Text Mining - Term Paper

Twitter 中英文情感分析之研究：以現任台灣總統與香港時事為例

610721204／陳克威／2019/06/24

GitHub Repository URL:
https://github.com/D1034181036/TextMining_TermPaper

1. Twitter API, Twitter Packages 簡介

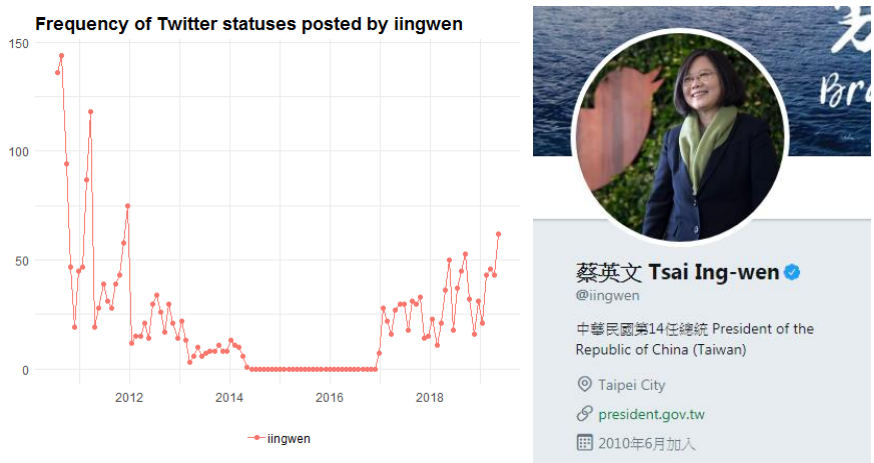
(1) twitteR： 需要申請 Twitter Developer API 後才能使用，申請過程需提供詳細資料，如學生版本必須提供學校名稱、課程名稱、教師姓名、分析方式與使用範圍等詳細資料。

(2) rtweet： 僅需要 Twitter 帳號即可使用，不需申請 Twitter Developer API，本次實驗多數使用此套件。

2. 蔡英文 Twitter 之中英文情感分析

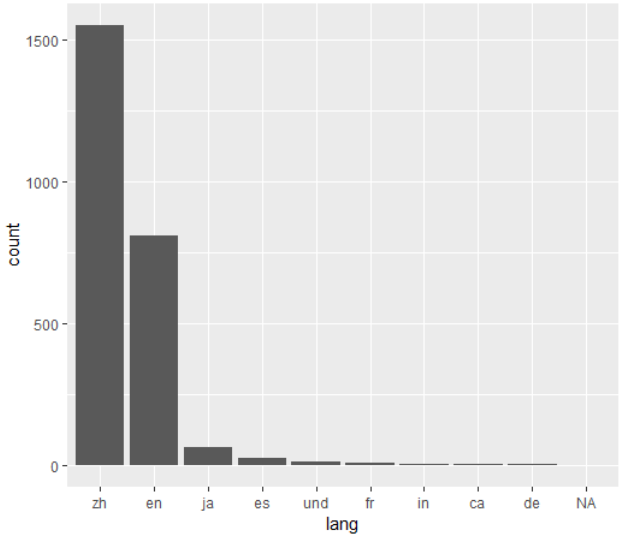
(1)蔡英文 Twitter 資料集

首先使用 rtweet 中的 get_timelines()取得 iingwen(蔡英文)的貼文(post)，總共有 2469 筆貼文，我們將其以時間軸畫出，如圖(一)所示，從圖中可以發現此帳號大約從 2010 開始經營，在 2014 年初停止發文，而約在 2017 又重新開始經營。

