PTT 各版文字分析及鄉民愛用詞彙之研究

Extended Abstract†

胡弘林
巨資四 B, 04170217

楊文瀚
巨資四 B, 04170228

ABSTRACT

本篇文獻在研究如何在 PTT 中了解網路使用者的留言用字愛好以及看出各版之間的差異性以及相同之處，本篇以 PTT 為例 (Figure 1)，在台灣人生活中最具影響力的一個公開網站，是一個台灣電子佈告欄（BBS）採用 Talnet BBS 技術運作，且平均同時在線人數高達 8 萬多人的社群平台。
本篇文章使用 Jieba 中文斷詞技術做文字切割並使用 Ngram 及自定義辭典做斷字的正確與否檢驗。接著使用 Word2vec 詞語項量化產出相似語並加以判斷出各版之間的相似用詞，以及抓出各版常共同出現之不好詞語以便版主管理，方便其訂定相關規章。

KEYWORDS
Web Crawler, Word Segmentation, Text mining

1 INTRODUCTION

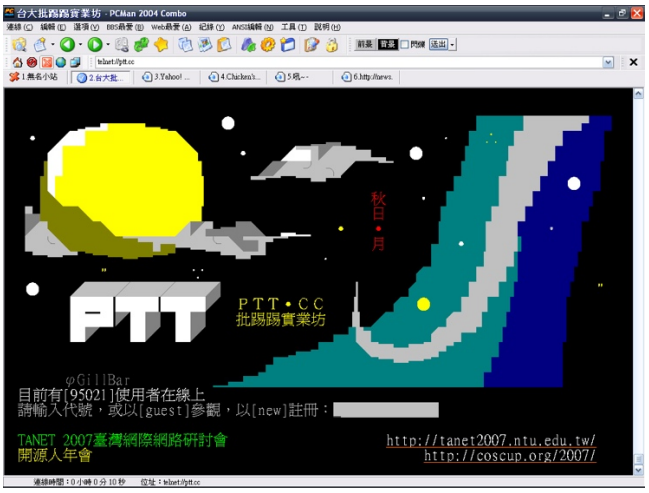


Figure 1: PTT 實業坊的實際畫面

由於近期社群網路的興起，越來越多的使用者花許多的時間在使用社群媒體的服務之中 Ex. Facebook, Instagram, Twitter 等，若以台灣為例，PTT 社群便是許多年輕人喜愛使用的社群媒體之一，而這些在 PTT 上的年輕人以鄉名自稱。PTT 中存在著許多不同的分類板塊，若是喜好政治的鄉名容易前往八卦版、政黑版，喜愛體育的鄉名容易前往籃球版、棒球版，其他還有許多具有人氣的版塊，像是女孩版、電影版、汽車板、股票版等等，因此我們想藉由文字探勘的技術來探討說 PTT 鄉名的愛用詞彙分別有哪些，以及在不同版的鄉名除了版上的專業議題外，是否用語詞彙也大致相同呢？

2 Methods

2.1 PTT 網路爬蟲

我們使用 Python 中的 requests 和 Beautifulsoup 套件進行網路爬蟲，一共抓取了四個 PTT 版塊，八卦版、棒球版、西洽版和汽車板，選取這四個版當作資料來源是因為每日在版上的人數皆有一定數量，且橫跨政治、體育、娛樂以及產業等不同領域，鄉民討論的內容擁有一定差距，因此會較符合我們想進行研究的題目。此外每個版所爬取的資料為作者、標題、時間、內文、推文數量、箭頭數量、噓文數量以及留言(Figure 2)，每個版各自抓取了約 10000 篇文章，共 40000 篇文章，此外爬取時有做了簡單的篩選，當留言數至少大於 20 個人的我們才進行爬取，因此可以有效忽略那些沒有意義 (鄉民稱之廢文) 的文章。

