第一屆金股獎 - PTT股版績效回測及分析

好想吃午餐

2021/6/21

一、研究動機

- 2020年歷經股市還有台積電的大漲，許多菜籃族以及學生迫不及待的想要進入股市撈錢，但我們想要了解，大部分的普通人進入股市真的會賺錢嗎？難道網路上的梗圖都是假的？所以我們希望透過台灣最大交流論壇 - PTT的stock版去模擬大家的投資情況，並分析文本內容與報酬之間是否有關聯。

二、實踐方法

1. 爬蟲

- 我們使用python爬取PTT股版自2018.01.01-2021.06.01分類為[標的]的文章，包含標題、作者、URL、日期、推數、噓數、留言數、內文以及圖片網址。

    1. 首先利用python的BeautifulSoup從PTT的標題頁面爬取標題以及網址：
        - 詳見 title&url.py
    2. 從網址爬取作者、URL、日期、推數、噓數、留言數、內文以及圖片網址：
        - 詳見 pttstock.py

- 爬下來的文件分別為 2018stock.csv、2019stock.csv、2020stock.csv、2021stock.csv

2. 資料清洗

- 再來我們利用python、excel、R語言同時做資料的清理，將資料以以下的標準輸出為統一格式：
    - 增加標的(股票代號.csv檔相互對照標的名稱)、分類、分析/正文、進退場機制四個欄位
    - 清除雜字
    - 依據分類判斷多空，依據關鍵字判斷長短，優先判斷長線，剩餘則為短線
    - 長線判斷標準：關鍵字含長線、長期投資、長期、不停利、不停損、長投、放（但不是“放空”）
    - 短線判斷標準：關鍵字含短線、（有“停利”“停損”但沒有“不停利”“不停損”）、放空、退場、出場

- 清洗完的文件分別為finalclean2018.csv、finalclean2019.csv、finalclean2020.csv、finalclean2021.csv：
    - R部分：cleanr.rmd
    - Python部分：cleanpy.py
    - 統整：unify.rmd

三、資料分析與發現

最後我們將乾淨的資料做資料分析，分述如下：

名詞解釋

- 年化報酬率：因不同時間長短的報酬率無法互相比較，所以將投資期間都變成以「一年」為單位計算報酬率
- 累積報酬率：投資期間的總報酬率，不考慮資金投入時間的長短效益
- 敏感度分析：透過更改不確定的變量以測試模型的穩健性
- 大盤：大盤一般泛指在股票市場的主要指數，以台股而言，最具代表性的指數即為台股加權股價指數

1. 簡單敘述性統計

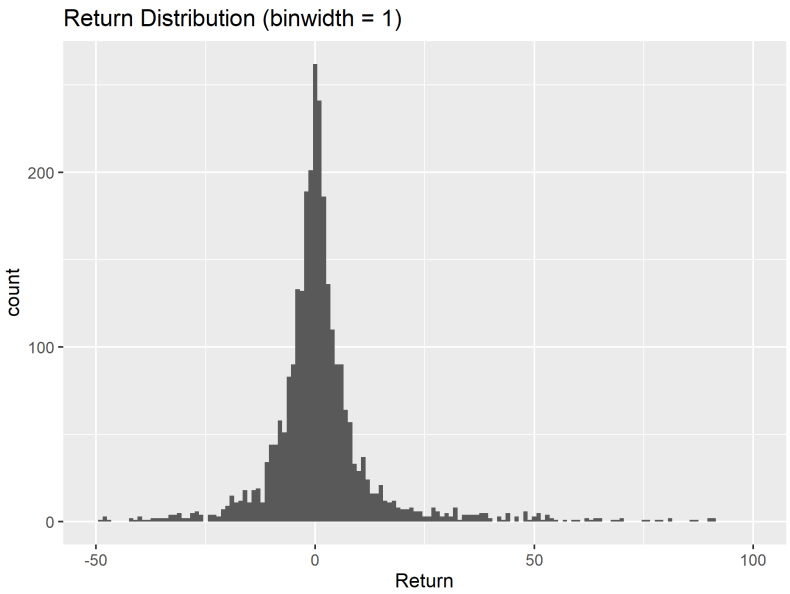
- 清洗完成的資料總計**5012**篇標的文

分類	看多	看空	長線	短線
數量	4058	954	888	4124

- 平均每位鄉民發了 **2.09** 篇標的文
- 平均每篇（可計算績效的）標的文的年化報酬率為 **15.65%**，篇數(n) = 2907
- 發文數量前10名的作者，其平均年化報酬率為**26.18%**
- 各年度發文篇數：

