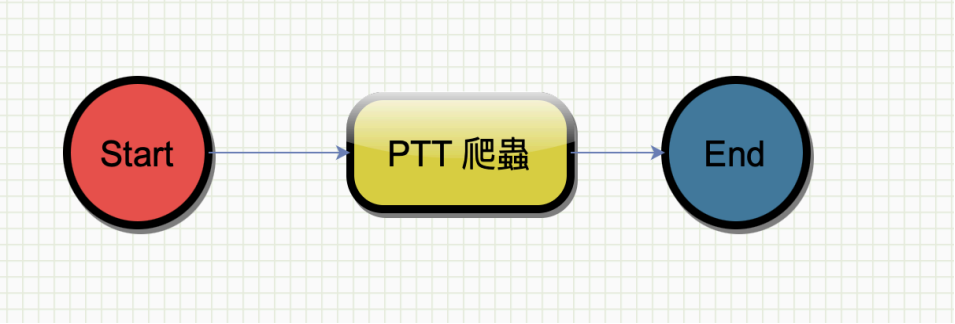
# 第一次讀書會報告(第四組)

**B104020009** 鄔仁迪 **M124020035**王唯宇 **M124020036**吳景煥 **M124020005**蔡秉祐
**B094020038**曹丹柔 **B094020041**吳芃儀 **M124020001**莊家綺

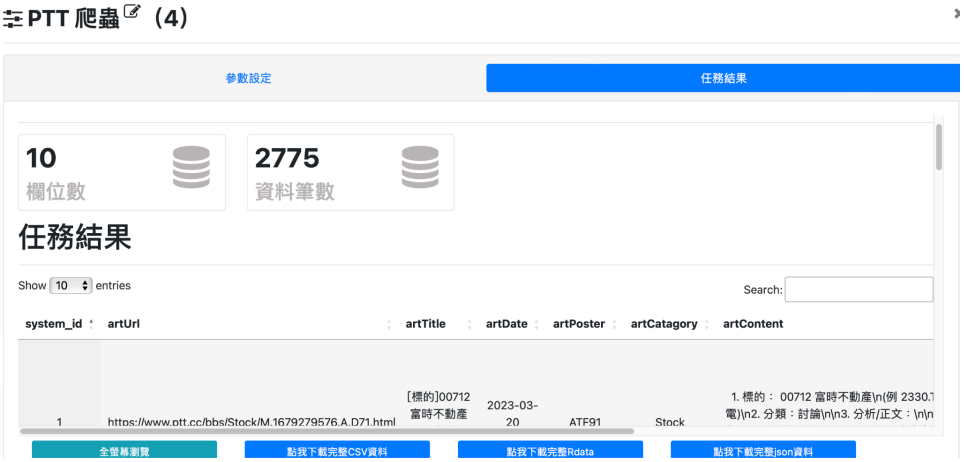
## 一、資料介紹

- 資料來源：Tarflow工作平台蒐集PPT股票版(Stock)文章
- 關鍵字：台積電、輝達
- 時間：2023-03-20~2024-03-20
- 資料筆數：2775

1. Tarflow 爬蟲流程圖







2. 資料資訊

```
RangeIndex: 2775 entries, 0 to 2774
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   system_id     2775 non-null   int64
 1   artUrl        2775 non-null   object
 2   artTitle      2775 non-null   object
 3   artDate       2775 non-null   object
 4   artPoster     2775 non-null   object
 5   artCatagory   2775 non-null   object
 6   artContent    2771 non-null   object
 7   artComment    2775 non-null   object
 8   e_ip          2619 non-null   object
 9   insertedDate  2775 non-null   object
 10  dataSource    2775 non-null   object
dtypes: int64(1), object(10)
memory usage: 238.6+ KB
```

## 二、資料前處理

1. 首先先初步繪製文字雲


2. 發現內容中一些數字可能是股價的資訊，但這並不是我們想分析的內容所以將其替換掉，下面是處理後的結果。

```
sent_df['sentence'] = sent_df['sentence'].str.replace(r'[^\w\s]+|[\d]', '', regex=True).astype(str)
```

3. 我們主要要分析的是大眾對於台積電的想法，但我們發現文字雲中出現了很多公司名稱，如國泰、富邦等，翻找文章之後發現文章很多都提到台灣50成分股的股價資訊，以及一些沒有意義的字詞，如表示、內容等。而這些都不是我們所關心的，因此我們一併將其移除。下圖是手動加入的停用字以及公司名稱。

```
stopwords_manual = ["原文", "標題", "今年", "來源", "公司", "評論", "署名", "內容時間", "現在", "連結", "目前", "股票",
                    "股價", "台積電", "台積", "台股", "台灣", "台灣股市", "台灣股票", "台灣股價", "台灣股票市場",
                    "台灣股市", "內容", "評論", "文章", "內文", "記者", "指出", "心得", "來源", "市場", "表示", "去年",
                    "預期", "億元", "TW", "持續", "未來", "產業", "券增", "排行", "證券", "今天", "報導", "分類",
                    "網址", "時間", "客戶", "認為", "討論", "張數", "相關", "影響", "機制", "億美元", "美元", "企業",
                    "金額", "資料", "發布", "格式", "投資"]

stopwords.extend(stopwords_manual)
```

| | |
|---|---|
| 3279 | 葡萄王 |
| 3280 | 世紀鋼 |
| 3281 | 群電 |
| 3282 | 世芯-KY |
| 3283 | 智伸科 |
| 3284 | 洋基工程 |
| 3285 | 大聯大 |
| 3286 | 美利達 |
| 3287 | 裕隆 |
| 3288 | 豐泰 |
| 3289 | 智原 |
| 3290 | 川湖 |
| 3291 | 旺宏 |
| 3292 | 聯強 |
| 3293 | 東哥遊艇 |

| | |
|---|---|
| 3266 | 鈺齊-KY |
| 3267 | 東和鋼鐵 |
| 3268 | 合勤控 |
| 3269 | 聯電 |
| 3270 | 立積 |
| 3271 | 同欣電 |
| 3272 | 晶技 |
| 3273 | 力積電 |
| 3274 | 高力 |
| 3275 | 麗豐-KY |
| 3276 | 大立光 |
| 3277 | 臺企銀 |
| 3278 | 全新 |

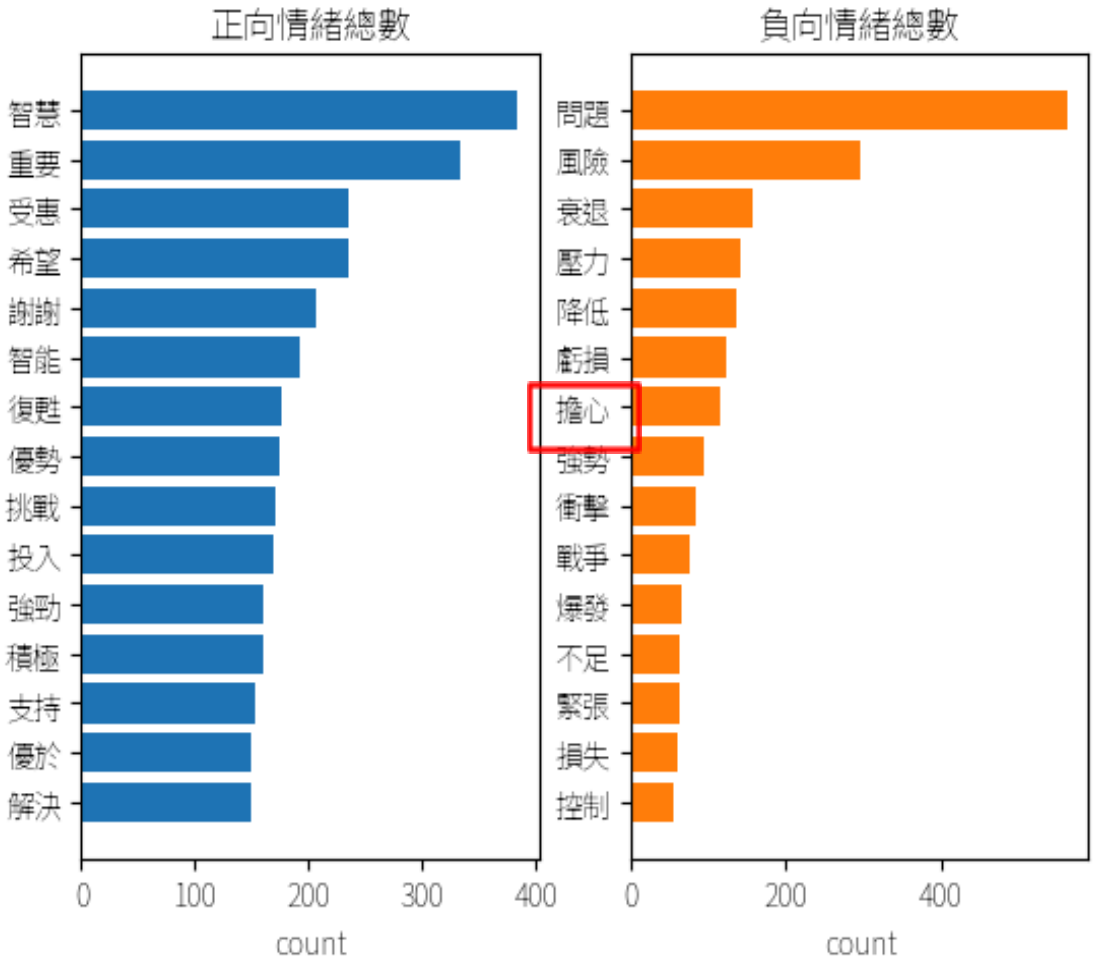
4. 另外，因為我們分析的主題是和金融相關的，但有很多金融相關的術語並沒有被收入到jieba的字典當中，如成交量、技術面等，可能斷詞會不太精準，因此我們將一些金融的專業字詞加入詞典中，並重新繪製文字雲如下圖。

