第四組

PTT TAG

會計三 陳韋傑　　會計四 江采嬪
化工四 石子仙　　財金三 王博奕
工管三 莊啟宏

動機與目標

動機

PTT可查詢特定文章，
但無文章推薦與分類系統。

目標

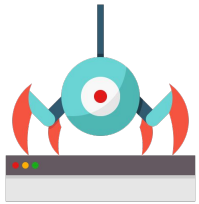
為PTT文章進行分群，並幫文章加上Tag，讓使用者可以瀏覽特定主題，或找到與特定企業相關的文章。

實作方法

01

使用爬蟲抓取資料

資料來源:Ptt stock板



02

文章前處理

斷詞

計算tf-idf

03

使用K-Means進行分群

文章數:2445

群數:21

04

替每篇文章加上tag

使用feature selection

選擇每群中較重要的tags

01

資料爬取

PTT Stock板

資料爬取

利用套件進行ptt 爬文:

- import requests
- from bs4 import BeautifulSoup

看板:Stock
時間範圍:2020/3/19~2021/1/14
共30904篇文章

	author	title	time	content	url
0	louis960126 (JLingogo)	Re: [標的] 鴻海 多	Thu Jan 14 22:48:05 2021	https://i.imgur.com/dGgzk3w.jpg\n去年四月的時候用博士獎學金...	https://www.ptt.cc/bbs/Stock/M.1610635687.A.5D...
1	Arizona989 (心平氣和，和氣生財)	[新聞] 氣候變遷題材發酵 美銀看好銅價突破10,000	Thu Jan 14 22:56:51 2021	\n1.原文連結：\n\nhttps://reurl.cc/3NDGEj\n\n2.原文內容...	https://www.ptt.cc/bbs/Stock/M.1610636213.A.1B...
2	CORYCHAN (CORY)	[心得]台積電7nm & HPC Q4季減的解讀	Thu Jan 14 23:03:35 2021	台積電的Q4營收達到財測高標，表現亮眼。\n\n但身為同時持有台積電跟AMD的小戶，\n比較...	https://www.ptt.cc/bbs/Stock/M.1610636617.A.A3...
3	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...
4	a025892000 (DK)	[其他] 台積電ADR目前已大漲9%	Thu Jan 14 23:05:27 2021	\n乳題，今天法說會剛結束\n晚上ADR馬上漲給你看\n還一舉突破前高\n果然只要拉回就是最...	https://www.ptt.cc/bbs/Stock/M.1610636729.A.B6...
...	...	...	...	...	...
38302	luckystrike5 (霸王鮮果汁)	Re: [請益] 交割戶重複交割	Mon Aug 17 22:26:04 2020	先告訴你結論\n很遺憾你拿不到什麼賠償\n台新最多手續費再給你一點點折扣\n下次記得打金管會...	https://www.ptt.cc/bbs/Stock/M.1597674366.A.C2...
38303	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...
38304	angelllll (長尾巴的天使)	Re: [新聞] 外資：聯發科5奈米晶片明年擬獲華為旗艦	Mon Aug 17 22:32:01 2020	※ 引述《charloette (三宅樹理)》之銘言：\n: https://news.cn...	https://www.ptt.cc/bbs/Stock/M.1597674724.A.A0...
38305	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...

資料爬取

為找出網友對個股公司的主觀見解，我們選出文章分類屬於「標的」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鬼神氣漲，\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...

02

文章前處理

斷詞、計算tf-idf

文章前處理–斷詞

- 由於分析中文文章，因此首先刪除非中文語詞、標點符號及語助詞等。
- 利用「，」切分句子，針對每一句話進行斷詞，留下長度≧2的詞彙。
- 採用monpa罔拍中文斷詞系統進行斷詞，可以切出專有名詞及特殊字詞。

	tokens
0	[站上, 那, 初次文, 至今, 來到, 元想必, 大家, 開始, 考慮, 該, 獲利, 以...
1	[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	[相信, 今天, 應該, 賺錢, 現在, 鬼神, 飆漲, 明天, 空單, 著, 停力, 回補...
3	[今日, 進場, 股, 總共股, 知道, 看, 感覺, 可笑, 最, 希望, 相隨, 台, ...
4	[引述, 銘言標佳凌, 分類, 分析, 正文, 很多, 覺得, 值得, 個, 價位, 因此,...