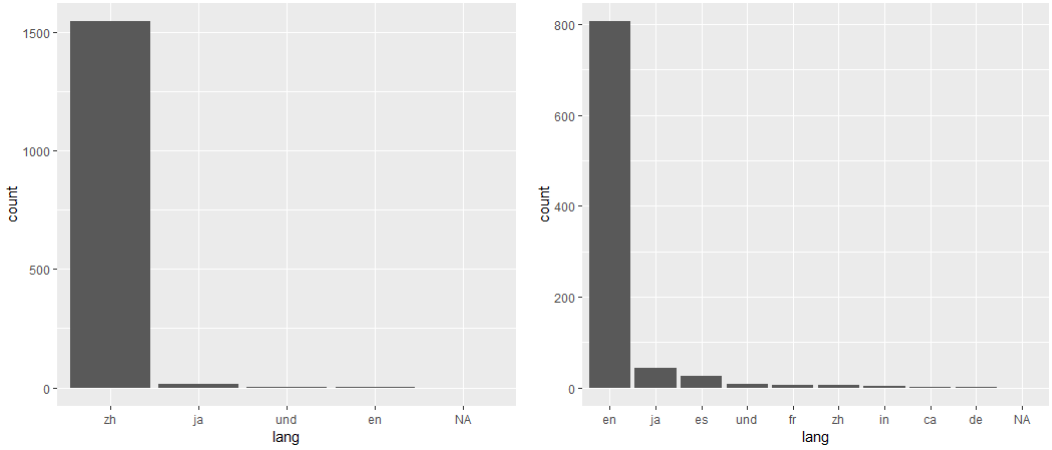


圖(一)　推特帳號 iingwen 貼文(post)時間軸。

接著使用 language 欄位來篩選貼文的語言，從圖(二)可以看出貼文以中文(zh)與英文(en)為主。我們試著將貼文從時間點 2016-01-01 做切割，再以語言做分類排序，如圖(三)所示，可以發現在 2010 至 2014 主要以中文貼文為主，而 2017 以後則多為英文貼文，因此我們將資料以 2016-01-01 做切割，分別進行中文與英文之情感分析。



圖(二)　推特帳號 iingwen 貼文語言類別。



圖(三)　推特帳號 iingwen 貼文語言類別，2016 前(左圖)、2016 後(右圖) 。

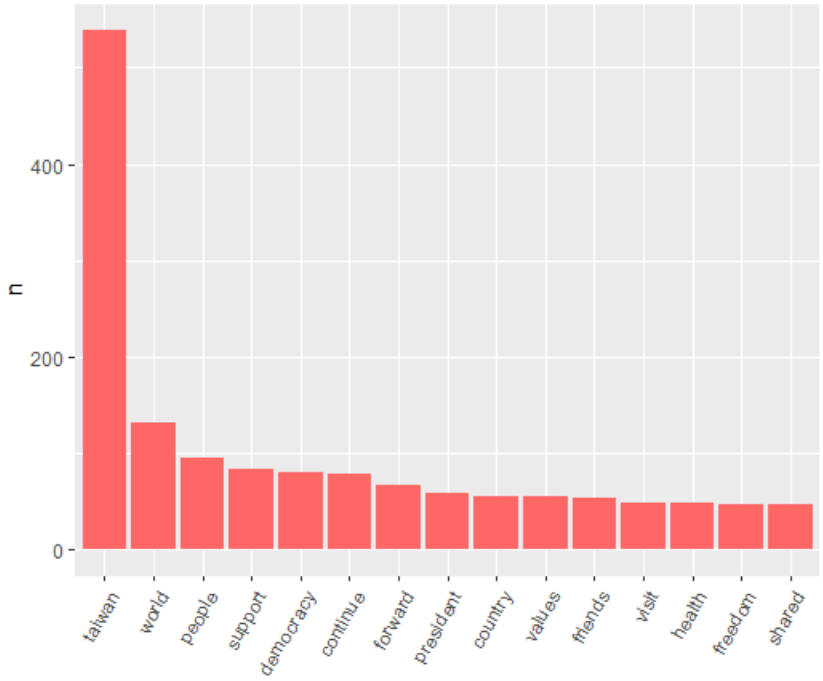
(2) 蔡英文 Twitter 英文貼文前處理

我們取出 2016-01-01 以後的貼文，將其中為英文分類的貼文取出，總共有 900 筆貼文，接著進行以下前處理步驟：

1. 使用 tidy text 將其進行斷詞，並且將非英文的 term 去除。

2. 將停詞與網址去除，如"to", "the", "https"等字詞。

3. 將 term 依照出現次數做排序並畫出，如圖(四)所示，總共有 3379 種不同的英文字詞，貼文中常出現"Taiwan", support", " democracy"等字詞。



