

第一組 - HW1

主題

這次的主題鎖定在過去一年內ptt股票版的文章

由於我對股票並沒有太多研究，所以我想看看大家是如何建議新手或是學生族群進行股票投資

資料來源：

- 工作流程平台蒐集PTT 股票版(Stock)文章
- 關鍵字：選股、進場、新手、學生、菜鳥、小白、入場、建議、年輕、年青
- 時間區間：2023-03-21 ~ 2024-03-21
- 資料筆數：共 2867 篇文章

選擇看板 *

- rabbit(兔)
- Reptile(兩棲爬蟲)
- Salary(職場)
- SENIORHIGH(高中)
- Soft_Job(軟體工作)
- Steam(Steam)
- Stock(股票)

搜尋關鍵字 ⓘ

- 選股
- 進場
- 新手
- 學生
- 菜鳥

排除關鍵字 ⓘ

以換行區隔，e.g.
壽山動物園
猴子
...

搜尋起始日期

2023/03/21

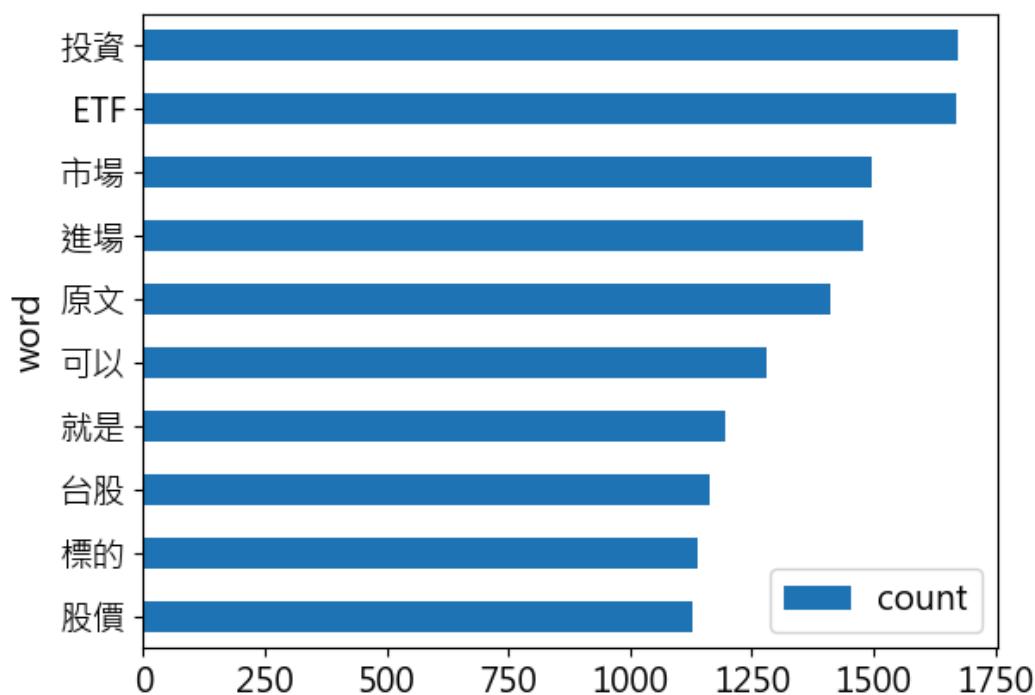
搜尋結束日期

2024/03/21

資料前處理

1. 將抓取的資料，同上課內容進行移除網址、空值、\n...等內容
2. 接著進行斷句斷詞等操作

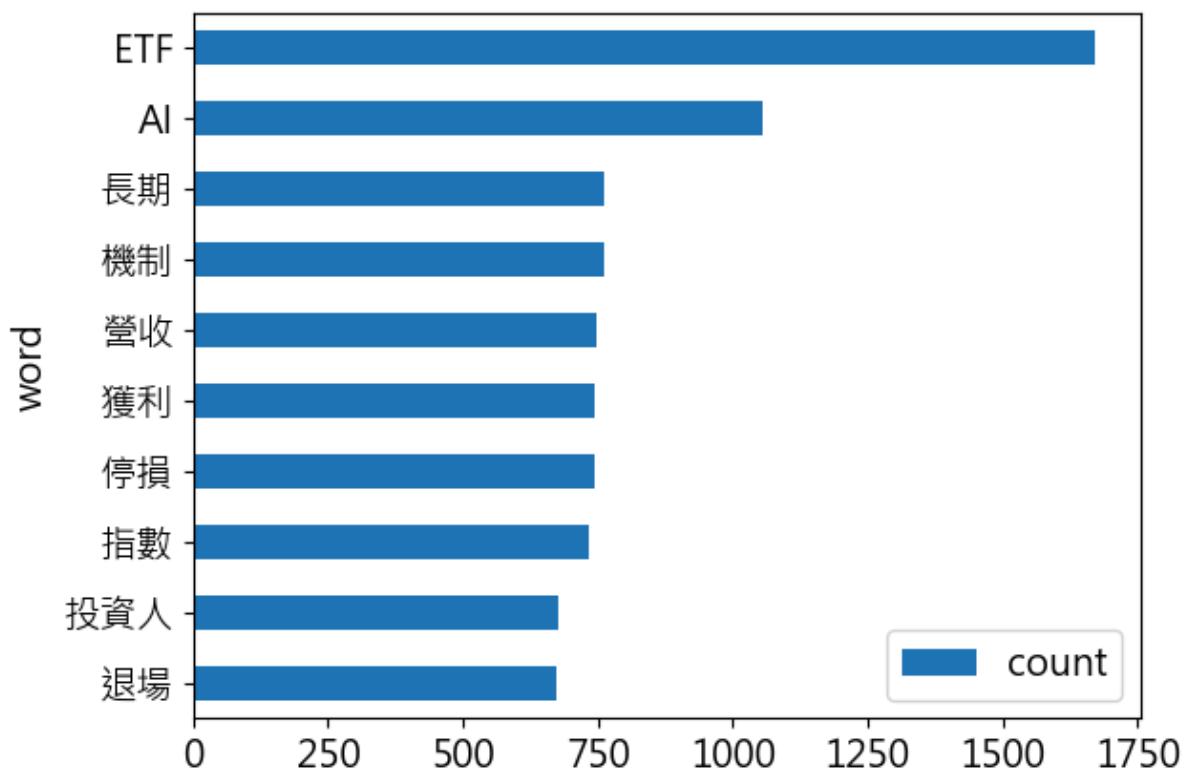
3. 下一步列出目前頻率前10的詞，並製作文字雲



4. 根據斷詞字典結果，我將自己認為不重要的詞加入停用辭典，以下為部分內容

3195	原文
3196	可以
3197	就是
3198	今年
3199	目前
3200	現在
3201	自己
3202	還是
3203	沒有
3204	可能
3205	這裡
3206	單純
3207	投資

