PTT Tag

IRTM Final Project

Group 4

陳韋傑	江采嬪	石子仙	王博奕	莊啟宏
會計三	會計四	化工四	財金三	工管三
B07702011	B06702046	B06504104	B07302230	B07701222

一、動機與目的

我們希望能夠了解網友對台灣企業或個股的見解，但我們觀察到目前 PTT 的搜尋功能，僅能依照文章標題進行搜尋，無法根據文章及其內容進行搜尋；同時，PTT 現存的文章分類僅仰賴作者自行在標題中進行人工標記的方式（例如心得、新聞……），相較其他新聞網站或是網路論壇，缺少自動主題分類、相關文章推薦等等功能。因此，我們希望能夠替 PTT 的文章進行分群，並替文章標上與內文相關的主題標籤。我們預期能夠建立主題式的文章瀏覽，以及在使用者看完文章之後，能夠顯示該文章的主題標籤與相關企業標籤，為 PTT 增添更具人性化、深入搜索的擴充設計。最後以 tkinter 設計簡易的介面，以實現上述功能。

二、程式設計邏輯

（一）資料爬取

使用 bs4 的 BeautifulSoup 套件來爬取 PTT 的 stock 看板，文章選取的時間範圍為 2020/03/19 ~ 2021/01/14，共 30904 篇。根據 Heap's Law 與 Zipf's Law：文章數量越多，dictionary size 呈指數增加，也造成更多的冗詞，造成運算的遲滯與低效率，因此我們決定刪減文章集的數量。考量 ptt 原分類中(創作、請益、其他...)，「標的」分類最能表達網友對個股公司的主觀見解，因此我們鎖定「標的」分類中，文章長度 >= 300 字的文章，共計 2445 篇。

	author	title	time	content	url
0	justicide (有話直說)	Re: [標的] 美股 AMCX (AMC Networks) 不是AMC	Thu Jan 14 23:03:59 2021	從站上30元那天初次PO文至今，AMCX來到了4X元。想必大家也開始考慮是否該獲利了結。\n...	https://www.ptt.cc/bbs/Stock/M.1610636643.A.CB...
1	MiniArse (型男)	Re: [標的]：美國石油公司 OXY.US	Thu Jan 14 23:28:28 2021	來到 24.x 囉...\n\n WTI 西德洲原油期貨已經差不多回到去年 1...	https://www.ptt.cc/bbs/Stock/M.1610638110.A.21...
2	poopooTaiwan (便便台灣)	Re: [標的] 台積電 逃命空	Thu Jan 14 23:31:12 2021	相信今天有空的應該都有賺錢，\n但是現在ADR鬼神飆漲，\n明天空單還留著的快停力回補，\n...	https://www.ptt.cc/bbs/Stock/M.1610638278.A.16...
3	kksis (流浪人生)	Re: [標的] 台積電 追高多	Thu Jan 14 23:38:57 2021	今日又進場買了100股總共600股，我知道上看1千感覺很可笑，但有夢最美希望相隨，買\n台積...	https://www.ptt.cc/bbs/Stock/M.1610638739.A.6F...
4	agqee (ptt)	Re: [標的] 4976 佳凌 嘎空多	Thu Jan 14 23:44:28 2021	※ 引述《agqee (ptt)》之銘言：\n: 1. 標的：4976 佳凌\n: 2. 分...	https://www.ptt.cc/bbs/Stock/M.1610639070.A.44...
...	...	...	...	...	...
2440	amin82 (三省吾身)	Re: [標的] 3406玉晶光，強力買進	Thu Jan 2 21:20:01 2020	回顧原po所說的，真的很準，短時間就上605元了，而且外資的目標價\n\n日系：668元\n...	https://www.ptt.cc/bbs/Stock/M.1577971203.A.01...
2441	a777starmy (呼力贏大師(大輸))	[標的] 2201裕隆，買進後會立即賺錢嗎?(立積)	Thu Jan 2 14:11:52 2020	1. 標的：2201 裕隆\n2. 分類：多\n3. 分析/正文：\n\n籌碼愈來愈穩定，...	https://www.ptt.cc/bbs/Stock/M.1577945519.A.F7...
2442	Sunrisesky (伴讀書僮)	[標的] 合晶 長多	Wed Jan 1 12:01:22 2020	1. 標的：6182合晶\n\n2. 分類：多/空/請益/心得\n長多\n\n3. 分析/正...	https://www.ptt.cc/bbs/Stock/M.1577851284.A.C8...
2443	Jamire (^^)	Re: [標的] 7566 亞果遊艇	Mon Aug 17 22:30:33 2020	董事捲入全台最大洗錢案 亞果遊艇發聲明澄清\n2020-08-17 15:39 經濟日報 /...	https://www.ptt.cc/bbs/Stock/M.1597674636.A.37...
2444	idome (你住海邊是不是?)	[標的] 1799 易威 藥證多	Mon Aug 17 23:14:05 2020	\n1. 標的：1799 易威\n\n2. 分類：多\n\n3. 分析/正文：\n\n8/1...	https://www.ptt.cc/bbs/Stock/M.1597677248.A.45...