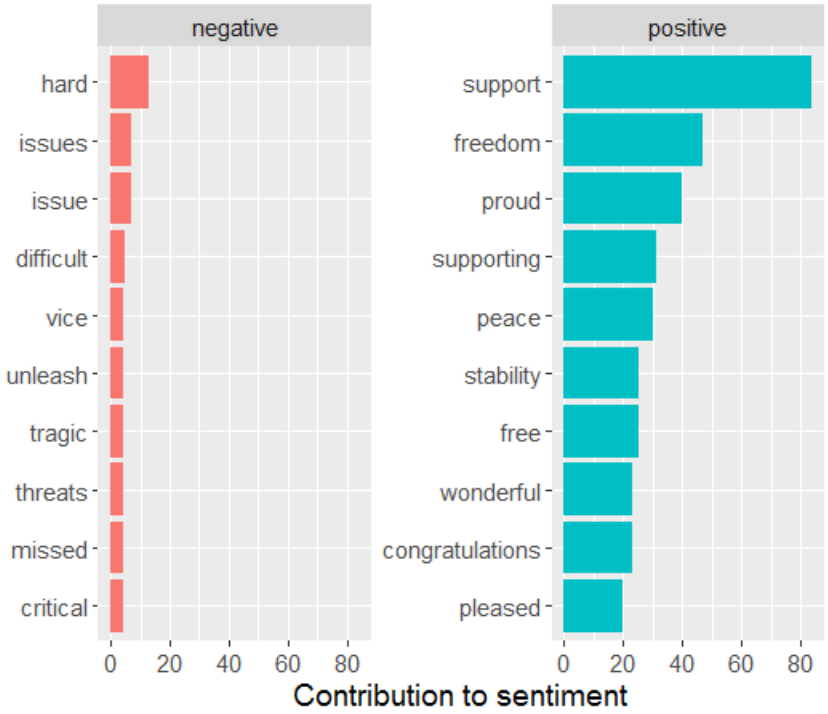
圖(四)　推特帳號 iingwen 英文貼文字詞頻率

(3) 蔡英文 Twitter 英文貼文情感分析，其步驟如下：

1. 使用 tidy text 中的 get_sentiments 取得 bing 情緒字典。

2. 使用 inner_join 將 term 與情緒字典進行比對，其中 positive 的詞共出現 1350 次，而 negative 的詞則出現 210 次。

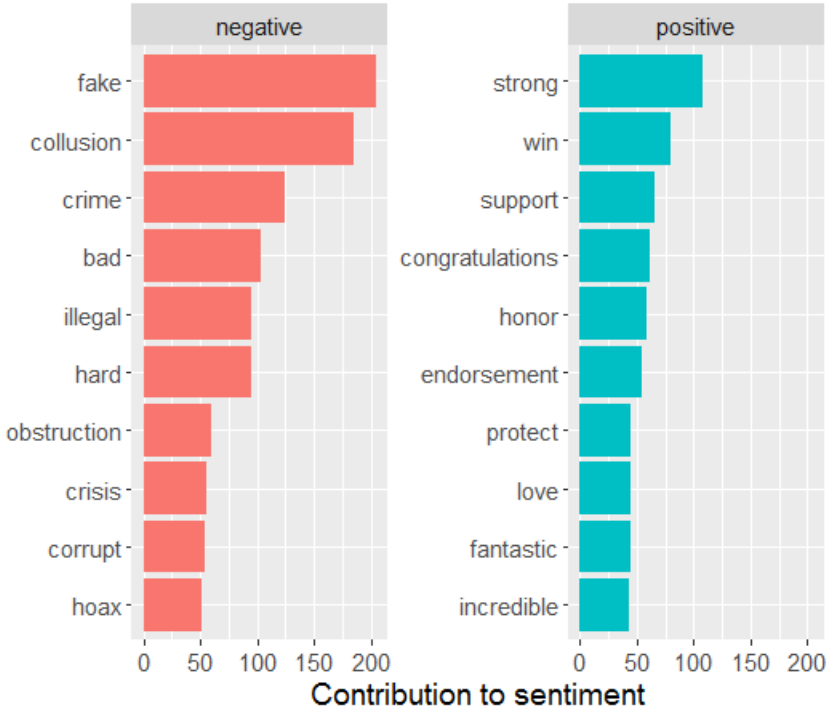
3. 根據貼文正負向詞分別依照頻率列出，如圖(五)所示，可以看出正向詞常常強調的"support", "freedom", "free"等詞，而負向詞多為"hard", "issues"等詞。



