊分享，如圖 34 所示。

表 21 資訊收集主題

激萃	**DM**	脫皮	門檻	購入
官網	保養	舒敏	賦活	泡泡
HR	引力	**推薦**	**建議**	狀況

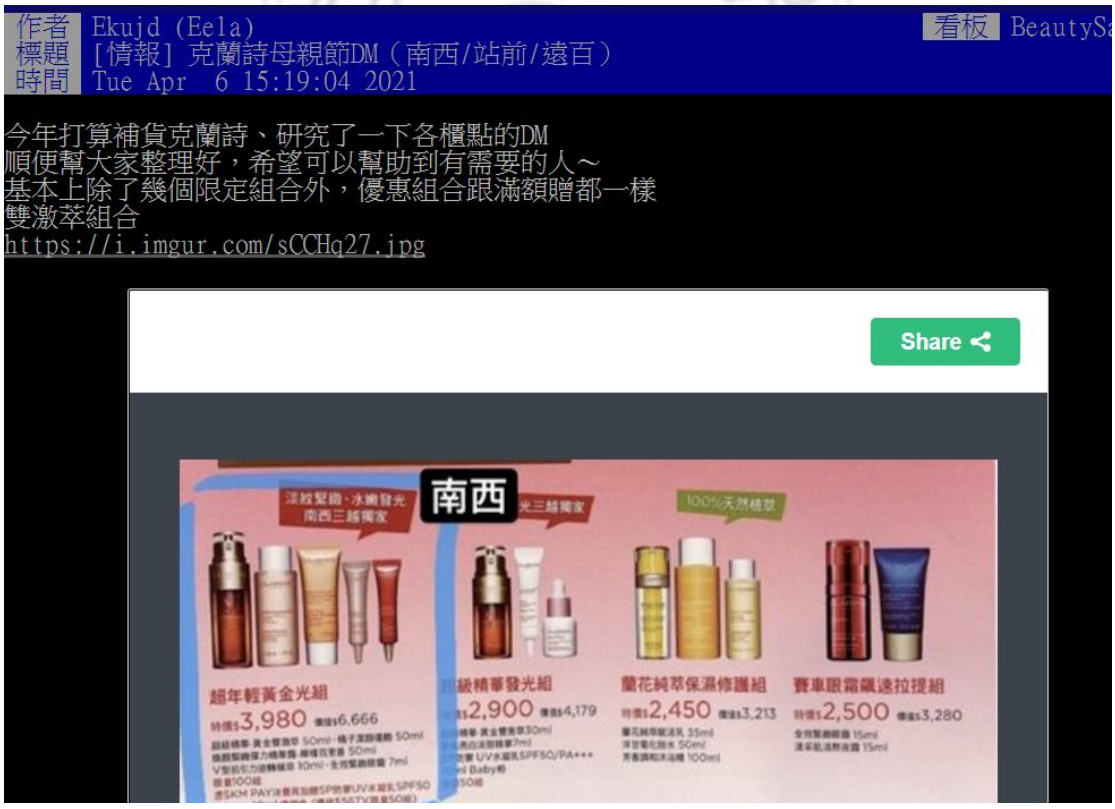


圖 34 PTT 資料收集

四、碧兒泉

(一)、產品購買

根據表 22 發現主題(一)中出現「入手」、「下單」、「正貨」、「購物」、「划算」和「百貨」等詞彙，從中可以發現消費者在 PTT 美容版中會分享自己購買碧兒泉的產品之經驗，消費者會利用品牌推出的優惠方案，以及搭配一些節慶做的促銷活動，使產品價格比以往更加優惠時，消費者會利用這段時間大量添購產品，並且會讓消費者對於品牌釋出的優惠價格和而外贈送的產品禮盒會，使其他有潛在的消費者產生想購買的動機，從銷售的層面來消費者在社群論壇當中分享購買的產品，看對於品牌形象是有正向的影響，如圖 35 所示。

