大數據分析語言 Python 網路社群文字探勘(Text Mining)-以 PTT 論壇網購版為例

李書政[1]
樹德科技大學 行銷管理系 助理教授
sclee99@stu.edu.tw
楊凱文[2]
樹德科技大學 行銷管理系 學生
s16111124@stu.edu.tw
蘇培元[3]
樹德科技大學 行銷管理系 學生
s16119256@stu.edu.tw

摘要

本研究希望利用大數據分析中相當熱門且功能強大的 Python 語言，撰寫網路爬蟲擷取 PTT 論壇中，日本網購版的精華文章內容，並進一步使用文字探勘技術，嘗試在 PTT 論壇網購版的使用者心得文章中，尋找出人們進行網購時討論的重要議題焦點，和現象背後隱含的重要知識與訊息，透過程式處理大量文字和數據的優勢，高效率地處理解讀海量訊息的問題，期望產生多重管理實務的啟發，並為未來研究提供參考建議，且可以透過文字探勘、關鍵字萃取、文字雲的視覺化呈現，幫助使用者能快速且清楚地了解到日本網購消費者的購物心得和文章談論重點，提供給使用者更豐富與精確的購物參考資訊，且能夠透過網路論壇的評論了解到各個網路店家的信用問題及品質參考，減少網路上消費者被黑心店家欺騙的案例，使得消費者在網購時能夠更安全並準確的購買到其所需的物品。

關鍵詞：文字探勘、文字雲、網購、大數據

壹、前言

一、研究背景

在現代，網路普遍，在日日更新的網路文章中，我們要找出所需的知識有一定的困難，文字探勘技術也因應此需求而逐漸發展，特別是因為多數的社群網站興起，產生了大量貼文，且使用者能隨心所欲發表意見，若能在使用者們的回應中加以萃取有用的知識與建議，對於企業在行銷或管理各方面，都會產生有用的決策建議，提升企業經營的效率及效能，甚至也可成為消費者網路購物的參考因素。但是，電腦無法理解文字這個非結構化資料，該如何在短時間內，有效率地在大量的文章中萃取出想表達的重要訊息，就需要使用到文字探勘的技術。透過這樣子的技術，能夠幫助電腦去學習人類的文字，更能夠幫助未來人類在使用電腦時可以更方便的去閱讀或是搜集所需要的資訊，使得我們在這大數據時代能夠更有效的將資料整理為有效數據。

二、研究動機與目的

文字探勘技術雖然有許多開源碼工具語言（例如、R、Weka、OpenNLP、LingPipe、FreeLing），Python 語言是近年成長最快速的大數據分析語言，在處理網路巨量資料上有其優越性。

本次研究撰寫 Python 語言，藉由其強大的網頁擷取與分析能力，結合文字探勘分析相關的功能模組，透過文字截取以及資料視覺化的方式能夠將網路上的資料系統性的整理並顯示出來，使得人類在蒐集資訊時能夠更清楚。

我們先以對 PPT 論壇網購版來當作測試目標，期望能透過文字探勘的技術並藉由目標網壇相關討論及評論的文章，來加以分析出所需要的有效數據，來產生相關建議，最後利用資料視覺化的文字雲呈現結果，此結果有助於成為未來使用者決策參考的因素之一。

貳、文獻探討

一、文字探勘資訊

也被稱為文本挖掘、文字採礦、智慧型文字分析、文字資料探勘或文字知識發現，一般是指從非結構化的文字中，萃取出有用的重要資訊或知識。資料探勘（Data Mining）與文字探勘（Text Mining）的關係緊密，前者是處理結構化的數值型資料型態，而後者是針對文字進行分析，在處理非結構化與半結構化的資料型態中，挖掘出隱含在文字中有用的訊息。

文字探勘（Text Mining） 是一種跨領域的應用，結合資料探勘技術與自然語言處理、資訊檢索技術，使大量的文字資訊能經由電腦分析歸納，主要的應用有自動分類、自動摘要、文件檢索、知識管理等。用以因應今日因網際網路(Internet) 興起，而造成的龐大的數據洋流。文字採礦之核心技術，大多來自於資料採礦技術，將藉助案例分析與文件資料之相互查詢與交叉比對，產生經驗與文件報告之交互參考對應。該技術整合了許多傳統資訊檢索技術，包括了關鍵字萃取、全文檢索、文件自動分類、自動摘要等等，以提供文字處理更強大的功能。

