

社群媒體分析

期中專案報告

PTT 八卦版文本分析：台積電高雄設廠

2021/11/30 第 12 組

組員 陳雅詩 N094220026
李明章 N094220023
洪嘉駿 N094220020
林志勳 N094220014



PTT 八卦版：台積電高雄設廠

研究動機和分析目的

台積電預計在高雄設廠的消息在今(110)年 9 月初，即在天下雜誌及 ptt 房屋版、八卦版網友間流傳，11 月 9 日經台積電正式證實，為了因應市場需求，將在高雄楠梓的中油高雄煉油廠舊址設立 7 奈米、28 奈米製程的晶圓廠，預計 2022 年動工、2024 年投入生產。

在台積電確定於高雄設廠的新聞正式發布之後，引起了許多新聞媒體、社群媒體的討論聲浪，因此本組選定 ptt 八卦版(Gossip)作為分析資料集的主要來源，並且也透過觀察 ptt 其他相關的版(Kaohsiung 高雄版、home_sale 房屋版、Tech_job 科技工作版及 Stock 股票版)，深度探討社群使用者對於本分析主題的態度及看法。



資料集描述

主要文本資料來源：ptt **Gossip**

其它文本資料來源：ptt **Kaohsiung**、ptt **home_sale**、ptt **tech_job**、ptt **Stock**

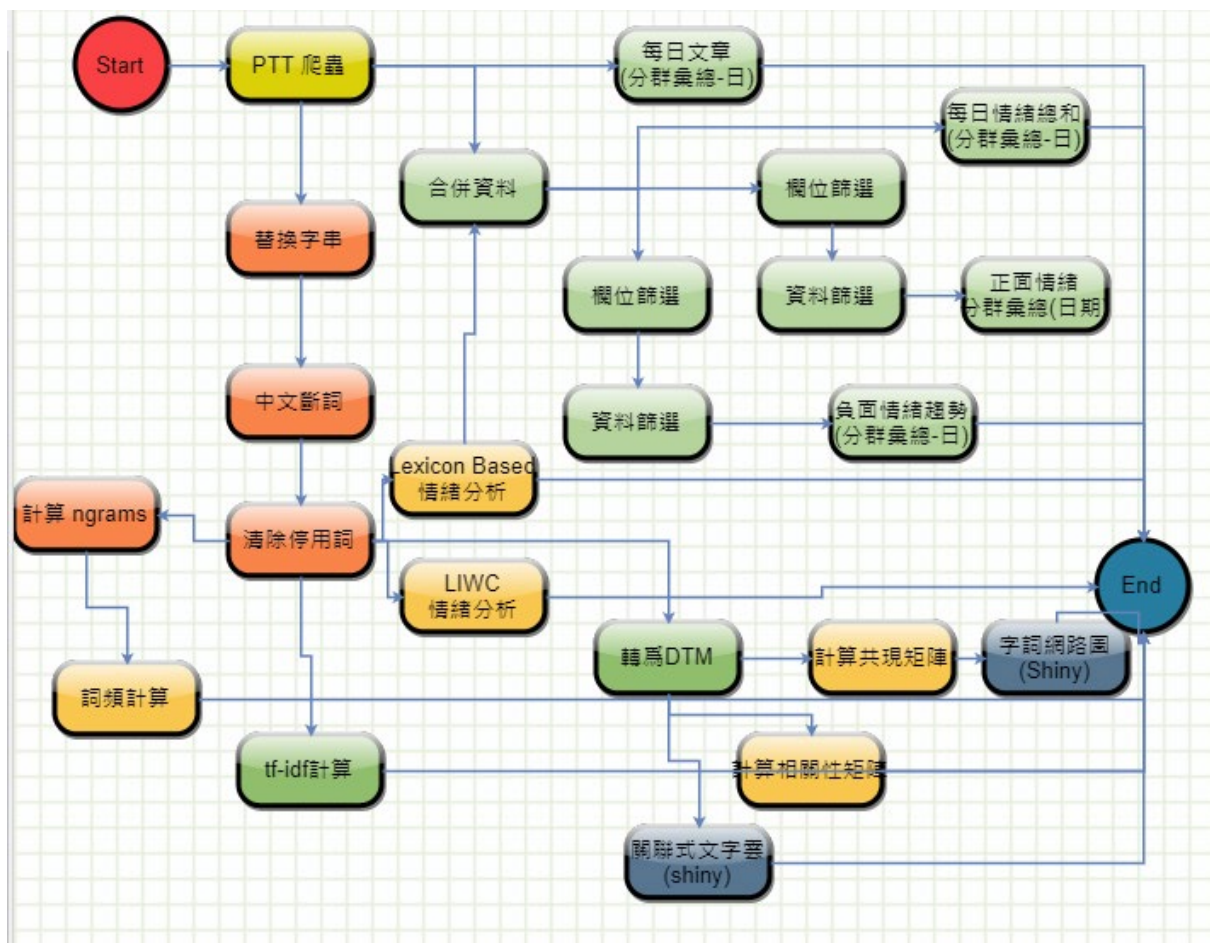
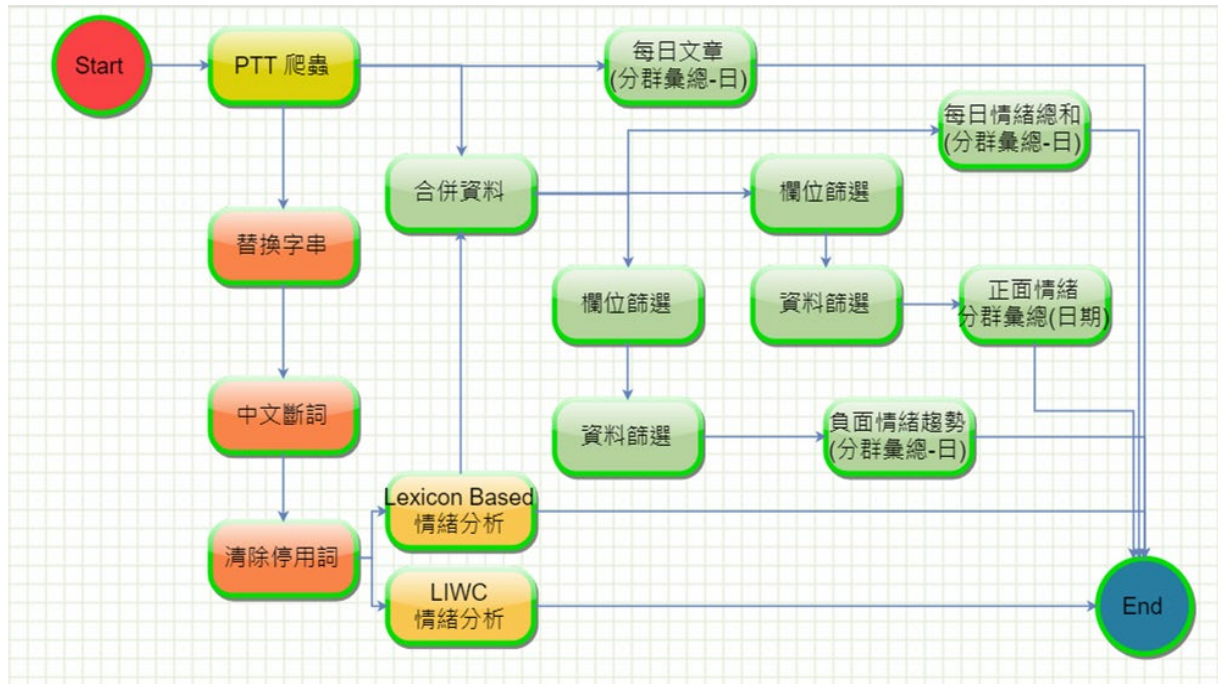
時間：2021/09/01 - 2021/11/20