表 22 產品購買主題

入手	**下單**	**百貨**	**正貨**	肌膚
磨砂膏	透露	修護	按摩	**購物**
眼霜	鎖定	男仕	**划算**	邁入

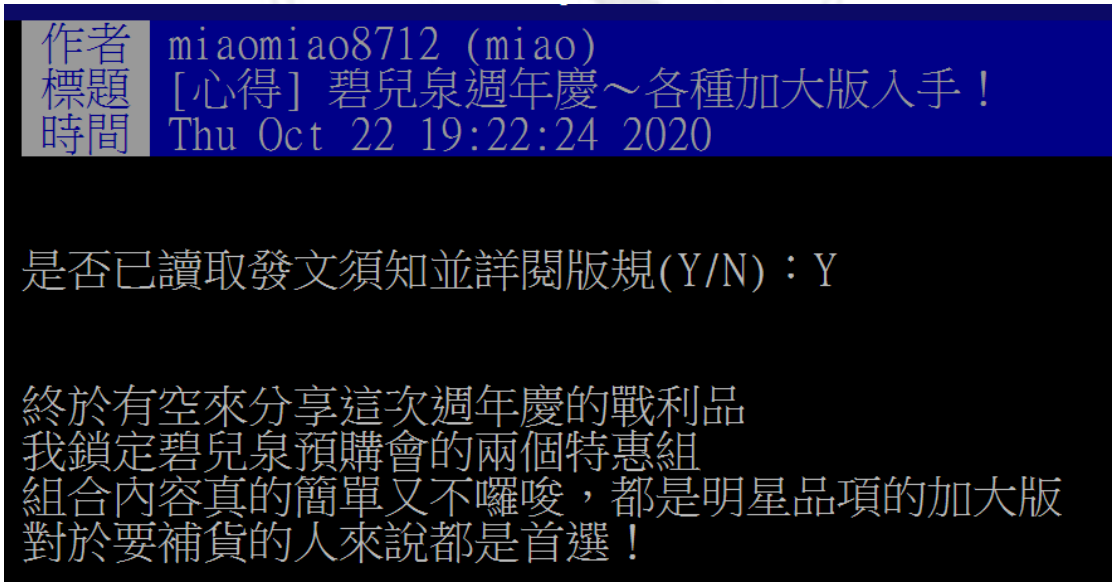


圖 35 PTT 產品購買

主題(二)、產品介紹

根據表 23 發現主題(二)中出現「奇蹟」、「水光」、「發光」和「cc」等詞彙，這些在這個模組中有大量出現一些特殊名詞，經過查詢後得知，相關詞彙為碧兒泉的產品名稱，包含有「奇蹟修護系列」、「水光保濕系列」、「奇蹟角質發光乳」、「cc 隔離乳」等相關產品。如圖 36 所示，使用者會在發文時會打上產品名稱，讓搜尋產品文章的消費者可以立即找到想瞭解的品牌產品，並且在品名底下分享使用心得，消費者可以依據自己需求做更深入的資訊收集。