圖(五)　推特帳號 iingwen 英文貼文正負向詞頻率

4. 我們在這邊以美國總統川普的貼文進行情緒詞的比較，將川普近 3200 則貼文取出，同樣經過斷詞、停詞、比對情緒字典等動作，將其常出現的正負向詞列出，如圖(六)所示，可以發現川普常使用"fake", "collusion", "crime"等詞，甚至負向詞的頻率多於正向詞，我們將兩位總統的貼文做情緒詞數量的比較，如表(一)所示，可以發現兩者貼文用詞的差異非常大。

	Iingwen(蔡英文)	realDonaldTrump(川普)
Positive	1350(86%)	2780(42%)
Negative	210(14%)	3849(58%)



圖(六)　推特帳號 realDonaldTrump 英文貼文正負向詞頻率

(4) 蔡英文 Twitter 中文貼文前處理

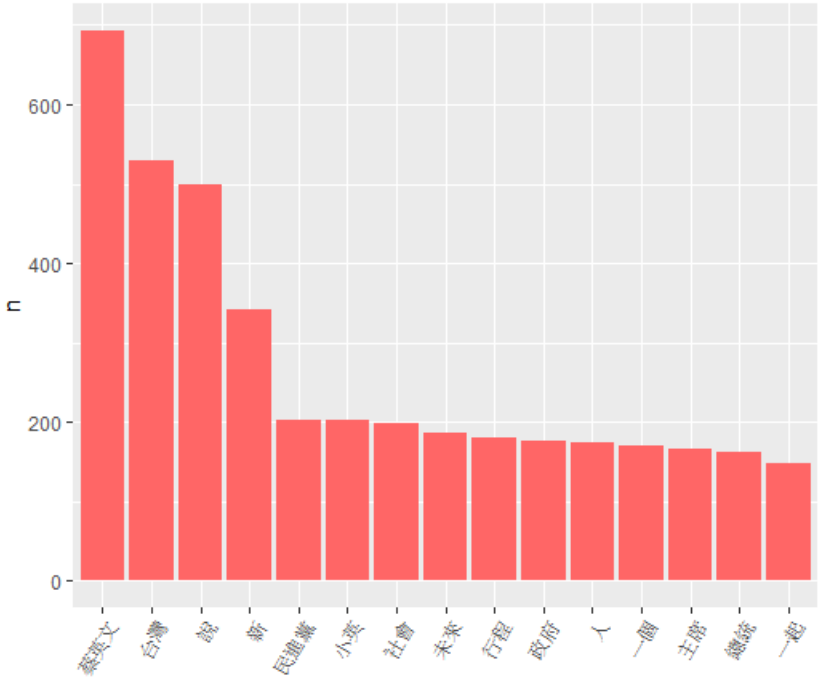
我們將 2016-01-01 後的貼文取出，並且將中文分類的貼文取出，總共有 1569 筆文章，接著進行以下前處理步驟：

