語言所「R語言與資料科學導論」

爬文PTT八卦版留言之情緒分析

徐榆婷 林蒨安 黃欒雅 劉育致 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。而爬蟲的核心任務可以簡單區分為兩個：請求資料（requesting data）與解析資料（parsing data）；其中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PTT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

```
#讀入json資料  jlist <- read_json("gossip.json")
#取出前面欄位
jdataframe <- fromJSON("gossip.json") %>% select(日期)
#讀入情緒辭典 library(readr) emotion <- read_csv("emotion.csv")
#把negative值改成-1
emotion$kind[emotion$kind %in% 0] <- -1
#建立斷詞 seg <- worker()
#f1與f2用來取出jlist中的推文內容
f1 <- function(x){
  x2 <- x$推文
  sapply(x2,f2)}
f2 <- function(c){
  c2 <- c$留言內容
  segged <- segment(c2,seg)}
#各貼文留言list(已斷詞)
comlist <- sapply(jlist,f1) %>% sapply(.,unlist)
#f3與f4用來取出comlist中符合emotion裡的詞語
f3 <- function(x){
  c <- emotion$kind[emotion$WORD %in% x]
  c2 <- sum(c==1)}
f4 <- function(x){
  c <- emotion$kind[emotion$WORD %in% x]
  c2 <- sum(c==-1)}
#各貼文的留言情緒正負向次數，存在向量中
pos <- sapply(comlist,f3)
neg <- sapply(comlist,f4)
#加到jdataframe中
jdataframe <- jdataframe %>% mutate(positive = pos,negative = neg)
# 修正日期，變為年/月/日格式
d <- jdataframe$日期
mon <- str_extract(d,"(?<=\\s)[a-zA-Z]*(?=\\s{1,2}[0-9])") %>% match
(.,month.abb)
date <- str_extract(d,"(?<=\\s{1,2})[0-9]{1,2}(?=\\s)")
time <- paste(mon,date, sep="-")
jdataframe$日期 <- as.Date(time, format = "%m-%e")
#summary要的資料(日期、pos頻、neg頻)，篩掉不要的日期
summar <- jdataframe %>%
  group_by(日期) %>%
  summarise(fre_pos = sum(positive)/n(), fre_neg = sum(negative)/n())
#作圖
ggplot(summar[1:31,])+
  geom_line(aes(x = 日期,y =fre_pos),color = "red")+
  geom_line(aes(x = 日期, y =fre_neg), color = "black")+
  labs(x="Date", y="Frequency",
       title="八卦版情緒正負頻率")+
  xlim(summar$日期[1], summar$日期[31])+
  ylim(7.5, 10.5)+
  theme(text=element_text(family="宋體-繁 標準體", size=14))
```

載入套件

require(jsonlite)
→ JSON 格式資料處理

require(dplyr)
→處理資料框(dataframe)