5. 清理前後的文字雲比對圖

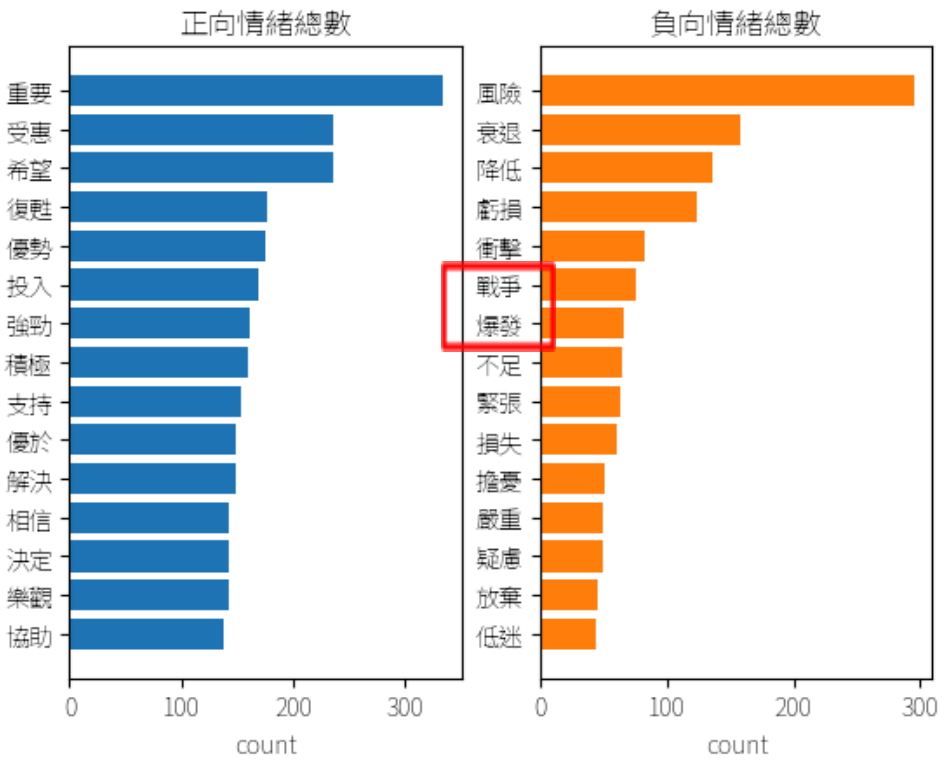


## 三、情緒分析 (字典法)

1. 「強勢」這個詞彙在LIWC字典被列入負向情緒詞，但我們討論後覺得「強勢」用在金融、股市領域應該代表著正向情緒，因此我們將「強勢」的情感改成positive



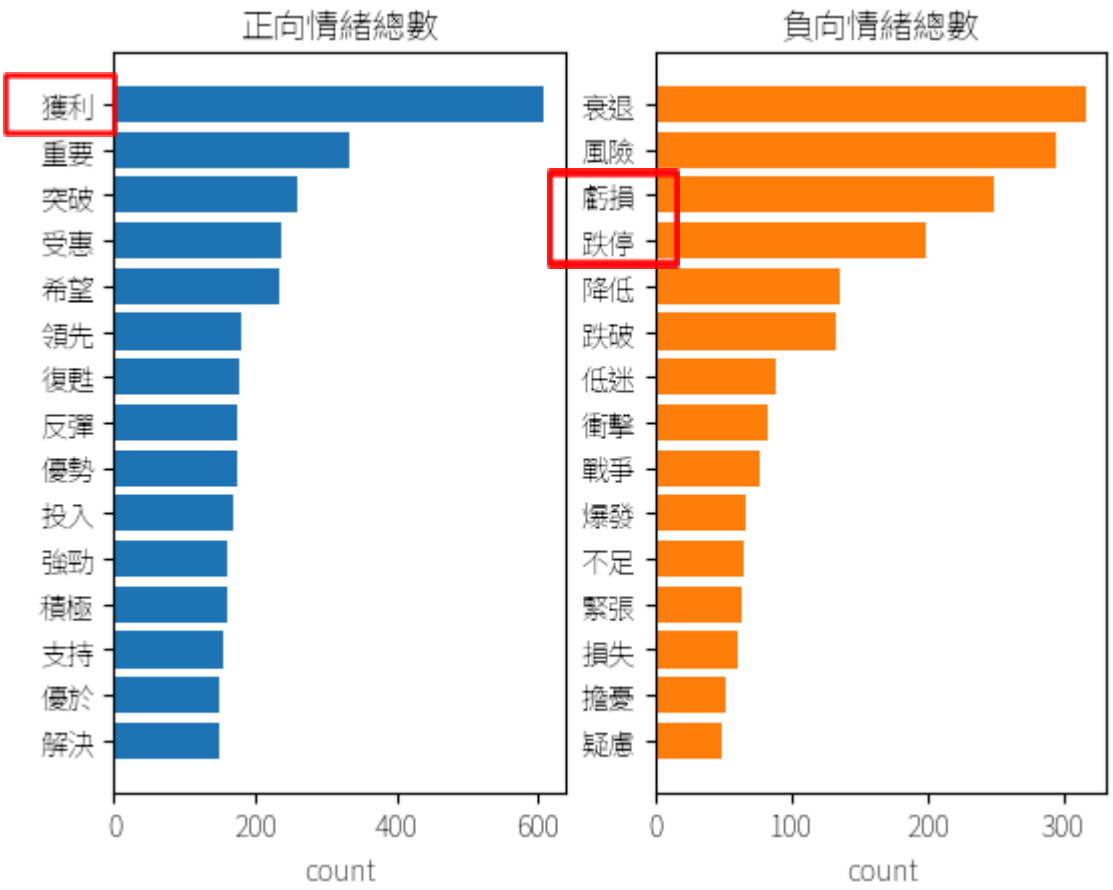
2. 「戰爭」、「爆發」等詞彙出現在負向情緒詞，這很可能與過去所爆發的以巴衝突、烏俄戰爭有關




3. 該資料詞典較少關於金融領域的情緒詞彙，因此我們追加金融分析術語至該詞典

```
4683  鼻酸,sad
4684  牛市,positive
4685  反彈,positive
4686  突破,positive
4687  加速上揚,positive
4688  回升,positive
4689  創新高,positive
4690  穩健,positive
4691  上揚,positive
4692  飆升,positive
4693  熱門,positive
4694  活躍,positive
4695  獲利,positive
4696  領先,positive
```

```
4697  優質,positive
4698  超買,positive
4699  賺錢,positive
4700  熊市,negative
4701  回調,negative
4702  暴跌,negative
4703  跌破,negative
4704  弱勢,negative
4705  壓力,negative
4706  創新低,negative
4707  下挫,negative
4708  跳水,negative
4709  冷淡,negative
4710  滑落,negative
4711  虧損,negative
```




