金融大數據期末報告 組員: 陳瑜臻(411336163)、吳東衡(410145141)

**a.資料背景；變數定義；相關結論**

本研究資料來源係以python爬蟲方式抓取2023年5月14至31日於鉅亨網之新聞資料，共計1537筆。紀錄於CSV之五項變數包含新聞標題、連結、類別、內容及時間，其中類別變數為鉅亨網各篇新聞之來源種類，包含台股、台股新聞、國際股、基金等等共33項類別如圖1所示。

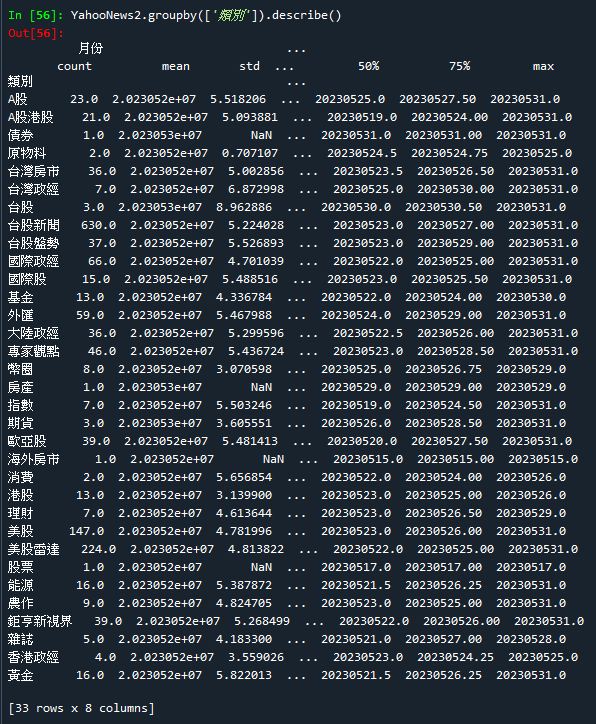


圖1. 類別變數

且在本研究結果部分有用到台積電及大盤指數之數據資料，其資料來源係台灣經濟新報資料庫(TEJ)所紀錄之收盤價日資料。藉由台積電與大盤指數做圖像化後以折線圖呈現如圖2所示，可以發現大盤指數與台積股價呈現高度連動性，然而在25號時台積電出現較大幅度上漲，因此於d章節將利用關鍵字出現次數來探討與股價是否也有相關連動性

由於鉅亨網都是個提供財經資訊與投資決策為主要服務內容之金融科技公司，因此新聞皆以金融與經濟為主要導向，其中更以股票類新聞佔絕大多數，由於在鉅亨網中若提到股票代號時，以台積電為例會寫為「2330-TW」做表示，因此在分析token時「TW」會出現非常多次，繪成文字雲如圖3所呈現TW會非常大，但TW僅代表股票代號表示法而沒有其他意義，所以本研究中將TW關鍵字拿掉以免妨礙分析。



圖3. 研究資料文字雲

**b.資料探索**

**1.觀察資料型態:發現欄位都為類別型態**

一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述

**2.檢查資料有沒有遺失值**

一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述

**3.處理時間欄位: 將時間拆成日期跟時間**

#將DataFrame中名為**"**時間"的欄位轉換為日期時間格式。

一張含有 文字, 螢幕擷取畫面, 軟體, 陳列 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述一張含有 文字, 軟體, 電腦圖示, 網頁 的圖片

自動產生的描述

**4.使用 groupby 方法分組並統計每天的新聞數量**

可以從表格中觀察到5月24號的新聞數量為最多，接著可以發現5月14、20、21、27、28號因為是周休二日，所以新聞數量較少

一張含有 文字, 螢幕擷取畫面, 陳列, 軟體 的圖片

自動產生的描述

**5. 使用 groupby 方法分組並統計不同類別的新聞數量**

可以從表格觀察到類別為**"**台股新聞**"**的新聞數量為最多

一張含有 文字, 螢幕擷取畫面, 軟體, 陳列 的圖片

自動產生的描述

**6. 接著觀察哪個類別的新聞標題最常提到AI**

可以觀察到類別為**"**台股新聞**"**的標題最常提到AI

一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述

**c.資料的整理與清理,欄位的新增**

* **資料的整理與清理**

鑒於資料抓取方式係以爬蟲的方式抓取，因此抓到的內文會有許多網頁結構的符號，因此我們需要先進行清理

1. **identify noise**

* 計算文本中雜訊的比例，並將結果存儲在impurity欄位中。

一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述

* 根據 impurity欄位的值對 DataFrame 進行排序，並選取前五行數據，以列出雜訊值最高的前五個文本。

一張含有 文字, 螢幕擷取畫面, 軟體, 陳列 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述

**2.Noise removal**

* import html: 引入 Python 的 html 模組，用於處理 HTML 轉義字符。
* text = html.unescape(text): 使用 html.unescape() 函式將 HTML 轉義字符轉換為相應的字符，例如 &amp; 轉換為 &。