文章前處理–計算tf-idf

- 計算每個term在每篇文章當中的tf及df
- 建立vocabulary詞集，把df與posting存成dataframe，並給定term_index

	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

- 合併tf_dict與df_dict，建立每篇文章的tf-idf集

	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

03

文章分群

K-Means clustering

文章分群

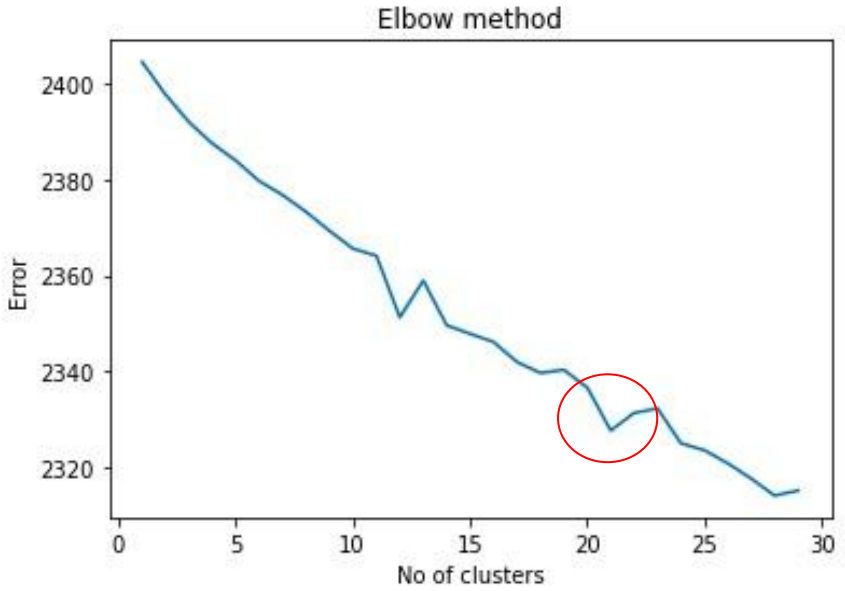
爲2445篇文章建立size為|V|的 tdidf vector，缺值補0，作為餵入Kmeans的參數。

	0	1	2	3	4	5	6	7	8	9	...	31974	31975	31976	31977	31978	31979	31980	31981	31982	K-means
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2440	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2441	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2442	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2443	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2444	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0

文章分群

使用方法：K-Means

- 使用sklearn.cluster中的K-Means實作
- 嘗試1~30分群數，計算每一種的Errors (Errors=文章與重心的距離加總)
- 以Elbow Method輔助判斷合適群數
- 比較各群的feature words，人工判斷分21群時，每群選出的feature words較具代表性。



