社群媒體分析

第四次讀書會報告_第十二組

一、分析主題

透過分析 PTT 股票版面文章主題分布。

二、組員

N114320002_李湘怡

N114320008_陳家卉

N114320009_林育伈

N114320017_黃敏喩

N114320020_曾子瑋

N114320027_吳政翰

N114320030 吳郁文

N114320031_洪宜綾

三、分析工具

• 中山大學工作流程平台: Tarflow

• 工作流程名稱:第四次讀書會

• 主題模型

四、動機目的

小組從 PTT 爬蟲股票版面蒐集新聞內容,並運用自訂的詞彙字典以及 Tarflow 平台提供的 Bertopic 主題模型,來進行相關主題的內容分布分析。

針對近期總統就職等重大事件可能會導致股票市場波動加劇,總統就職期間 (2024/05/02~2024/05/22), 最終目的為識別在關鍵時期內 PTT 看板討論的主題偏重,以獲得有價值的資訊和洞察,進而更深入了解台灣在政治、財經上的動態。

五、資料來源

● 資料來源: PTT 爬蟲資料庫之股票看板。

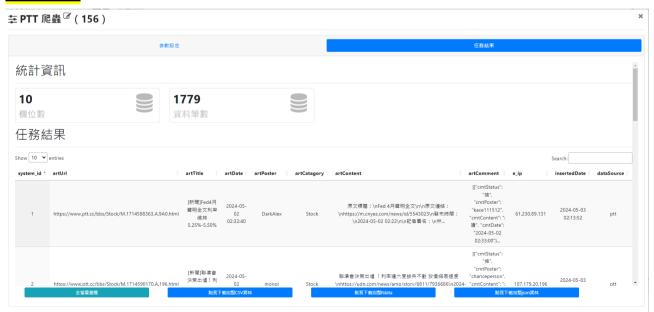
● 關鍵字:無

• 搜尋日期: 2024/05/02~2024/05/22

● 資料筆數: 1779

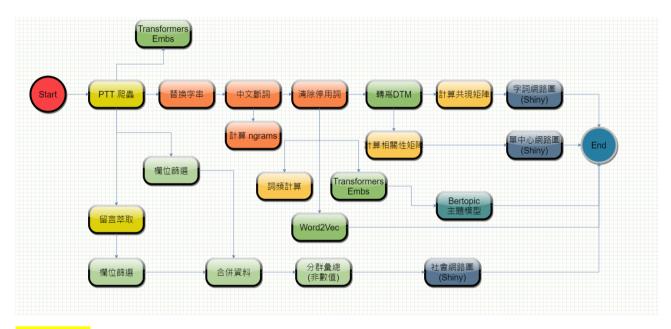
幸PTT 爬蟲[☑](156)





六、系統流程圖

- 6-1.資料前處理:以 PTT 爬蟲資料庫,進行資料觀察(各版文章數量)和前處理(斷詞、停用詞)。
- 6-2.分群彙總: 了解 PTT 原資料集之新聞其各類別分別為多少數量。
- 6-3.將資料進行預處理(替換字串、中文斷詞、清除停用詞)
- 6-4.透過兩種方式訓練與判斷該文章為何主題。(自建字典轉 DTM、Bertopic 主題模型)



任務結果

七、資料預處理-替換字串、中文斷詞、詞性篩選、保留重要字

7-1 將 PTT 爬蟲透過正規表達式把 URL 都替換



7-2 賦予專有名詞權重提升中文斷詞的精準度

升息 200

成交量 200

台股 200

台積電 200

上市 200

漲停 200

跌停 200

收盤價 200

走高 200

走低 200

面板 200

加權指數 200

萬點 200

停損 200

外資 200

三大法人 200

超買 200

超賣 200

投信 200



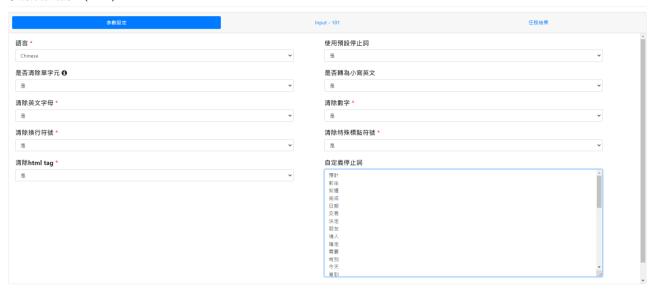
7-3 清除停用字

將以下文字設為停用字:

預計、前往、知道、完成、日期、交易、決定、朋友、情人、確定、需要、特別、看到、今天、八卦、寶貝、溫泉、早上、晚上、下午、之後、現在、提供、目前、現在、的、113、