關鍵字：台積電、高雄、設廠

資料筆數：5786 筆

【板主:ubcs/q347/ffooxx/arsonlo..】 [八卦] Omicron 歐米可龍 系列《Gossiping》					
[←]離開 [→]閱讀 [Ctrl-P]發表文章 [d]刪除 [z]精華區 [i]看板資訊/設定 [h]說明					
編號	日期	作者	文章	標題	人氣:12880
2647	11/17	CircusWorld	[問卦]	為什麼高端沒有像台積電一樣到處擴廠	
2648	13/18	thickbanana	[問卦]	寄台積電的掛號求職信沒下落	
2649	38/18	kiwibee	[新聞]	台積電設廠房價飆，高雄社宅能救火？實	
2650	13/18	shangguan125	[新聞]	台積電工程師哀號45歲退休難 過來人曝做	
2651	15/18	ccugoo	[新聞]	台積電高雄設廠 聯發科有意跟進	
2652	6/18	peterlin495	[新聞]	文組生必看！如何進台積電工作？過來人	
2653	6/19	a245065789	[問卦]	台積電學長674T 有學弟要來找學長玩嗎？	
2654	25/19	WaWa3	[問卦]	請問台積電的黑科技是什麼？	
2655	64/19	wahaha99	[問卦]	為什麼不讓台積電蓋核5就算了	
2656	1/20	mean	R: [問卦]	為什麼不讓台積電蓋核5就算了	
2657	3/21	gundamguy	[問卦]	台積電吃電怪獸知所進退 滾出台灣	
2658	6/21	andandandand	R: [問卦]	怎麼好像沒人在意台積電賣的八卦呀	
2659	17/22	taeyeon56	[新聞]	聯發科攜手台積電 全球首顆7奈米8K數位電	
2660	15/22	alwang	[新聞]	台積電攜手聯發科 推全球首顆7奈米8K電視	
> 2661	+X/23	fur	[新聞]	台積電決定高雄設廠 蔡英文竟罵陳其邁「	
2662	+ 4/24	applegg	[問卦]	資動部有沒資檢過台積電	
2663	+2/24	joster9527	[問卦]	如果沒有果凍、慈濟、台積電過年會怎樣？	
2664	+ 2/25	Qunnine	R: [問卦]	為啥媒體沒有像捧台積電那樣捧鴻海啊？	
2665	+ 6/25	ccugoo	[新聞]	台積電首設廠務學院 今年已培訓超過500	
2666	+ 11/25	Irene534	[問卦]	打BNT的怎麼不感謝政府跟台積電	
文章選讀 (y)回應(X)推文(^X)轉錄 (=[]<>)相關主題(/?a)找標題/作者 (b)進板畫面					

使用「工作流程平台」分析資料



資料前處理：替換字串、中文斷詞、清除停用字

去除換行符號、網址、html 標籤、底線、多餘的空格

≡ 替換字串 (7)

參數設定

Input - 4

任務結果

選擇處理欄位 *

artContent

替換字串設定 ⓘ

```
\n>>
https?:\\(www\\.)?[-a-zA-Z0-9@:%_\\+~#={1,256}\\.[a-zA-Z0-9]{1,6}\\b([-a-zA-Z0-9@:%_\\+~#?&//=]*)>>
\\W>>
```

選擇替換規則檔案 ⓘ

dictionary_example.csv

儲存更改

進行中文斷詞、清除可能干擾分析結果的停用字

≡ 清除停用詞 (15)