年份	2018	2019	2020	2021（上半年）
篇數	1342	1045	1503	1122

- 累積報酬率趨近常態分配，右尾多極端值



- 發文標的數統計

Target_name <chr>	n <int>
台積電	174
大盤	154
長榮	76
鴻海	59
國巨	54
穩懋	49
聯電	44
聯發科	40
華新科	35
群創	35

程式碼：return.rmd、statistics.rmd

2. 投資期間敏感度分析

- 原先設定投資期間並計算績效方式為長線時間250天，短線時間5天，截至6/10為止。
- 將短線及長線的時間分別縮短及增加，觀察年化報酬率以及累積報酬率是否有明顯差異。

結論：短期年化報酬率有顯著差異、長期則無顯著差異

短線	3天	5天	7天
平均累積報酬率	0.26%	0.33%	0.21%
平均年化報酬率	**21.89%**	16.42%	7.43%

長線	250天	300天	350天
平均累積報酬率	11.52%	15.33%	**18.08%**
完整天數	10.86%	**14.35%**	13.32%
平均年化報酬率	11.52%	12.77%	**12.92%**
完整天數	10.86%	**11.96%**	9.52%

*註：完整天數為可計算250天之標的文的報酬率
程式碼：return.rmd

3. 績效與大盤的比較

結論：股版平均報酬率未必能超越大盤

- 績效前50名的作者平均報酬顯著超越大盤，發文數前50名的作者則不一定

	2018	2019	2020	2021（至6/10）
大盤	-8.6%	**23.3%**	22.8%	**16.57%**
PTT股版	**23.78%**	8.5%	**24.65%**	1.7%
發文前50名平均報酬	32.58%	7.5%	32.31%	17.8%
績效前50名平均報酬	460.21%	304.41%	591.45%	689.71%

4. 股王排名

- **最佳績效獎（累積報酬率）**

時間	標的	作者	操作	績效
2020/11/23	長榮	cinghsang	長多	287.67%
2020/5/8	圓剛	ALAN	長多	255.31%
2020/4/28	聯電	dust	長多	247.62%
2020/12/31	高端疫苗	米食主義者	長多	226.15%
2020/3/17	群創	南茂哥	長多	188.65%

- **最佳效率獎（年化報酬率）**

時間	標的	作者	操作	績效
2020/4/27	康那香	還敢皮啊	短多	2391.17%
2020/6/17	愛普	麻瓜	短多	2299.84%
2020/7/17	美德醫療-DR	高級本省人	短多	1859.67%
2021/5/30	高端疫苗	酵公菌	短空	1833.62%
2020/4/20	寶齡富錦	股票戰船	短多	1828.61%

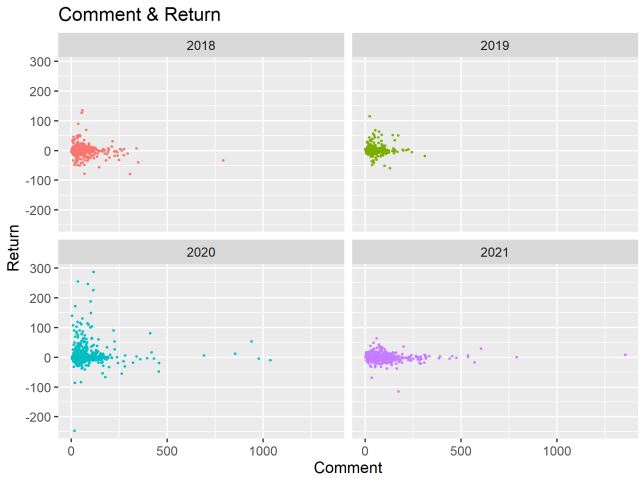
- **股海明燈獎（發文數超過5篇且累積報酬率最高）**

作者	篇數	平均績效
南茂哥	9	42.87%
Mr.King	11	29.73%
妤欣	13	23.04%
Chupei	9	19.56%
馬芙丸5566	7	18.09%