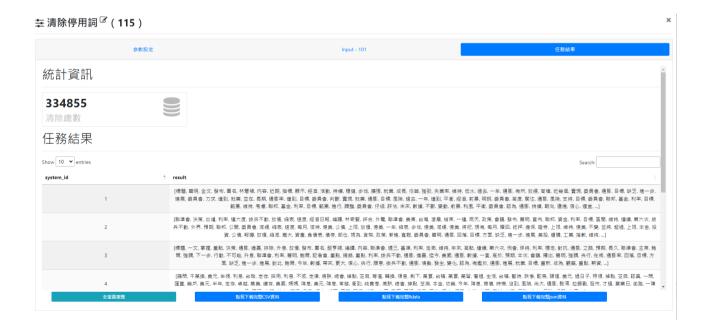
記者、至本、心得、内文、期止、表示、時間、歸屬於、指出、來源、事項、公司、董事、連結、原文、千元、營收、市場、資訊、業主、決議。

幸清除停用詞 ☑ (115)



任務結果

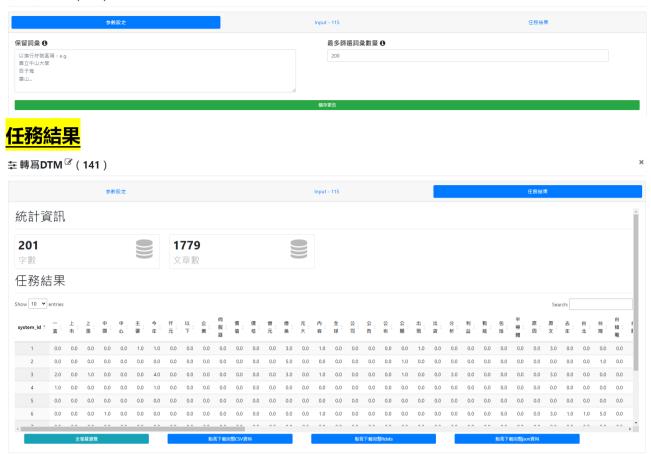
停用字 334855 個



八、DTM

8-1 保留重要字詞後轉為 DTM 作詞頻計算

幸轉爲DTM[♂] (141)



九、Bertopic 主題模型

9-1 編碼





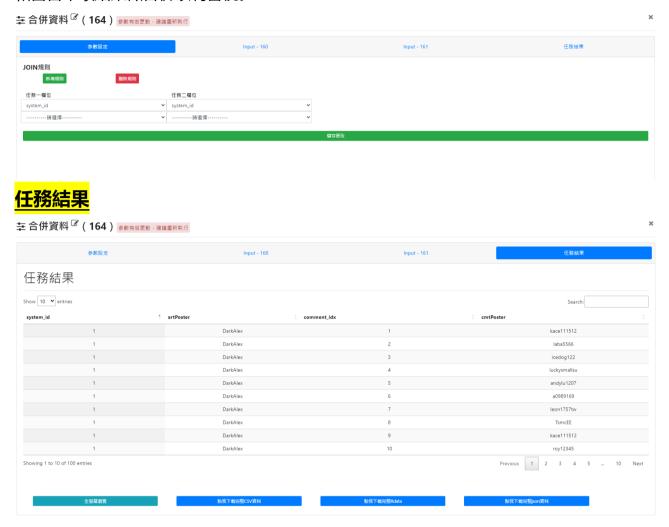
9-2. 建立 Bertopic 參數

設定 5 個主題;文章向量壓縮維度 5;主題保留關鍵字數量 20



9-2 資料合併

為幫助了解 rawdata 與留言萃取模型分類是否符合預期,以 system_id 為主,將原始資料內容和留言萃取結果做合併以利審視。



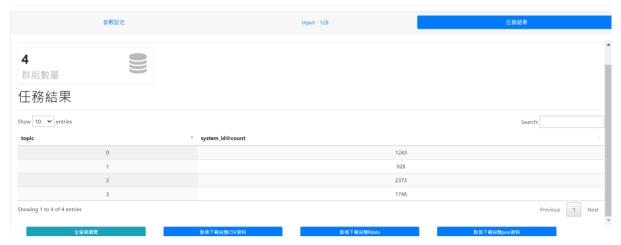
9-3 分群彙總

根據發文及底下回覆文章者做分群彙總。

⇒ 分群彙總 (非數值) ³ (126)

参數設定	Input - 128	任務結果
使用欄位進行分群(按住ctrl(Windows)或command(MAC)可以複選) *	匯總函數 * 1	
system_id 0 1 1 2 3 3 topic 計算欄位(按住ctrl(Windows)或command(MAC)可以複類)*	count nunique min max first last sum	
system_id 0 1 2 3 topic	A	
	儲存更改	