require(jiebaR)
→進行斷詞

require(stringr)
→字串處理

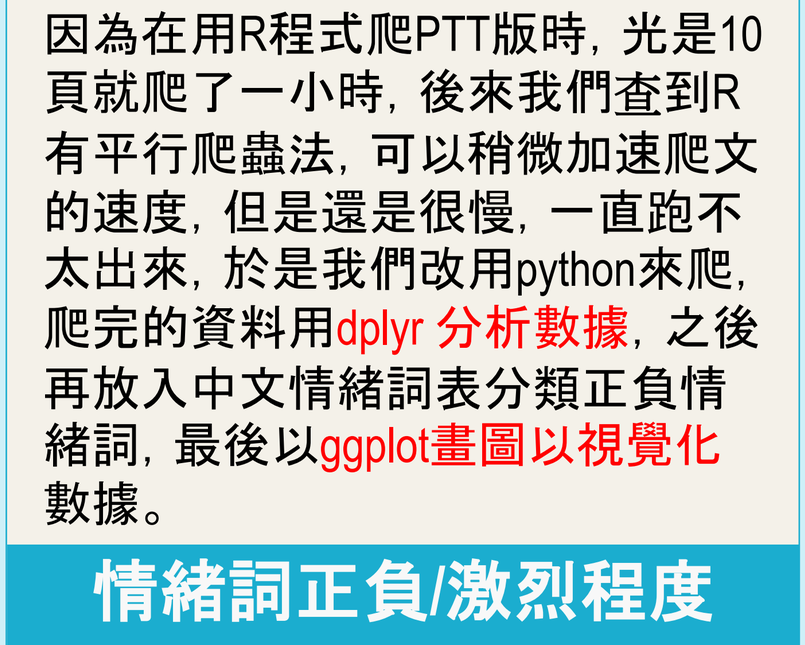
require(tidyr)
→資料合併與分離、變長或變寬的小技巧

require(ggplot2)
→繪製討論數量時序圖

Methods and R function we use

因為在用R程式爬PTT版時，光是10頁就爬了一小時，後來我們查到R有平行爬蟲法，可以稍微加速爬文的速度，但是還是很慢，一直跑不太出來，於是我們改用python來爬，爬完的資料用dplyr 分析數據，之後再放入中文情緒詞表分類正負情緒詞，最後以ggplot畫圖以視覺化數據。

情緒詞正負/激烈程度分析圖

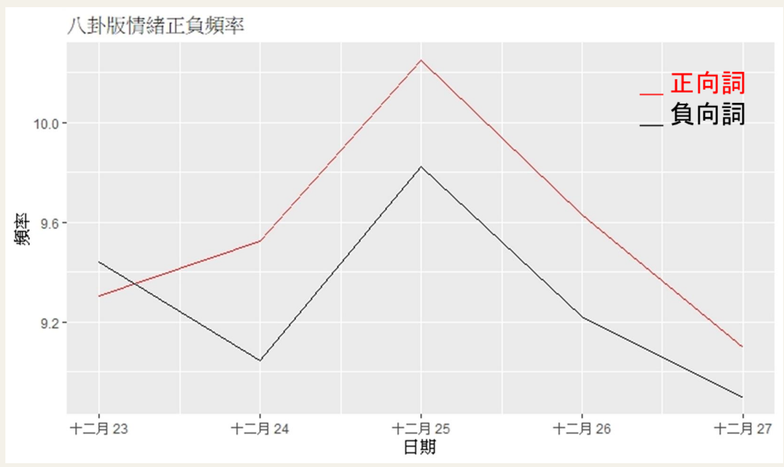


CVAW- 中文維度情感辭典

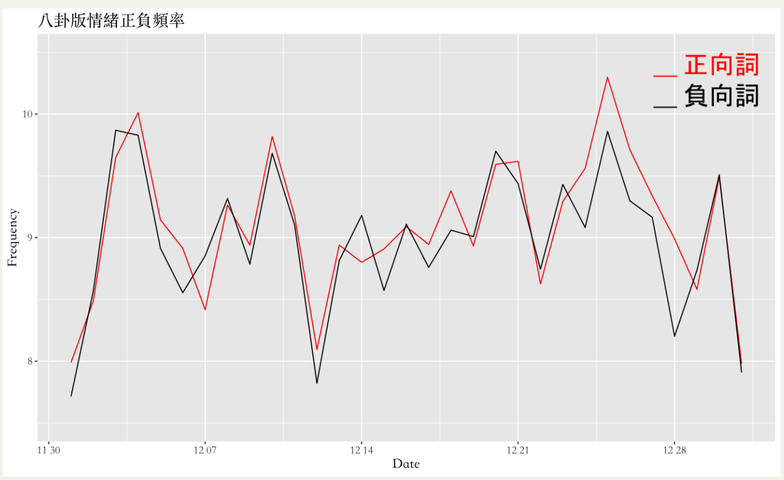
Chinese Valence-Arousal Words (CVAW)

- 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，我們可以推測其原因是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

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

Poster template by ResearchPosters.co.za