表 23 產品介紹主題

品項	**奇蹟**	新光	**水光**	**發光**
評價	狀況	天氣	滿額	**CC**
保養品	包裝	**兒泉**	專屬	類似

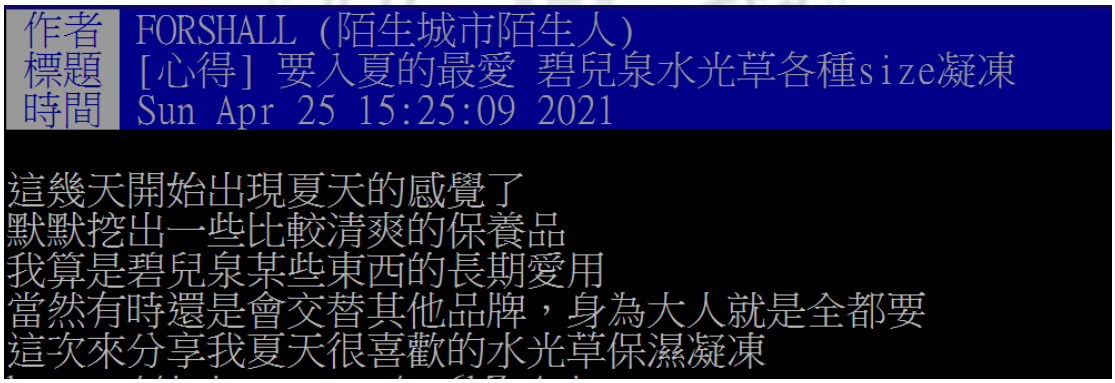


圖 36 PTT 產品介紹

主題(三)、優惠價格

根據表 24 發現主題(三)中出現「小資」、「點數」、「原價」和「折價券」等詞彙，從圖 37 可以發現消費者除了到百貨公司購買產品外，部分消費者會到 momo 購物網進行購買，誘因可能會是因為在網路購買可以累積會員點數以及獲得划算的折價券，透過這樣的方式除了提供實體店面的販售外，也增加了網路通路，並且在網路通路中給予相關優惠價格，會使喜歡網路購物的族群進行品牌產品的購買，因此可以觀察到網路購物搭配優惠方案是可以得到部分消費者的青睞，藉此也會分享至社群論壇中使消費者可以獲得更多優惠價格的資訊。

表 24 優惠價格主題

大滿貫	櫃姐	精露	滴管	**小資**
味道	質地	凝凍	回購	參加
大罐	印象	**點數**	**原價**	**折價券**


作者 junfanray (迷茫)
標題 [閒聊] momo 買蘭蔻/碧兒泉的會員點數
時間 Tue Oct 13 21:03:04 2020

先說我不是幫momo打廣告 XD

蘭蔻 和碧兒泉 在momo是官方旗艦店
所以在momo買不怕假貨又一樣可以累積會員點數
偶有划算的折價券
很適合想小補貨又不想跑百貨的我

最近碧兒泉出了新的app疑似更改會員規章
我記得以前點數效期是2年
但是在他app顯示
我的所有點數居然是到年底過期2020/12/31
（ps包括我今年3月入帳的）

圖 37 PTT 優惠價格

主題(四)、產品資訊

根據表 25 發現主題(四)中出現「DM」、「資訊」、「戰利品」和「容量」等詞彙，會發現消費者會在社群論壇分享自己購買的產品品項，以及分享自己是如何使用這些產品，像是不同季節或時段需要如何使用這類產品，和塗膜在臉上的用量多寡，從關鍵詞中也可以得知產品的容量也是消費者關注的一環，因此在提供產品資訊時會標註上產品的 ml 數，讓其他消費者可以作為一個參考，如圖 38 所示。