結論：K = 21 時，有不錯的分群效果

04

Feature Selection and Tagging

Chi-square Feature Selection

Feature Selection

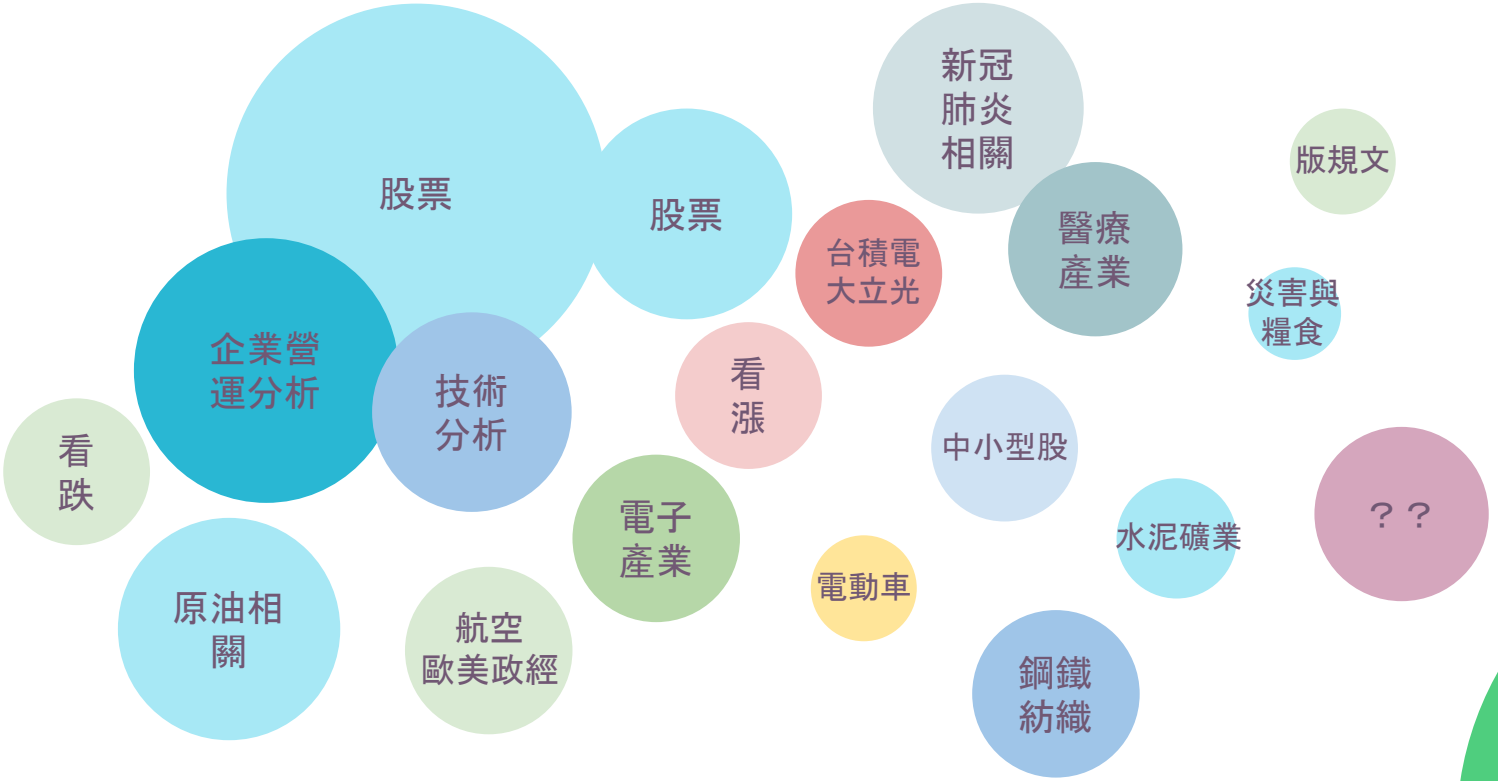
使用方法：Chi-Square

- 為每個Token計算21個Class的Chi-Square分數
- 取每個Class前100高的詞，並進行人工篩選
- 挑選出具有意義的Feature words

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

	0	1	2	3	4	5	6
站上	0.533367	25.333828	1.126380	2.531687	0.638040	0.004096	0.000005
初次文	0.020041	0.026585	0.011043	0.244581	0.004499	0.040491	0.029448
至今	0.034087	18.314323	2.699614	2.194495	0.269939	2.429448	1.766871
來到	3.126380	33.875050	6.234808	0.452389	0.701840	0.849587	0.553000
元想必	0.020041	0.026585	0.011043	0.244581	0.004499	0.040491	0.029448
...	...	...	...	...	...	...	...
出血	0.020041	0.026585	0.011043	0.244581	0.004499	22.737460	0.029448
核可	0.020041	0.026585	0.011043	0.244581	0.004499	22.737460	0.029448
賭聲	0.020041	0.026585	0.011043	0.244581	0.004499	22.737460	0.029448
利多連發	0.020041	0.026585	0.011043	0.244581	0.004499	22.737460	0.029448
醒進	0.020041	0.026585	0.011043	0.244581	0.004499	22.737460	0.029448

Features - Cluster Results



Features - Cluster Results

電動車相關

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

Features - Cluster Results

新冠肺炎相關

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

Features - Bad Example

難以決定
相關主題

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

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

Tagging

1. 依照分群、Feature Selection的結果：
    - 針對文章所在該群之feature，替此篇文章標上內文有出現的feature
2. 依照上市公司名單：
    - 除了上述的tag以外，也為每篇文章標上內文有出現的上市公司名稱