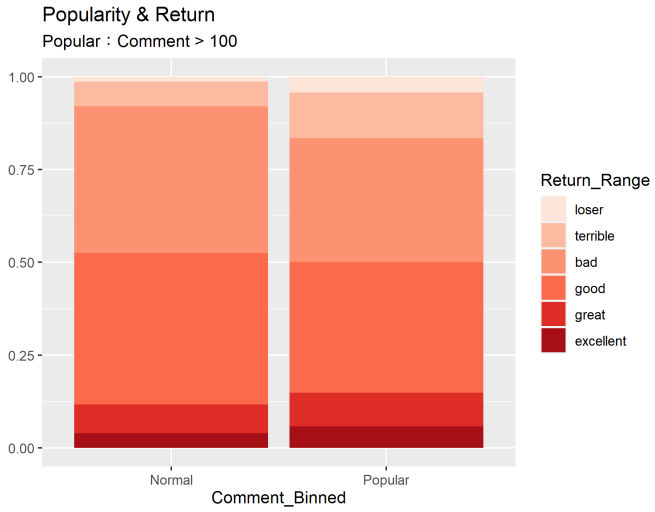
程式碼：statistics.rmd

5. 績效與留言數

- 2020年後標的文熱度較高
- 2020年高報酬標的文比其他年多，我們推測與**疫情**有關：
    - 美聯儲推出提振經濟的政策刺激股市
    - 大家都從股市獲得高報酬因此更願意貼標的文



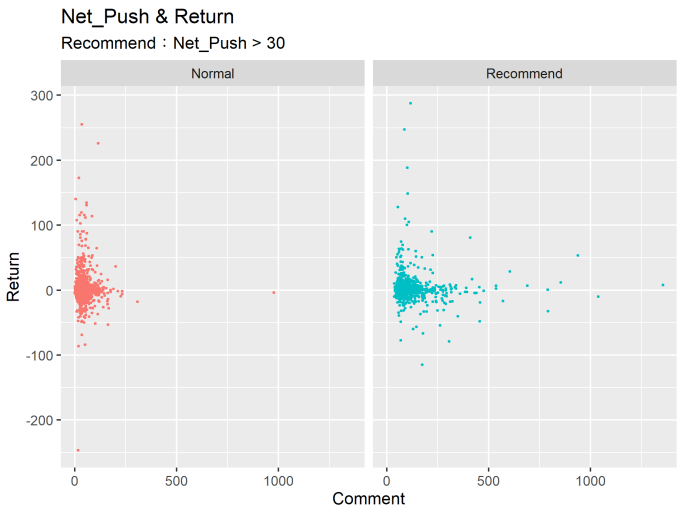
- 高人氣標的文極端報酬比例較高
- 圖中報酬率以30%、10%、0%、-10%、-30%為界



