PTT 八卦版的發文情緒檢視

以 PO 文種類、頻率進行分析

組員：

陳星丞, 鄧禮頡, 張震奕, 丁子翔

2021-06-20

目錄:

簡介:

將 PTT 的文字抓下來再進行整理，對文字進行分析，再對文章種類進行分類，記錄其頻率。最後把三種變因進行視覺化與統計分析。

方法:

一、 抓下 ptt 八卦版的發文標題:

(套件使用:requests、bs4)

Code:

```
### 資料爬取
```python=
#載入套件
import requests,re,threading
from bs4 import BeautifulSoup as bs
#設定cookie
cookies={
    "__cf_bm":"b061f4d11494eb98c01e49950891fa97dc746683-1620981046-1800-AauhlnO0M9nCanP6vCyG9DD1VZlcWfe1fBc7L0xv+BO6F/ISc3qfjbHCfZCXvkGUiVTFZhJEjHSYoLmDYsSy9vk",
    "_ga":"GA1.2.197183732.1620494134",
    "_gid":"GA1.2.986367917.1620886017",
    'over18':1,
    "__cfduid":"dda0debaa78515d636afe69ca1066d3871618801296"
}
```

先設定 cookie 以免無法瀏覽網站

```
r = requests.Session()
payload ={
    "from":"/bbs/Gossiping/index.html",
    "yes":"yes"
}
f=open("ptt_data_mult",'w')
temp="0/0"
r = requests.Session()
backup=[]
r1 = r.post("https://www.ptt.cc/ask/over18?from=%2Fbbs%2FGossiping%2Findex.html",payload)
```

進入 PTT 網頁時有 18 歲認證，為此必須要有這段 Code 進入網頁

```
def process(i):
    place="https://www.ptt.cc/bbs/Gossiping/index"+str(i)+".html"
    text=r.get(place)
    content=bs(text.text,features='html.parser')
    titles=content.find_all('div',attrs={'class':'title'})
    date=content.find_all('div',attrs={'class','date'})
    print(i)
    try:
        for k in range(len(titles)):
            """
            if str(date[k].text) == ' 1/01' and str(temp) == ' 12/31':
                j = j + 1
            temp = str(date[k].text)
            """
            lock.acquire()
            #以csv格式輸出
            f.writelines(str(i)+","+'"'+re.sub('"',"",re.sub(",","/",str(re.sub("\n","",titles[k].text))))+'",'+str(re.sub(" ","",date[k].text)+"\n"))
            lock.release()
    except:
        print("errrrrrrrow")
```

依照該網站的 html 碼進行搜索，再以 csv 檔匯出
```

二、 進行資料篩選與分類:

a.(套件使用:tidyr、dplyr)

Code:

```
        thread[i].join()
```

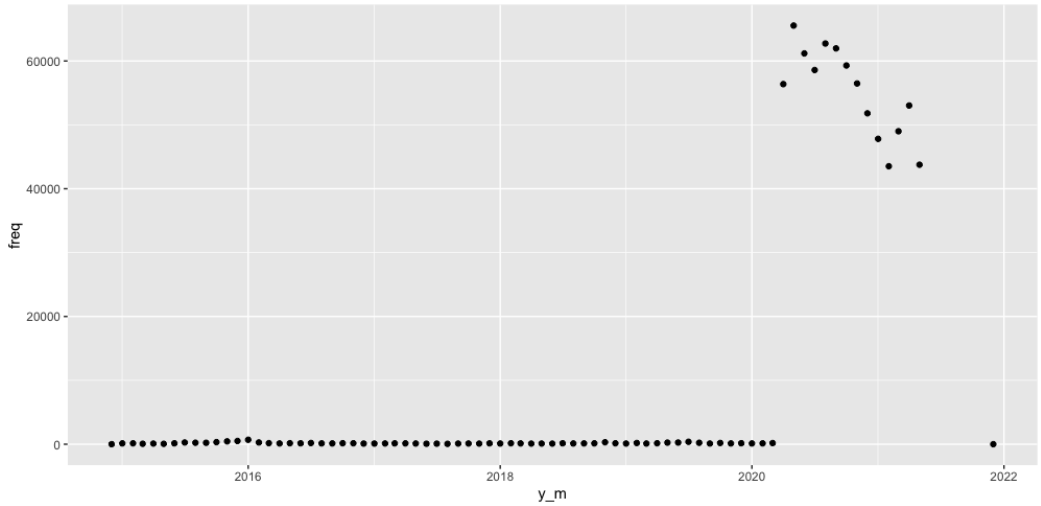
資料篩選

```r
library(tidyr)
library(dplyr)
library(ggplot2)
library(utils)
library(lubridate)
ptt_2014to2021_year<-read_csv("2014to2021_year.csv")

ptt_2014to2021_year%>%filter(cata %in% c("問卦","新聞","爆卦"))%>%group_by(year(time))%>%summarise(n())->tt
mutate(ptt_2014to2021_year,year=year(time),month=month(time))->tt
mutate(tt,ym=paste(year,month,"1",sep="-"))->tt
tt%>%group_by(ym)%>%summarise(freq=n())->tt
```