參數設定

Input - 28

任務結果

清除英文字母 *

是

清除數字 *

是

清除換行符號 *

是

清除特殊標點符號 *

是

清除html tag *

是

自定義停止詞

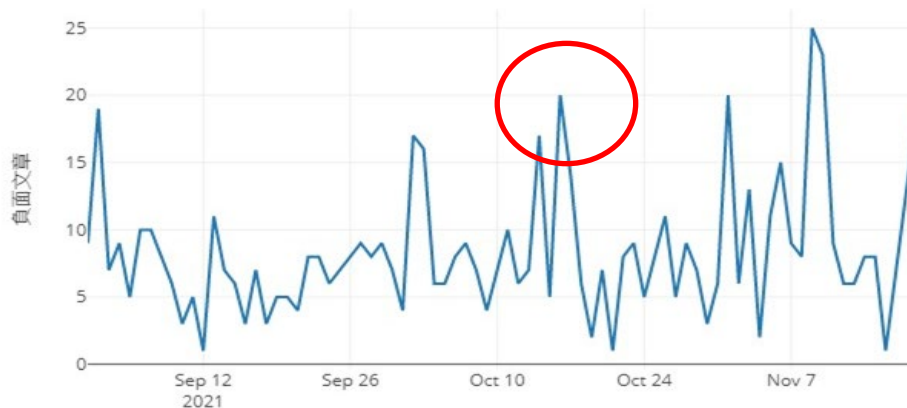
你好
不要
問題
八卦
疫苗

資料前處理的發現

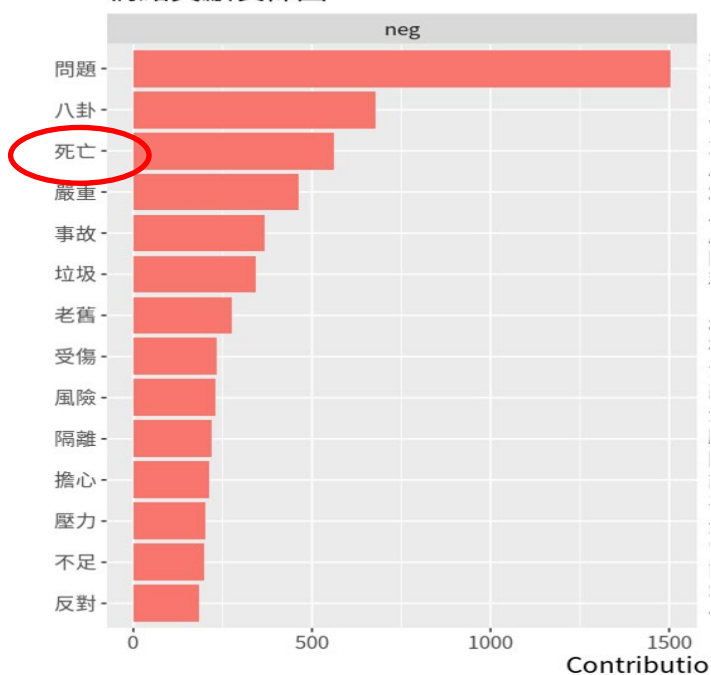
在完成基本的資料前處理後，進行情緒分析及負面情緒趨勢圖觀察時，我們發現情緒分析的負面詞中出現「死亡」這個字，且在 10/14 出現負面文章數量高峰，但是查詢網路上有關「台積電高雄設廠」的相關討論，在 10/14 這一天並沒有與此分析主題相關的特殊事件或新聞發生。

因此我們推測，可能是當天所發生的高雄城中城大火事件導致 ptt 八卦版負面聲量增加，為了避免干擾聚焦於本主題的分析結果，我們將原始資料集中的關鍵字「高雄」兩字拿掉，並重新進行爬蟲及資料前處理

負面情緒趨勢(八卦版)