5.新的辭典及文字雲如下





總結: 可以看出有不少網友推薦買ETF、台積電，也看好AI產業等等，投資時間則推薦長期存放

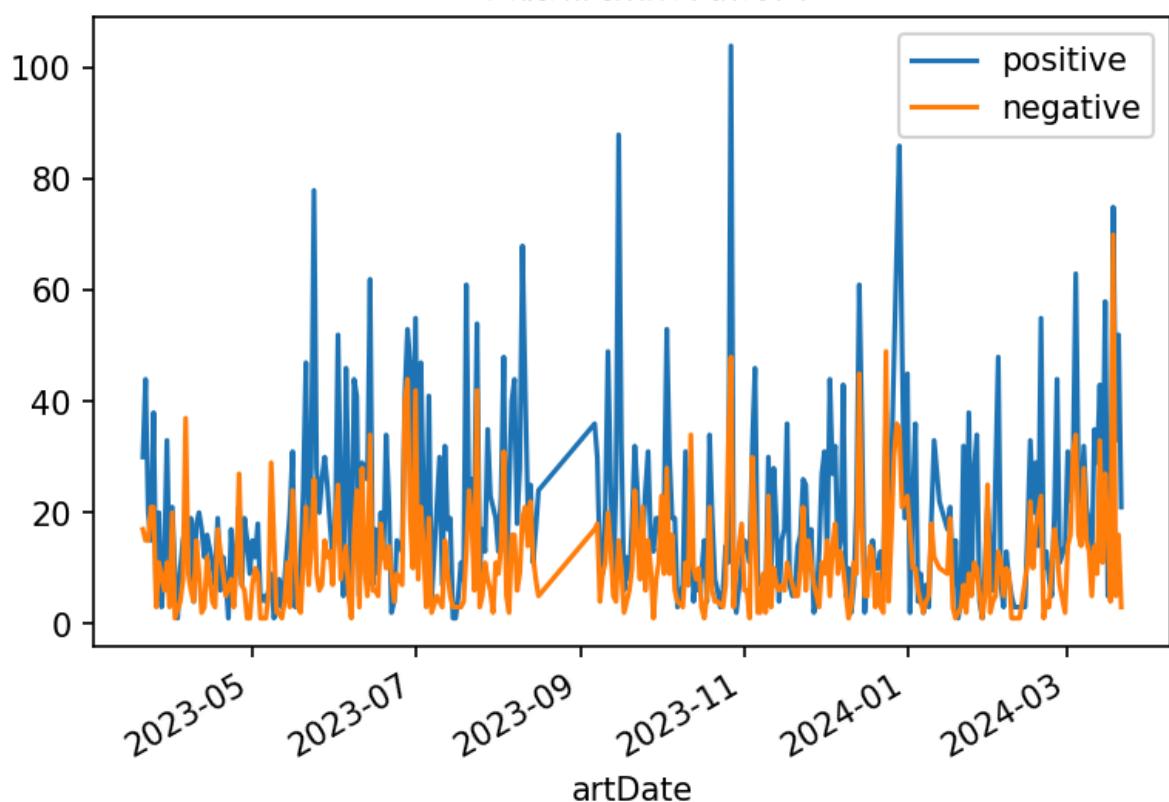
情緒分析1

1. 首先分析每篇文章的情緒

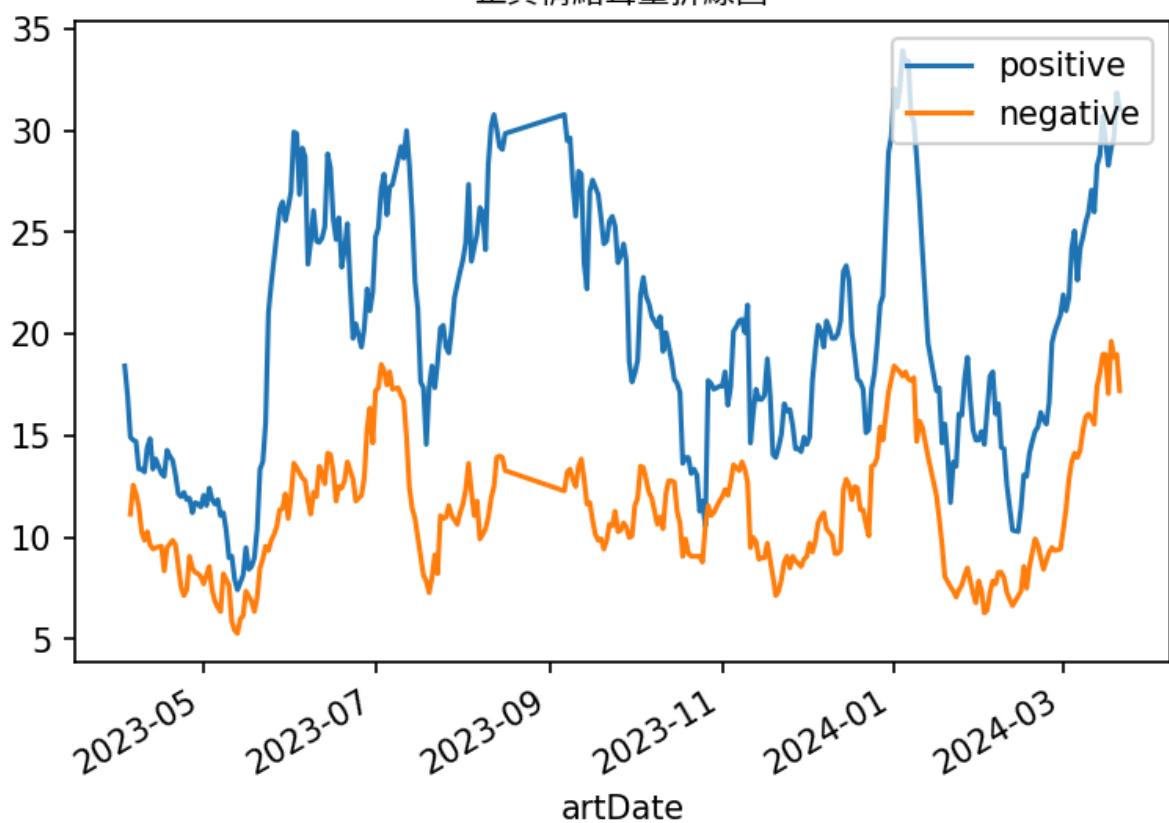
	artDate	sentiments	size
3	2023-03-21	negative	17
4	2023-03-21	positive	30
9	2023-03-22	negative	15
10	2023-03-22	positive	44
15	2023-03-23	negative	15
...
1687	2024-03-19	positive	33
1691	2024-03-20	negative	16
1692	2024-03-20	positive	52
1695	2024-03-21	negative	3
1696	2024-03-21	positive	21

