

# 國立臺北大學電機工程學系專題報告

運用論壇文章作情緒文字分析系統

組員：康富城

指導老師：林嘉淦 老師

執行期間：2014 年 9 月至 2016 年 6 月

## 1. 摘要

在較以前的時代裡，人們購買商品的評價，大多透過鄰居，人傳人傳達到整個社區，企業難以透過有效且便利的方式取得消費者對於該產品的評價，企業因此失去改善產品的機會。近年來，由於網路科技發達，使得消費者習慣更容易從各種不同管道抒發自己對於產品的看法，如撰寫在線上網站，或是透過填寫商品的回饋單，甚至是透過客服信箱進行客訴等等。於現今通訊發達，網路無所不在的環境中，消費者對於商品的資訊可以說是隨手可得。其中，在消費者的眾多評論中，有大多數的消費者會從中留下各種情緒句子，像是喜歡、討厭等形容詞。在上述的情形中，每一個句子都有隱含正面、中立、負面評價的情緒字眼。而企業為了獲取更多的商機，則想辦法收集分析這些消費者的評價後，做出商品的改良與推廣。基於我去年跟同學一起發表的”建構情緒分析系統-以車電產業為例”後，今年將整個系統變更環境做出改良，呈現更有效率、更美觀的介面。結合文字探勘與輿情分析，對分析的結果做出更精準的判斷，另外建構一個實質網站，方便使用者簡單閱讀判斷。並且利用定期更新資料的方式隨時提供當下情資。

關鍵字：評價、情緒字眼、文字探勘、輿情分析、網路輿情

## 2. 設計動機

現今 google 與百度等大企業提供了相關的圖表分析功能系統，是簡單、快速且容易操作。並且在平時開會時，常常有聽聞有關於中小企業對於蒐集網路上情報與資訊的需求，並且是要求更優的精準度，而現今網路上的系統比較難於提供這類服務。於是我將嘗試做出一個能夠提供企業想要的相關分析系統，新增情緒分析的正負評價，能夠提供字詞對於消費者是好還是壞。

## 3. 系統特色

- a. 建構一個可以有效分析論壇文章的情緒分析系統。
- b. 本次取材的文章來自台灣知名論壇 mobile01。
- c. 以 python 程式語言為基礎，使用其下 django 骨架構築可操作全端的網頁。
- d. 引用超過 10 萬篇文章與留言，分析其作者，時間，內容，正負詞語。

## 4. 程式語言的選擇

去年使用 R 語言後，今年我改用 Python 去做這次系統的開發，他們各有其中的優缺點，而我選擇 Python 語言作為系統基底的理由是：

- 學習容易，如果要借接系統的話，也容易傳接。
- 與其他語言的互動性很高，本身也有套件製作網頁系統。
- 旗下套件多，可以使用的 function 方便且容易看懂。
- Python 使用者是多於 R 的使用者的。

## 5. 文字分析邏輯(word2vec 套件)

本文使用 python 套件 word2vec 去作分詞後的相關性分析，其步驟為下：

- 給予已經分詞斷句好的訓練集，利用 python 結合套件，對訓練集做多重向量字詞分類。
- 處理完後，系統內部會做出一個霍夫曼樹做判斷依據。
- 使用者可以輸入字詞，系統將顯示相關度高的詞，並顯示其相關程度。