表 25 產品資訊主題

是	這	效果	**DM**	**資訊**
搭配	**容量**	天使	**戰利品**	極光
會員	口罩	角質	預購	精華液

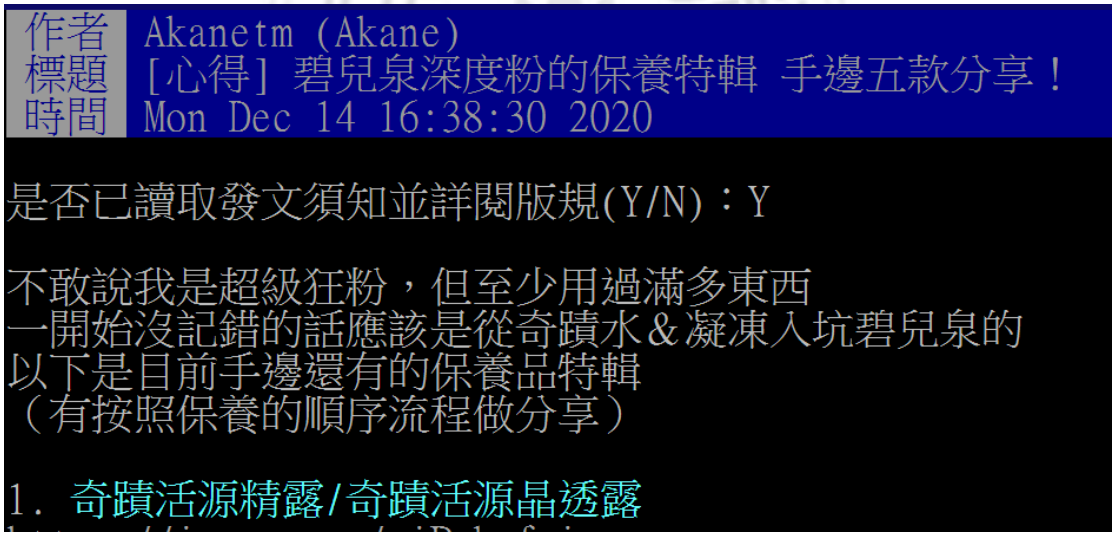


圖 38 PTT 產品資訊

主題(五)、產品推薦

根據表 26 發現主題(五)中出現「喜歡」、「體驗」、「推薦」、「產品」和「男士」等詞彙，產品推薦是消費者藉由自己本身的產品體驗後，把使用的效果以及是否喜歡的程度進行分享，如圖 39，除此之外，關鍵字的男士為一大亮點，有少部分男性也會在社群論壇中進行產品的使用分享和推薦，經由查詢後可以發現碧兒泉有一系列產品是專門提供給男性的，並且在早期利用藝人金城武來為男性產品做代言，以自信、成功、耀眼的風采，來詮釋「男人魅力」，別於其他品牌著重在女性市場，可以發現開始有男性注重皮膚保養，而且讓使用保養品不再局限於女性市場。