2.2 使用 Jieba 進行中文斷字斷詞

在進行文字的分析之前，首先第一步要做的是斷字斷詞的處理，目前處理中文語言的 Python 套件不外乎就是 Jieba、Stanford CoreNLP 以及中研院的 CKIP 模組，在這裡我

	作者	標題	時間	內文	推文	箭頭	噓文	回文
0	WindAragon (亞拉岡)	[新聞] 超跑再撞！全台限量的海神瑪莎拉蒂 北宜	Tue Dec 18 16:30:36 2018	\n\n原文連結：https://udn.com/news/story/7320/35438...	30	11	5	: 護欄家屬表示：他是孝子：沒害到其他孝子 了不起 給推：超跑家屬：一定是超跑帶壞...
1	simon87410 ()	[Trash] 國產70萬級距唯一全景天窗，Luxgen U5	Tue Dec 18 12:39:52 2018	\n國產70萬級距唯一全景天窗，Luxgen U5「2019年式」全新上市！\n\n新聞來源...	54	26	20	: 我先：XDDDD：30萬我考慮買：讚！多了一個地方要修理！：。。。:...
2	Scape (缺鈣缺很大)	[影片] Lexus 內鬥! LFA vs LC500	Tue Dec 18 12:59:35 2018	\nhttps://youtu.be/4t5a7cPtWac\n\n說是內鬥其實也鬥不起來，...	24	34	3	: LFA還是厲害不少：韓哥開lfa在德國 帥爆：買L牌就是開帥 不開快啊：L...
3	simon87410 ()	[超跑] 貧富差距擴大賠不起 民間發起制定「超跑條款	Tue Dec 18 14:26:12 2018	\n貧富差距擴大賠不起 民間發起制定「超跑條款」\n\n2018-12-18 11:02經濟...	38	57	14	: 超額險那麼便宜也不保.....：這什麼鬼啊：什麼叫看到名車就要讓三公...
4	ChrisDavis (工業電風扇)	Re: [超跑] 貧富差距擴大賠不起 民間發起制定「超跑條款	Tue Dec 18 14:46:01 2018	\n	43	108	4	: 千萬超跑紙糊的，貨車可是有大樑，不同：差在哪，裝上去如果貨物毀了不用賠是不是：...

Figure 2: 汽車版爬取資料樣態示意圖

們選用 Jieba 套件進行斷字斷詞，原因為其方便使用，且容易透過自定義的字典新增原始無法斷開的字詞，此外我們也透過 Python NLTK 的套件去除中文的停止詞 Stopwords，而不論是斷字斷詞的字典亦或是停止詞字典我們皆使用網路上公開有人整理過的字典資料，以下附上其來源。

https://raw.githubusercontent.com/fxsjy/jieba/master/extra_dict/dict.txt.big

https://github.com/chdd/weibo/blob/master/stopwords/中文停用词库.txt

2.3 使用 Ngram 檢查斷字斷詞

為了解決 Jieba 斷字斷詞不精確的問題，因此我們簡易的使用 Ngram 的模型幫助我們篩選有無出現頻率很高，但卻沒有被 Jieba 套件成功斷出的字詞。Ngram 的意思是斷完詞後找尋前後同時出現的字詞，再重新計算加總，因為若是沒有斷出的字詞一定也會出現在該字詞的前後 Ex. ('安全', '距離')，因此我們查看前後兩個字詞的出現頻率 (Bigram) 以及前後三個字詞的出現頻率 (Trigram)，在進行人工檢查，最後將應該是字詞的組合加回 Jieba 自訂義的字典裡面。(Figure 3)