二、文字探勘進行程序

(一). 資料檢索與蒐集

在文字探勘的應用中，具有一項優勢在於其分析的對象多屬電子化的文本，在蒐集資料上比紙本方便許多，若是需要更新或是插入新資料，更可以即時處理新資料並產生新結果。資料蒐集不限定內容，在研究中會因應其需求而採用不同類型的文本，不同的文本在處理前應先注意其環境而隨時更動程式碼。

(二). 文本前處理

首先文本前處理要先將句子明確的做出分隔，通常使用標點符號作為其分隔符號即可。其次則是進行斷詞作業。為了瞭解中文文章之意義，必須對文章進行斷詞(喻欣凱，2008)。中文斷詞系統主要有兩種，分別是結巴斷詞(Jieba)與中央研究院詞庫小組(Chinese Knowledge Information Processing Group，簡稱 CKIP)開發的斷詞系統。

斷詞所遇到的最大瓶頸在於未知詞的判讀，詞庫中若沒有其提到的字詞，就難以進行處理，也容易造成錯誤的斷詞或是詞彙無法說明意思的情況。結巴斷詞為 Python 程式語言中的中文斷詞套件，其程式碼的開源與可以自訂辭典的特性讓使用者在斷詞上有較高的彈性。Jieba 中文斷詞套件為中國百度公司員工孫君意在 2012 年 10 月 7 日首次於開放原始碼社群平台 Github 上發表，而後也有許多人陸續加入改善套件並開發不同程式平台的 Jieba 套件。本次研究中則是使用 Python 的 Jieba 套件繁體中文版本詞庫，對所蒐集到的文本進行斷詞。

(三). 核心挖掘作業

完成斷詞後則進行核心挖掘作業，此作業包含特徵萃取、分類與集群、詞頻等，透過核心挖掘作業可將已完成斷詞的資料進一步統整與分析。特徵萃取用以找出文本中的詞彙特徵，並區別出文本的類別與屬性，使文本更加結構化。分類與集群則是透過詞彙出現次數的計算，對主題進行分類並產生關連，或是尋找詞彙之間的群集關係。

(四). 結果呈現

在完成分析後，將結果清楚的呈現出來才可讓他人明確的了解分析的內容。視覺化工具為此程序中的主要類別，最後呈現出的結果必須包含前文所提到的核心挖掘作業之結果(分類、群集等)，並針對需求特殊化讓結果得以更清楚的被表現，若有圖像不足呈現的部分則可以使用表格和文字加以解說。

(五). 詮釋

不論挖掘、分析的結果為何，若不對其進行人為的詮釋往往難以將結果化為有用的訊息。因此對分析結果加以解釋及歸納為最後步驟，同時必須檢視討論中的研究目標與假說，提出缺陷與調整的方向，以望在未來的研究能發展為一個更完整的文字探勘工具。

三、文字探勘相關研究

(一). 林名彥(2015)－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 PTT 論壇的網購版的使用者文章進行文

字探勘，以尋找文中的關鍵字詞，並瞭解網友們經常關注的主題和關聯的字詞。

(二). 鄭凱文(2014)－本研究樣本為 2011 年中國大陸所有上市公司所揭露的 MD&A 及相關財務資訊，MD&A 非量化資訊係運用 Stanford Word Segmenter 斷詞資料庫、正負向詞典、TFIDF、K-means 等技術進行群集分析，並結合財務資訊的 K-Means 群集分析，分析出中國大陸 2011 年上市公司 MD&A 揭露是否誇大。

(三). 劉育華(2014)－本研究以兩家宮廟的 Facebook 粉絲專頁的官方貼文為分析標的，以文字探勘(Text Mining)的工具從貼文中找出最常出現的關鍵字，分析其詞頻以文字雲 Word Cloud)來呈現，並依照內容分析法(Content Analysis)將關鍵字貼文做性質分類，進一步使用社會網絡分析(Social Network Analysis)，來探討哪些類型及性質的貼文會吸引較多使用者回應，期望藉由文字探勘及社會網絡的結合，找出使用者最感興趣的主題與溝通方式。