1.使用正則表達式 re.sub() 函式將 Markdown 格式的 URL [Some text](https://....) 轉換為純文本的形式，只保留連結文字部分。

2.使用正則表達式將方括號中的文字或代碼替換為空格。

3.使用正則表達式將一些特殊字符或符號替換為空格。這些字符可能是獨立的，如 &#，但不包括像 #cool 這樣的字符。

4 使用正則表達式將多個連字符或加號替換為空格，例如 --- 或 ==。

5.使用正則表達式將連續的空格替換為空字符串，從而刪除多餘的空格。

6.使用正則表達式刪除文本中的斜線字符 /。

7.使用正則表達式刪除文本中的 HTML 標籤，例如 <div>、<p> 等。

8.return text.strip(): 返回經過清理和預處理的文本，使用 strip() 函式刪除文本前後的空格。

最後，df["clean\_text"] = df["內容"].map(clean) 將 clean 函式應用於 DataFrame 中的"內容"欄位，並將清理後的結果存儲在 "clean\_text" 欄位中。

**3.Data msking**

使用了 textacy.preprocessing.replace 模組中的一些方法來進行文本預處理。

以下是對每行文本進行的處理步驟的解釋：

1.使用 replace.urls 方法來替換文本中的網址（URL）。

2使用 replace.emails 方法來替換文本中的電子郵件地址。

3.使用 replace.hashtags 方法來替換文本中的主題標籤（hashtags）。

4.使用 replace.user\_handles 方法來替換文本中的使用者帳號（user handles）。

上述步驟主要是為了移除文本中的一些特殊字元或符號，例如網址、電子郵件地址、主題標籤和使用者帳號等，以便進一步進行文本分析或其他處理。

**4.文章清理後，可以重新計算文章雜訊**

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

**5.重新命名欄位，並且將計算雜訊值得欄位移除**

一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

自動產生的描述

* **欄位新增**

**1.新增長度欄位**

**2.接著進行描述性統計**

一張含有 文字, 軟體, 電腦圖示, 網頁 的圖片

自動產生的描述

**3.觀察每個類別新聞內容長度平均值**

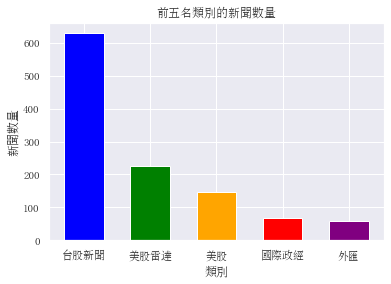
一張含有 文字, 螢幕擷取畫面, 軟體, 陳列 的圖片

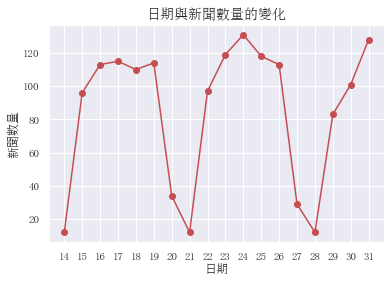
自動產生的描述

一張含有 文字, 軟體, 電腦圖示, 多媒體軟體 的圖片

自動產生的描述

**4.資料探索性分析**





一張含有 文字, 螢幕擷取畫面, 行, 鮮豔 的圖片

自動產生的描述

**d.資料內容探勘分析**

表1為本研究分析結果所紀錄之資料，包含關鍵字「台積電」、「輝達」、「上漲」及「下跌」於全部新聞之內文中所出現次數，並且藉由關鍵字「上漲」及「下跌」出現次數相減後得出「上漲下跌相差次數」以做後續研究結果比較。另外再藉由TEJ抓出2023年5月15至5月31之台積電股價及大盤指數，其中週末未開盤之股價係以禮拜五之收盤價做代替。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **表1 相關研究結果紀錄** | | | | | | | |
| **日期** | **台積電**  **出現次數** | **輝達**  **出現次數** | **上漲**  **出現次數** | **下跌**  **出現次數** | **上漲下跌相差次數** | **台積股價** | **大盤指數** |
| **15號** | 14 | 8 | 37 | 33 | 4 | 495.5 | 15475 |