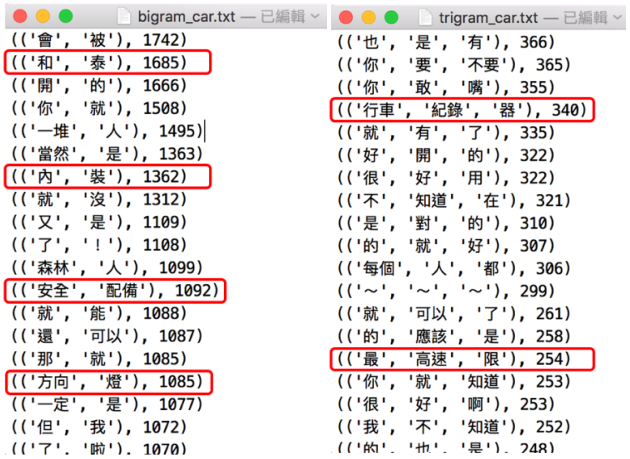


Figure 3: 汽車版 Bigram 結果 (左)，汽車版 Trigram 結果 (右)

2.4 加入自定義字典再次斷字斷詞

從 2.3 Ngram 的結果中我們進行了人工的校正，查看一起出現的次數至少大於 100 次的文字，在建立自定義的字典供 Jieba 重新斷字，附圖為重新加入 Jieba 的字典 (Figure 4 圖片擷取部分內容)。

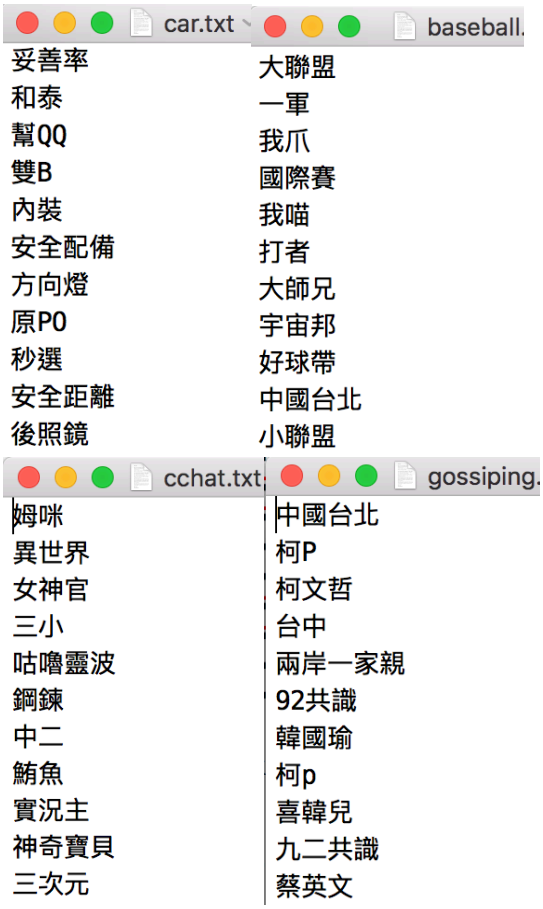
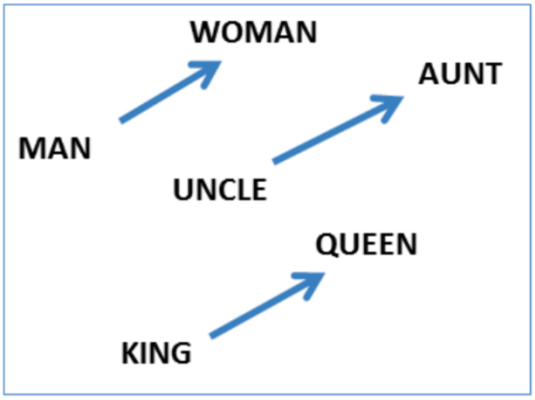


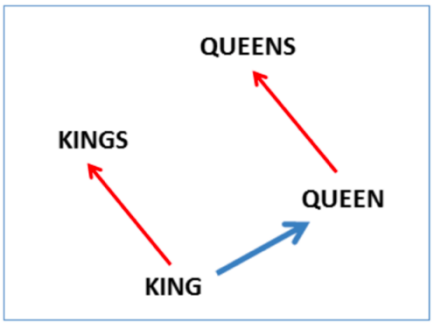
Figure 4: 汽車版 (左上)，西洽版 (左下)，棒球版 (右上)，八卦版 (右下) 其中一部分自定義字典