(四). 楊正銘(2004)－國際疾病分類系統是全球公共衛生界用以描述疾病、分析病歷及訂定衛生政策時經常使用的溝通工具。而疾病分類的工作是由醫院病歷單位的疾病分類人員閱讀病歷，並評估疾病資料之適當性，以確認主要診斷、次要診斷、處置及併發症，並依照 ICD-9-CM 手冊上準則轉換成適當的數字代碼，而早期記載的出院病歷摘要文件需靠疾病分類人員解讀進行編碼，然而經由疾病分類人員解讀病歷不但耗時也容易造成編碼錯誤或遺漏。因此，本研究動機是希望結合文字探勘技術與國際疾病分類編碼規則，試圖從出入院病歷摘要文件資料中找出描述疾病與編碼所隱藏的規則，並將此規則應用於疾病分類系統中，並探討此輔助系統與疾病分類人員編的疾病碼差異性及改善自動文件分類的準確性，藉以協助疾病分類人員能提升編碼工作效率及節省編碼時間。

(五). 柯秀奎(2005)－本研究希望透過一自動客訴留言分類的機制，將會員的留言藉由文字探勘的技術和分類機制將其自動的分門別類，使客訴能有效率且較準確地配屬於合適的客服人員或資訊工程師，透過客訴留言的自動分類機制來加速處理會員在使用上不懂的疑問或客戶抱怨。而透過此系統不但能提升顧客的滿意度也可降低人工瀏覽客訴進行問題配屬的人力，在網站經營有限的人力上，提高人員的生產力及工作效率。此外，我們也利用自動客訴留言分類機制，來管理網站服務的品質，透過管制圖概念及文字探勘技術的結合，針對客訴留言內容進行監控，藉由分類後屬於顧客抱怨的留言數量，繪製成管制圖，以用來維持及改善社群網站的品質，達到持續吸引新會員加入及避免舊客戶流失的目的。

四、PTT 論壇網路社群介紹

全名批踢踢實業坊，簡稱批踢踢、PTT，是一個臺灣電子布告欄（BBS），採用 Telnet BBS 技術運作，建立在台灣學術網路的資源之上。擁有超過 2 萬個不同主題的看板，每日超過 2 萬篇新文章及 50 萬則推文被發表，從八卦、政治、運動、娛樂、文學、網購……應有盡有，是台灣使用人次最多的網路論壇之一，也是台灣最有影響力的網路社群之一。根據 Ptt 的官方說明，它的目標是「建立一個快速、即時、平等、免費，開放且自由的言論空間」，且絕不商業化、絕不營利。雖然這種網路匿名式討論空間隨處可見，但 PTT 擁有幾項特色，情報流通快速、資訊容易搜尋、討論能見度高、匿名性帶來真實想法、鄉民力打造群眾智慧，使得 PTT 始終是人氣最集中的空間。

本次研究擬採用 PTT 論壇網購版精華區日貨買家心得分享文章做為文字探勘分析的研究對象，以分析出網友們的談論焦點及相關評論重點，藉此發現網購消費者心中真正想關注的議題重點，以對網購業者提出行銷與管理實務上有價值的相關建議。

圖一、PPT 論壇 Logo

五、Python 介紹

Python，是一種廣泛使用的高階程式語言，由 Guido van Rossum 創造。Python 的設計哲學強調程式碼的可讀性和簡潔的語法（尤其是使用空格縮排劃分程式碼塊，而非使用大括號或者關鍵詞）。相比於 C++或 Java，Python 讓開發者能夠用更少的代碼表達想法。不管是小型還是大型程式，該語言都試圖讓程式的結構清晰明瞭。Python 擁有動態型別系統和垃圾回收功能，能夠自動管理記憶體使用，並且支援多種編程範式，包括物件導向、命令式、函數式和程序式編程，經常被當作腳本語言用於處理系統管理任務和網路程式編寫，Python 也非常適合完成各種高階任務。

六、Anaconda 介紹

Anaconda 中文是森蚺(ㄖㄢˊ)，是一種非常肥大的蟒蛇。簡單來說，就是把 Anaconda 當作是 Python 的懶人包，除了 Python 本身(python2,3)還包含了 Python 常用的資料分析、機器學習、視覺化的套件。

本次研究使用 Anaconda 內含套件中的 Notebook 功能，一個輕量級 web-base 撰寫 Python 的工具，在資料分析這個領域很熱門，雖然功能沒有比 Pycharm,Spyder 這些專業的 IDE 強大，但只要 code 小於 500 行用此功能撰寫非常方便。