2.各種情緒分析繪製折線圖如下

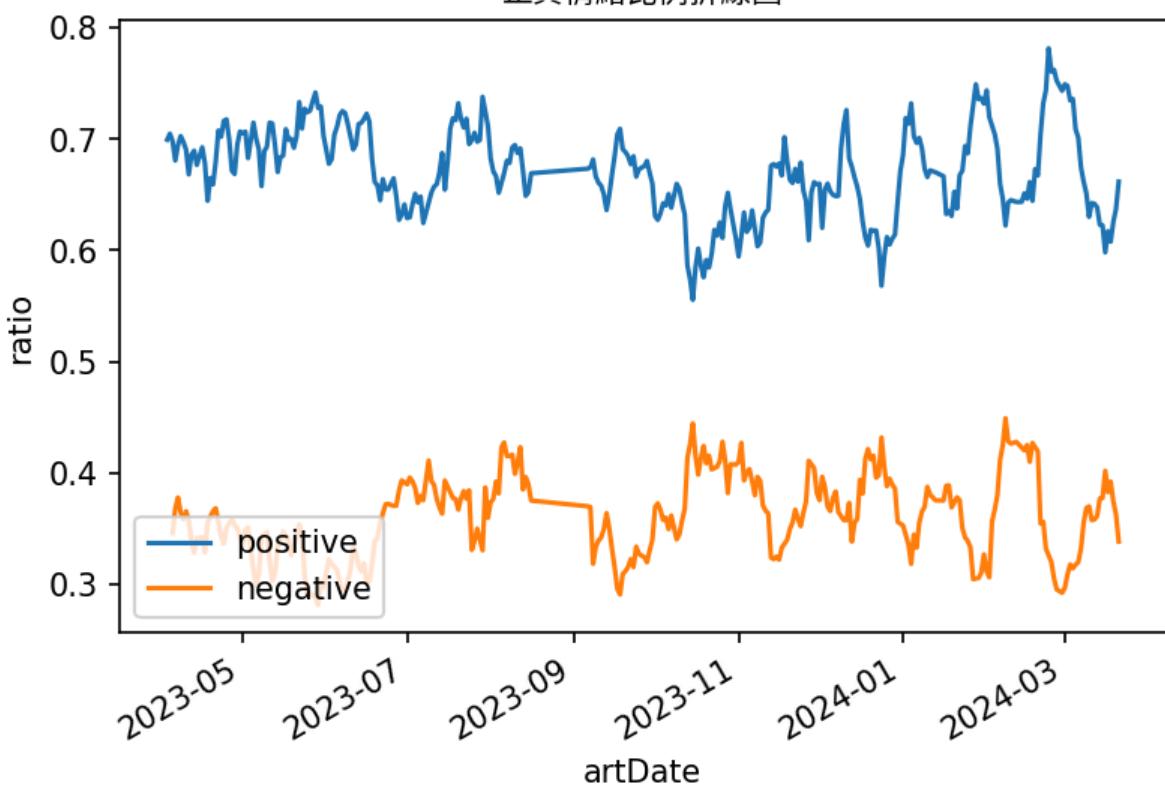
正負情緒詞彙頻率折線圖



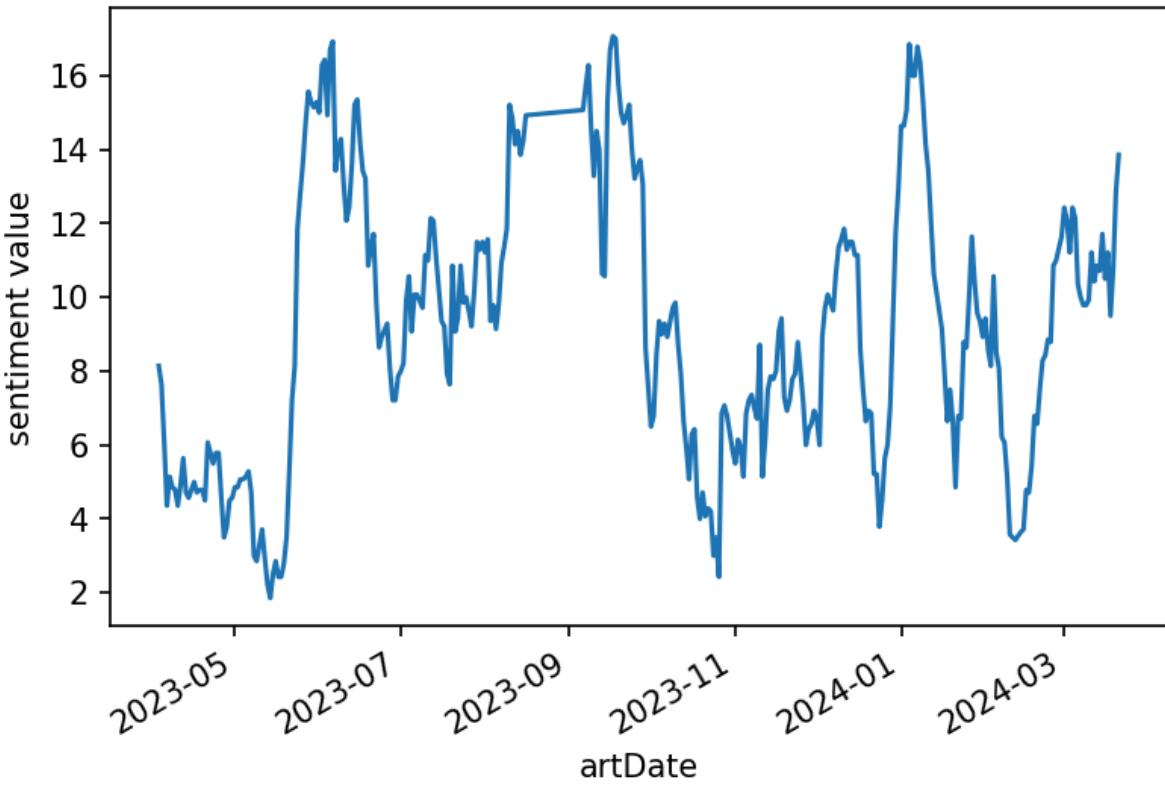
正負情緒聲量折線圖



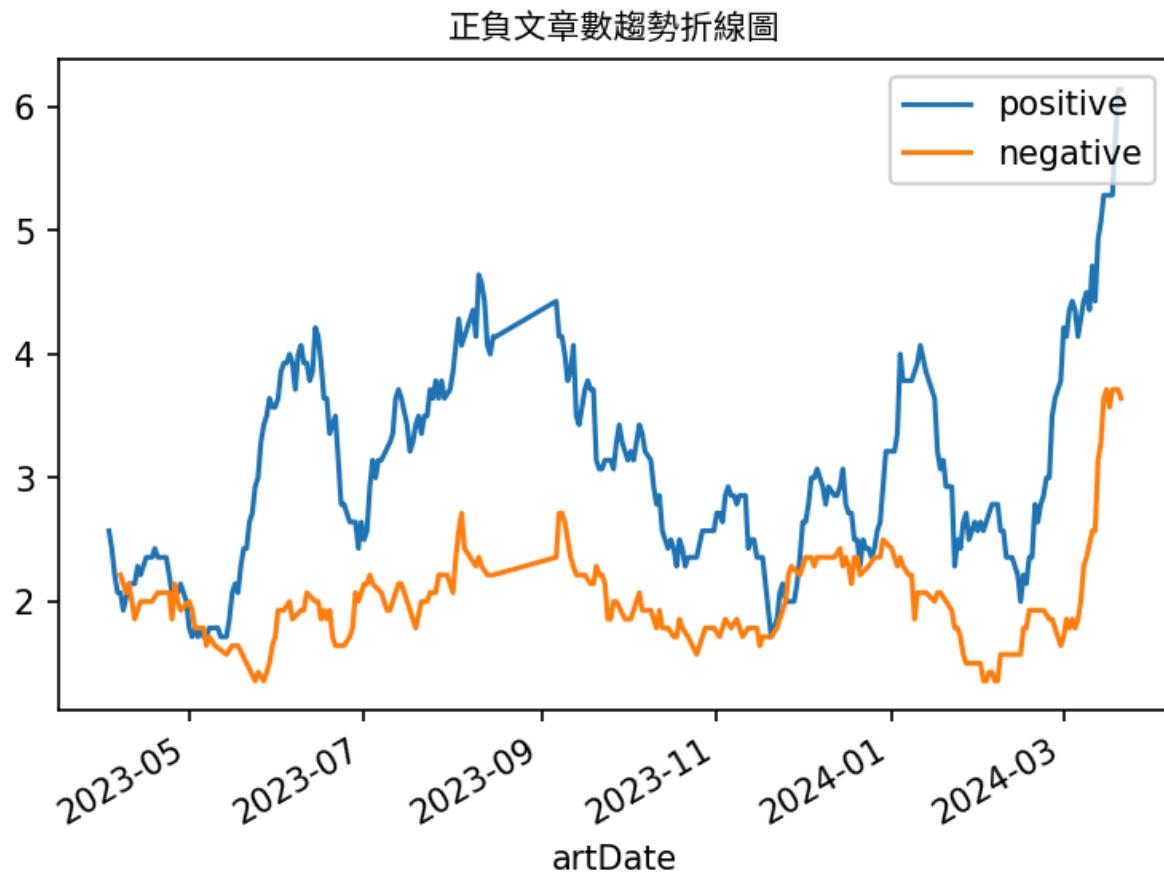
正負情緒比例折線圖



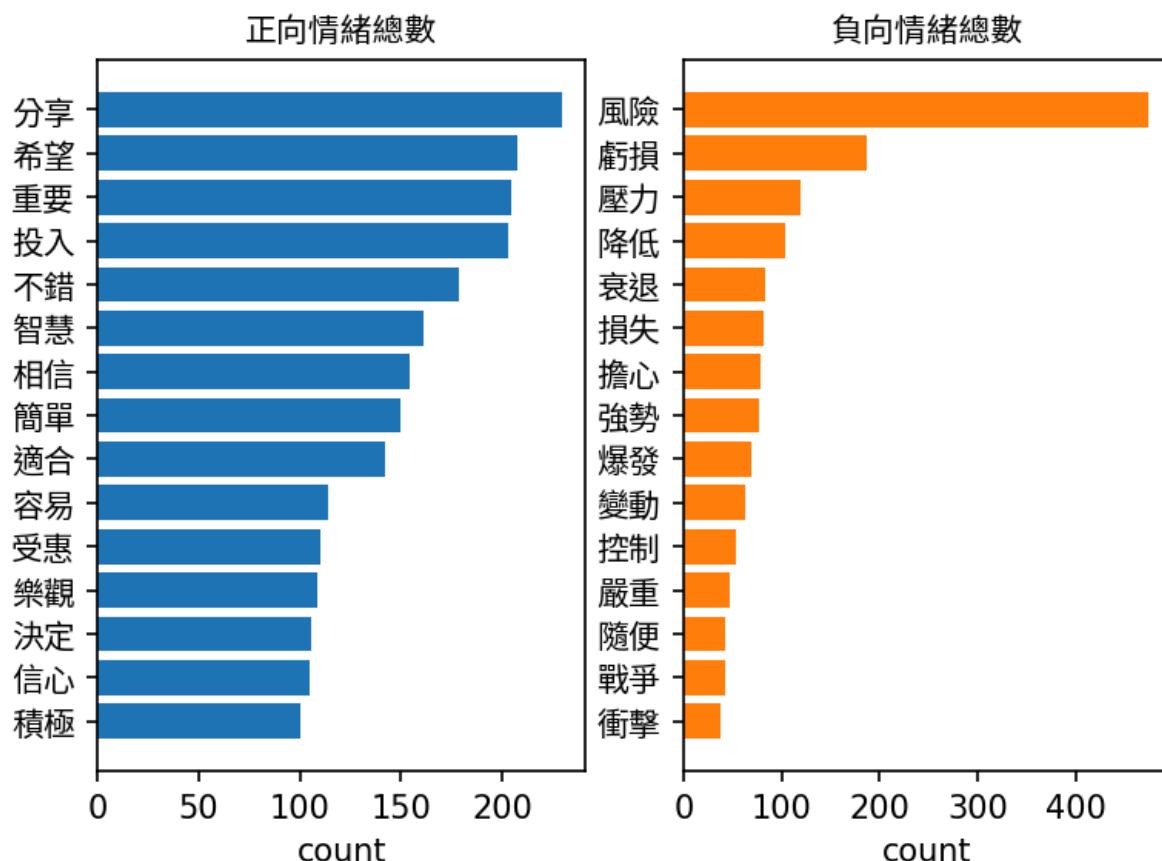
正負情緒分數趨勢折線圖



3.接著用文章數分析文章的趨勢，可以看出在23年末的時候正面聲量低迷，負面聲量也相對高，近期則是正負聲量都很高



4.分析正負情緒詞的出現次數，我認為正向部分的分享、希望、重要...等詞都不是正面，以及負向的風險、降低...也不一定是負面，這些詞應該更趨向於中立，另外像爆發這個詞，也有疫情爆發、商機爆發等兩種不同情緒



5.進行文本檢查，可以確定這些詞的確更趨向中立，以下以"分享"作為範例

```
for sentence in filtered_df['sentence'].to_list():
    print(sentence)
    print("====")
```