4. 可以看出金融相關詞彙已出現在關於「挑戰」的句子有正向情緒也有負向情緒，難以判定，故選擇刪除該詞彙

之後納指一定要死守以前必定要完成上述這個過程

==========

如果整體量能不想這樣發展納挑戰

==========

但仍面臨艱巨挑戰

==========

營運面臨較大挑戰

==========

有挑戰萬六的機會

==========

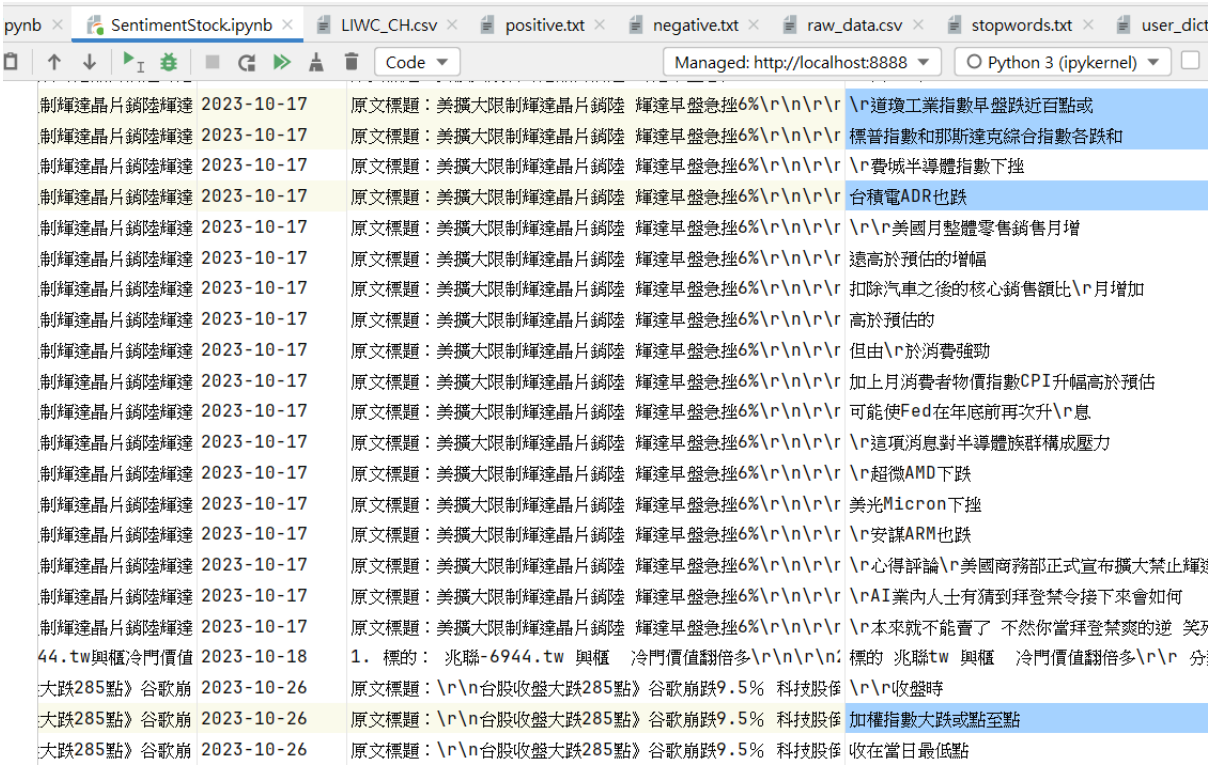
供應鏈傳出濟日報 記者李珣瑛挑戰

5. 我們發現到2023的10月中到11月情緒有顯著下降趨勢


6. 我們進一步分析該時間段文字雲

7. 發現有大量提及美債與下跌的資訊，我們進一步查看與美債相關負面詞


8. 發現文章提及道瓊指數下殺，與台股重挫



pynb | SentimentStock.ipynb | LIWC_CH.csv | positive.txt | negative.txt | raw_data.csv | stopwords.txt | user_dict

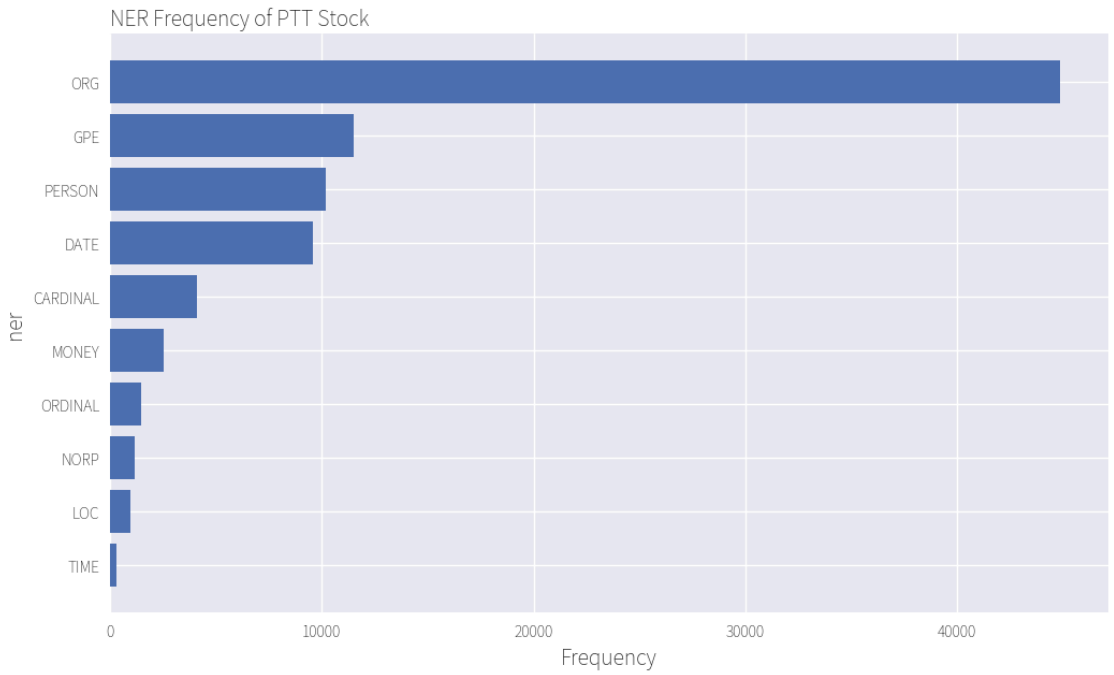
Code | Managed: http://localhost:8888 | Python 3 (ipykernel)

| | | | |
|---|---|---|---|
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | \r道瓊工業指數早盤跌近百點或 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | 標普指數和那斯達克綜合指數各跌和 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | \r費城半導體指數下挫 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | 台積電ADR也跌 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | \r\r美國月整體零售銷售月增 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | 遠高於預估的增幅 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | 扣除汽車之後的核心銷售額比\r月增加 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | 高於預估的 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | 但由\r於消費強勁 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | 加上月消費者物價指數CPI升幅高於預估 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | 可能使Fed在年底前再次升\r息 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | \r這項消息對半導體族群構成壓力 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | \r超微AMD下跌 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | 美光Micron下挫 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | \r安謀ARM也跌 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | \r心得評論\r美國商務部正式宣布擴大禁止輝達 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | \rAI業內人士有猜到拜登禁令接下來會如何 |
| 制輝達晶片銷陸輝達 | 2023-10-17 | 原文標題：美擴大限制輝達晶片銷陸 輝達早盤急挫6%\r\n\r\r | \r本來就不能賣了 不然你當拜登禁爽的逆 笑死 |
| 44.tw與權冷門價值 | 2023-10-18 | 1. 標的： 兆聯-6944.tw 與權 冷門價值翻倍多\r\n\r\n\r\n | 標的 兆聯tw 與權 冷門價值翻倍多\r\r 分 |
| 大跌285點》谷歌崩 | 2023-10-26 | 原文標題：\r\n台股收盤大跌285點》谷歌崩跌9.5% 科技股係 | \r\r收盤時 |
| 大跌285點》谷歌崩 | 2023-10-26 | 原文標題：\r\n台股收盤大跌285點》谷歌崩跌9.5% 科技股係 | 加權指數大跌或點至點 |
| 大跌285點》谷歌崩 | 2023-10-26 | 原文標題：\r\n台股收盤大跌285點》谷歌崩跌9.5% 科技股係 | 收在當日最低點 |