1. 首先使用 tmcn 將編碼轉成 UTF-8 並且將所有中文皆轉為繁體，其目的為避免將同一個詞區分成簡體與繁體兩種。

2. 使用 gsub 將英文字、數字去除，僅留下中文的部分。

3. 使用 jiebaR 進行斷詞與去除停詞。

4. 同樣我們將 term 依照頻率做排序並畫出，如圖(七)所示，總共出現 9034 種不同的字詞，文章中常使用如"蔡英文", "台灣", "民進黨"等字詞。



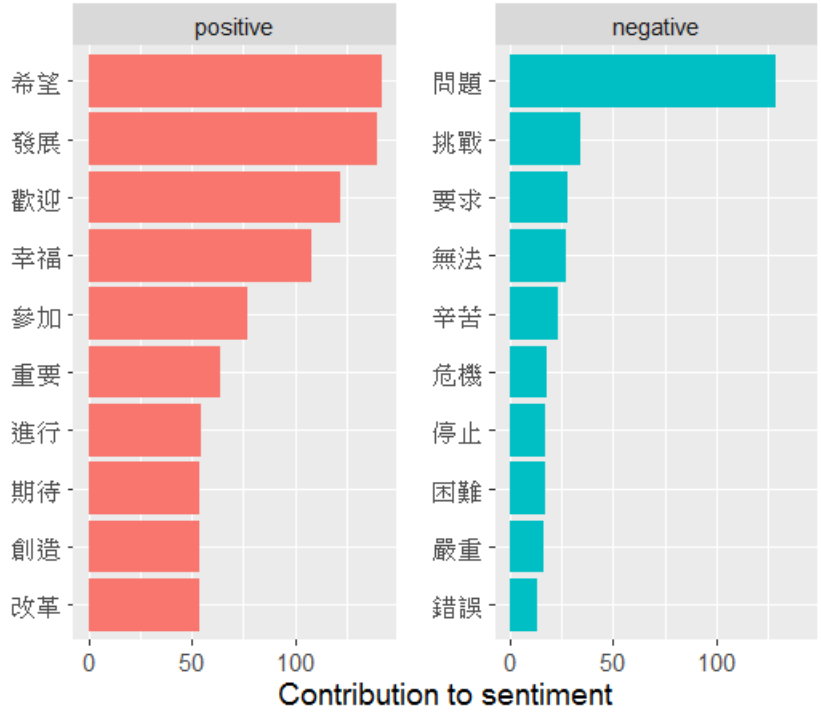
圖(七)　推特帳號 iingwen 中文貼文字詞頻率

(5) 蔡英文 Twitter 中文貼文情感分析，其步驟如下：

1. 我們使用 ntusd 的正負向字典建立中文情緒字典。

2. 使用 inner_join 將 term 與字典進行比對，positive 的詞共出現 3430 次，而 negative 的詞出現 1319 次，可以發現與英文的情感分析相近，正向詞的使用的次數較負向詞多。

3. 我們同樣將其正負向分別列出來，如圖(八)所示，可以觀察出正向詞常出現"希望"、"發展"、"歡迎"等字，而負向詞則出現"問題"，"挑戰"，"要求"等字，這些負向字在部分情況下可能並不完全代表負向，如"問題"可能為"解決問題"、"挑戰"可能為"面對挑戰"，若未來要更進一步分析，可以透過自訂字典來更準確的對應到的不同情況下的情緒字詞。