參、研究方法

一、研究架構

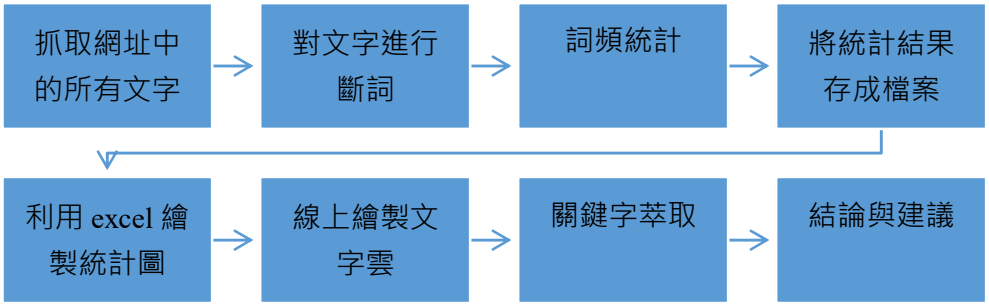
本次研究架構分成三個階段五步驟，第一階段為資料蒐集處理，第二階段為資料分析，第三階段為資料呈現與詮釋。五步驟依序為：1.資料檢索與蒐集　2.內容斷詞處理　3.重點挖掘作業　4.結果呈現　5.詮釋。此外，每個步驟皆有相對應文字探勘擬採用的 Python 功能模組。

表一研究架構

研究架構	研究步驟	Python 相關功能模組
第一階段 （資料蒐集處理）	1.資料檢索與蒐集 2.文本前處理	網頁資料擷取 （requests、Beautifulsoup）
		中文斷詞 （Jieba）
第二階段 （資料分析）	3.核心挖掘作業	詞頻統計 （Counter）
		關鍵詞萃取 （sklearn）
第三階段 （資料呈現與詮釋）	4.結果呈現 5.詮釋	分析圖表呈現 （Matplotlib） 視覺化呈現 (Wordcloud、Word Art)

二、研究流程

以下為實際研究執行所進行的流程步驟：



圖二、研究流程圖

三、母體資料來源

本研究所擷取的母體資料來源，是針對 PTT 論壇網購版精華區日貨買家心得分享文章為對象，透過 Python 爬蟲程式抓取，並以此資料經由本次研究中的文字探勘流程分析後，探討詮釋其中之結果，以探索發掘出隱含在這些文章中的重要知識與訊息。

四、文字探勘相關技術

(1) 文本前處理

從網路上所爬下來的資料可能包含了很多贅詞、亂碼與廣告，或是雜亂無章的篇幅，因此必須先對所要處理的文字進行分割整理，將研究的目標進行分類，這樣可以有效的進行後續的分析。

(2) 中文斷詞

又稱為單詞，是能獨立運用並含有語義內容或語用內容（即具有表面含義或實際含義）的最小單位。任何語言處理的系統都必須先能分辨文本中的詞才能進行進一步的處理，例如機器翻譯、語言分析、資訊萃取。當處理不同領域的文檔時，領域相關的特殊詞彙或專有名詞，常常造成分詞系統因為參考詞彙的不足而產生錯誤的切分。為了解決這個問題，最有效的方法是補充領域詞典加強詞彙的搜集。中文與英文最大差異在於詞與詞之間不像英文有明顯的區隔。此時我們就需要透過中文斷詞的過程來將文章中的句子分成詞。本次研究採取在 Python 語言中，屬於開放原始碼，比較廣泛被採用的 Jieba 斷詞模組來進行斷詞。

(3) 詞頻統計

一個字詞在某個文件，或是語料庫中出現的次數，稱之為詞頻(Term Frequency，TF)。在詞頻的基礎上，要對每個字詞分配一個權重，例如在中文中最常見的詞，如「的」、「是」等無法表達出意思的詞給予較小的權重，而較少出現的詞，如「電腦」、「手機」等表面或內含意義的詞則給予較大的權重。這樣的權種稱為逆像文件頻率(Inverse Document Frequency，IDF)，而逆向文件頻率的大小與一個字詞的常見程度成反比。