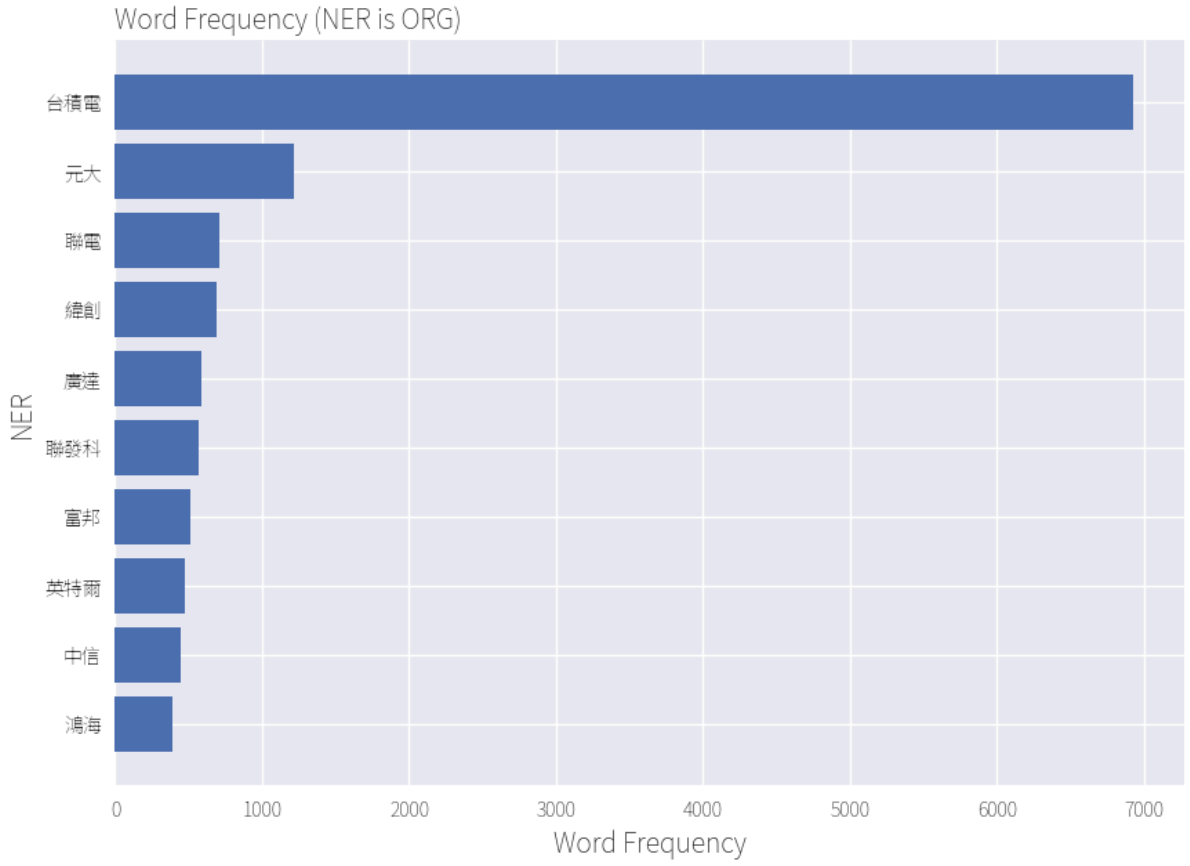
## 四、NER分析(CKIP)

CKIP是由中研院所開發出來的中文自然語言處理套件，效果較jieba好，故我們使用它作為接下來分析的工具。

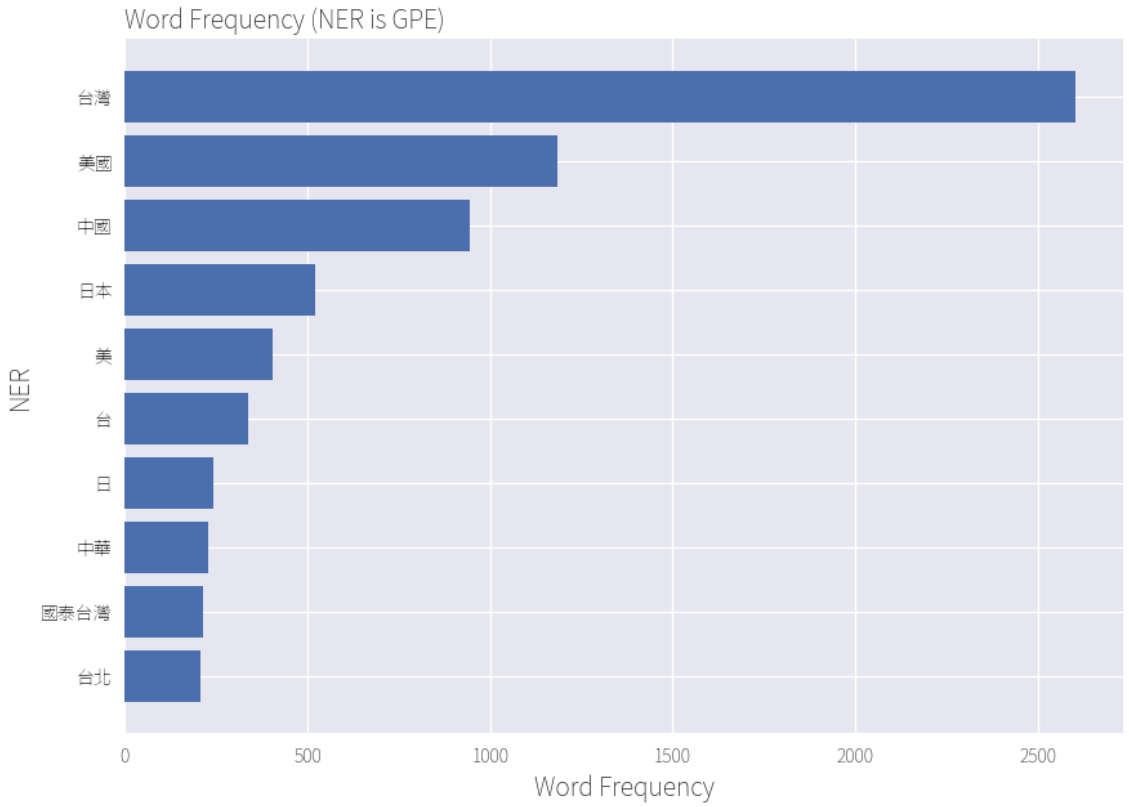
1. 首先，使用NER分析用來辨識實體種類。發現組織(ORG)出現頻率最高，第二高為地緣政治實體(GPE)，接下來依序為人物、日期、數字(CARDINAL)。