2445 rows × 5 columns

（二）文章前處理

1. 由於我們處理的是中文文章，我們將數字、英文字及助詞、連接詞剔除，並以逗號分句，再使用 monpa 罔拍中文斷詞套件進行斷詞，此套件的資料庫涵蓋巨量專有名詞及特殊字詞，能夠更精確地做出斷詞。

	author	title	time	content	url	tokens
0	justicide (有話直說)	Re: [標的] 美股 AMCX (AMC Networks) 不是AMC	Thu Jan 14 23:03:59 2021	從站上30元那天初次PO文至今，AMCX來到了4X元。想必大家也開始考慮是否該獲利了結。\n...	https://www.ptt.cc/bbs/Stock/M.1610636643.A.CB...	[站上, 那, 初次文, 至今, 來到, 元想必, 大家, 開始, 考慮, 該, 獲利, 以...
1	MiniArse (型男)	Re: [標的]：美國石油公司 OXY.US	Thu Jan 14 23:28:28 2021	來到 24.x 囉...\n\n WTI 西德洲原油期貨已經差不多回到去年 1...	https://www.ptt.cc/bbs/Stock/M.1610638110.A.21...	[來到, 西德洲, 原油, 期貨, 已經, 差不多, 回到, 去年, 月底, 水準, 當時,...
2	poopooTaiwan (便便台灣)	Re: [標的] 台積電 逃命空	Thu Jan 14 23:31:12 2021	相信今天有空的應該都有賺錢，\n但是現在ADR鬼神飆漲，\n明天空單還留著的快停力回補，\n...	https://www.ptt.cc/bbs/Stock/M.1610638278.A.16...	[相信, 今天, 應該, 賺錢, 現在, 鬼神, 飆漲, 明天, 空單, 著, 停力, 回補...
3	kksis (流浪人生)	Re: [標的] 台積電 追高多	Thu Jan 14 23:38:57 2021	今日又進場買了100股總共600股，我知道上看1千感覺很可笑，但有夢最美希望相隨，買\n台積...	https://www.ptt.cc/bbs/Stock/M.1610638739.A.6F...	[今日, 進場, 股, 總共股, 知道, 看, 感覺, 可笑, 最, 希望, 相隨, 台, ...
4	agqee (ptt)	Re: [標的] 4976 佳凌 嘎空多	Thu Jan 14 23:44:28 2021	※ 引述《agqee (ptt)》之銘言：\n: 1. 標的：4976 佳凌\n: 2. 分...	https://www.ptt.cc/bbs/Stock/M.1610639070.A.44...	[引述, 銘言標佳凌, 分類, 分析, 正文, 很多, 覺得, 值得, 個, 價位, 因此,...

2. 掃過每篇文章，建造一個 term dictionary，並記錄每個詞出現的文章編號，長度約為 30000 字。

```
df_df
```

	t_index	term	df	posting
0	1	丁丁	2	[795, 1260]
1	2	丁二	1	[272]
2	3	丁二烯	1	[270]
3	4	七傷拳	1	[2166]
4	5	七八	7	[706, 712, 849, 1374, 1961, 2239, 2324]
...	...	...	...	...
31978	31979	龐大千	1	[1300]
31979	31980	龜密	3	[89, 92, 119]
31980	31981	龜山	1	[879]
31981	31982	龜毛	1	[329]
31982	31983	龜笑	2	[1026, 1031]