匯入套件，對爬下來的資料進行分組。

對時間的格式也進行修改

再來以視覺作圖對資料的頻率進行檢視:



作圖完後，發現以頻率去製作的圖形極端的在 2020-03 月之後，因此開始篩選 2020-03~2021-05 之間的資料

```r
select(ptt_2014to2021_year,time,cata,title)->dt
filter(dt,cata %in% c("問卦","爆卦","新聞"))->dt
filter(dt,as.Date(time)>as.Date("2020-03-31") & as.Date(time)<as.Date("2021-06-01"))->dt
dt%>%group_by(time,cata)%>%summarise(titles=paste(title,collapse = " "),freq=n())->ptt_2020to2021_daily
write.csv(ptt_2020to2021_daily,'./ptt_20202021_daily.csv')
```

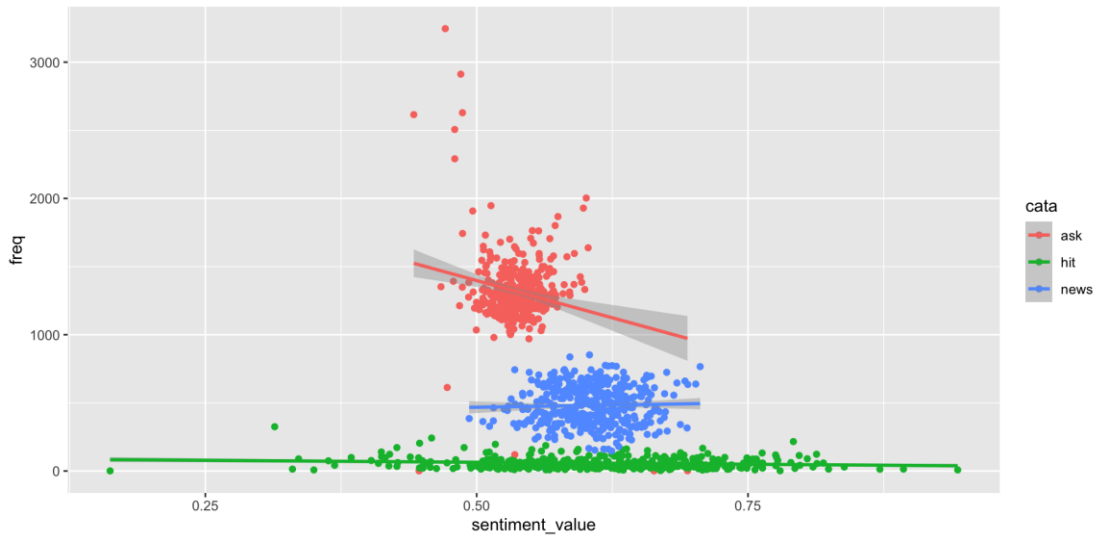
對資料進行清理與時間的文字格式一致性

三、 進行情感正面程度分析:

（套件使用:snowlp）

Code:

```
### 進行情感正面程度分析
```bash=
pip install snownlp #下載snownlp
```
```python=
import csv
from snownlp import SnowNLP,sentiment

csvfile=open('ptt_20202021_daily.csv', newline='')
rows=csv.DictReader(csvfile)
data=[]
for i in rows:
     data.append(i)

x=0
for i in data:
     x=x+1
     lis=i['titles'].split()
     soc=0
     for j in lis:
          snow=SnowNLP(j)
          soc=soc+snow.sentiments
     i['sentiment_value']=(soc/len(lis))
     print(x/len(data))
     print((soc/len(lis)))
with open('sentiment_ptt.csv',"w",newline='') as csvfile:
     tname=['','time','cata','titles','freq','sentiment_value']
     writer = csv.DictWriter(csvfile, fieldnames=tname)
     writer.writeheader()
     for i in data:
          writer.writerow(i)
```
```

情感分析後將匯出的值放入其中一欄再將 data 匯出。

結果:

一、作圖:

-情緒正面程度與發文頻率

Code:

```
### 對情緒正面程度與發文數做分析
```r=
ggplot(data=sentiment_ptt)+
geom_point(mapping = aes(x=sentiment_value,y=freq,color=cata))+
geom_smooth(method = "lm",mapping = aes(x=sentiment_value,y=freq,color=cata))
```
```

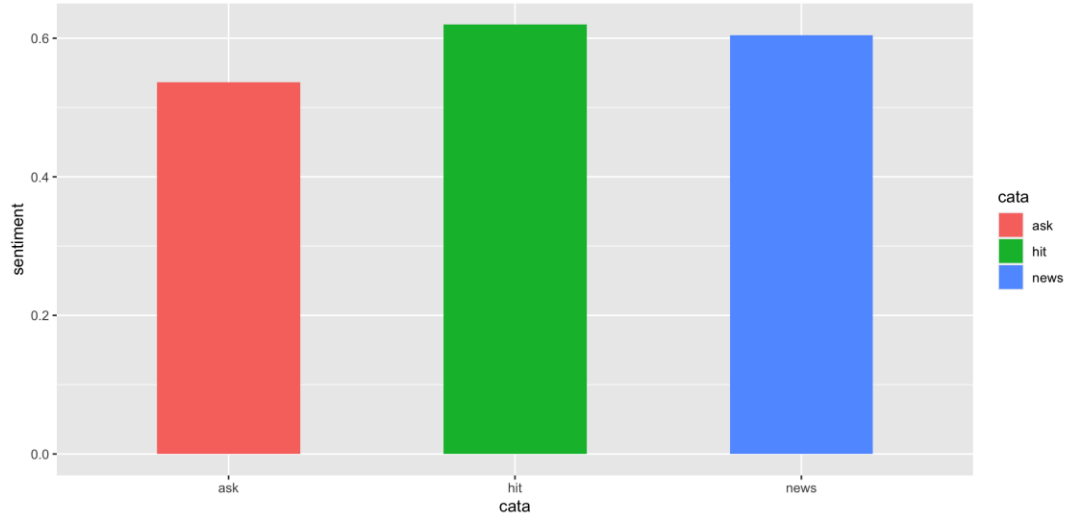


作圖分析:1.唯有問卦類別有明顯的線性關係，負相關 2.爆掛的情緒展幅最大 3.問卦的 PO 文數較高

-各類別平均情緒

Code:

```r
### 各類別的平均情緒
sentiment_ptt%>%group_by(cata)%>%summarise(sentiment=mean(sentiment_value))->cata_sent
ggplot(data=cata_sent,aes(cata,sentiment,fill=cata))+
geom_bar(stat='identity',width = 0.5)
```