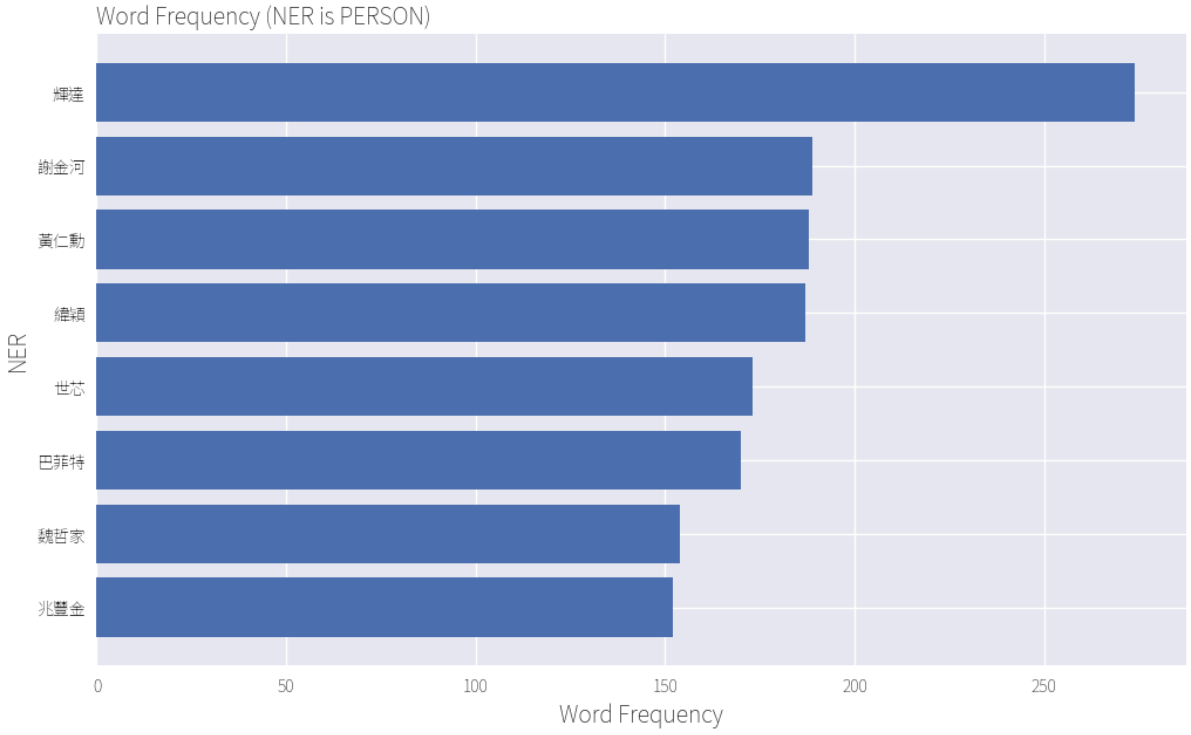

2. 針對上述結果，我們深入將出現頻率最高的組織(ORG)中的字篩選出來，可以看到「台積電」、「元大」、「聯電」為最常出現的關鍵字。



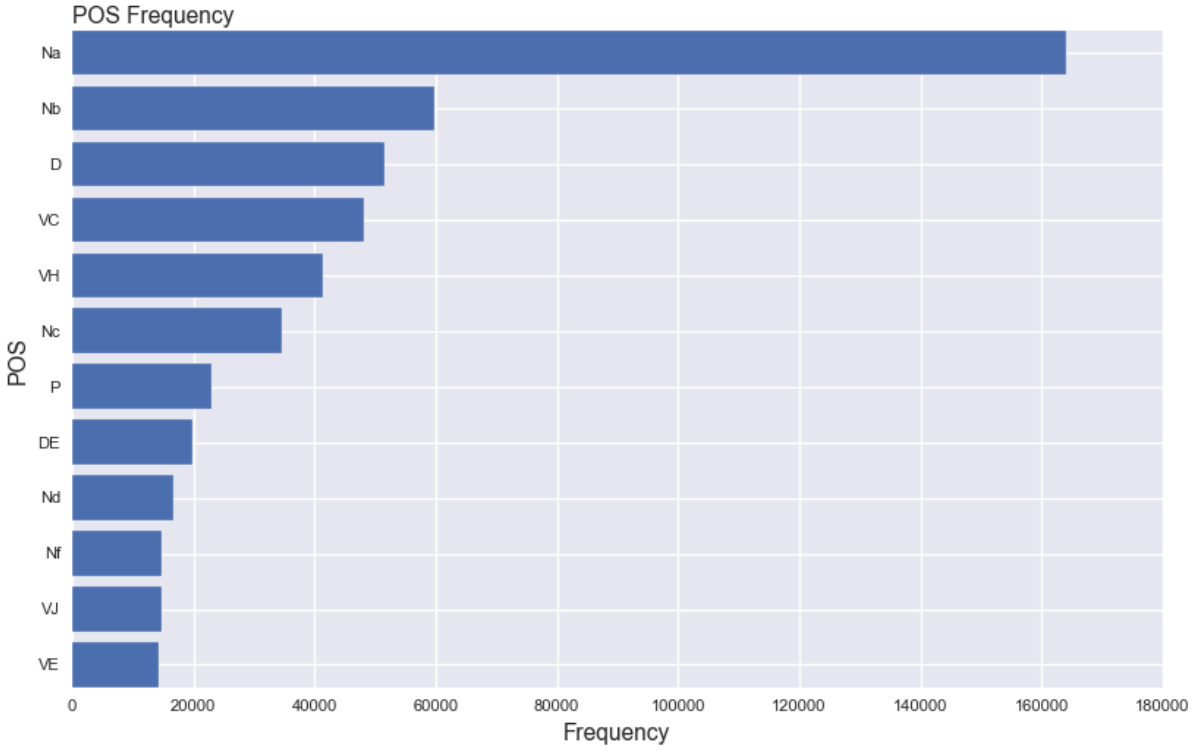
3. 再針對第二高的地理政治實體(GPE)進行分析，可以看到「台灣」、「美國」、「中國」為最常出現的關鍵字。




4. 針對第三高的人名PERSON進行分析，可以看到「輝達」、「謝金河」、「黃仁勳」為最常出現的關鍵字。但「輝達」、「緯穎」、「世芯」這些並不是人名，這是因為PERSON也有可能找到不是人名的詞，這是模型上的問題，若要進行後續分析的話，不建議以不正確的結果做。