| **16號** | 35 | 6 | 82 | 44 | 38 | 505 | 15673 |
| **17號** | 44 | 24 | 80 | 58 | 22 | 519 | 15925 |
| **18號** | 19 | 18 | 84 | 39 | 45 | 530 | 16101 |
| **19號** | 12 | 2 | 84 | 35 | 49 | 532 | 16174 |
| **20號** | 8 | 1 | 31 | 38 | -7 | 532 | 16174 |
| **21號** | 1 | 5 | 0 | 5 | -5 | 532 | 16174 |
| **22號** | 25 | 4 | 20 | 28 | -8 | 531 | 16180 |
| **23號** | 14 | 15 | 70 | 44 | 26 | 530 | 16188 |
| **24號** | 16 | 12 | 46 | 84 | -38 | 525 | 16159 |
| **25號** | 23 | 79 | 73 | 56 | 17 | 543 | 16292 |
| **26號** | 32 | 59 | 81 | 51 | 30 | 566 | 16505 |
| **27號** | 3 | 51 | 39 | 10 | 29 | 566 | 16505 |
| **28號** | 2 | 5 | 2 | 1 | 1 | 566 | 16505 |
| **29號** | 15 | 42 | 32 | 11 | 21 | 568 | 16636 |
| **30號** | 13 | 57 | 64 | 26 | 38 | 566 | 16622 |
| **31號** | 20 | 44 | 43 | 78 | -35 | 558 | 16578 |

經由關鍵字「台積電」在研究資料中不同日期出現的次數以長條圖繪製，並對應到折線圖為台積電股價於圖4所示，其中橫軸20、21及27、28為周休二日，因此新聞量較少，股價部分也持平不動。然而可以看出在16至17號台積電股價的上漲，與新聞出現次數的提高有較強連動性。不過在24至26號台積股價的上漲並未明顯與新聞出現次數有連動關係，因此續作往後探討。

關鍵字「輝達」在研究期間各天期出現的次數與台積電股價對應繪於圖5所示，可以看出在25號出現次數突然大幅提升，同時疑似有帶動台積股價上漲之跡象，因此回頭查看25號的新聞，經查發現，當日多數新聞提到有關輝達的事件「昨晚美股下跌，然而輝達盤後公布營收優於市場預期，因此盤後股價逆勢上漲24%。並且宣布AI發展帶動高階處理器需求暴增，使第二季銷售有望增加64%至110億美元，遠超過市場預估的72億美元」由於當日該利多消息傳遍而激勵台積電股價大幅上漲。

圖6中直條圖為「上漲下跌相差次數」，並將其與大盤指數對應比較，結果發現上漲與下跌相差次數跟股價有極度明顯連動關係，主要原因係當天若大盤上漲，新聞自然會提到較多跟「上漲」有關的字眼，大盤下跌則新聞也是提到較多「下跌」字眼，從圖中可以看出只有在24及31號大盤是下跌，因此關鍵字「下跌」在新聞中也明顯出現較多次。

圖7至8係本研究中先藉由找出有關鍵字「台積電」、「輝達」的新聞共185篇、136篇，並分別針對這185篇與136篇進行Token分析得出之結果，研究發現台積電Token前五名關鍵字為「AI」、「美國」、「上漲」、「指數」、「億元」；輝達Token前五名關鍵字為「AI」、「美國」、「指數」、「上漲」、「市場」。兩者Token前五名中有四個關鍵字相同，表示「台積電」與「輝達」在本研究資料中具有高度相關性，且其Token第一名皆為關鍵字「AI」，也可以驗證AI與台積電及輝達的密切相關性。最後再將該Token作為文字雲繪於圖9及圖10

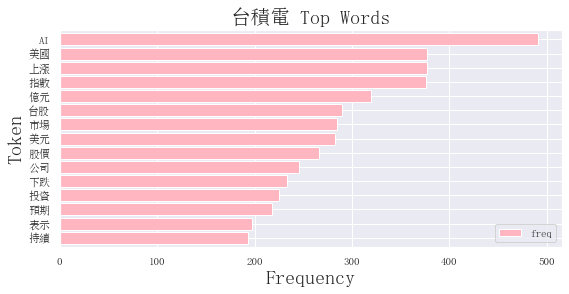


圖7. 台積電Top Words

一張含有 文字, 螢幕擷取畫面, 繪圖, 數字 的圖片

自動產生的描述

圖8. 輝達Top Words



圖9. 台積電新聞文字雲



圖10. 輝達新聞文字雲