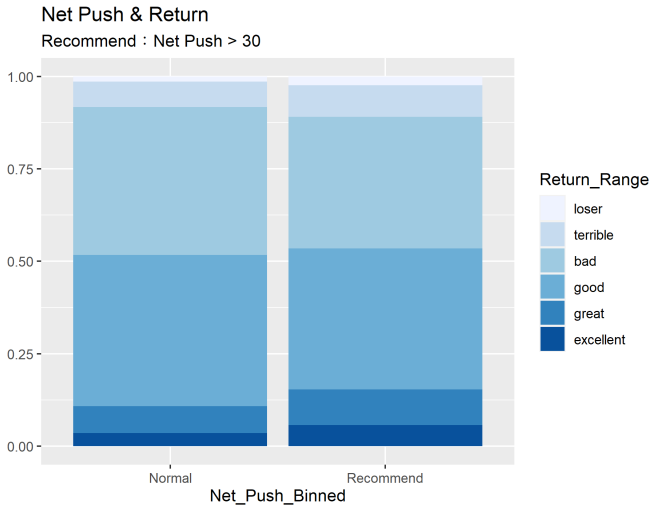
程式碼：statistics.rmd
*註：人氣標準：留言數100以上

6. 績效與推&噓

- 高淨推數文的資料點較為分散、報酬範圍較大



- 高淨推數文極端報酬比例稍高

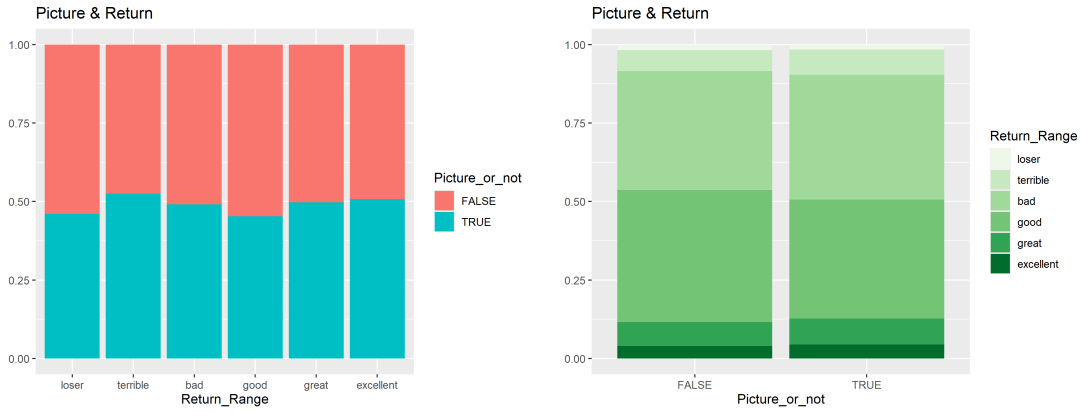


*註：淨推數 = 推數-噓數 > 30

程式碼：statistics.rmd

7. 績效與有無附圖

- 無顯著差異

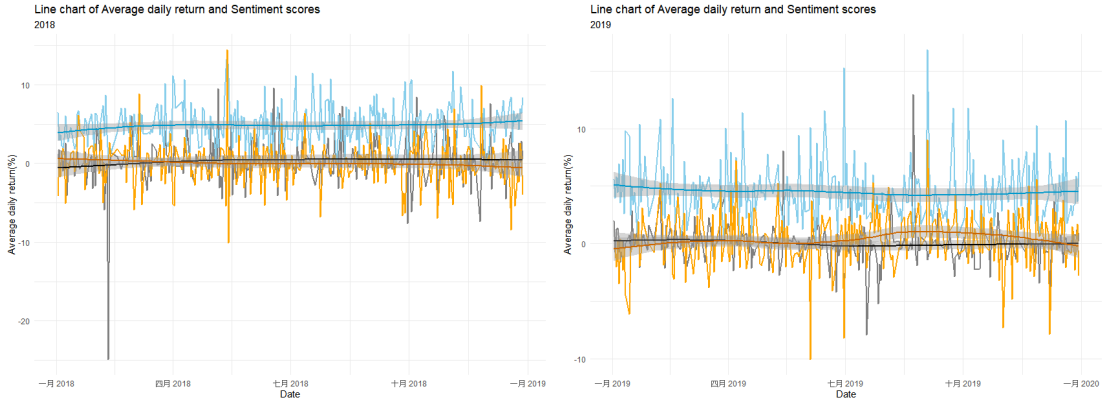


程式碼：statistics.rmd

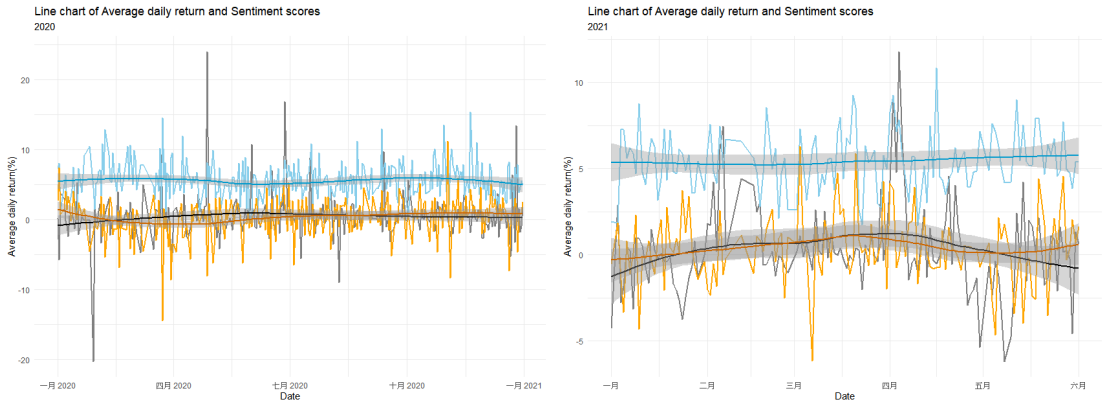
8. 績效與情緒分數

結論：情緒相關分數變動與績效變動趨勢相近

- 用詞強烈程度同時包含正向及負向情緒，因此分數較高
- 使用中文維度型情感字典
- 2018/2019



- 2020/2021



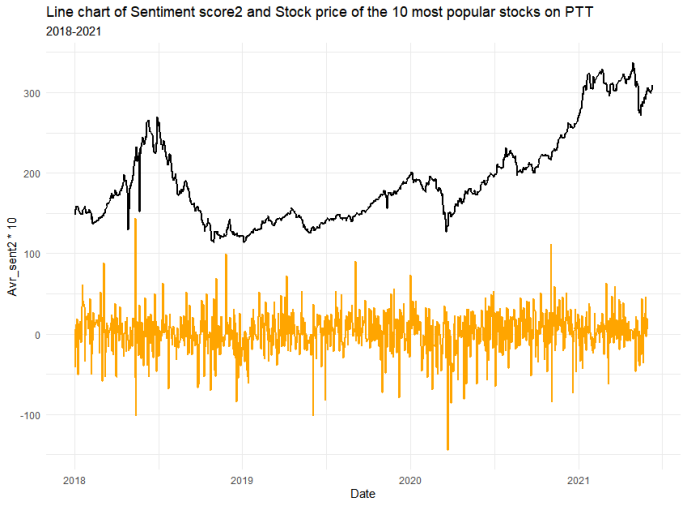
  - 淺藍：用詞強烈程度 = Arousal mean x abs（Valence mean - 5）