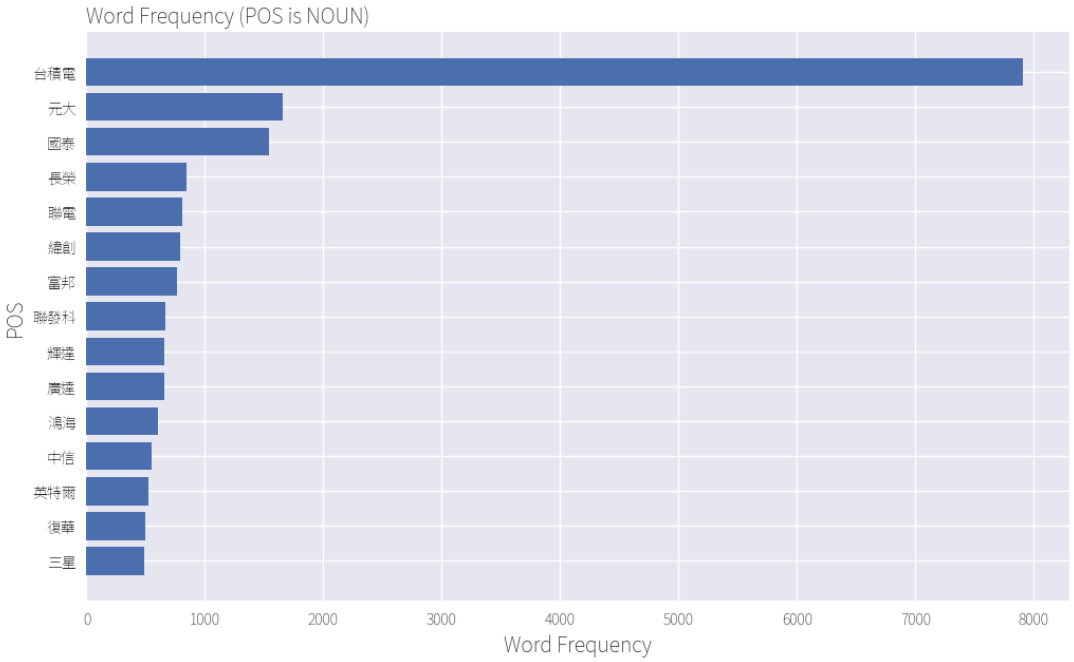


5. 接下來我們進行POS分析查看詞性頻率。可以看到「Na: 普通名詞」、「Nb: 專有名詞」、「D:副詞」為最常出現的詞性。

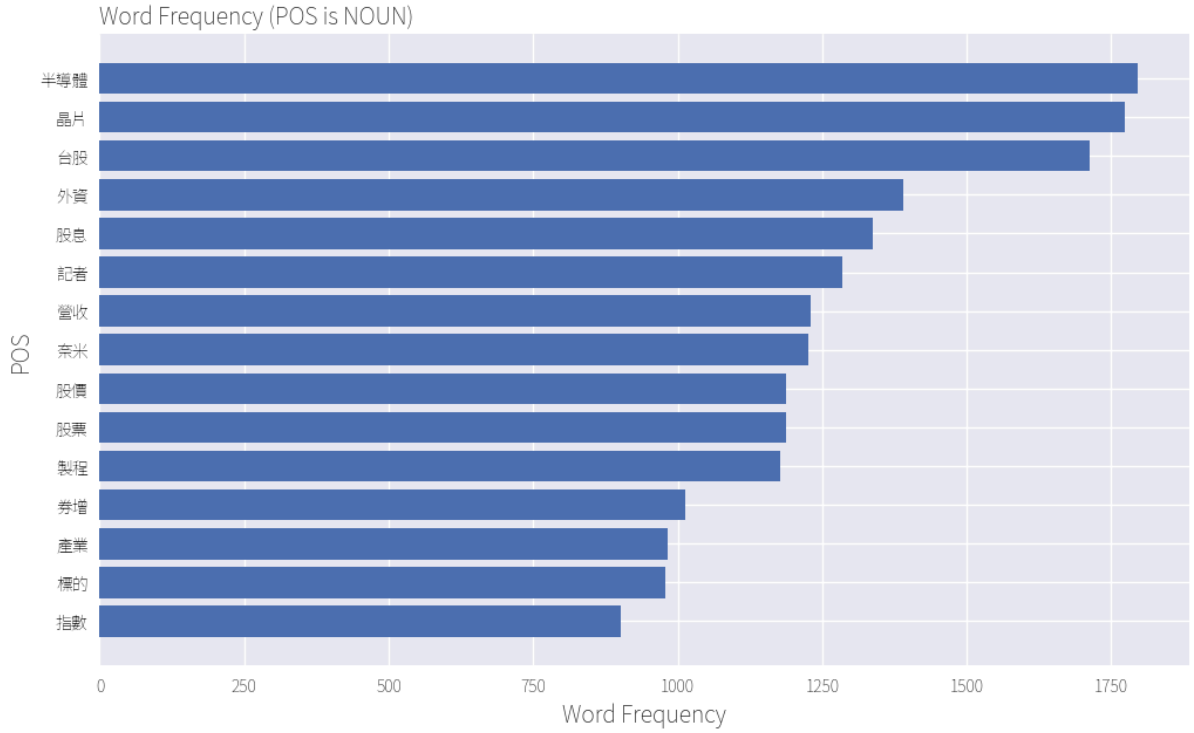



6. 再針對名詞進行分析，名詞分為「Na: 普通名詞」、「Nb: 專有名詞」等。

首先，我們針對專有名詞進行分析。發現談論台積電相關的PTT文章中，最常提到的除了本身關鍵字「台積電」以外，依序為「元大」、「國泰」等，推論「元大」、「國泰」是股票版文章時常附註在下面的相關股票，或者是相關ETF，因其常和台積電一起被提出討論，才會出現頻率那麼高。

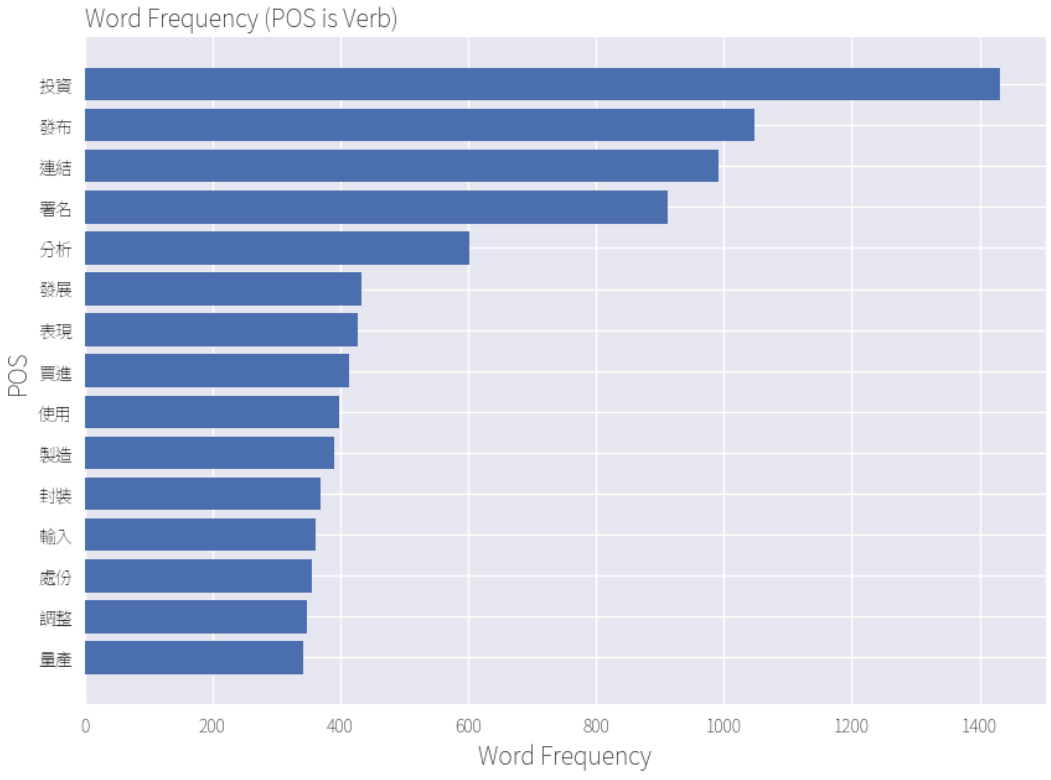


針對普通名詞分析，最常被提到的關鍵字為「半導體」、「晶片」、「台股」、「外資」等。

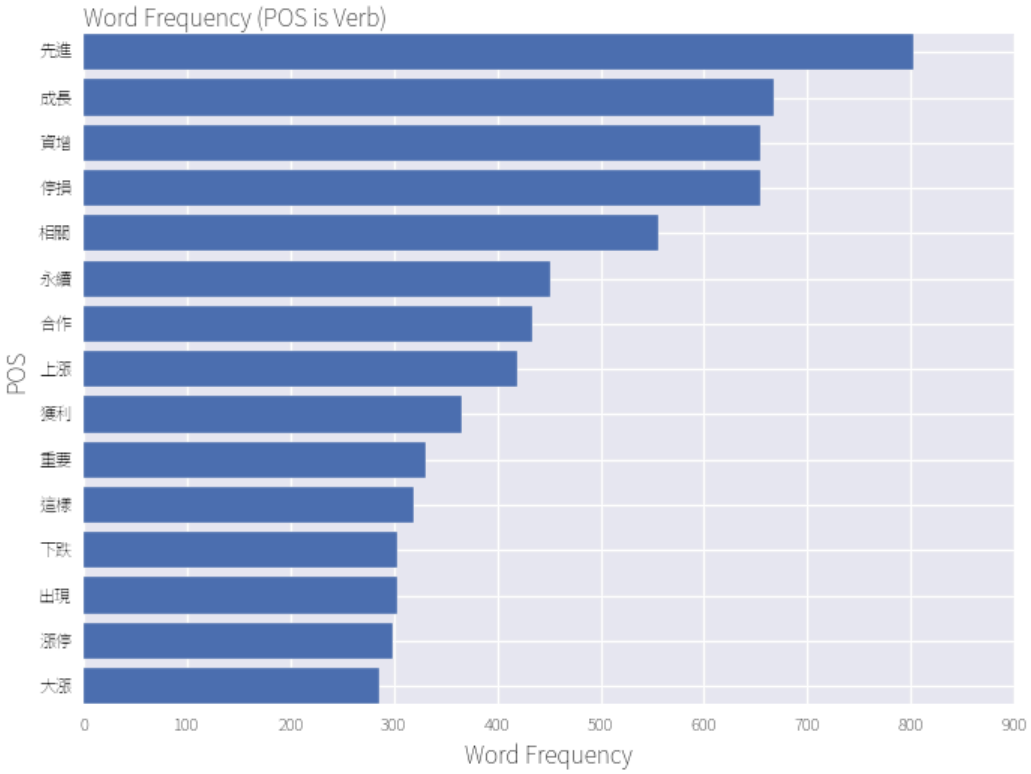



7. 針對動詞進行分析，動詞分為「VC: 動作及物動詞」、「VH: 狀態不及物動詞」、「VA: 動作不及物動詞」等。

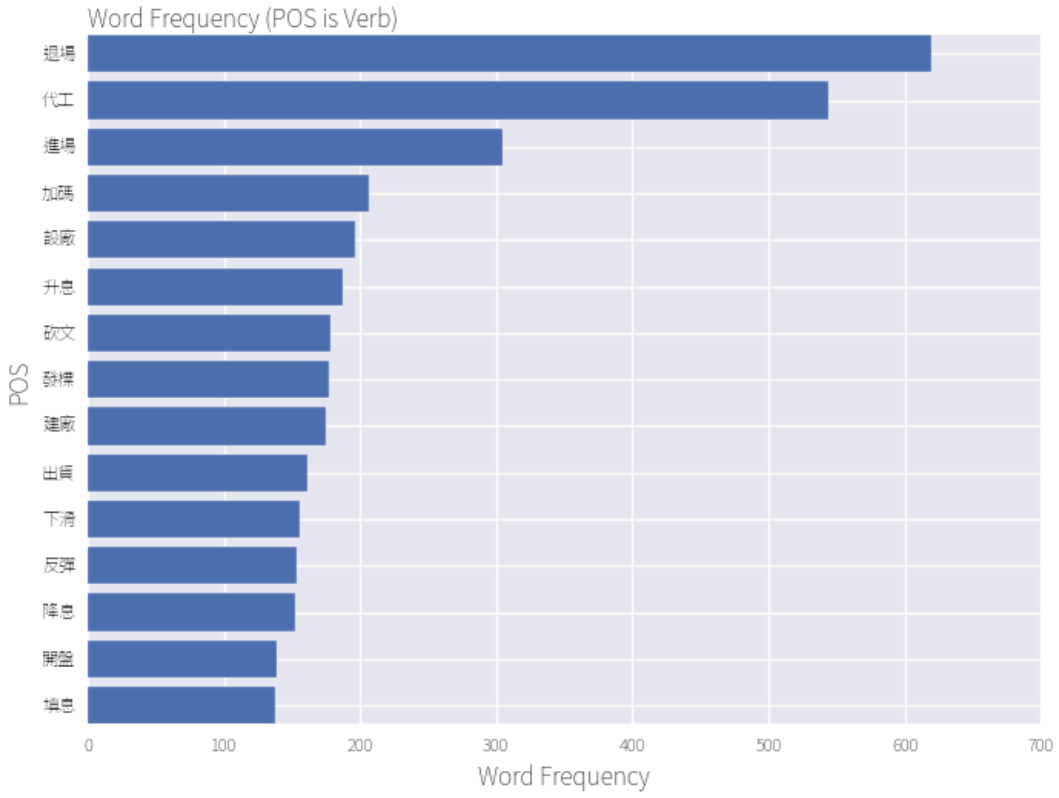
首先，針對動作及物動詞分析，最常被提到的關鍵字為「投資」、「發佈」、「連結」等。