✓ 0.0s

分享小弟我對技術分析的一些淺見

還是一堆人搶著買版上分享的人也不曾停過

小弟曾經分享清流君說明高股息金融商品缺點的影片

板上前輩可以分享一下看法嗎

但這篇只是一種觀念的分享而已

是因為我的選股方式在版上都有分享過了

1標的 6196 帆宣2分類多3分析正文昨天分享的廣積追高有風險

今天想和大家分享一下個人如何透過新聞抓到IGBT 概念股的方法

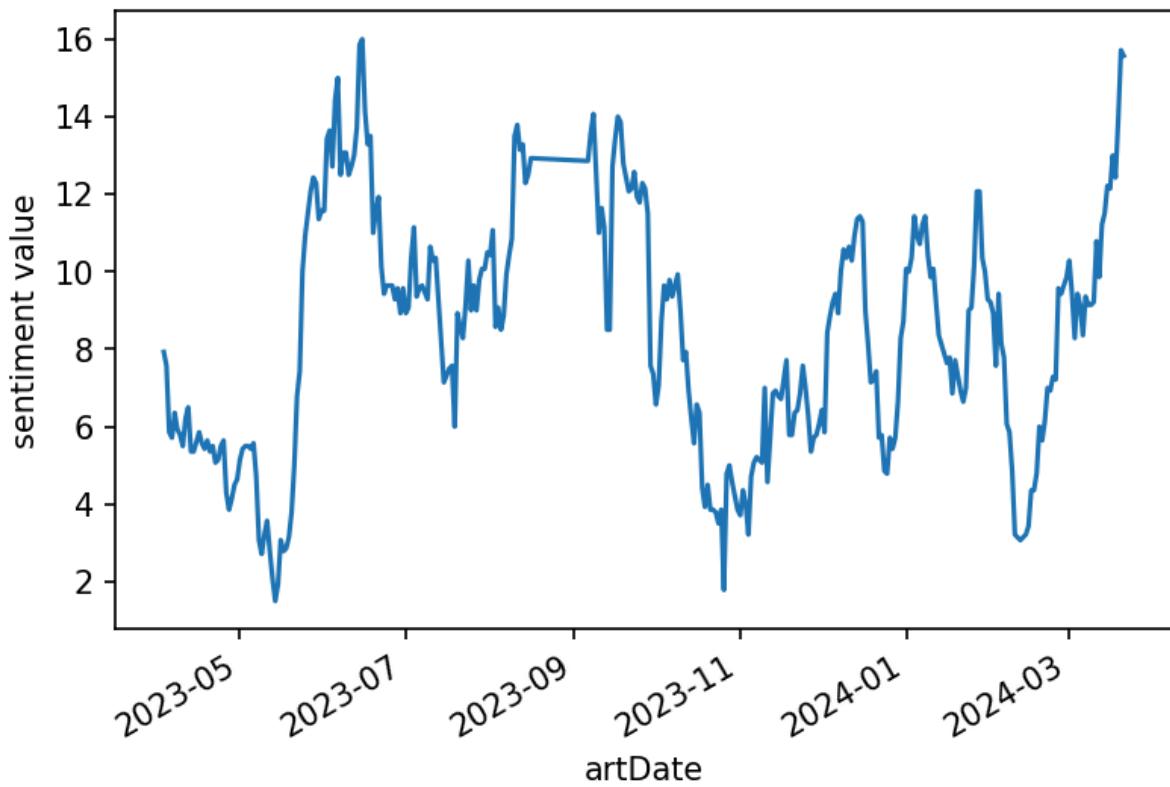
内心真的非常感謝這些前輩的無私分享及教學

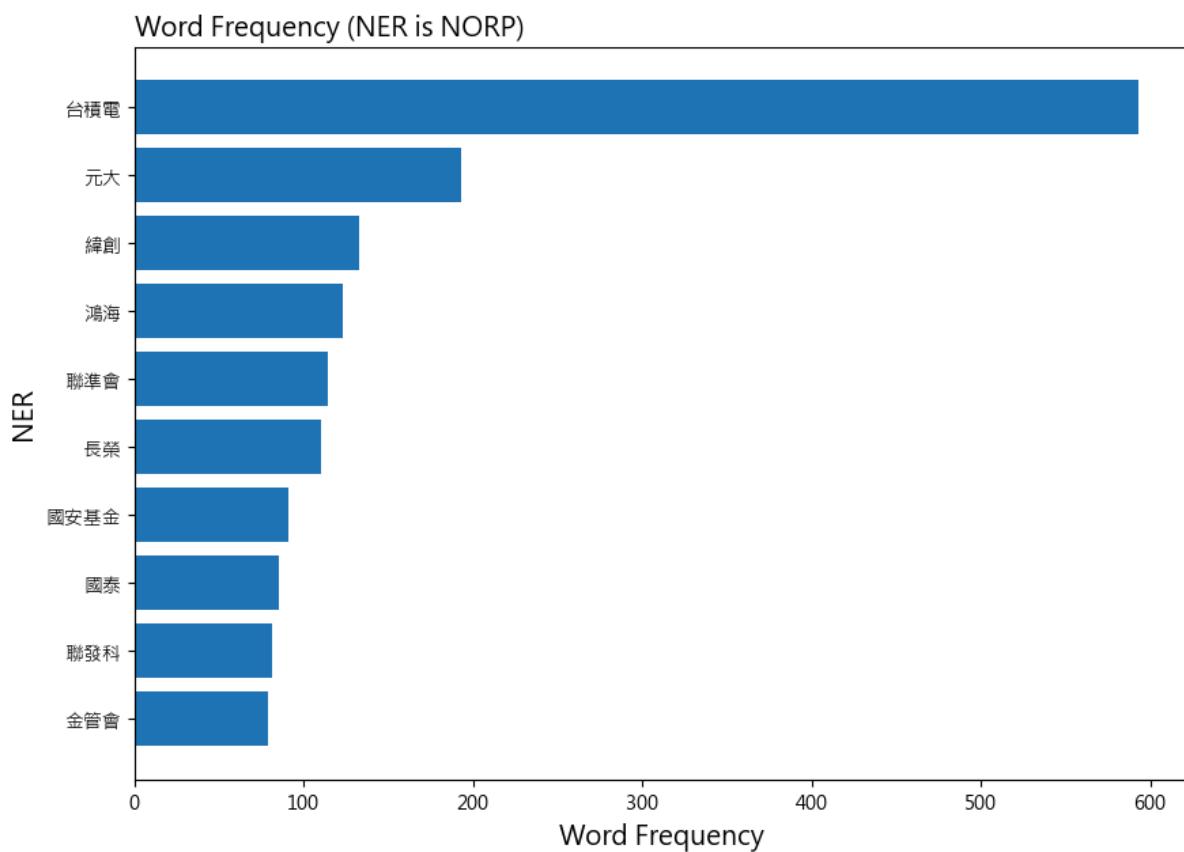
6.最後移除以下詞彙，

'分享', '希望', '重要', '決定', '隨便', '風險', '降低', '強勢', '爆發', '控制', '投入', '智慧', '情緒'