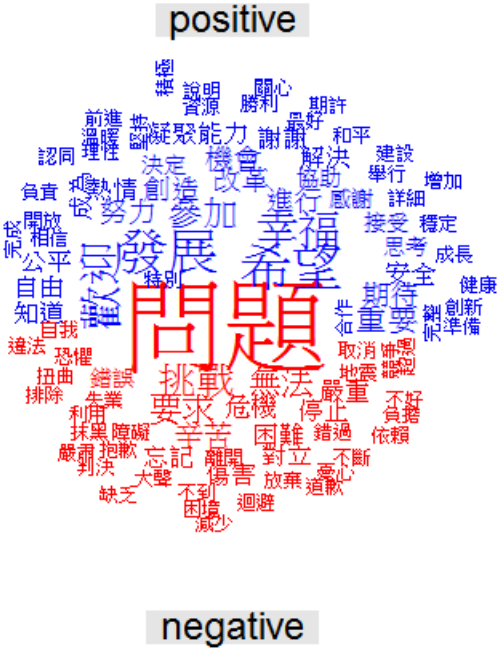


圖(八)　推特帳號 iingwen 中文貼文正負向詞頻率

(6)以文字雲呈現常用字詞頻與常用正負向字詞

圖(九)　推特帳號 iingwen 中文貼文字詞文字雲

圖(十)　推特帳號 iingwen 英文貼文正負向字詞文字雲

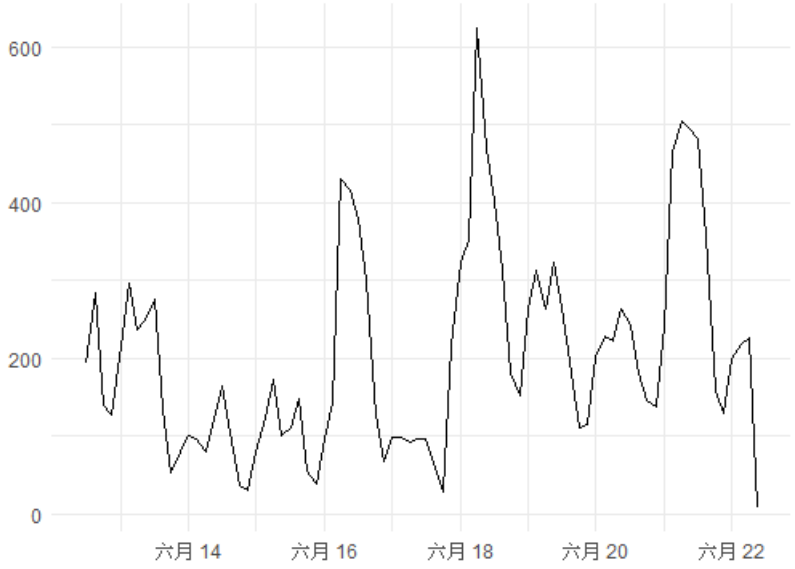


圖(十一)　推特帳號 iingwen 中文貼文正負向字詞文字雲