  - 橘：情緒正向/負向程度 = Arousal mean x（Valence mean - 5）
  - 黑：每日平均績效

程式碼：Sent_related_graph.r

9. 股價與情緒分數

結論：放大情緒波動後可見與股價存在關係

- 在2018年股價兩次急跌以及2020年疫情爆發時最為明顯



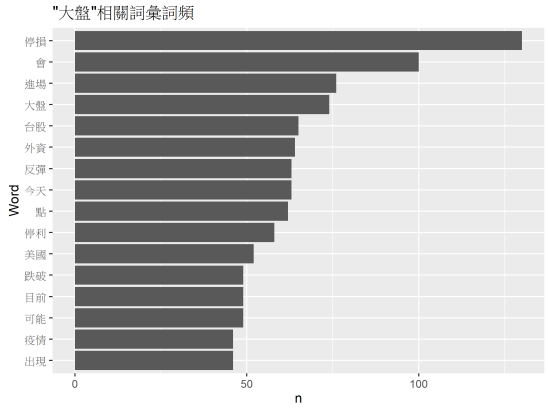
  - 黑：股價
  - 橘：情緒正向/負向程度

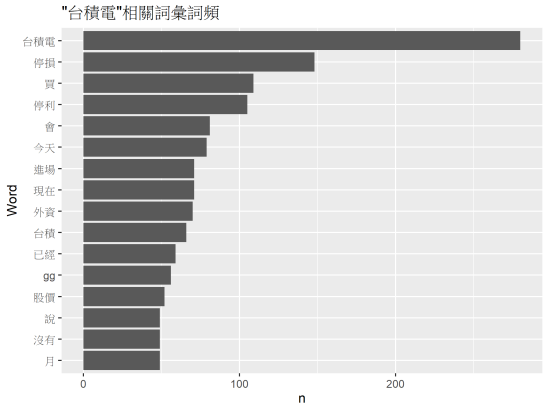
註*1*：股價方面為了方便計算，選擇前十大熱門的股票的價格，*(*這些標的佔了所有資料筆數的*1/3)*。 註2：情緒分數因為和股價的數值差距較大，為了讓波動幅度更明顯所以乘以10倍

程式碼：Sent_related_graph.r

10. 詞頻表

- 檢視和「大盤」相關的詞彙，可以看到**外資**、**美國**、以及**疫情**是這幾年標的文討論的重點
- 檢視和「台積電」相關的詞彙，可以看到**外資**、**gg(台積電的暱稱)** 常出現在標的文中