針對狀態不及物動詞分析，最常被提到的關鍵字為「先進」、「成長」、「增資」等。




針對動作不及物動詞分析，最常被提到的關鍵字為「退場」、「代工」、「進場」等。



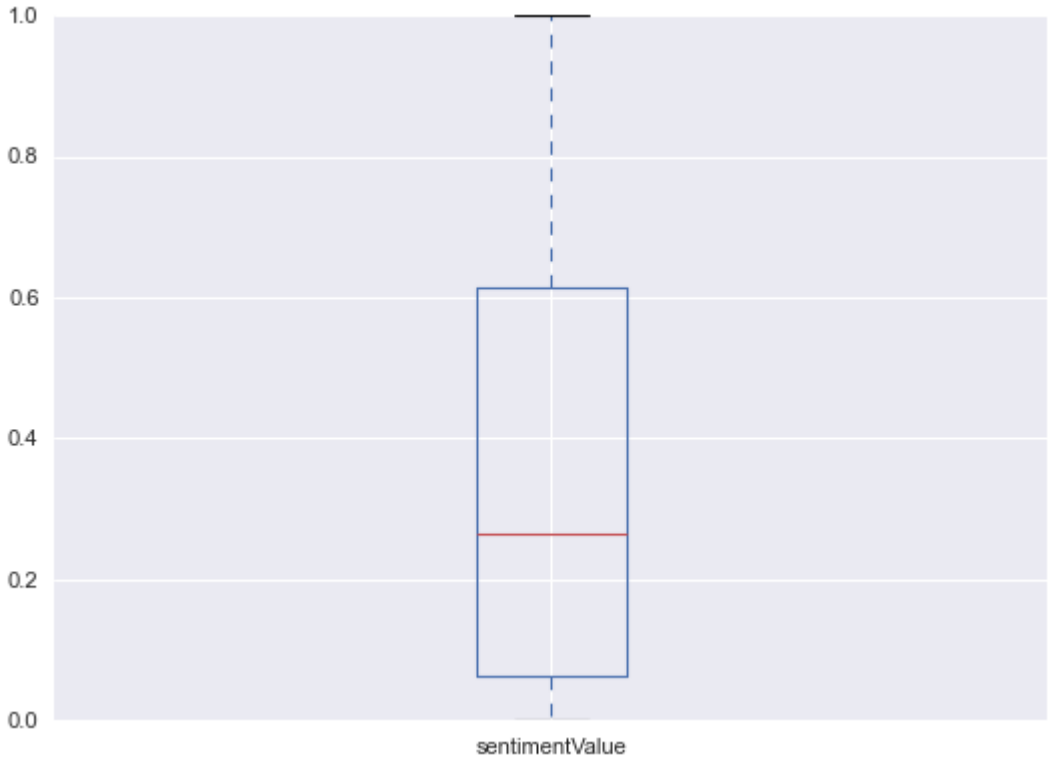
## 五、情緒分析(SnowNLP)

**輝達(但他不是PERSON)**

利用SnowNLP進行情緒分析，因SnowNLP為針對中文進行分類，並提取句子的情緒分類，情緒分數從負面到正面為0~1。

使用先前NER得到的「輝達」相關文章，看看輝達相關文章的情緒表現。

1. 情緒分布的盒狀圖

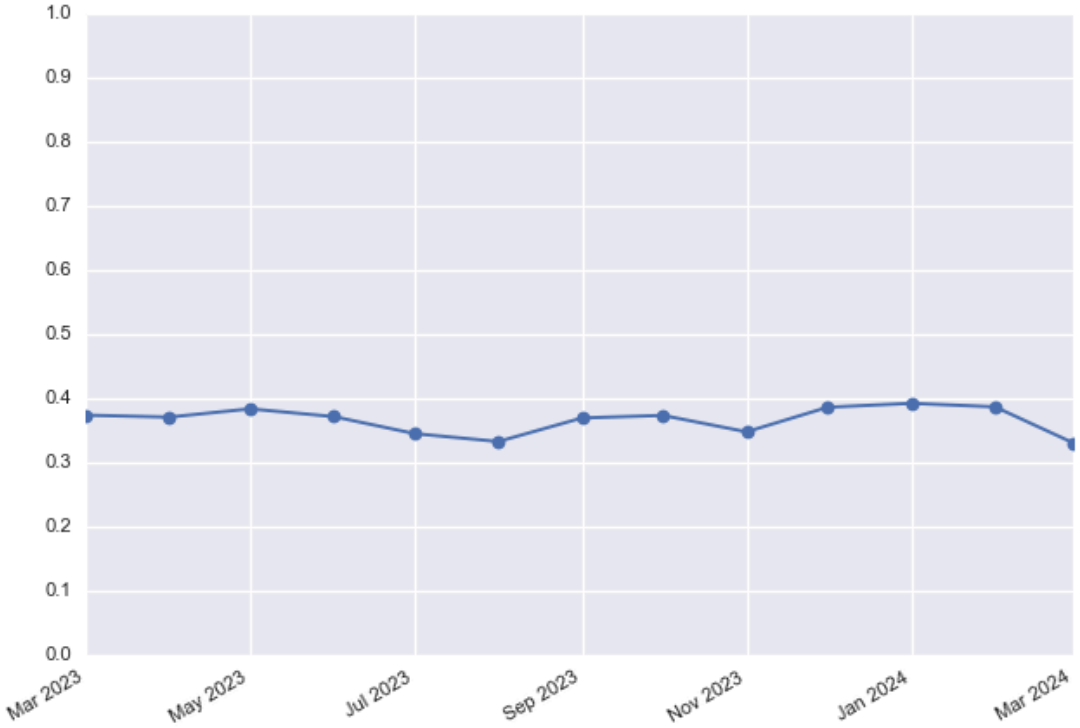



情緒分佈數值

| count | mean | std | min |
|---|---|---|---|
| 9105.0 | 0.3686392 | 0.321305 | 0 |

| 25% | 50% | 75% | max |
|---|---|---|---|
| 0.06236859 | 0.2636034 | 0.6148531 | 1 |