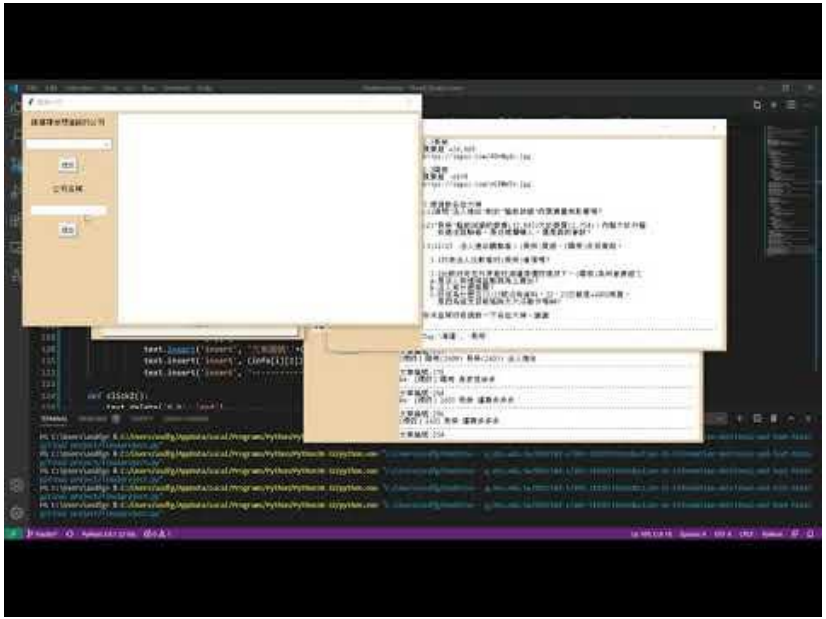
05

Results and Demonstration

User interface and document reader

成果:使用者介面

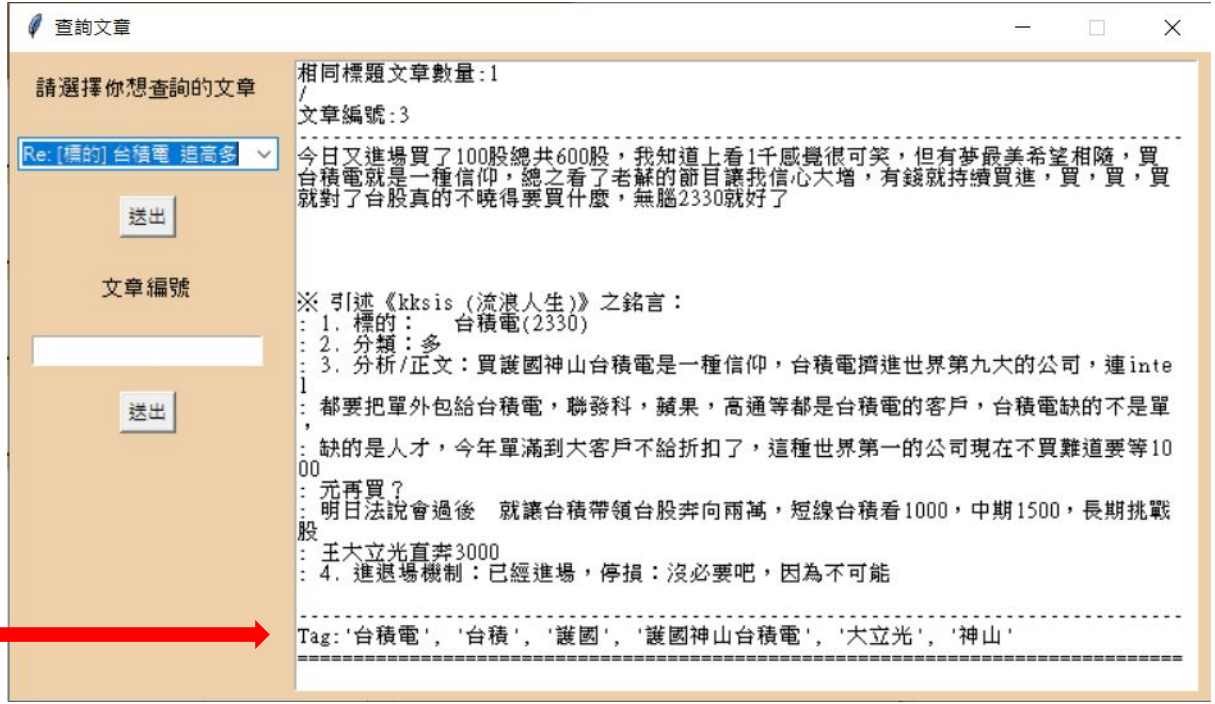
https://youtu.be/3PJanHvEdhE



成果:查詢文章

依照使用者需求
選擇滑單式或
直接輸入編號
(在查詢文章時會
附上編號)

TAG



輸入公司名稱得到文章標題

輸入Tag得到文章標題

06

Problems Encountered

Problems

- 資料量過於龐大，運算資源不足
- 難以決定適當的分群數量
- Feature Selection之結果並非完全可解釋，
  可能是作者當初發表了與企業標的無關的內容。
- 無法隨時間更新分群結果

07

Conclusion

Conclusion

- 分類：以內文相關性建立不同主題之文章集
- 標籤：以內文包含之重要詞語為文章做Tagging

-> 為PTT現有文章進行分類與標籤
使用者看完文章後，能夠觀看相關推薦文章
亦能針對不同感興趣之議題或公司進行主題式搜尋

THANKS