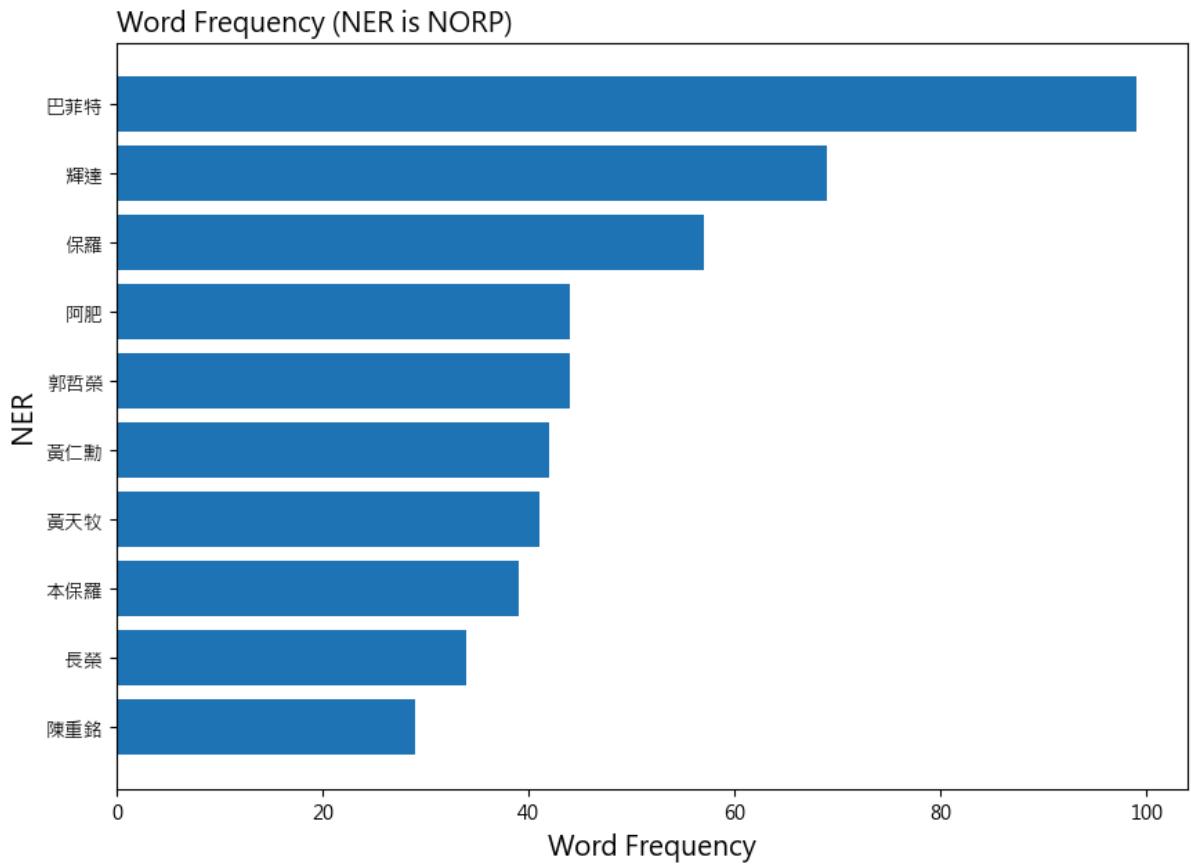
7.去年五到七月情緒變化非常極端，讓我很疑惑發生甚麼事，所以接下來會集中分析該區段

正負情緒分數趨勢折線圖



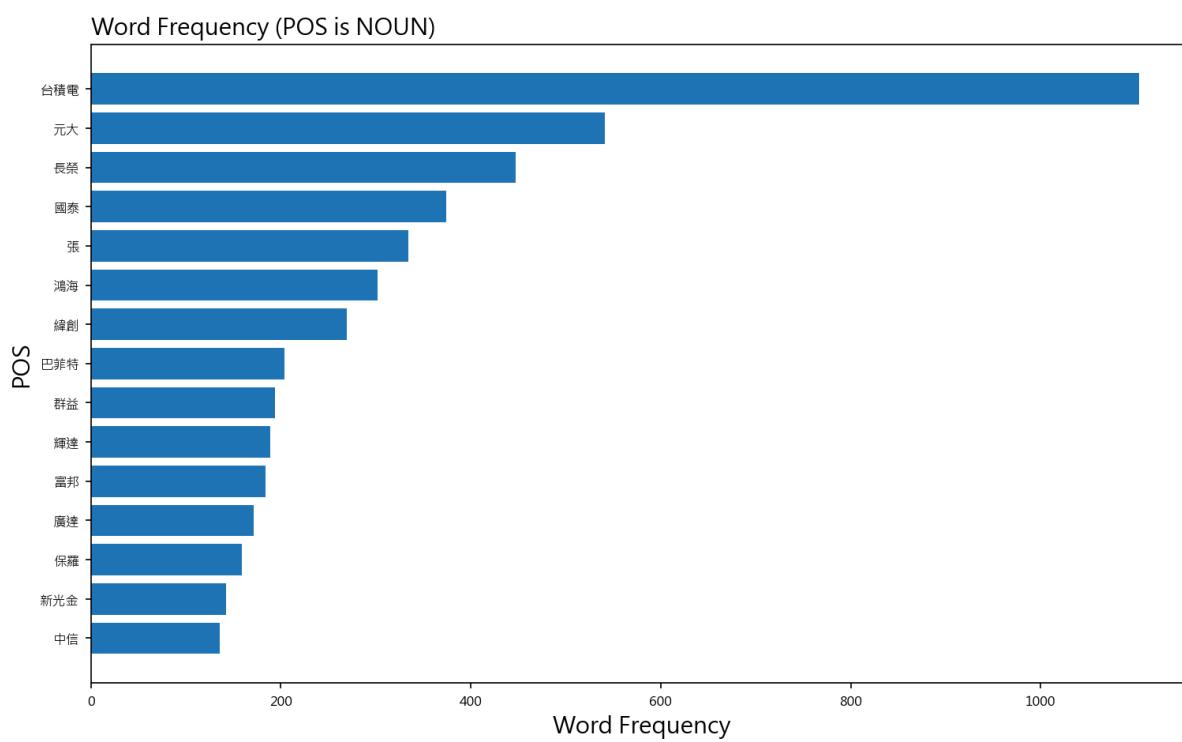
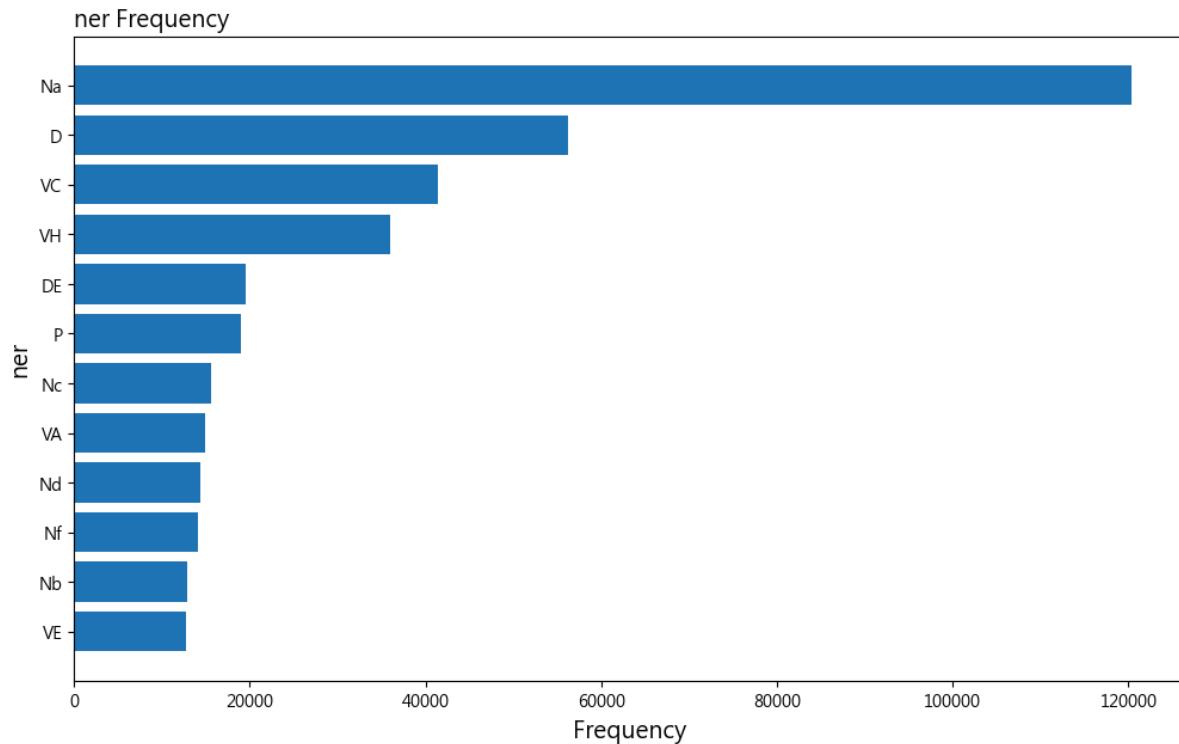