31983 rows × 4 columns

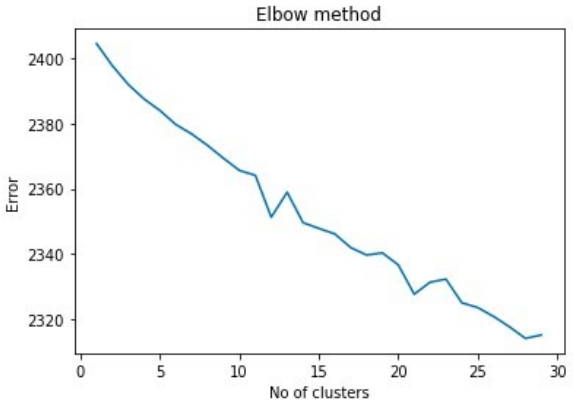
3. 替每篇文章建立 tf-idf vector，紀錄各詞的 unit tf-idf。

	term	tf	t_index	df	tf-idf
21	下跌	1	423	182	0.068113
60	主力	1	1059	268	0.057967
42	之前	1	1146	427	0.045754
22	今天	1	1622	929	0.025372
8	似乎	1	1829	116	0.079923
...	...	...	...	...	...
32	開始	1	29658	708	0.032495
58	階段	1	30238	79	0.089995
20	隨即	1	30266	14	0.135365
51	靠攏	1	30722	6	0.157581
37	馬上	1	31304	99	0.084078

73 rows × 5 columns

（三）文章分群

將所有文章的 tf-idf vector 丟入 K-means 套件，並算出分成 1~30 群時的 Error。由於人不傾向使用過多的群數，因此我們將範圍限定在 1~30 群。利用 elbow method 找出 error 逐漸平緩或低谷的點，我們主要有三個候選的群數：12, 14, 21。我們將 2445 篇文章分別分成 12, 14, 21 群，利用 chi-square feature selection 替每群選出 100 fearure words。最後我們發現 21 群的分群結果所選出的特徵字組內相關性最高。因此我們決定以 21 群進行分群。

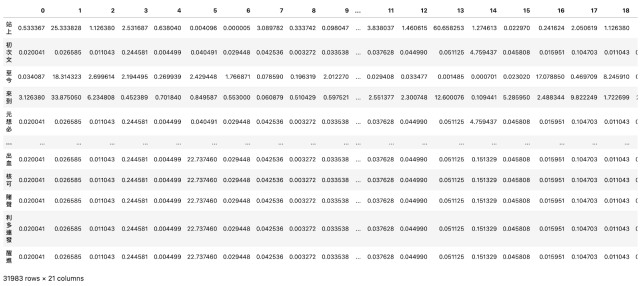


此方法為 hard clustering，即每篇文章僅會對應到一個分類，而之後的文章標籤僅會包還該分類當中的特徵詞。

（四）文章標籤

利用 chi-square feature selection 替每群選出 100 feature words，並由人工篩選，剔除重複及不具意義的詞彙，為每群分類留下 20-50 個標籤。最後我們針對每篇文章貼上標籤，該標籤是屬於該文章所在的類別中的特徵字，且該文章有出現的詞彙。

$$\chi^2(D,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$



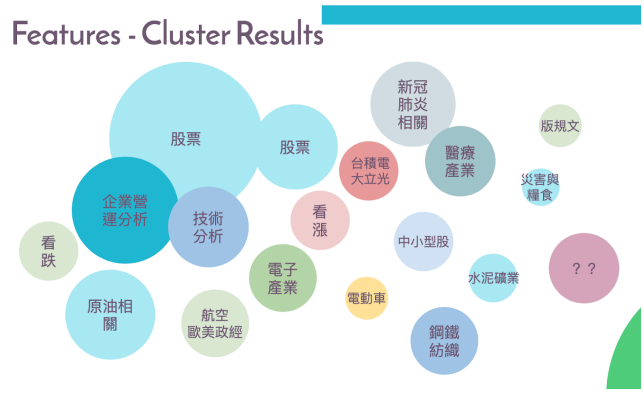
31983 rows × 21 columns

（五）企業標籤

除此之外，我們建立了台灣所有上市公司的名單，並替文章標上內文中有出現的公司，以便使用者針對感興趣的企業做進一步的查詢。

三、分群結果

依照每群文章選出的特徵字，可以替文章分類取名，分類結果如下：股票、企業營運分析、技術分析的文章集較多，原油、新冠肺炎、醫療產業的討論文章也較多；惟其中兩群的特徵字相關性較不明顯，無法概括該群體，因此歸類為「未知」。