2. 平均情緒分數時間趨勢




從上圖中發現大部天數的情緒值都在0.3~0.4之間，最高值為2024-01的0.392412，但可以看出情緒都是趨於中性。

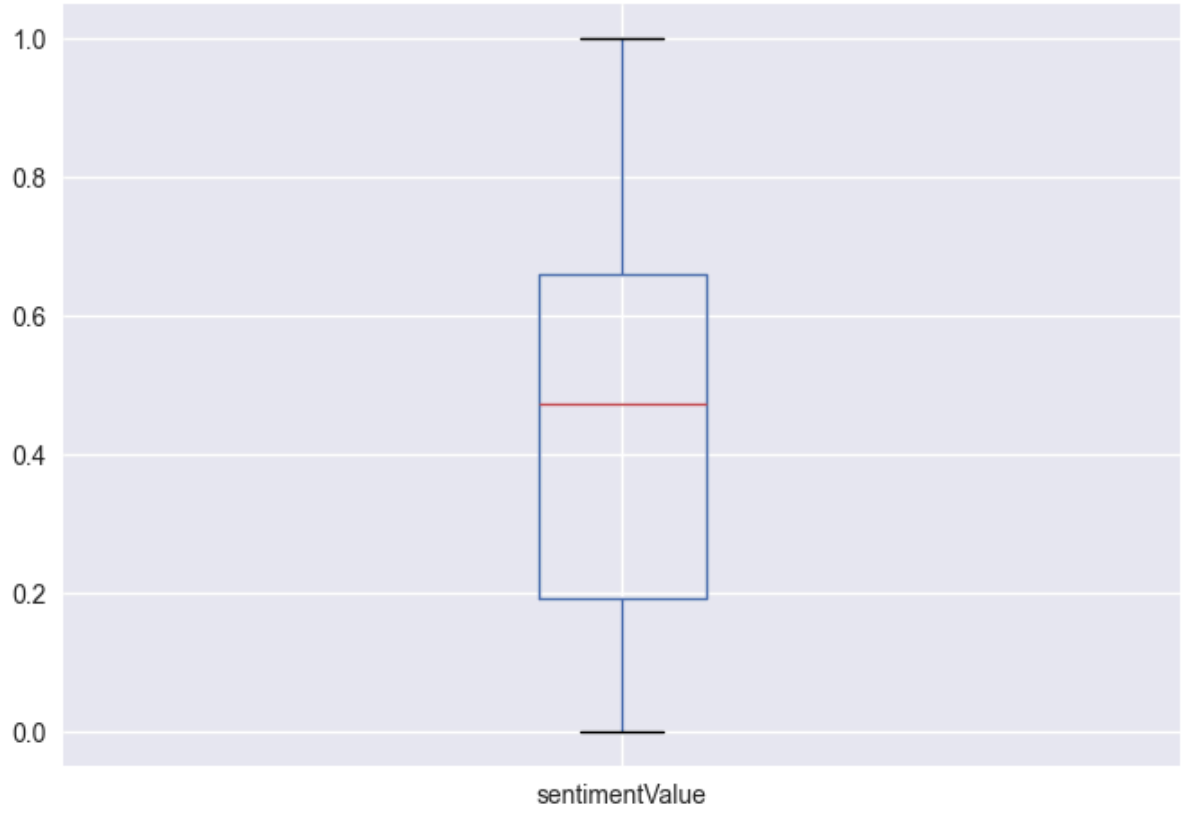
3. 最後，將一些多餘的字刪掉，並繪製正面詞彙文字雲，下圖得出最後的結果。可以看出和輝達有關的正向文章中，有「半導體」、「黃仁勳」、「台灣」、「中國」等詞彙常常被提到。


(設定：輝達相關文章，2024-1月(情緒分數最高月份)，max_words：30字)

## 台積電(ORG實體頻率最高)

使用先前NER得到的「台積電」相關文章，看看台積電相關文章的情緒表現。

1. 情緒分布的盒狀圖

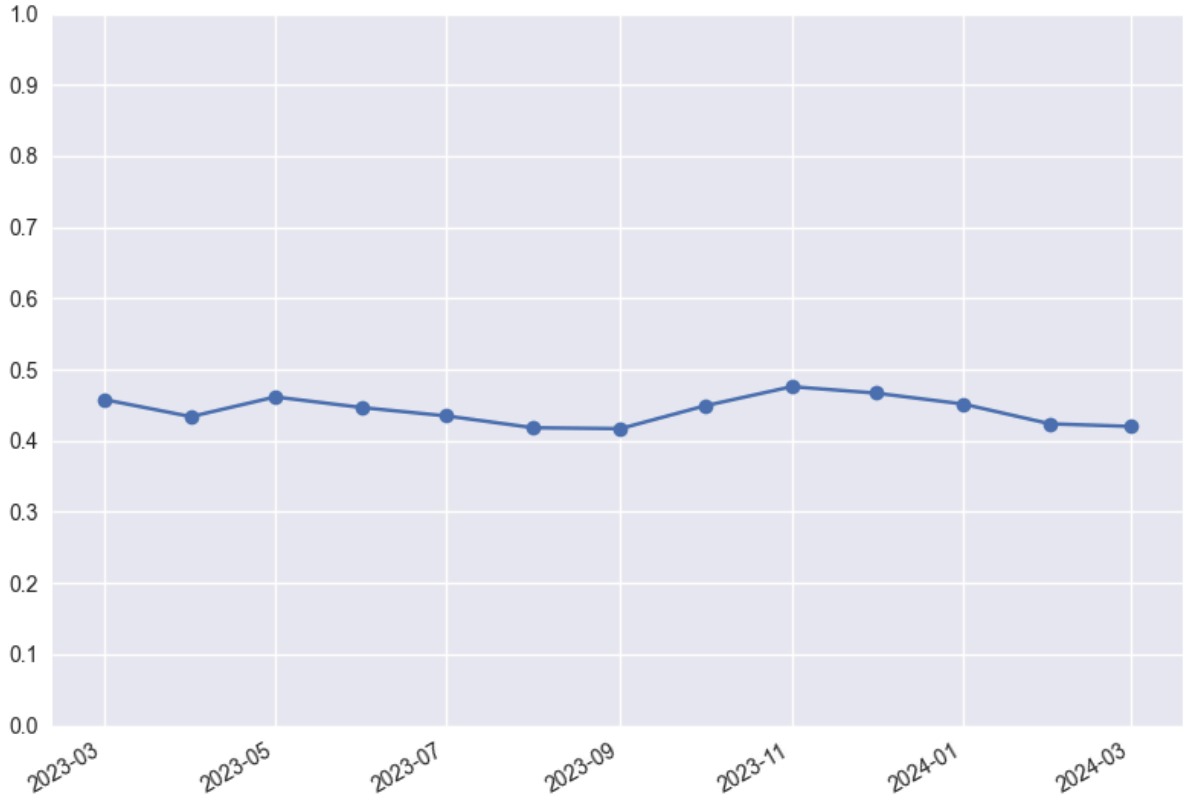



情緒分佈數值

| count | mean | std | min |
|---|---|---|---|
| 115913 | 0.4443745 | 0.2903989 | 0 |

| 25% | 50% | 75% | max |
|---|---|---|---|
| 0.1930021 | 0.4737672 | 0.6585312 | 1 |

2. 平均情緒分數時間趨勢




從上圖中發現大部天數的情緒值都在0.4~0.5之間，最高值為2023-11的0.476，但可以看出情緒都是趨於中性。

3. 最後，將一些多餘的字刪掉，並繪製正面詞彙文字雲，下圖得出最後的結果。可以看出和台積電有關的正向文章中，有「賣超」、「美債」、「股息」等詞彙常常被提到。

總結以上SnowNLP分析的結果:

## 輝達vs台積電

1. 公司性質說明
    - 輝達: 是一家以設計和銷售圖形處理器(GPU)為主的無廠半導體公司，目前為全球最大的GPU龍頭公司，執行長為黃仁勳。

    - 台積電: 為臺灣一家從事晶圓代工的公司，是全世界最大的專業積體電路公司，董事長為劉德音，總裁魏哲家。

2. 文字雲比較

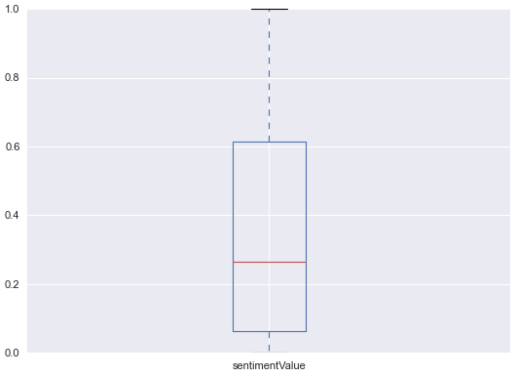
    雖然兩家都是半導體公司，但是可以明顯看出文章討論重點差異，輝達提到更多關於科技相關人名與技術，並多次提及執行長，而台積電的文章更多提到股票與美債相關。


輝達

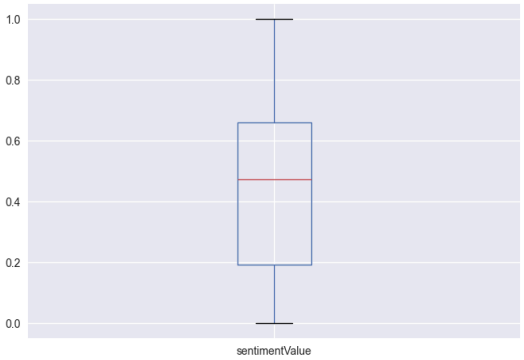

台積電

3. 情緒分布比較

    台積電情緒分布較輝達正面。輝達呈現右傾趨勢，台積電則趨近常態分佈。




輝達情緒分布




台積電情緒分布