表 26 產品推薦主題

ml	活源	**喜歡**	**體驗**	面膜
購買	出現	**推薦**	按壓	momo
使用	**男士**	組合	**產品**	克蘭

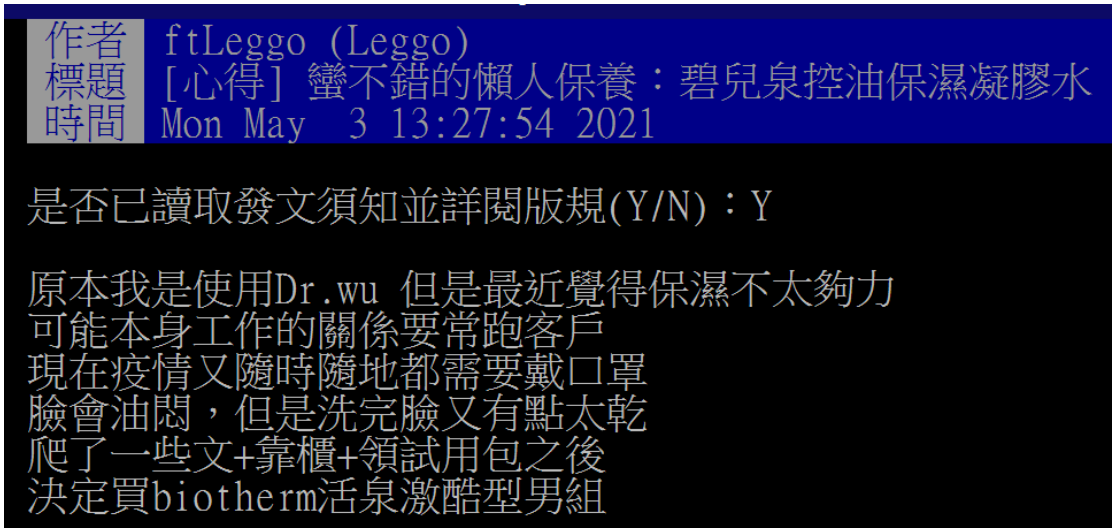


圖 39 PTT 產品推薦

以上研究透過無監督式的機器學習所產生各品牌 LDA 模型，每個品牌總共有 5 個模型，模型裡共有 15 個詞彙，並從關鍵詞中觀察到不同品牌中的關鍵詞之間和主題的相關性，經由整合和分析後如表 27 所示。可以發現每個品牌的文章中，使用者所分享的內容有所差異，並且關注的內容中可以透過主題模型的歸納後發現各品牌的特點。「雅漾」的主題模型中以「防曬功效」為特點之一，使用效果注重防曬係數、保濕、控油等功效，而負面關鍵詞為使用後產生肌膚過敏以及泛紅等狀況，是部分消費者使用完產品後經常出現不滿意的詞彙，雅漾在銷售層面中以連鎖店面以及 MOMO 購物平台為主要通路，並會搭配醫美節和購物平台上的折扣給予產品優惠價格來提高消費者購買意願。

在「蘭蔻」的主題模型中，發現使用者最注意的是產品優惠議題，包含「檔期優惠」、「節日折扣」、「官網購物」主題模型，經由統整後發現蘭蔻的銷售通路和促銷方案非常多元，包含版百貨公司的專櫃周年慶、Line 購物、官網購物、特賣會，此外消費者會非常關注產品的 DM 內容並且會研究如何購買才會最划算，可見消費者對於價格上的優惠非常重視，並且也經常分享蘭蔻的贈品和回饋以及折扣代碼給其他消費者，利用主題模型可以發現蘭蔻的文章中，以如何購買才能最划算以及產品價格優惠方案為文章最常被討論的議題。

在「克蘭詩」的模型當中發現「護理療程」和「使用分享」的主題模型中為主要特點，從中發現除了販售產品外，還提供了不一樣的服務內容，像是建立美妍中心 SAP 館，利用專業的按摩技術以及提供多元化的服務，讓部分消費者會收到吸引和關注，別於其他品牌所提供的服務項目。除此之外，使用分享內容主題中，可以觀察到克蘭詩擁有專業的售後服務諮詢，有專門的客服人員處理消費者購買完產品後若臉部產生一些不良狀況可直接供意見和協助，並且使用者會把收集到的客服資訊分享在社群論壇平台中，讓有相同狀況的消費者可以當作參考。

在「碧兒泉」的主題模型中，發現「優惠價格」和「產品推薦」，為碧兒泉最常被關注的主要內容，在價格中可以發現為碧兒泉雖為專櫃產品，但部分消費者認為這是購買專櫃產品初階可以入手的品牌，因外該品牌也推出小資系列，可以吸引無法購買太高單價的消費族群，不僅如此品牌也推出累積點數和折價券等優惠方案提高消費者購買意願。在產品推薦中除了女性產品外還推出一系列專屬於男性可以使用的產品種類，並且以金城武作為碧兒泉男性產品代言人，這部分引起男性對於保養的關注度外，讓碧兒泉除了擁有女性市場外，也能擴張到男性市場，是其他品牌中較無觀察到的主題內容。

表 27 LDA 主題模型彙整表

品牌名稱	主題模型	詞彙
雅漾	產品介紹	SPF30、SPF50、太陽、乳液、舒護
	防曬功效	皮膚、保濕、潤色、控油、品牌
	心得分享	膚況、狀況、缺點、心得、美白、滋潤、吸收
	使用狀況	防曬、曬黑、過敏、討論
	購買組合	ml、組合、分享、購入
品牌名稱	**主題模型**	**詞彙**
蘭蔻	檔期優惠	回購、購物、回饋、贈品、百貨
	節日折扣	母親節、補貨、組合、購入、櫃姐
	使用方式	步驟、參考、用量、產品