情緒貢獻長條圖



重新爬取資料及資料前處理

依據前述發現，我們重新調整搜取資料集的關鍵字，

處理前

資料來源：PTT 八卦版

時間：2021/09/01 - 2021/11/20

關鍵字：台積電、高雄、設廠

資料筆數：5786 筆

處理後

資料來源：PTT 八卦版

時間：2021/09/01 - 2021/11/20

關鍵字：台積電、設廠

資料筆數：1218 筆

統計資訊

10

欄位數



1218

資料筆數



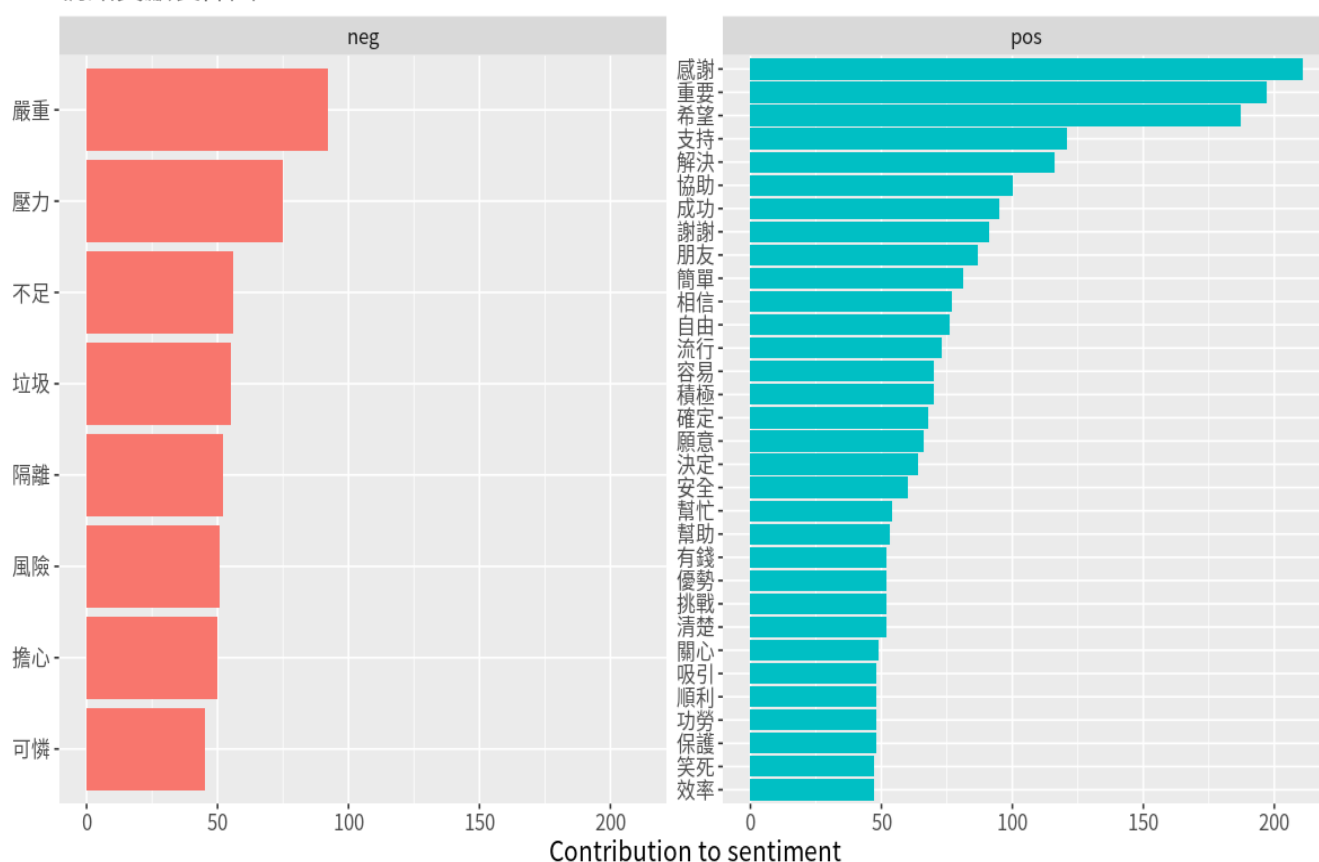
使用 Lexicon Based 情緒分析

將無法明確確定意義的負面字詞例如「問題」、「八卦」、「網址」、「內文」等字做停用字去除，另將會影響詞頻計算的「台積電」、「設廠」等也做停用字後再做一次分析，得到的：