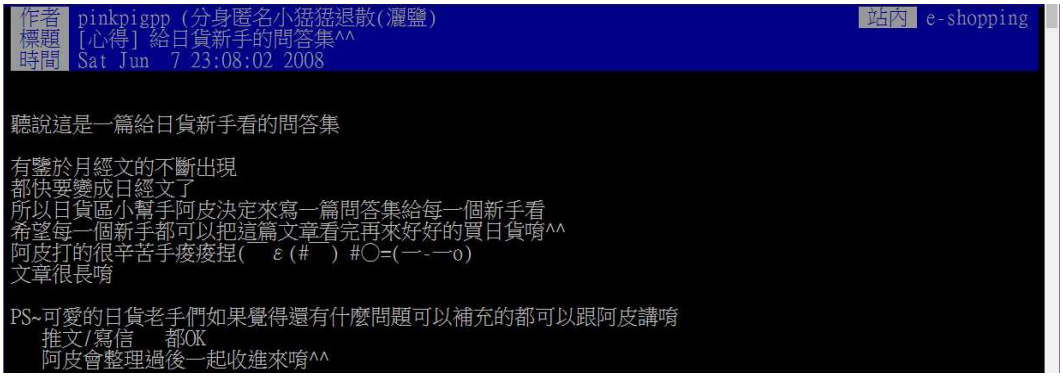
(4) 資料視覺化－文字雲(Wordcloud、Word Art)

在前述中文斷詞、詞頻統計、關鍵字萃取等過程中，本研究擬將結果以圖表呈現，另外也採用文字雲的方式，呈現出網購相關重要詞組。文字雲是文字探勘一個很重要的環節，因為它是最直接也最清楚的一種文字探勘表現方式，圖的構成是由許多的文字組合而成，一塊一塊的文字塊看起來就像雲一樣，故稱為文字雲。文字雲可以提供研究者一個快速而整體的視覺印象，讓閱讀者很快地掌握特定議題相關的關鍵字、並有助於相關概念的啟發。

肆、資料分析

本研究以 PTT 社群網站上面的日本網購版精華區的文章為標的，先利用爬蟲程式，抓取原始頁面內容，經過 python 的 Beautiful Soup 模組解析成 HTML5 格式，從中再擷取出原始純文字內容，我們模擬網頁瀏覽器正常瀏覽 PTT 網站的方式，以防止爬蟲程式被阻擋而無法抓取網頁內容，但是此程式碼是針對 PTT 網站格式而設計，並不適用於其他網站(格式類似的網站例外)，如：蘋果日報、YAHOO 新聞等。

以下是擬抓取的文章範例畫面：



圖三、本次研究採用之範例文－PTT 論壇網購版(部分文章)

一、程式碼說明

首先，程式先匯入本次研究任務所需要的 python 模組，以方便後續程式引用當中的功能作分析。

```
import requests
from bs4 import BeautifulSoup as bs
import pandas as pd
import numpy as np
import re   #沒有匯入也可執行正則表達式
import jieba
import jieba.analyse
import operator     #一定要裝才能用operator功能
#from operator import itemgetter, attrgetter    若安裝operater若仍無法執行再加上去
```

圖四、程式碼－匯入必要的程式套件

以下程式的內容主要在於模擬一般瀏覽器抓取網頁資料的方式，取得網頁上特定位置中的網頁內容，解析成 html5 的網頁標籤格式，加以下載。

```
url='https://www.ptt.cc/man/e-shopping/DB9A/D40A/D5E8/M.1213031940.A.4C5.html'
request_headers={
'user-agent':'Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/63.0.3239.132 Safari/537.36'
}
r=requests.get(url, headers=request_headers)
soup=bs(r.text, 'lxml')
category=soup.find('div',{'id':'main-content'}).text
print(category)
```

圖五、程式碼－抓取網路文章文字內容

```
Building prefix dict from C:\Users\rb891\pythonwork\dict.txt.big ...

作者pinkpigpp (分身匿名小豬豬退散(灑鹽)站內e-shopping標題[心得] 給日貨新手的問答集^^時間Sat Jun  7 23:08:02 2008

聽說這是一篇給日貨新手看的問答集

有鑒於月經文的不斷出現
都快要變成日經文了
所以日貨區小菊手阿皮決定來寫一篇問答集給每一個新手看
希望每一個新手都可以把這篇文章看完再來好好的買日貨唷^^
阿皮打的很辛苦手痠痠捏(￣ε(#￣) #o=(一-一o)
文章很長唷

PS~可愛的日貨老手們如果覺得還有什麼問題可以補充的都可以跟阿皮講唷
   推文/寫信   都OK
   阿皮會整理過後一起收進來唷^^
```

圖六、程式碼執行結果(1)(部分文章)