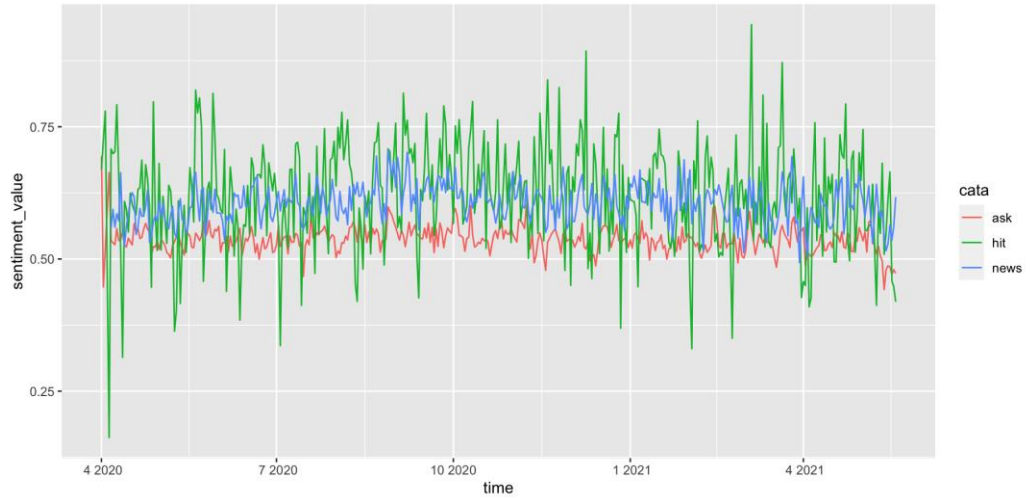


作圖分析:報卦和新聞的情緒正向程度都較高且相近

-時間與情緒正面程度

Code:

```r
### 對時間與情緒正面程度做分析
ggplot(data=sentiment_ptt,mapping = aes(x=time,y=sentiment_value,color=cata))+
geom_line()
```



作圖分析:1.爆卦以時間分析情緒起伏最大 2.情緒起伏高低和 PO 文分類無關

二、以 po 文種類和情緒程度進行分析

-樣本數變異數分析:

Code:

```
157  bartlett.test(sentiment_ptt$sentiment_value~sentiment_p
     tt$cata)
```

```
        Bartlett test of homogeneity of variances

data:  sentiment_ptt$sentiment_value by sentiment_ptt$cata
Bartlett's K-squared = 903, df = 2, p-value < 2.2e-16
```

分析判斷:1.p-value<0.05,拒絕 H0,各組變異數有顯著差異,此測驗決定能否進行 F 檢定

-樣本數平均數分析:

Code:

```
161  Ft<-
     aov(sentiment_ptt$sentiment_value~sentiment_ptt$cata,se
     ntiment_ptt)
162  summary(Ft)
```

```
                    Df Sum Sq Mean Sq F value Pr(>F)
sentiment_ptt$cata    2 10.314   5.157    1405 <2e-16 ***
Residuals          1230  4.515   0.004
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

分析判斷:1. p-value<0.05,拒絕 H0,各組變異數有顯著差異,平均數不相同

-事後分析:

Code:

```
166  pairwise.t.test(sentiment_ptt$sentiment_value,sentiment
     _ptt$cata, p.adjust.method="bonferroni")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  sentiment_ptt$sentiment_value and sentiment_ptt$cata