負面字詞有嚴重、壓力、不足、垃圾、隔離、風險...等

正面字詞有感謝、重要、希望、支持、解決、協助...等

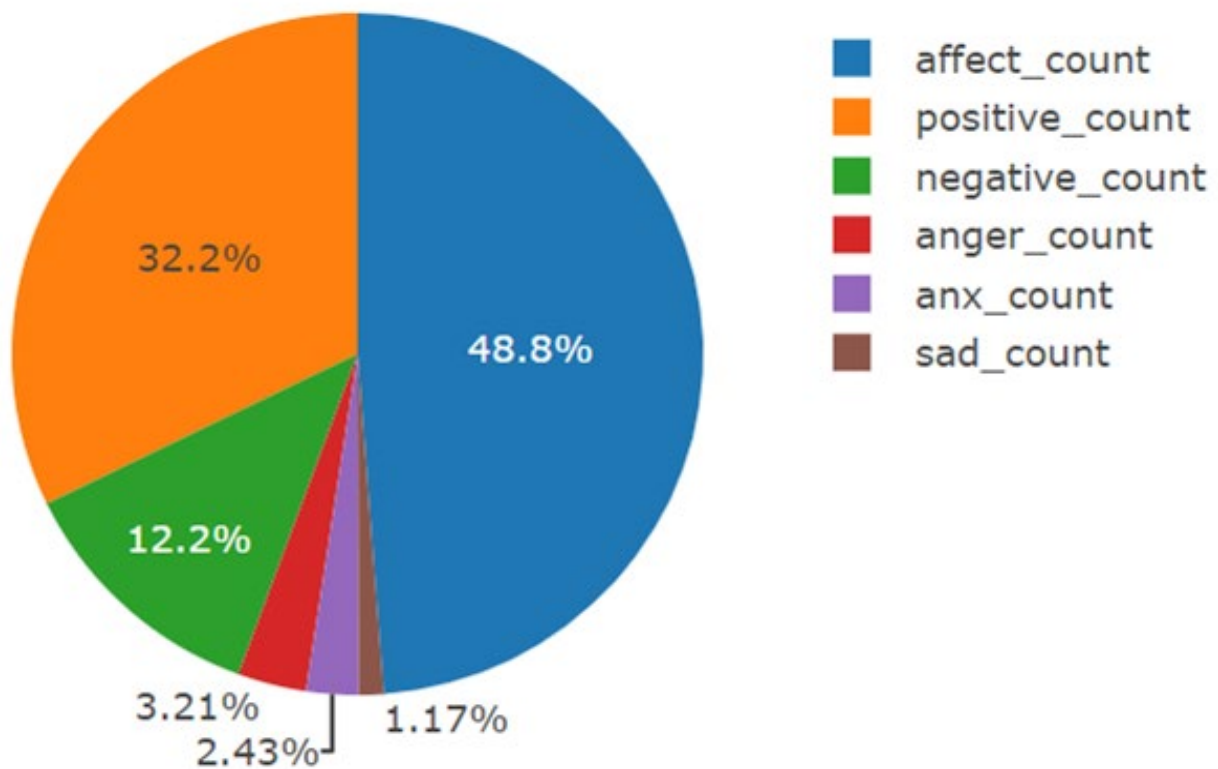
情緒貢獻長條圖



使用 LIWC 情緒分析

觀察實際的正負面文章數量，正面文章數量略多於負面文章數量
但是並沒有太過懸殊的明顯差距

推測本分析主題在這段分析期間的正、負面聲量(情緒)沒有太顯著的差異



將情緒分析與斷詞結果進行合併

合併 ptt 八卦版和 Lexicon Based 情緒分析，兩個任務結果的資料進行欄位篩選，參數設定：選取 artDate 和 sentiment_value，任務結果列出分析資料期間，每日的 sentiment_value，進行負面資料篩選(參數設定: \$sentiment_value < 0)及正面資料篩選(參數設定: \$sentiment_value > 0)

☰ 正面情緒 分群彙總(日期) (84)

參數設定

Input - 82

任務結果

76

群組數量

任務結果

Show 10 entries Search:

artDate	sentiment_value
2021-09-01	8
2021-09-02	56

全螢幕瀏覽

點我下載完整CSV資料

點我下載完整Rdata

☰ 負面情緒趨勢 (分群彙總-日) (48)

☰ 負面情緒趨勢 (分群彙總-日) (48)

參數設定

Input - 54

任務結果

73

群組數量

任務結果

Show 10 entries Search:

artDate	sentiment_value
2021-09-01	7
2021-09-02	15

全螢幕瀏覽

點我下載完整CSV資料

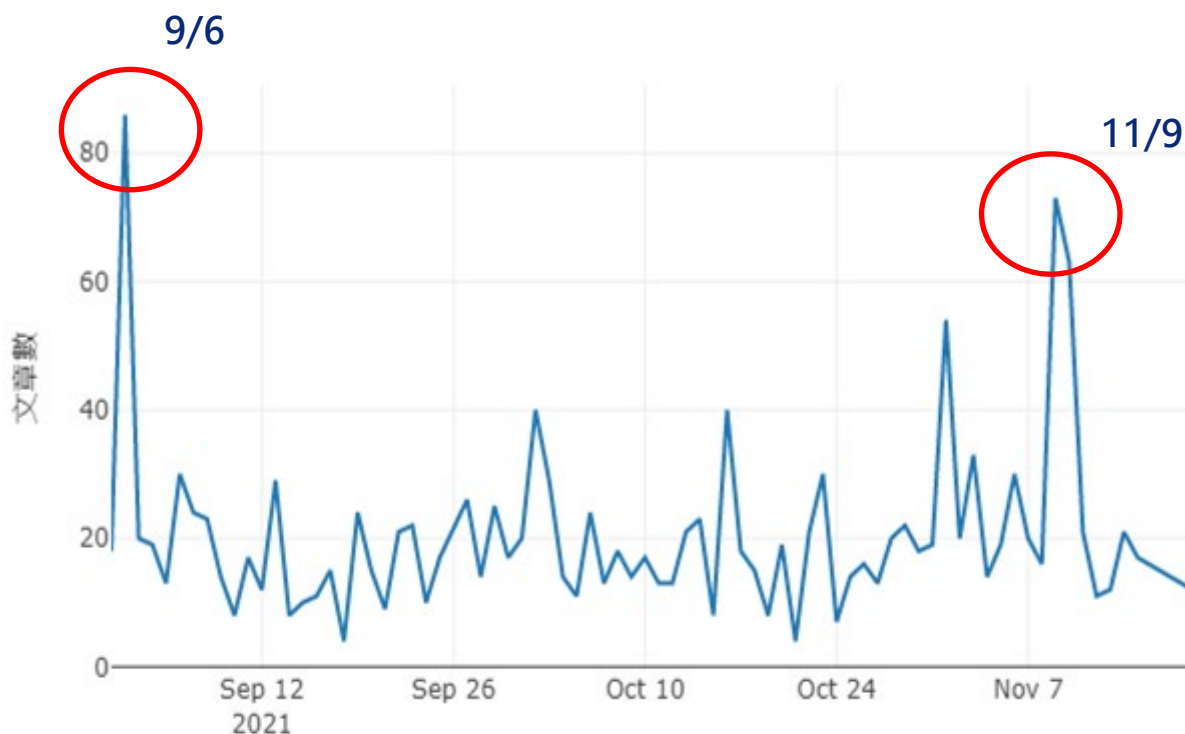
點我下載完整Rdata

分析結果解釋 - 每日文章統計

八卦版有關台積電的討論文章數量最高的時間有出現兩個高峰，分別為 **9月6日** (天下雜誌及網路新聞相關報導) 及 **11月9日** (台積電對外正式宣布證實高雄設廠訊息) 兩日。