利用 chi-square feature selection 的結果當中，大多類別的相關性明顯，以「電動車」、「新冠肺炎」兩群為例：

電動車
特斯拉,電池,電動車,車廠,續航,馬斯克,閉環,車主,**蔚來**,車子,車企,換電,二手車,**寧德**,特斯拉股東,降臨,租用,底盤,**蔚能**,**秦力洪**,逍遙法外,充電,戒慎,上海,駕駛,加碼股,里程,**李斌**,車輛,保量,拆股,馬達,車價,變故,耐用度,主觀,車系,二蔚,大昌,賤價,燃油車,充電樁,元合,感知,自研,昂貴,衰減,電站,標普,能源,國產化,二手,回推,運營,砍價,使命,藍圖,傳奇,用車,液冷,打氣,康友

新冠肺炎
疫苗,口罩,**國光生**,國家隊,恆大,確診,**康那香**,試驗,臨床,高端,國光,感染,如火如荼,生化,國家,人數,成果,封鎖,病毒,防疫,歐美,幹桿,物競天擇,英國人,封國,延續,台康,流感,防禦,**解藥**,動員,生物,炒完,微型,感冒,死亡,死亡率,菲律賓,對抗,解盲,入境,疫苗廠,**陳時中**,侵犯,索羅斯,上台,提款機,鑽木取火,以物易物,義大利,小道瓊,案例,研發,壓驚,美國,試劑,人權,日本,生技展,石器,細胞,免疫,錯覺,醫藥,管制,北韓,增添,疫情,現代,抗體,衛生,網購化,實踐期,註冊中心,端午節,金融周刊

而少數特徵詞匯相關性不明顯的類別，舉例如下：

未知群一
華安,伴讀,讀書,母單,穩懋,合晶,點位,二哥,滾量,雷同,破底,隊長,軍團,薩諾斯,散亂,借券,戰線,吃貨區間,短彈,抱單,毛手毛腳,耳提面命,排擠,休息段,界線,認養

未知群二
放款,適足率,資本,分類,組成,鑽戒,試妝,覆蓋率,必要性,下載年報,遭受,當場,英尺,入手價,大潤發,小三美日,鑽研,猶疑,三年級,準備金,低潮期,攻守,現金流充裕,收掉,兼備,梅西,備抵,姪子,全食,街邊店,同時線,奢華,商場,小學,免稅店,銷售中心,高鑫,嚴選,資產,變差,和會,提撥,減免,實體店,精品,分店,債券,上半年稅,產險,驚奇,新臺幣,風險性,簡稱,平方,比率,化妝品,人流,標榜,入股,出處,房租,逾期,巨擘,存款,銷售額,玩法,階層,績效,提列,試駕,概況,隔年,物業,逛街,證券,擔保,表格,大標,股債

六、應用介面設計

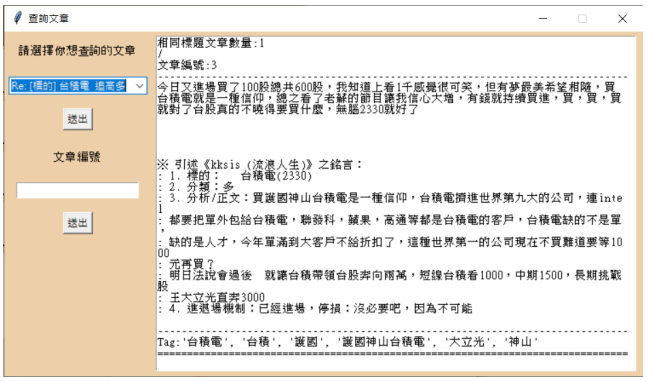
我們將上述針對 ptt stock 版的分類功能做成介面，以達到下列幾個目標：

1. 使用者可以下滑式選單、或輸入文章編號來查詢相關文章
2. 文章下方顯示該文的標籤與企業標籤
3. 透過特定標籤、企業標籤，來搜尋所有「文章內文」中有出現該詞彙的相關文章