以下程式內容主要是啟動 jieba 斷詞模組，將文章中的句子斷開成一個個有意義且常見的中文詞彙，同時也設有排除字的功能，排除一些無意義或不重要的字詞，最後將結果存到陣列中。

```
jieba.set_dictionary('dict_big.txt')  #切換台灣繁體版dict.txt斷詞效果優於內建簡體詞庫
jieba.load_userdict("userdict.txt")  #附加載入使用者自訂詞庫,補上柯文哲字詞
words=jieba.cut(category,cut_all=False)
break_words=[]   #宣告一個串列,將words所有斷詞結果放入
for j in words:
    break_words.append(j)
stopwords=[]
for word in open('stopwords.txt','r',encoding="utf-8-sig"):
    stopwords.append(word.strip())
del_stopwords=[]
for k in break_words:
    if k not in stopwords:
        del_stopwords.append(k)
```

圖七、程式碼－利用 Jieba 斷詞模組進行斷詞

以下程式將陣列中的斷詞文字存成資料框架格式(dataframe)，並利用正則表達式過濾文字內容，只選取出中文漢字斷詞內容，並去掉換行符號

```
df=pd.DataFrame(del_stopwords)  #將headline陣列轉為dataframe
data_clean=df[df[0].str.match('^[\u4e00-\u9fa5]{0,}$')]
#利用正則表達式抓取words欄位是"漢字"的資料
data_clean.columns=['words']
data_clean = data_clean[data_clean.words != '\n']
#去掉含有換行符號的斷字
kk=[]
for i in range(len(data_clean)):
    kk.append(data_clean.values[i][0])
```

圖八、程式碼－抽取漢字，過濾無意義之符號(如：標點符號）

以下的程式碼內容，主要是執行斷詞後的詞頻統計，統計出各個詞彙出現的次數頻率，扣除一些無意義的語助詞，出現較多次數的字，通常是相對比較重要的字。

```
word_count = dict()  #宣告一個字典型態變數,含鍵key和值value,放字詞和出現次數
for word in kk:
    if word in word_count.keys():
        word_count[word] +=1      #每一個在words中的元素,若在word_count中發現相同的鍵key名稱,就在該鍵值values累加1
    else:
        word_count[word] = 1      #若沒在word_count中發現相同的鍵key名稱,鍵key名稱就是元素名稱,值values=1
    sorted_wc =sorted(word_count.items(), key=operator.itemgetter(1), reverse=True)
    #利用word_count.item()指令取出wordcount所有元素進行排序,再以operator.itemgetter(1)取出word_count的第二欄值value來作排序的依據
    #reverse=Ture反向排序,由大排到小
```

圖九、程式碼－詞頻統計(部分程式碼)

以下的程式碼，主要是將詞頻統計的結果，分別存成.csv 檔和.txt 檔，以方便後續分析使用。

```
df=pd.DataFrame(xfile)  #將xfile陣列轉為dataframe
df.to_csv('ppt_count_japan_update01.csv',encoding="utf-8-sig")  #將結果轉存為csv
f = open("ptt_count_japan_update01.txt","a")    #用附加的方式a,開新檔案,一筆筆寫入
for t in range(0,a):
    f.writelines(xfile[t][0]+","+str(xfile[t][1])+"\n") #用str轉字串,加入逗號分隔",",加入換行符號\n
f.close()     #請養成檔案開啟後要關閉的習慣
```

圖十、程式碼－將結果存進 csv 檔和 txt 檔

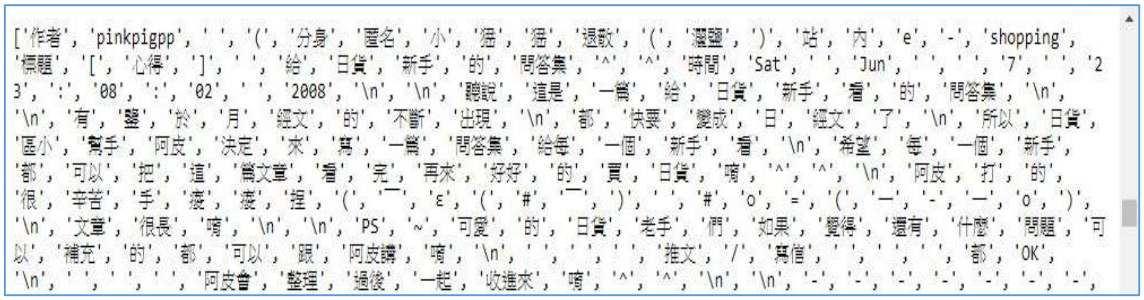
以下的程式碼，主要的功能在於，利用 jieba 模組內建的關鍵字萃取功能，快速地萃取出特定文章內容的關鍵字，本例當中萃取出 50 組文章關鍵字(原本預設值為 20 組)，並且將找出的這 50 組關鍵字分別以”|”符號，將它們串連在一起。

```
keywords = jieba.analyse.extract_tags(category, 50)  #keywords是關鍵字串列,預設值為20
print("jieba內建關鍵字萃取方法extract_tags取出前50個關鍵字內容如下(可能包含標點符號與數字)：\n",keywords)
print("用中分符號將50個關鍵字串接在一起，內容如下：")
print("|".join(keywords))  #用"|"將串列keywords中的所有元素連成一個字串
```