     ask    hit
hit  <2e-16 -
news <2e-16 1

P value adjustment method: bonferroni
```

分析判斷:1. 依前檢驗行事後分析, alpha 值經過調整, 各變數以雙獨立變數之分析, 測試結果:只有爆卦與新聞的情緒程度平均值有顯著差異

三、頻率與情緒正面程度的數據分析:

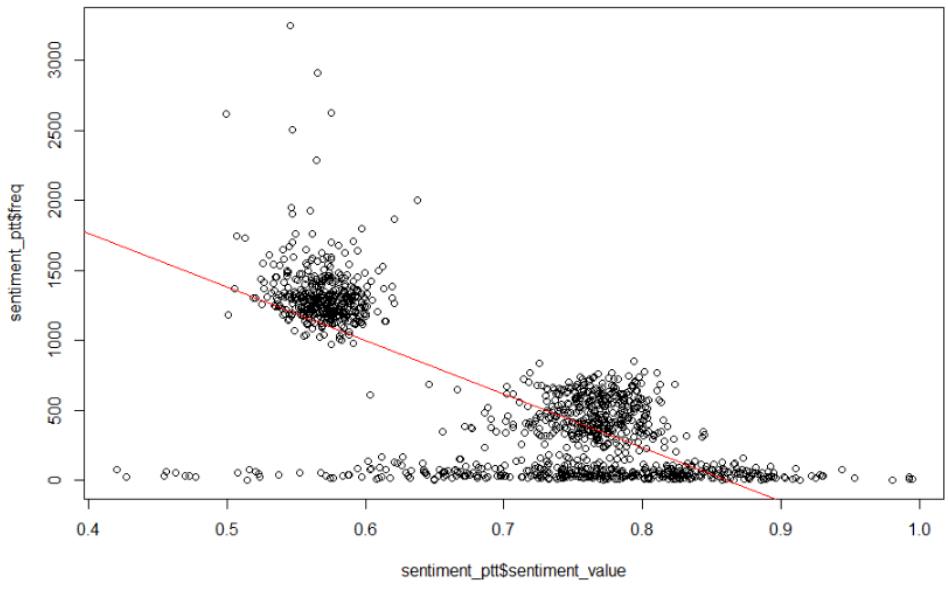
-頻率與情緒正面程度的相關係數:（不依時間, 類型分類）

```
192 cor(sentiment_ptt$sentiment_value,sentiment_ptt$freq)
193 cor.test(sentiment_ptt$sentiment_value,sentiment_ptt$fr
    eq)
```

```
194 #Pearson's product-moment
195 #correlation
196 #data:  sentiment_ptt$sentiment_value and
    sentiment_ptt$freq
197 #t = -40.9, df = 1231,
198 #p-value < 2.2e-16
199 #alternative hypothesis: true correlation is not equal
    to 0
200 #95 percent confidence interval:
201 #-0.7817021 -0.7342845
202 #sample estimates:
203 #       cor
204 #-0.7589979
```

作圖: 圖型:趨勢線，資料視覺化

```
208 plot(sentiment_ptt$sentiment_value,sentiment_ptt$freq)
209 abline(lm(sentiment_ptt$freq~sentiment_ptt$sentiment_va
    lue), col = 'red')
```



先進行線性回歸分析，未進行資料統整時，頻率和情緒進行相關

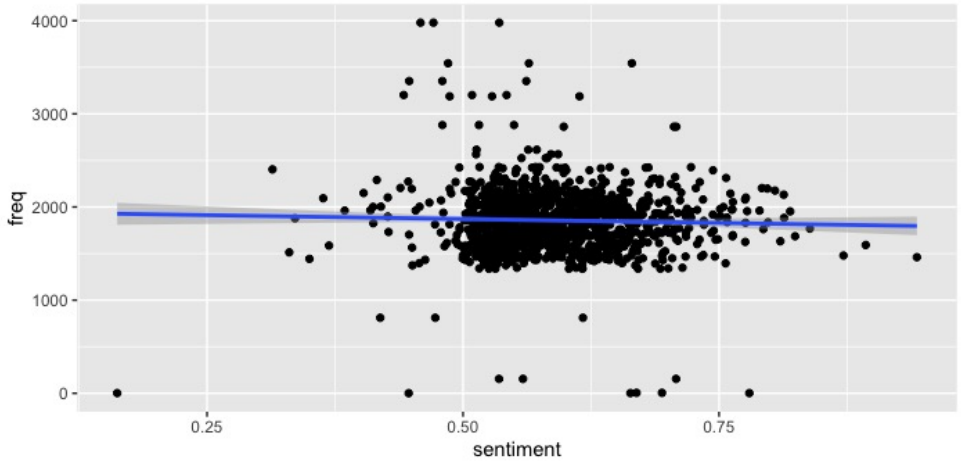
係數的假說檢定。

分析結果:p-value<0.05，拒絕虛無假說，故頻率和情緒有線性關係

而其相關係數的樣本估計值:-0.759

-合併類別進行分析