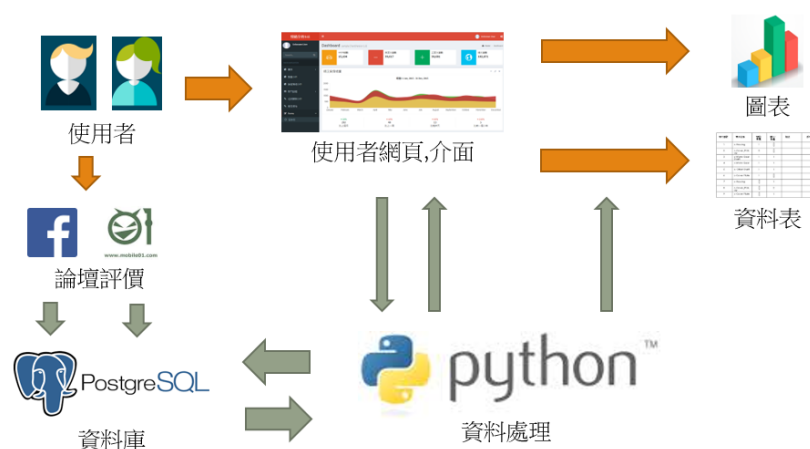
## 6. 專題進行方式

**步驟 1：** 消費者將自身使用商品經驗寫在臉書、微博、論壇等。透過搜尋引擎搜尋關鍵字，並且找出網頁連結將文章、評論等一切重要資訊儲存在資料倉儲。(另外，藉由 crontab 定時取資料更新)使用 python 套件 requests 與 BeautifulSoup4 做抓取網路論壇文章內容的動作(分成時間，作者，內容等…)，最後存入預設好的資料庫，我使用 Postgresql 資料庫作為基底。

**步驟 2：** 透過 python 套件將句子切開(斷詞斷句)，並做出情緒字詞依據和發言人、時間綁定，作為分析與繪圖的前置動作。利用套件 jieba(TF-IDF)去做文章內容的分詞斷句，用空格隔開。利用套件 word2vec 去對字詞預設給定的訓練集做訓練，作為正負詞彙分析的依據，另外，加上基本的正負情緒字典(NTUSD)再一次做分析，增加準確率。

**步驟 3：** 創建網站，使用不同程式語言做出分析結果的長條圖，曲線圖，表格等。使用 python 的 django 架構去製作網站，包括前端的網頁界面外觀，與後端的資料處理與前端介面連接。結合 chart.js 與 javascript 去做繪製統計圖形、資料表的動作，並設定可以搜尋特定時間段，與最熱門的文章。

**步驟 4：** 佈署於雲端服務平台上面，方便建置與防止駭客，並持續開放方便使用。將此網站建立到遠端微軟 azure 雲端服務平台，方便操作與維護，並且防止駭客。並且使用 apache2 做佈署的動作，讓使用者可以隨時觀看。



圖一、操作程序之流程圖

## 7. 主要成果與評估

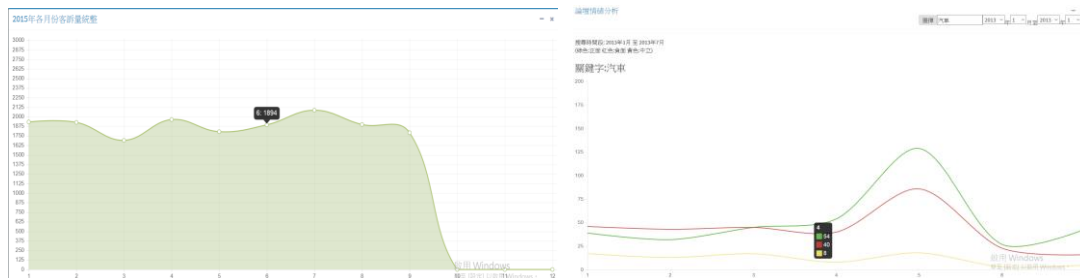


(網頁介面首頁展示)(圖二)



(會員登入系統)(圖三)

這個網頁系統有許多的功能可以使用，像是簡單的折線與曲線圖外，還有對搜尋特定日期之文章的資料表，並附上其相關的討論熱度與情緒判斷。另外不光只有對文章作分析，對於使用者在論壇上的留言次數，留言偏向，以及發表的文章數都有相對應的分析。最後也有加入保護系統的會員登入，方便管理使用者數與給予觀看的權限。



2015年各月份客訴量統整(整年每月份客訴量曲線圖)(圖四)

2015年關於“汽車”一詞分析(論壇情緒分析折線圖)(圖五)

X軸為月份，Y軸為數量。左圖為分析在2015一整年某家廠商蒐集到的客訴量分析。右圖則是分析在論壇裡有關“汽車”一詞在2015年1月~7月有關的正面、負面與中立文章量。