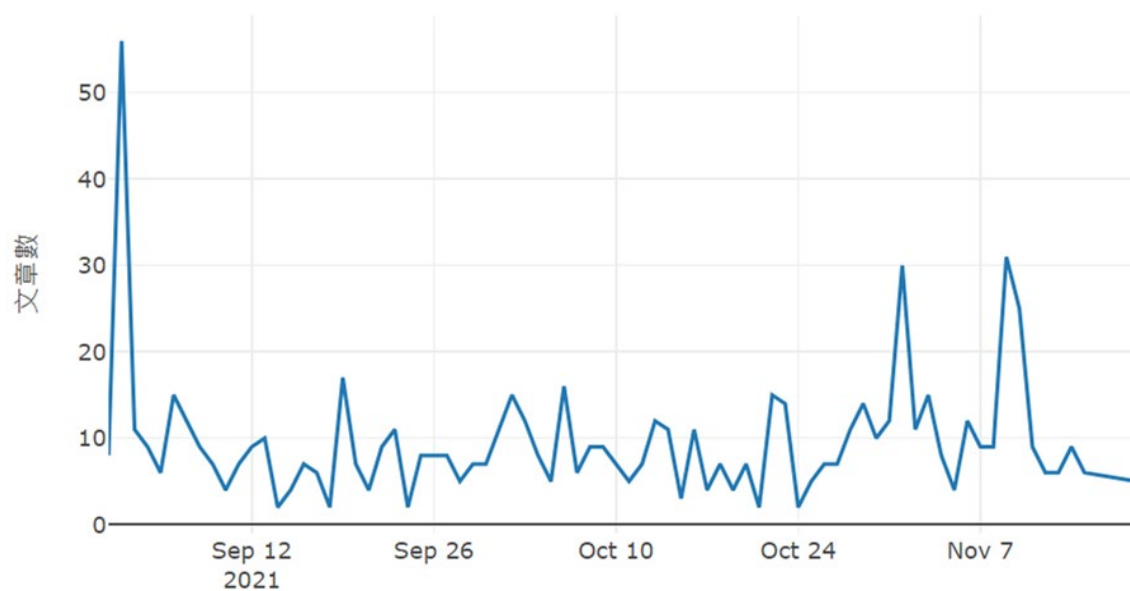


每日文章統計



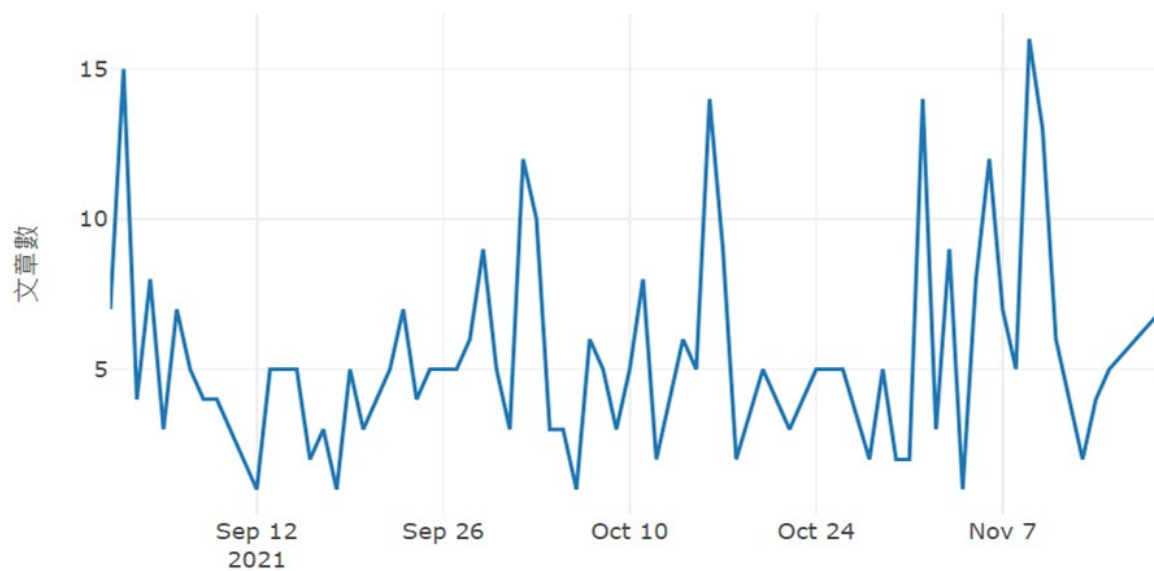
正面文章趨勢圖

正面文章趨勢



負面文章趨勢圖

負面文章趨勢

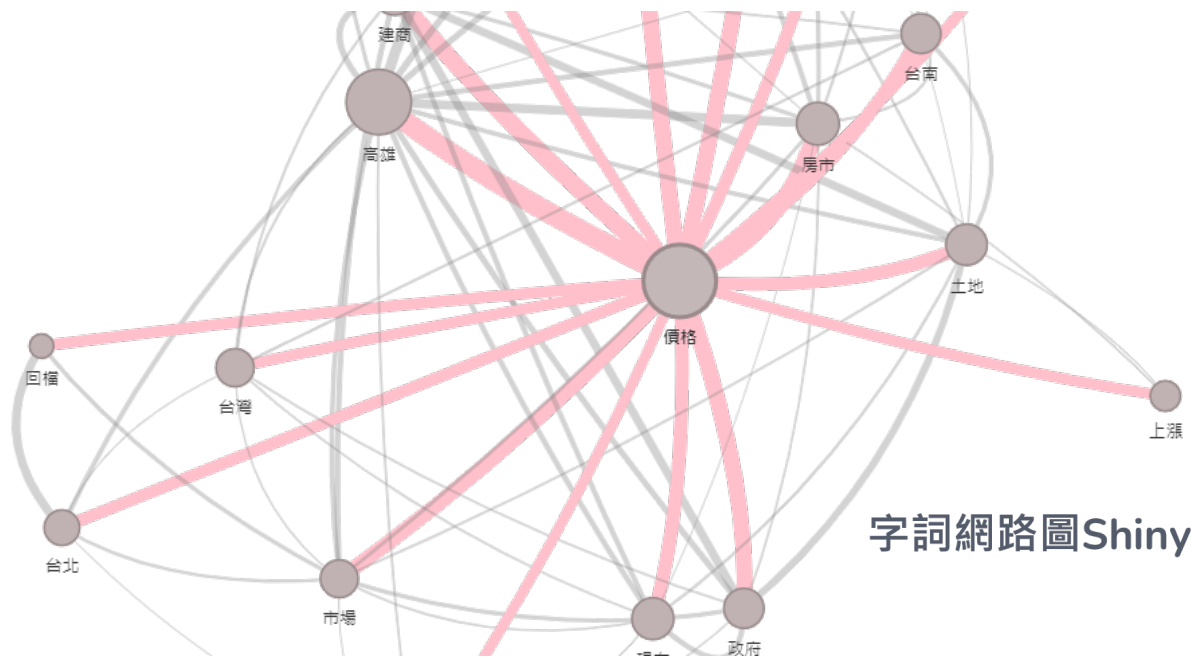


詞頻計算 ptt 八卦版 文字雲



進一步分析「home_sale 房屋版」的資料

在依據本分析主題 home_sale 所產生的文字雲中，我們發現了出現「價格」、「飆漲」等關鍵字，透過 TF-IDF 對文本重點字詞進行加權之後，我們畫出了在 home_sale 版裡針對本分析主題「台積電高雄設廠」的字詞網路圖及關連式文字雲，可以看出「房市」與「價格」兩個字詞的關聯程度很高，在本分析主題的資料集中出現次數也多



資料分析結果的發現與結論

透過相同的資料集，在不同 ptt 版別文字雲中可以看出，在不同的社群討論版中討論重點、熱度與情緒趨勢，即使是針對相同主題的分析，在不同類別的社群討論中，所著重關注的焦點亦可能有所不同。

本次分析主題原本預期會看見正面或負面情緒的明顯差距，正面(台積電帶來更多科技工作機會、高薪、人才流)，或者負面(帶來汙染、高房價)，但是我們在 ptt 八卦版裡的情緒分析結果發現，正反面情緒文章的篇數，並沒有顯著的差距。因此推斷有關本分析主題的正、負面評價略可稱衡平。

雖然分析主題和資料集皆相同，但是在不同社群版別進行分析時，關鍵字的提取、停用字的選擇，仍需依據不同社群討論主題的特性與趨勢，再透過觀察 n-gram 與詞頻計算之結果而隨之變動。