語言所「R語言與資料科學導論」

PTT 八卦版留言之情緒分析

徐榆婷 林蕎安 黃欒雅 劉育政 林家妤

The world's most valuable resource is no longer oil, but data.

分析背景

R 語言是一種能用來做統計和資料分析的語言，此外也能進行網路爬蟲。爬蟲的核心任務可以簡單分為兩個：請求資料（requesting data）與解析資料（parsing data）請求資料的運作就像我們在瀏覽器中輸入網址一般，只不過送出請求的管道由瀏覽器改變成為 R 語言程式碼；解析資料的運作則是將伺服器回傳的資料內容去蕪存菁，萃取必要的一小部分。

分析簡介

我們想利用 R 的網頁爬蟲技巧，尋找在 PPT 八卦版上留言的情緒有什麼趨勢，比如哪幾天大家的情緒有什麼起伏變動或留言的頻率高低有什麼走向，再加入中文情緒詞表分類正負情緒詞，藉此了解大家在留言時的情緒變動與日期有什麼關連性。

Coding

載入套件

require(dplyr)

→處理資料 (dataframe)

require(tidytext)

→文字探勘處理

require(jiebaR)

→進行斷詞

require(stringr)

→字串處理

require(tidyr)

→資料處理與分析

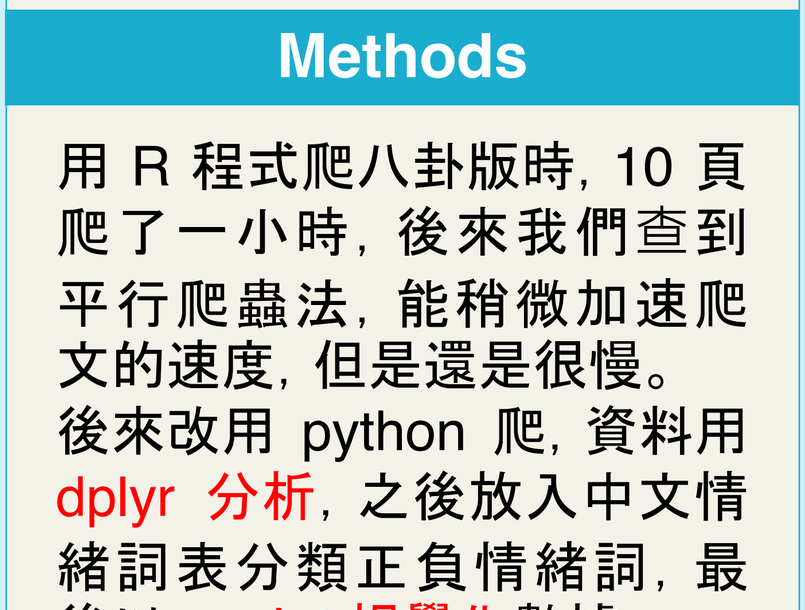
require(ggplot2)

→繪製討論數量時序圖

Methods

用 R 程式爬八卦版時，10 頁爬了一小時，後來我們查到平行爬蟲法，能稍微加速爬文的速度，但是還是很慢。後來改用 python 爬，資料用 dplyr 分析，之後放入中文情緒詞表分類正負情緒詞，最後以 ggplot 視覺化數據。

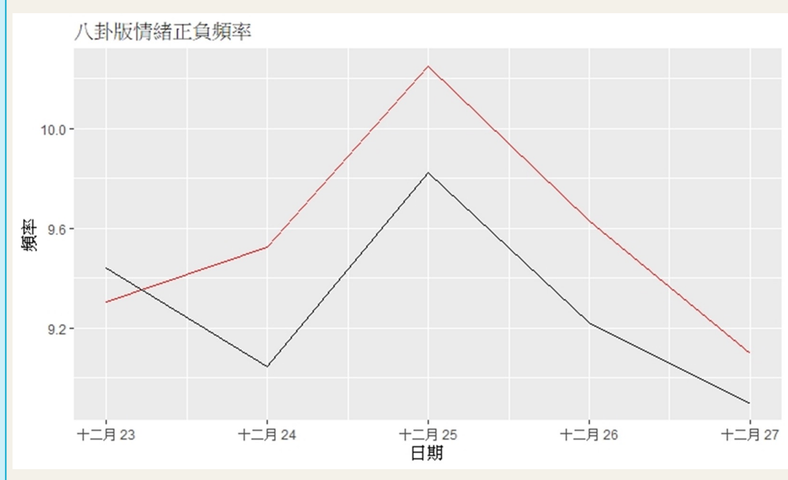
情緒詞彙分析圖



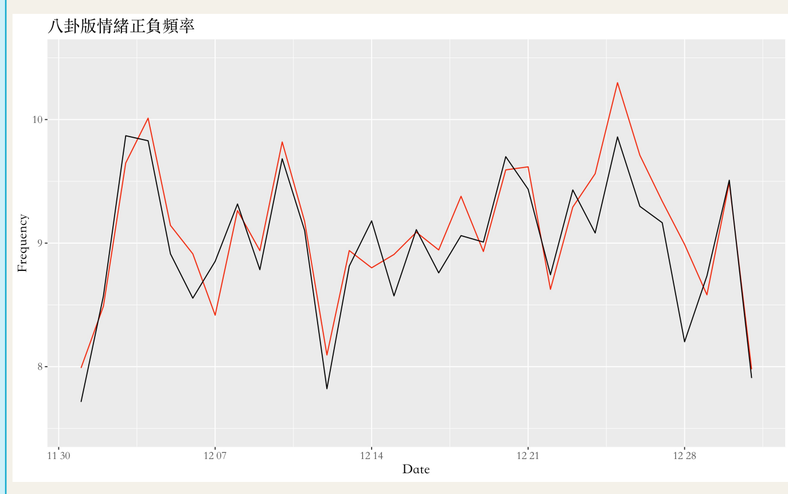
CVAW - 中文維度情感辭典
Chinese Valence-Arousal Words

- 橫軸:越右邊代表情緒越正向
- 縱軸:越上方代表情緒越激烈

Result & Discovery



從資料繪製的圖可以發現，在 12/23 到 12/27 期間，12/25 的情緒正負頻率都是最高的，代表此時的留言數很高。推測其原因為聖誕佳節大家分享喜悅或一起慶祝，正向詞頻率很高，但令人疑惑的是，負面情緒相關詞出現的頻率也很高。



為了找到更多的情緒趨勢傾向，我們進行了第二次爬蟲，爬了PTT 八卦版留言完整 12 月的情緒分析，從更長一段的時間來看，我們可以發現正負情緒頻率趨近重疊，而正向詞稍微多於負向詞，在聖誕節附近正向詞才明顯多於負向詞。

Reference

Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16), San Diego, California, USA, 12-17 June, 2016.

https://rlads2019.github.io/lecture/16/ch.senti.lex.csv

http://nlp.innobic.yzu.edu.tw/resources/cvaw.html