圖十一、程式碼－利用 Jieba 模組內建的關鍵字萃取功能（取前５０個的結果）

二、程式執行結果

（一）斷詞結果畫面：先將文章中的句子斷詞出來



圖十二、斷詞(部分)

（二）詞頻統計結果畫面：以詞頻統計的方式來做文字內容整理

圖十三、詞頻統計(部分)

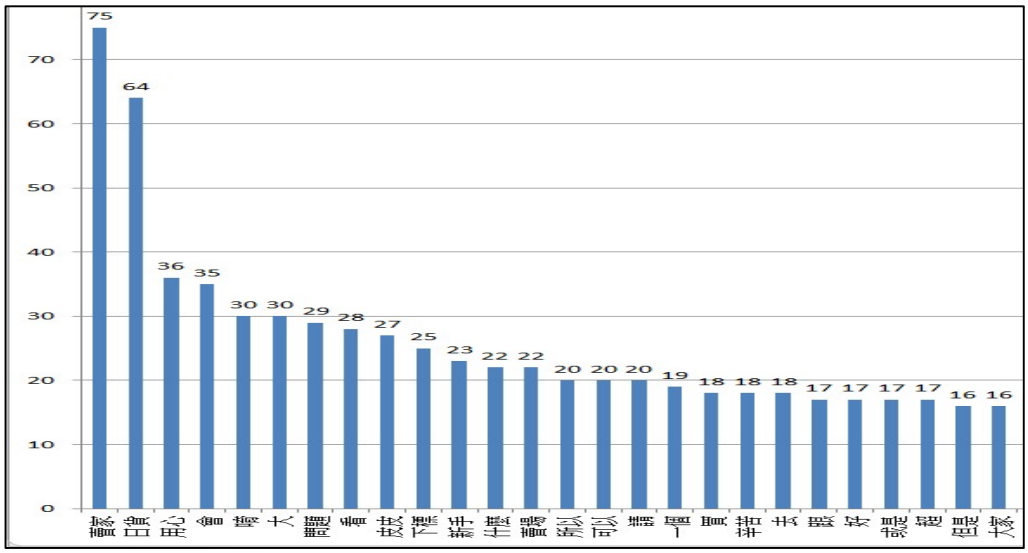
（三）關鍵字萃取畫面：以關鍵字萃取的方式來作文章重要內容萃取整理

圖十四、關鍵字萃取

（四）文字雲視覺化呈現：透過文字雲呈現，能夠讓使用者更清楚明白他所觀看的文章裡頭的重點以及關鍵字表達。

圖十五、視覺化呈現(線上繪製文字雲)

（五）詞頻統計圖：將詞組以在文章中出現的次數從多到少排列



圖十六、詞頻統計圖(部分)

伍、結論與建議

一、結論

在這個大數據的時代，資料的量是猶如大海般一樣的多，如果不能試著將資料整理成有效數據的話，將會有很多資料被浪費掉，但基於人體學習及吸收的速度和容量都有個限度，必須透過電腦這樣子的工具來幫助我們，且利用文字探勘的技術能夠彌補電腦不懂人類文字的這個缺點，將文章的關鍵字整理出來並統計，最後再以文字雲的方式呈現出來，使得我們能夠快速的了解每篇文章的重點及大意，這樣可有效率的幫助我們從大數據中得到資訊並記錄。

這樣子的重點整理方式，可運用在商業、學術、政策、醫學及日常生活，能夠使這個世代的生活更加的有效率且系統性，不會再盲目的搜尋資料和統計，並減少人們在交換資訊時資訊不對稱的發生機會，使得資訊的傳遞及分享能夠更準確更清楚。針對在 PTT 論壇的運用上，其論壇本身就是網路使用者發表心得及看法的論壇，透過文字探勘的技術可針對該文章的主題，了解到普遍網友對於該主題的看法及心得，再來綜合整理出該文章的重點以及網友們的心得看法，在網路購物上能夠發會非常大的用處!