2.5 訓練 Word2Vec 模型

Word2Vec 是一種詞向量的類神經網路模型，利用向量的概念將所有訓練的字詞投射到高維度空間當中，在利用輸入的字詞找尋相鄰位置的字詞。如下圖，代表不同身份的男生計算相同距離時分別也能對應到同樣身份的女生，而此圖也是類似的概念在 word2vec 中實際計算的是餘弦值 cosine，距離範圍會介於 0~1 之間，也就是 cos 值越大兩個詞的關聯度會越



高。而我們使用 Python 的 gensim 套件訓練我們的 word2vec



模型，訓練資料為資料中的回文欄位，而不同版的資料各自會產生一個模型，因此最後總共產生出四個 word2vec 模型 (汽車版回文、西洽版回文、棒球版回文、八卦版回文)。

3 RESULTS

我們使用訓練好的四個模型分別針對不同的字詞做相似詞的搜尋，每個輸入詞我們設定會產生前 40 個最鄰近的字詞，同時搜尋的結果以文字雲的方式呈現，而文字雲的文字大小權重使用的是 word2vec 產生的 cos 值，以下是我們使用幾種字詞輸入模型產生的結果 (Figure 5~11)，左上圖為汽車版，左下圖為棒球版，右上圖為西洽版，右下圖為八卦版

Figure 5: Input word 為"選舉"

Figure6: Input word 為"國際"

Figure7: Input word 為"約"

Figure8: Input word 為"燈"

Figure9: Input word 為"狗"

Figure10: Input word 為"台女"

Figure11: Input word 為"廢文"

由上面列出之搜尋的相鄰字詞的途中，我們可以發現出現相較之下不正向的詞彙會出現交集或者是跟政治相關的文字會也會出項交集，反倒是一些正向或無關景耀的詞語出現的交集的可能性反而不高。也看出台灣鄉民普遍有愛用負面字詞傾向並且在女性字眼中也存在很大的共識。

4 CONCLUSIONS

本篇論文我們選用 PTT 的八卦版、棒球版、西洽版和汽車板因為這四大版每日有大量的文章更新及留言。我們抓取了各版高留言數的文章各 10000 篇的留言並作分析。在技術方面我們發現在做中文詞彙的斷詞時，可以多方參考使用不同的方法讓斷詞結果更加完善，若有相關的語料庫亦會對斷詞結果有更大的幫助。而我們使用到的方式是先做 Jeiba 後接著再使用 Ngram 中的 Bigram 以及 Trigram 去觀察斷完結果出現在前後的字詞的出現頻率，得出之結果後在人工觀察出是否有斷詞不夠準確地部分，並再次做斷詞段字，並新建屬於這篇論文的語料庫來提升斷詞的準確率，我們也發現確實有實際的提升，最後匯入 Stopwords 的語料庫去除停止詞。

我們訓練了 Word2Vec 模型後繪出文字雲以便看出各個文字在各版中的結果並在文章中舉了一些例子，我們觀察出在有仇視的言語上得出了四版中有類似的詞語出現。在政治相關的文字上也得出了相同的結果。但在一些比較通俗常用的詞語中，各個版專有詞語上結果出現分歧，可能是因為我們抓取的版像是汽車版跟棒球版本身就難以存在共通點，除非是當今熱名話題。E.g 韓總，柯市長，柯 P 都會找到類似的文字，或者也可能是在於各版的人之間並不一定存在著許多跨版的使用者。

未來規劃，我們未來希望能更加精確的分析出各版留言的趨勢以及預測推噓文的可能性並做到不止四大板而是各版的留言的走向。最主要的目的在於分析鄉民的留言語句，若是能提早預測留言中吵架的事發生，或是鄉民集體鬧版的狀況，我們就可以推薦給各個版主做諮詢以及各版規的參考依據，讓版主們更加清楚各版的留言狀況，甚至是推薦給 PTT 站長了解各個版的共同狀況為何。