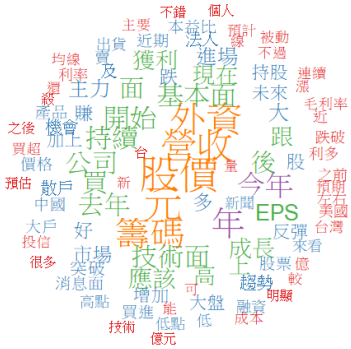


程式碼：frequency.rmd

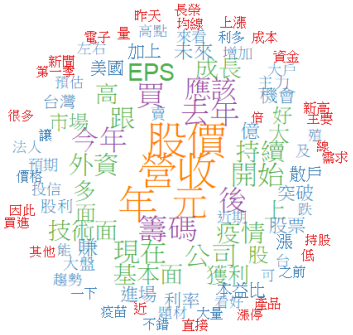
11. 文字雲

結論：歷年無明顯變動

- 2018(左)：「外資」詞頻進入前五、2019(右)：「美國」跌榜、「貿易戰」上榜



- 2020(左)：「疫情」始入榜、2021(右)：「疫情」持續在榜上



程式碼：wordcloud.r

12. 文本相似度分析

結論：無明顯差異

- 前五大熱門股票相似度波動無明顯規律

```
$台積電2330文本相似度
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.04042 0.09791 0.13345 0.18956 0.75425

$大盤50文本相似度
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.04922 0.10346 0.15493 0.21222 0.85853

$長榮2603文本相似度
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.07056 0.16397 0.20993 0.33630 0.77258

$鴻海2317
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.02916 0.08325 0.13694 0.19353 0.94668

$國巨2327文本相似度
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0481  0.1104  0.1677  0.2440  0.8081
```

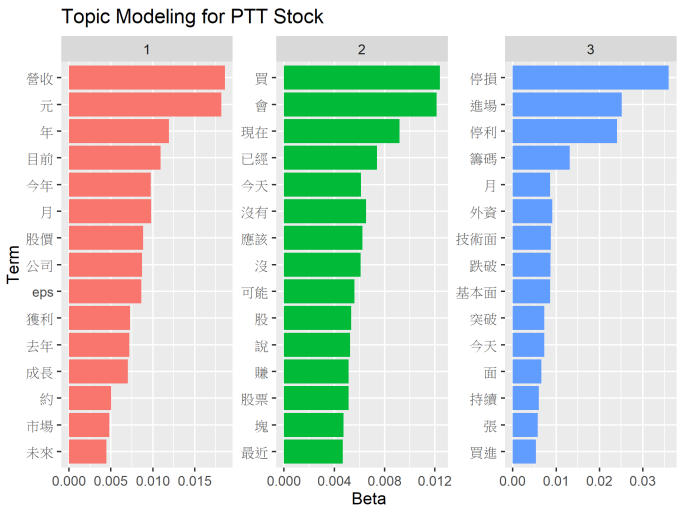
- 前六多發文作者相似度
- 其中排名二作者發了兩篇一模一樣的文所以作者二的文本相似度最大是1

```
$`Zyldr () 發本相似度`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.04885 0.10865 0.12098 0.17179 0.56390

$`kksis (流浪人生) 發文相似度`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2698  0.4622  0.5304  0.5280  0.5937  1.0000

$`drgon (蔡阿飛) 發文相似度`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.3356  0.5201  0.4664  0.6691  0.9202

$`hrma (資深象迷) 發文相似度`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2750  0.4119  0.3860  0.5255  0.7067

$`komica (糟糕島) 發文相似度`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.05634 0.15915 0.20526 0.35563 0.56941

$`mayingnine (個性決定能否投資賺大錢) 發文相似度`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01232 0.09016 0.14660 0.14878 0.19238 0.42357
```

程式碼：hot_5_stock.r/head_six_author.r

13. Topic Modeling

- 從左到右的主題分別和**公司面資料**、**股民主觀詞彙**以及**進退場判斷**的詞彙有關



程式碼：frequency.rmd

四、結論：To be 韭菜 or not to be, that is the question.

1. 股版年化報酬率每年**未必能超越大盤**
2. 累積報酬率趨近常態分配
3. 股市變動與PTT內文**情緒分數趨勢相近**，但目前**尚未觀察出因果關係**
4. 標的文最終的績效表現與文章熱門程度（推、留言）無顯著關係
5. 透過詞頻及文字雲觀察股市熱門字詞，歷年無明顯變動

五、分工表

組員	分工內容	投入程度
詩蘋	Google爬蟲、PTT文章爬蟲、資料清洗、書面報告	6
益嘉	股價爬蟲、資料清洗、績效計算分析、Topic Model、口頭報告	6
恩甜	詞頻表、文字雲、文本相似度分析	6
潔玟	情感分析、績效與發文數分析、口頭報告	6