日期	標題	發文人	點 讚	置 頂	回覆	作者	文章數	正面情緒字數	負面情緒字數	中立情緒字數
Aug 1, 2015, 5:39 p.m.	關於Lugen的賽車引擎動靜(再出聲量)	Via888	449	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=104892	4596	8825	3647	5402
July 9, 2014, 4:57 a.m.	別讓家裏人了解你大愛什麼(1)	yidai	407	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=97543	2305	2090	2387	4485
April 21, 2014, 2:11 p.m.	仔細看看你的一向所愛重機的缺點	subwing	413	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=94440	1740	1051	761	2942
Feb 11, 2015, 5:39 p.m.	出車買車上論壇！為力當家LUGEN的發動機主們擔心！	subdigen	314	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=94440	962	806	725	1388
Jan 10, 2014, 1:03 a.m.	動力隨傳隨到 Lugen S3 Turbo L&L	盧紹工	358	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=93334	899	0	0	0
Feb 9, 2015, 8:38 a.m.	Lugen 為何能成為台灣最夯車(與眾不同)	petrop433	357	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=93767	790	761	233	574
Oct 25, 2013, 3:28 a.m.	你了解你的家裏引擎Lugen 到底是.....	jeanyc	334	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=93853	471	3827	863	1491
Nov 24, 2014, 8:28 a.m.	別再Lugen不能說好壞	waltonw	300	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=94732	692	299	389	456
Nov 21, 2013, 1:59 p.m.	LUGEN的引擎聲量與引擎動靜	亞龍子龜	299	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=93923	524	627	464	970
Jan 21, 2015, 3:59 p.m.	關於你對Lugen的動靜了解	wooooo	295	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=94840	528	520	256	763
Jan 21, 2015, 3:59 p.m.	關於你對Lugen的動靜了解	wooooo	295	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=94840	518	489	233	674
Jan 21, 2015, 3:59 p.m.	關於你對Lugen的動靜了解	wooooo	295	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=94840	510	296	180	539
Jan 21, 2015, 3:59 p.m.	關於你對Lugen的動靜了解	wooooo	295	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=94840	507	338	177	493
Jan 21, 2015, 3:59 p.m.	關於你對Lugen的動靜了解	wooooo	295	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=94840	497	438	328	649
Jan 21, 2015, 3:59 p.m.	關於你對Lugen的動靜了解	wooooo	295	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=94840	484	273	213	565
Jan 21, 2015, 3:59 p.m.	關於你對Lugen的動靜了解	wooooo	295	0	0	http://www.motorsport.com.tw/forum.php?mod=viewthread&tid=94840	427	528	138	555

2015年1月分關於Altis文章分析(文字雲+堆疊圖)(圖七)

2015年1月分關於U6文章分析(文字雲+堆疊圖)(圖八)

由左圖可以看到資料表中，可以看到依據近幾年比較熱門的討論文章，他們的討論聲量與情緒偏向判斷，給予其 URL 方便使用者可以去查看其文章討論細節；而右圖則是對留言者與作者做出分析，依照他們最多發文留言數作排序，分析其個人發表的正面、中立以及負面的文章，來做為訪問或參考的依據，幫助企業判斷。

## 8. 結語與展望

這是我的實體網站連結：

<http://caranalysis.cloudapp.net/>

從以上研究成果中，我比去年改進了系統上速度的優化，使用者可以選擇的功能更多元，前端網頁的設計與資料處理，以及學到了 python 與 R 不同的優點與缺點。可以發現到這些方法的重點都是要嘗試從大量的文字中，辨識出正面與負面等情緒，好讓這些結果產生出商業價值，這也是口碑評價對於現在公司來說非常重要的一環，公司該在消費者評價這一部份的管理層面上多出一份力。其實並未完全做到最好，還有更多改進的空間。像是前端介面設計的更優化，使用者是否可以從系統中獲得自己想要的資訊等。另外，在情緒分析上面還有許多東西要獲得解決，像是嘲諷的語句，又或者去除大量無意義，無作用的用詞上，還得去專研在機器學習的精進。我期望在大學畢業後能夠繼續精進自己的知識，包括統計與商業企業分析，以及深度學習方面。

## 9. 銘謝

感謝林嘉淦教授給我到華創公司學習的機會，並且還認識了吳柏翰學長。在教授及學長的教導下，每星期都能接觸更新的資訊，覺得自己又學到的更多的東西和知識。生長在大數據時代，嘗試處理這些資料是一件非常新鮮的事情，也在寫程式當中，曾柏諺給予我撰寫論文上非常多的幫助，才能來完成這次的專題，真的非常謝謝。

## 10. 參考文獻

- [1] Bing Liu 著，2012，” Sentiment Analysis and Opinion Mining”。
- [2] T Mikolov 等人著，2013，Distributed Representations of Words and Phrases and their Compositionality
- [3]吳柏翰學長的 GITHUB:<https://github.com/rippleblue>
- [4]我本人的 GITHUB:<https://github.com/kangfizz>