```r
### 合併類別進行分析
entiment_ptt->ttt
ttt$cata<-NULL
head(ttt)
ttt%>%select(time,freq,sentiment_value)%>%group_by(time)%>%summarise(freq=sum(freq),sentiment=(freq*sentiment_value)/(sum(freq)))->nocata_value
ggplot(data=nocata_value,aes(sentiment,freq))+
  geom_point()+
  geom_smooth(method = "lm")
```




```r
cor.test(nocata_value$freq,nocata_value$sentiment)
```

```
#Pearson's product-moment
#   correlation
#data:  nocata_value$freq and
nocata_value$sentiment
#t = -1.4788, df = 1231,
#p-value = 0.1395
#alternative hypothesis: true correlation is
not equal to 0
#95 percent confidence interval:
# -0.09770684  0.01374965
#sample estimates:
#        cor
#-0.04210961
```

再以時間進行分類作加權平均數後，發現相關係數明顯和原資料有差別。p-value:0.1395>0.05 不拒絕相關係數為 0 的可能

-以簡單線性回歸分析頻率與情緒之關聯:

```
model_1=lm(freq~sentiment,data = nocata_value)
summary(model_1)
```

```
#output:
#lm(formula = freq ~ sentiment, data = nocata_value)
#
#Residuals:
#     Min       1Q   Median       3Q      Max
#-1925.35  -224.07    10.25   177.83  2114.04
#
#Coefficients:
#            Estimate Std. Error t value Pr(>|t|)
#(Intercept)   1954.5       82.7  23.634   <2e-16 ***
#sentiment     -167.2      139.8  -1.196    0.232
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
#Residual standard error: 357.4 on 1231 degrees of freedom
#Multiple R-squared:  0.00116,  Adjusted R-squared:  0.0003487
#F-statistic:  1.43 on 1 and 1231 DF,  p-value: 0.232
```

決定係數:0.00116，而上述分析方式為情緒能否有效解釋頻率，而其中有效解釋的佔比數即是 0.116%，推論出兩組的關聯性不高。

四、上述分析的結論:

-在情緒正面程度與發文量上並無明顯的相關:

這個結論來自於最後的線性回歸分析，而線性回歸的假說檢定目標即是兩者是否有因果關係，在做完 R 內部的分析後做出來的值極低，也顯示了頻率和情緒關聯性低

-在以爆卦、問卦及新聞的分類中的情緒正面程度高低

此結論來自於作圖中的直方圖，而經由 anova 的檢驗後，發現與直方圖結果相符，另外報卦與問卦的樣本平均值差異不大，測驗的結果也顯示其母體平均值是可能相等的(因所得資訊非完整的，僅有 2020-03~2021-05 期間的資訊)

-爆卦情緒正面程度差異大

推想主要是以爆卦為發文的主幹時，相較會以主觀方式述寫

討論:

一、數據分析中兩次相關係數分析差距大:

先分析兩圖差異，合併類別中將原先的分類合併，圖上每一點為一天。而另一張圖則是每篇文即一個點。推論原因為，以每天作為分類會下降極端值對於分析的影響，但也能看到不同天對於這項分析的影響不大

二、為何爆卦類項的 PO 文差異性大卻又相對正向呢?

這個問題似乎能從作圖中的點陣圖來解答，而答案也顯示了這次作圖的缺點，1.爆卦的分文頻率低於其餘兩項導致了極端值大大的影響平均值 2.爆卦的變異性大，將其分類出來再進行情緒分析的值的比較價值較低，更是顯示了分組方式的紕漏

工作分配表:

社會一 陳星丞:蒐集資料+分析
公衛一 鄧禮頡:整理內容+PPT 報告
公衛一 張震奕:分析+海報+書面
公衛一 丁子翔:海報