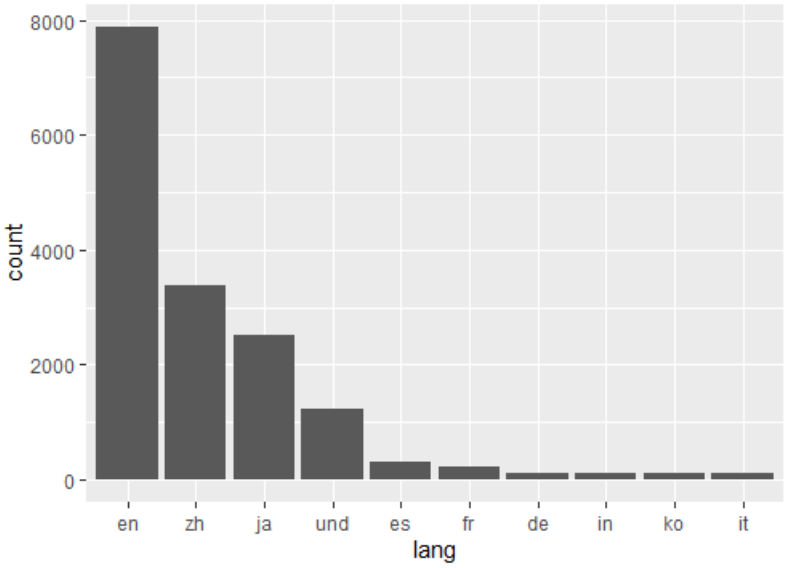
3. 香港時事(反送中)之中英文情感分析

因其分析方法與上一節相同，唯有資料集不同，因此這邊不重複說明處理步驟，僅提及資料集的取得方式。

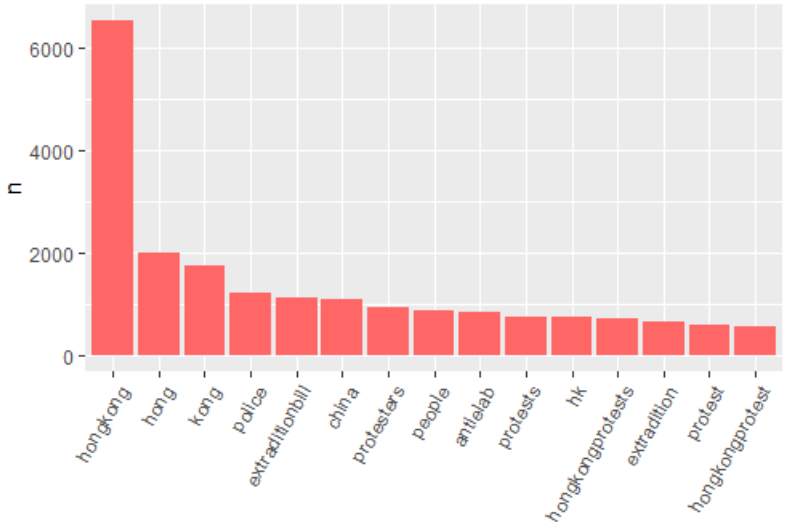
使用 rtweet 中的 search_tweets()取得#HongKong 與#反送中的推文(tweets)，總共取得 16347 筆推文，我們將其以時間軸畫出，如圖(十二)所示，可以發現近期討論度非常高。