對於只想先看重點或者是時間有限的現代人，就好像是能夠先整理出所謂的懶人包給大眾參考，這樣子的方式對於這個世代是非常的方便，但是整理出來的資訊也僅供參考，並沒有一定正確的資訊，還是要呼籲人們不能一昧的聽信結果，有時候詳細內容也有必要去親自蒐集並過目了解一下。

二、建議

本研究以 PTT 論壇網購版精華區日貨買家心得分享文章為研究對象，利用文字探勘技術找出文章中詞頻高的關鍵字，並將這些關鍵字以文字雲的方式呈現結果，讓我們可以在視覺化的圖形當中初步且快速地了解這篇文章主要想傳達的訊息，但研究結果找出的關鍵字並非都是具有其重要性或意義性，因此在未來的研究方向與建議提供下列數點建議：

1.未來可以嘗試進行多篇同類型文章的抓取，並利用文字探勘技術分析，可必較其文章多寡對關鍵字的完整性及重要性程度。

2.斷詞方面未來可以使用更好的斷詞系統或是修改程式，讓斷詞系統的效率運作更好，使斷詞系統的效果更佳，避免取得較沒有意義的字詞。

3.未來可以使用多個不同的關鍵字分析方法，以取得一篇文章的重要關鍵字，加以驗證關鍵字的重要性，以及避免漏掉未取得的重要關鍵字。

4.未來可以將文字探勘技術結合人工智慧，讓電腦不僅僅只是取得關鍵字，而是可以讓電腦更進階的直接代替人類讀懂內容繁雜的網路文章。

陸、參考文獻

1. 林名彥(2015)。應用文字探勘技術於客訴資料之研究－以台大 PPT 論壇為例。龍華科技大學資訊管理系碩士班論文。
2. 吳克洋(2016)，網路社群成員之生活型態分析－以 PTT 平台為例，東吳大學
3. 吳宜隆(2010)。建構於雲端運算之文字探勘服務系統。虎尾科技大學資訊管理研究所碩士論文。
4. 陳柏江(2014)。運用文字探勘與推薦系統之技術建置失智症患者照護導引平台。國立臺北護理健康大學資訊管理研究所碩士論文。
5. 喻欣凱(2008)。運用支援向量機與文字探勘於股價漲跌趨勢之預測。輔仁大學資訊管理學系碩士論文，新北市。
6. 柯秀奎(2005)。應用文字探勘技術於客訴留言品質及分類管理之研究。國立台北科技大學機構典藏平台
7. 楊正銘(2004)。以文字探勘技術應用於疾病分類之輔助系統-以出入院病歷摘要為例。華藝線上圖書館
8. [資料分析&機器學習]第 1.1 講:Python 懶人包 Anaconda 介紹＆安裝(https://reurl.cc/EnjDv)
9. 維基百科－Python(https://zh.wikipedia.org/wiki/Python)
10. 數位時代。台灣最有影響力的網路社群－https://www.bnext.com.tw/article/38609/bn-2016-01-29-161210-178
11. Churchill,G.A.(2007). Marketing research: Methodological foundations. Cengage Learning Thompson: London.
12. Feldman, R. & Sanger J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.*New York: Cambridge University Press.
13. Fayyad, U., Piatetsky-Shaprio,G & Smyth, P.(1996).For, Data Mining to Knowledge Discovery :An overview .In advances in Knowledge Discovery and data Mining,471-493.
14. Sullivan, A. (2001). Cultural capital and educational attainment. Sociology, 35(04),893-912.
15. Tan, A. H. (1999). Text mining: The state of the art and the challenges. InProceedings of the PAKDD 1999 Workshop on Knowledge Disocovery fromAdvanced Databases Vol. 8, 65-70).
16. Van Rossum, G., & Drake Jr, F. L. (1995). Python reference manual. Amsterdam: Centrum voor Wiskunde en Informatica.
17. Zikmund, W. G. (2009). Business research methods. 8th Edition. The Dryden Press: Harcourt Brace College Publishers