2.另外我也好奇PERSON的結果，但可以看到有很多像輝達、長榮等應該屬於ORG的詞反而出現在PERSON。

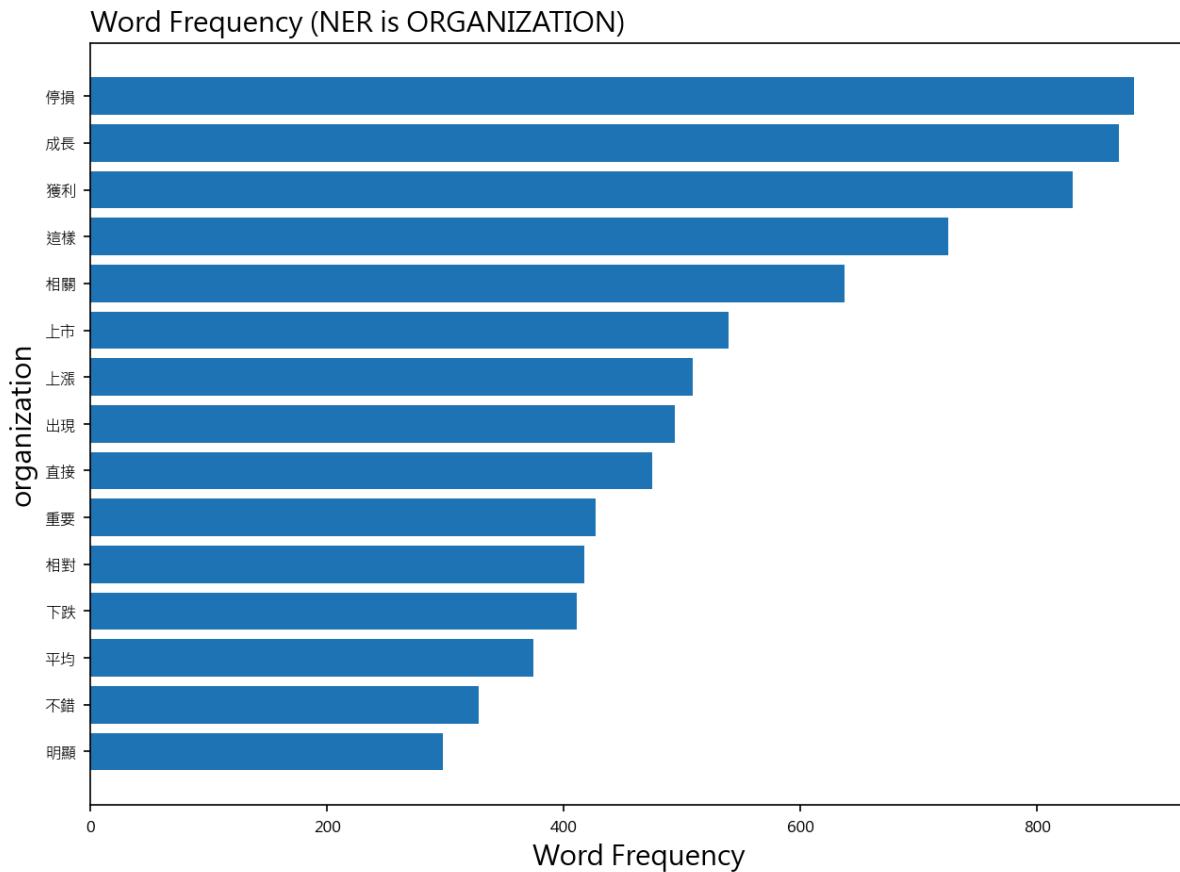


POS分析

1.名詞分析，可以看到台積電依舊是最多討論度的，果然是護國神柱！！

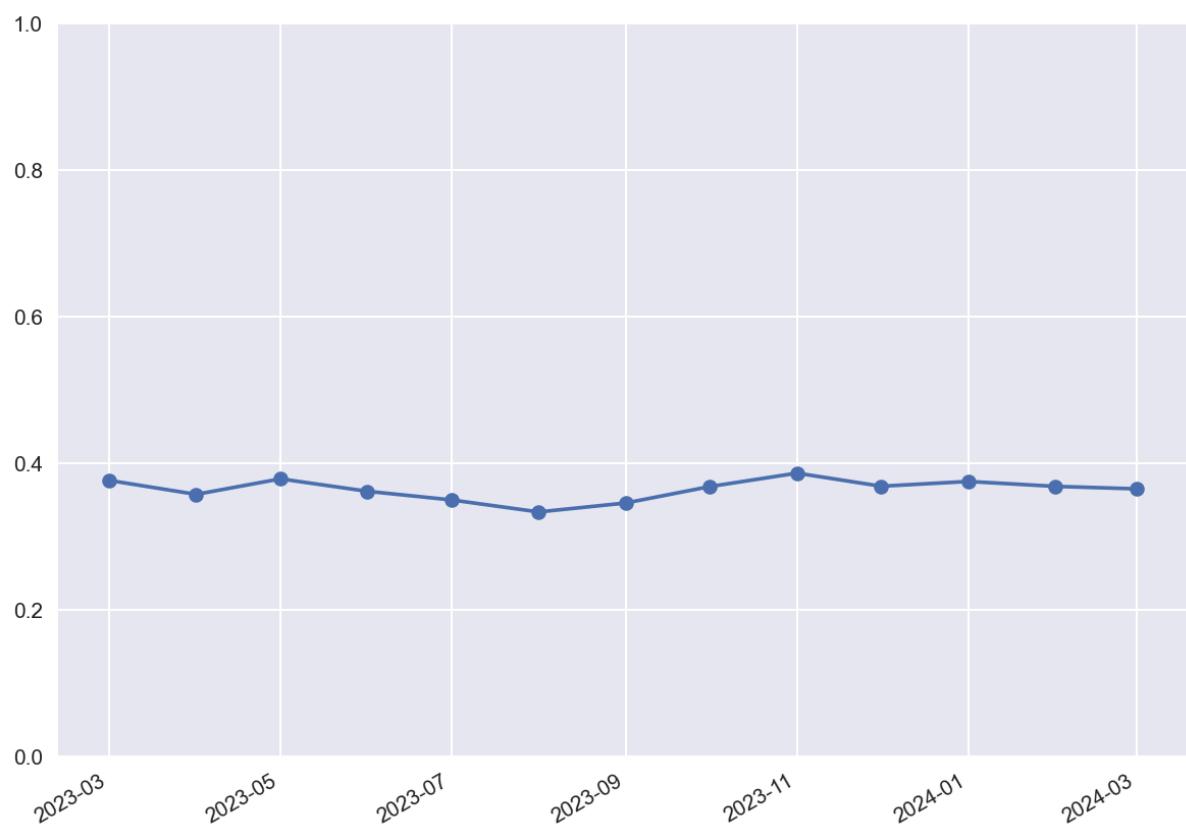
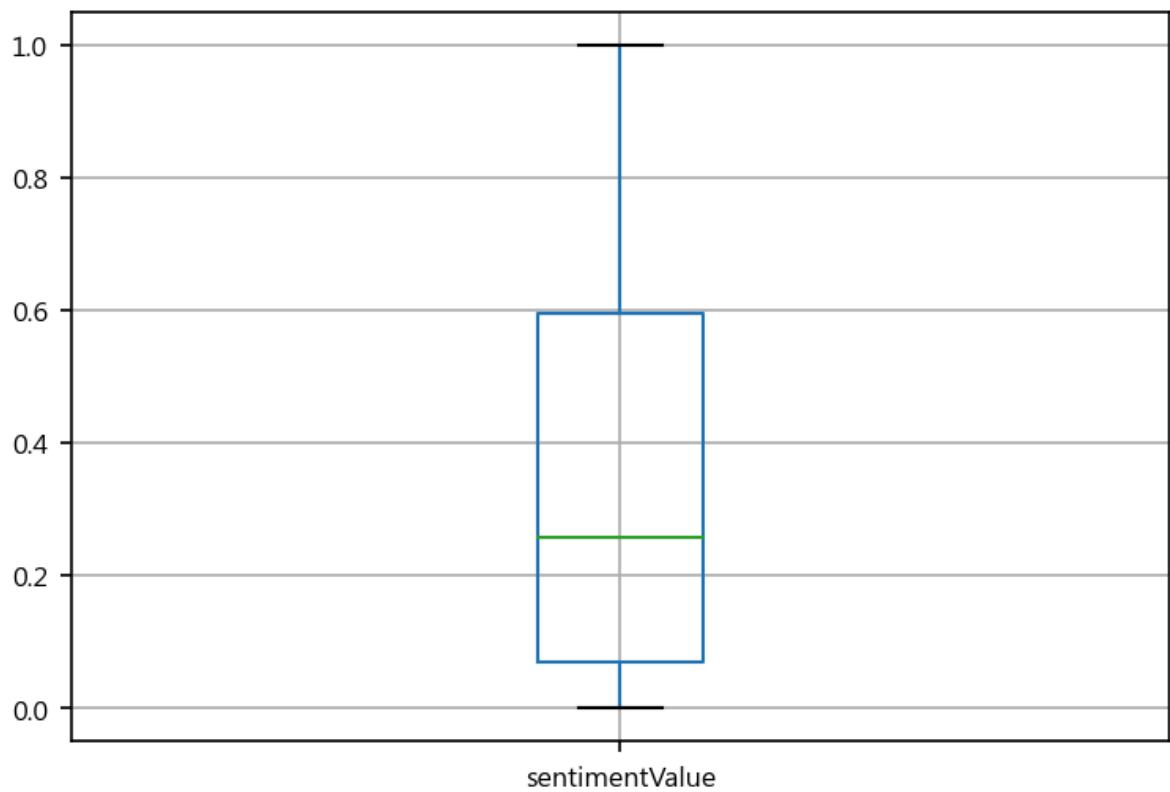


2.動詞分析，發現在貼文中的動詞，「停損」佔了相當高的頻率，其他還有出現「獲利」、「成長」等詞彙。



Snow NLP分析

1. 將文章的語句打上情緒分數，並繪製成盒狀圖和走勢圖。可以觀察到大部分天數的情緒值都在0.4左右，偏向中性，推測是因為PTT股市板上多數為請教投資相關問題的緣故，且根據自身觀察，有些用戶喜歡在虧損時發文宣洩，可能是造就分數較低的原因。



2. 進行詞彙探索，從上圖中可以發現 2024 年 3 月 19 日的情緒分數相對高，我們把這天的正面句子挑出來看看他們都用甚麼詞，並歸納成文字雲。藉由用水、水電，可以看出大多在講台電調漲電費對半導體產業的影響會不會波及到股市