	皮膚狀況	吸收、味道、皮膚、膚質
	官網購物	下單、DM、預購、小樣
品牌名稱	**主題模型**	**詞彙**
克蘭詩	產品效果	保濕、肌膚、吸收、護理
	使用分享	兩頰、臉部、過敏、詢問
	護理療程	SPA、身體、預約
	產品介紹	全效、蘭花、美白、乳霜、鼠尾草、精萃
	資料收集	DM、官網、推薦、建議
品牌名稱	**主題模型**	**詞彙**
碧兒泉	產品購買	入手、下單、正貨、購物、划算、百貨
	產品介紹	奇蹟、水光、發光、cc
	優惠價格	小資、點數、原價、折價券
	產品資訊	DM、資訊、戰利品、容量
	產品推薦	喜歡、體驗、推薦、產品、男士

第五章、結論與未來研究方向

第一節 結論

本研究擷取 PTT 美容版 4013 篇文章進行文字探勘研究，透過中研院辭庫小組研發的 Ckiptagger 和 Python 提供的 Jieba 進行中文斷詞作業，以文字雲來呈現文章中最常被提及的品牌為「雅漾」、「蘭蔻」、「克蘭詩」和「碧兒泉」，再以這四個品牌進行關鍵詞爬蟲，透過 TF-IDF 運算模式取得各品牌前 50 個關鍵詞彙，以及利用以上的關鍵字和主題模型進行分析，每個品牌共有五組主題模型，每組主題模型各由 15 個關鍵詞所組成，再針對關鍵詞的組成類別給予主題名稱，從中觀察各品牌文章的特點以及分析共同之處。

從上述的主題模型中可以發覺，品牌有各自的產品優勢，但同時也有共通點，在三個品牌的主題模型中，各有提到「保濕」和「吸收」的關鍵字，可得知「保濕」的功效意旨肌膚的保水功能，一直以來是消費者著重的效果；除此之外，皮膚對於產品的「吸收」程度，消費者也較為著重，由於台灣為海島並附屬於亞熱帶氣候的國家，天氣較為為悶熱，因此多數人並不喜歡皮膚上的黏膩感，更加喜歡使用吸收度較佳的產品，總和上述可發覺產品「保濕」和「吸收」效果是消費者考量的關鍵點之一。因此本研究建議各品牌可以著重研發「保濕度」和「吸收度」較佳的產品種類，並針對乾性肌、油性肌、中性肌、混和肌和敏感肌五大類肌膚的實際測量，採用數據化的方展現出產品的「保濕度」和「吸收度」，來吸引消費者進行產品購買。

此外本研究發現，在四個品牌的主題模型中，有三個品牌的關鍵字分別提到「DM」，「DM」為廣告文宣的意思，美國直郵及直銷協會（DM/MA）對 DM 的定義為「對廣告主所選定的對象，將印就的印刷品，用郵寄的方法傳達廣告主所要傳達的信息的一種手段。」在化粧保養品「DM」主要以產品的目錄方式呈現，

內容包含品牌販售的產品資訊，以及每季推出的優惠價格和推廣品牌活動之相關資訊，目的利用圖片和文字來傳達品牌資訊吸引消費者的眼球。過去多數「DM」是以紙本式作為行銷宣傳，而現今因為網際網路發達也開始發展出線上式「DM」。本研究發現在 PTT 美容版的文章中，消費者在平台上採用紙本式「DM」做資訊分享，可以觀察到紙本式「DM」對於消費者來說有重複翻閱和勾選預購買產品，以及計算產品價格等功用。因此本研究建議，品牌可以深入了解消費者使用「DM」的習慣，並在產品目錄設計上進一步符合消費者的需求，可以使消費者在閱讀紙本式「DM」更為便利。

第二節 未來研究方向

本研究方法，仍需改進建議未來若要針對特定主題進行文字探勘研究，在中文斷詞前需要事先加以瞭解專有名詞，並新增至自建詞當中，讓中文斷詞結果可以更為準確。此外，在取得高詞頻(TF-IDF)關鍵字的計算方式其中優點是快速且容易理解，缺點為精確度有限在衡量文章中詞的重要性不夠全面，建議未來可以使用 word2vec 和 Bert 等其他模型進行文本分析。