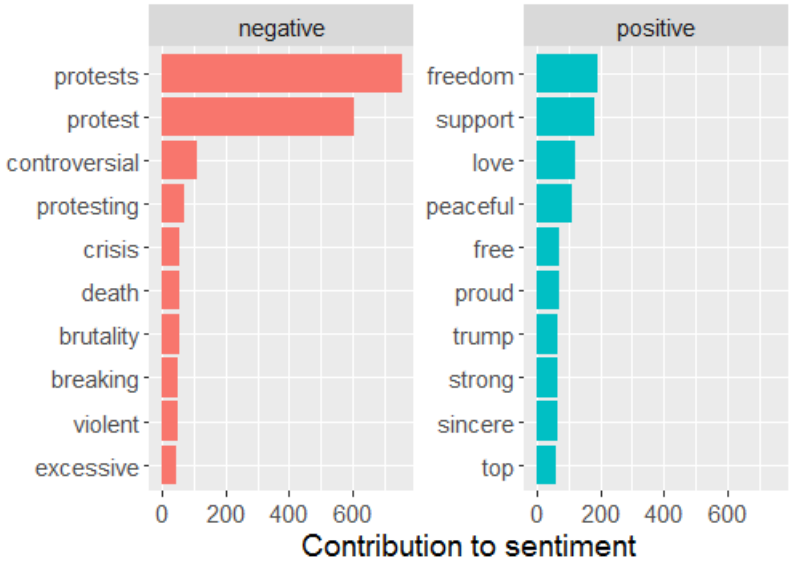
圖(十二)　#HongKong 與#反送中推文時間軸。



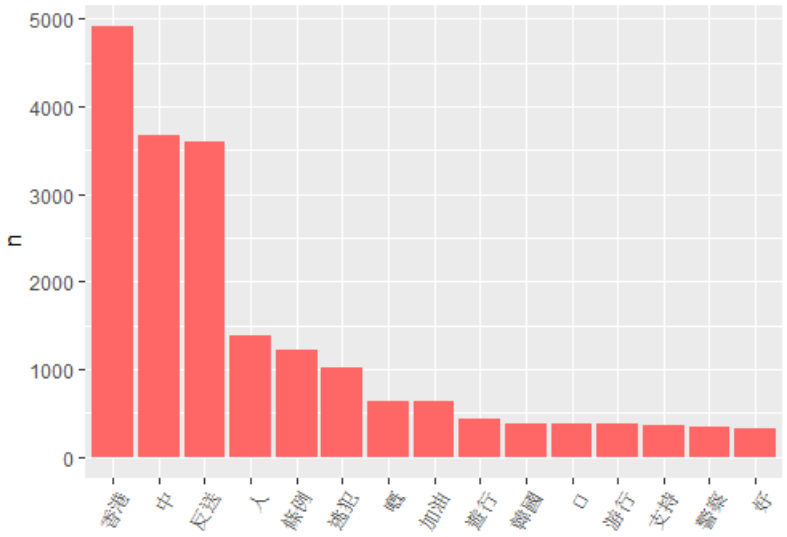
圖(十三)　#HongKong 與#反送中推文語言類別。



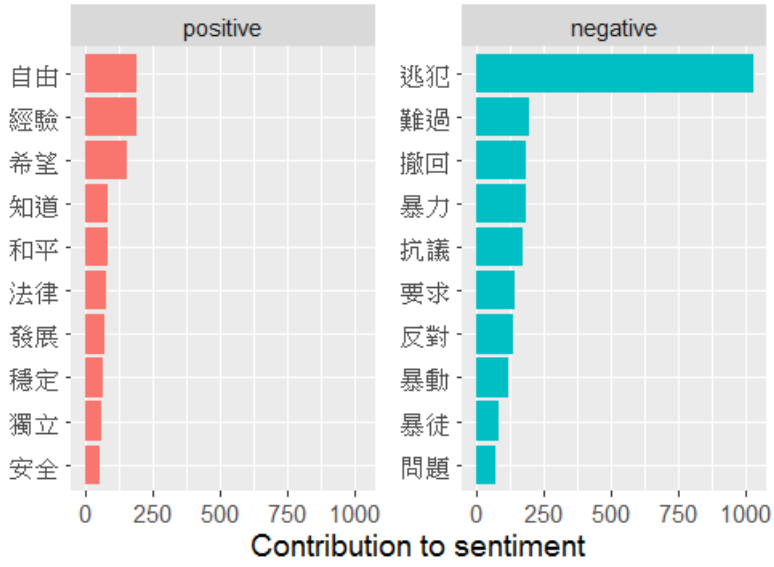
圖(十四)　#HongKong 與#反送中英文推文字詞頻率



圖(十五)　#HongKong 與#反送中英文推文正負向詞頻率



圖(十六)　#HongKong 與#反送中中文推文字詞頻率



圖(十七)　#HongKong 與#反送中中文推文正負向詞頻率

圖(十八)　#HongKong 與#反送中中文推文字詞文字雲

圖(十九)　#HongKong 與#反送中英文推文正負向字詞文字雲

圖(二十)　#HongKong 與#反送中中文貼文正負向字詞文字雲

4. 參考資料

1. David Robinson - Text analysis of Trump's tweets confirms he writes only the (angrier) Android half
http://varianceexplained.org/r/trump-tweets/ (last access:2019/06/24)

2. Kearney MW (2018). *rtweet: Collecting Twitter Data*. R package version 0.6.7, https://cran.r-project.org/package=rtweet. (last access:2019/06/24)

3. Scraping Twitter data and using it in R
http://utstat.toronto.edu/~nathan/teaching/sta4002/Class1/scrapingtwitterinR-NT.html (last access:2019/06/24)

4. 椰子笑 - R 学习整理笔记（五）——用 jiebaR 包进行中文分词
https://zhuanlan.zhihu.com/p/35846130 (last access:2019/06/24)

5. 劉育銘 - 中文資料使用不同情緒字典的情緒分析
https://rpubs.com/dcw102213006/chinese_senti (last access:2019/06/24)