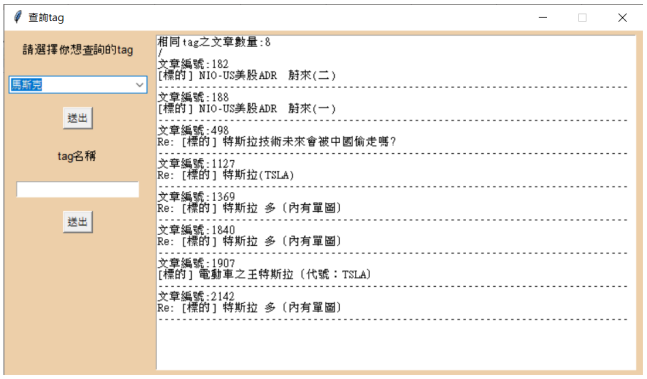
介面設計以圖文說明如下：

- 首先，使用者將會導入第一個介面，可以選擇查詢文章、查詢 tag、查詢公司以及離開

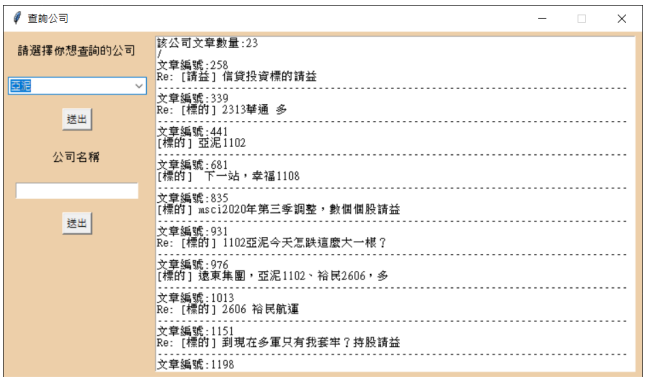
- 查詢文章：使用者有兩種方法來查找文章，用捲單或者用文章編號。其中文章編號的使用時機大多為查詢完 tag 和公司時得到標題時也會得到一組編號。



- 查詢 tag：一樣有兩種選擇，結果會吐回編號以及標題。



- 查詢公司：同查詢文章。



介面操作之介紹影片可參照連結：
https://youtu.be/3PJanHvEdhE

七、困難

1. 資料量過於龐大，運算資源不足

原始文章總數 30904 篇，在 K-means 中丟入文章數 x 字數的二維矩陣。大約為 30000x140000 的二維矩陣，非常巨大的資料量，無法負荷此計算量，因此才最後只挑出 2445 篇文章進行分群以及後續的實作

2. 難以決定適當的分群數量

我們使用 elbow method 挑選出較為不錯的 knee，並將每個 knee 實作一次，手動挑選出我們主觀認為較為好的分群數量，此動作需要透過我麼自己手動觀察，無法良好的透過程式分析選出一個最好的分群數量。

3. Feature Selection 之結果並非完全可解釋

雖然有些群 tag 的內容非常相近且符合某一種特定的主題，但也有少數群中的 tag 是人類難以辨識的內容，且彼此之間看似無關聯。

4. 可能是作者當初發表了與企業標的無關的內容

我們只透過文章標的進行篩選，沒有對其內容進行分析是否與其企業標的相關，因此若發文者所發的內容是無關的，我們無從得知。

5. 無法隨時間更新分群結果

每當有人新貼出文章時，都必須再將其爬下來，並重新計算 tf-idf、分群、tag 等，目前為止我們尚未達到此功能。

八、結論

我們針對目前觀察到 PTT 的功能不足，使用 Information Retrival 與 Text Mining 的技術，先將文章分群，為文章內容建立了 tag，並且使用不同 tag 來達到推薦文章、主題閱讀等功能的實現。

我們為 PTT 建立了主題式的文章瀏覽，以及在看完文章後，能夠顯示該文章的主題標籤與相關企業，為 PTT 增添更具人性化、深入搜索的擴充設計，讓 PTT 相較其他論壇或網站不足之處得以補強，更貼近使用者的需求。

九、組員分工

1. 陳韋傑：clustering、feature selection、簡報、上台、書面
2. 江采嬪：爬文、文章前處理、clustering、feature selection、簡報、上台、書面
3. 石子仙：建立企業與文章標籤、簡報、書面
4. 王博奕：統合結果、介面設計、簡報、書面
5. 莊啟宏：簡報、K-means elbow method、書面

十、參考資料

上市公司代碼一覽表：
https://www.tej.com.tw/webtej/doc/uid.htm