參考文獻

1. Arndt, J. (1967). Role of product-related conversations in the diffusion of a new product. Journal of marketing Research, 4(3), 291-295.
2. Abdous, M. H., & He, W. (2011). Using text mining to uncover students' technology-related problems in live video streaming. British Journal of Educational Technology, 42(1), 40-49.
3. Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. Management science, 57(8), 1485-1509.
4. Chowdhury, G. G. (2003). Natural language processing. Annual review of information science and technology, 37(1), 51-89.
5. Cheung, C. M., Lee, M. K., & Rabjohn, N. (2008). The impact of electronic word-of-mouth: The adoption of online opinions in online customer communities. Internet research.
6. Dichter, E. (1966). How word-of-mouth advertising works. Harvard business review, 44, 147-166.
7. Gode, D., & Mayzlin, D.,"Using Online Conversations to Study Word-of-Mouth ommunication", Marketing Science, 23(4), pp. 545-560, 2004.
8. Harrison-Walker, L. J. (2001). The measurement of word-of-mouth communication and an investigation of service quality and customer commitment as potential antecedents. Journal of service research, 4(1), 60-75.
9. Lafferty, B. A., & Goldsmith, R. E. (1999). Corporate credibility's role in consumers' attitudes and purchase intentions when a high versus a low credibility endorser is used in the ad. Journal of business research, 44(2), 109-116.
10. Podoshen, J. S. (2008). The African American consumer revisited: brand loyalty, word-of-mouth and the effects of the black experience. Journal of Consumer Marketing.
11. Sullivan, M., Lehane, C., & Uhlmann, F. (2001). Orchestrating anaphase and mitotic exit: separase cleavage and localization of Slk19. Nature cell biology, 3(9), 771-777.
12. Trusov, M., Bucklin, R. E., & Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. Journal of marketing, 73(5), 90-102.
13. Tang, C., & Guo, L. (2015). Digging for gold with a simple tool: Validating text mining in studying electronic word-of-mouth (eWOM) communication. Marketing Letters, 26(1), 67-80.
14. Kolter,P.(1997).MarkingManagement:Analysis,Planning,Implementation,and Control.New York:Prentice-Hall, Inc.

1. 江義平,蔡坤宏,& 黃耀德.(2015).網路口碑行銷效果探究-以經驗品為例.中原企管評論,13(2), 33-64.
2. 陳世榮.(2015).社會科學研究中的文字探勘應用: 以文意為基礎的文件分類及其問題. 人文及社會科學集刊,27(4),683-718.
3. 潘彩君.(2018).以文字探勘法檢證習近平時期之中共國際話語權.
4. 程婉婷.(2011).網路口碑對消費者購買決策影響之探討-比較 Facebook 與部落格.
5. 黃嚴弘,&黃瓊儀.(2018).企業形象,體驗價值,網站品質對滿意度與購買意願之影響-以台糖健康易購網為例.育達科大學報第 46 期.
6. 洪偉章、李金枝、陳榮秀(1997)。《化妝品原物料及功能》。台北市，藝軒圖書出版社。

網際網路

1. 財團法人台灣網路資訊中心. (2019). 〈2019 年台灣網路報告〉〈https://www.twnic.tw/doc/twrp/201912e.pdf3〉
2. 台灣經濟部統計處. (2021). 〈「宅經濟」發酵，帶動網路銷售額成長〉〈 https://www.moea.gov.tw/Mns/dos/bulletin/Bulletin.aspx?kind=9&html=1&menu_id=18808&bull_id=7590〉
3. Kantar. (2021)〈口罩必備的疫年後數位趨動保養市場復原力 2021 挑戰再起. 〉.〈https://www.kantarworldpanel.com/tw/news/20210602-beauty-digital〉