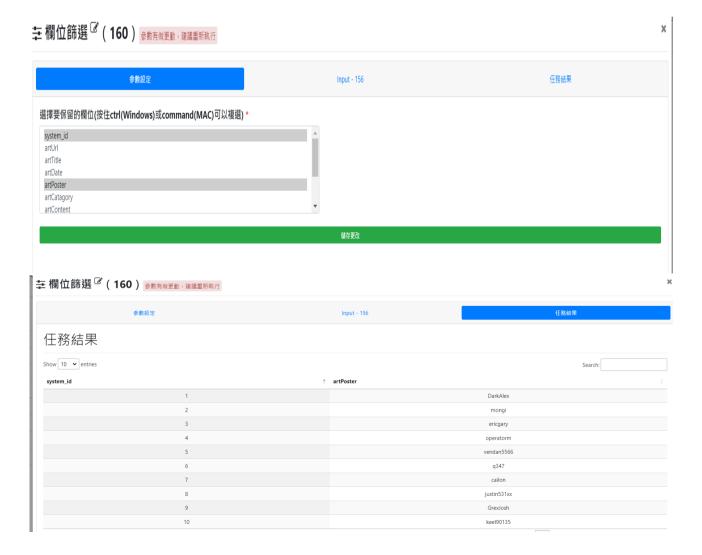
幸 分群彙總 (非數值) [☑] (126)



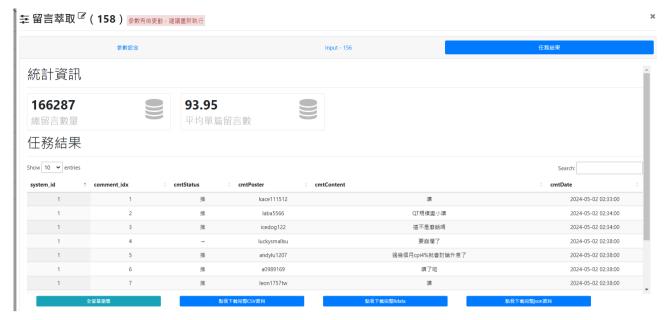
十、社會網路圖(Shiny)

10-1 欄位篩選

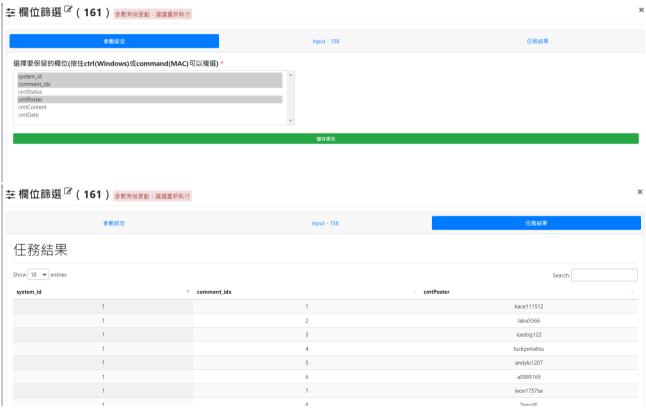
選 system id 與 artPoster



10-2 留言萃取

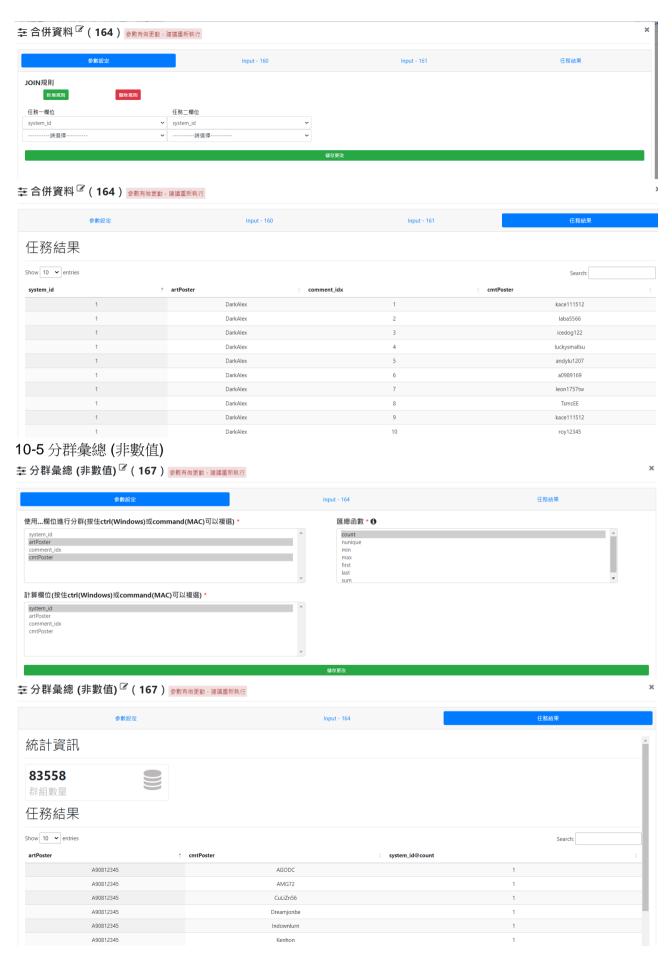


10-3 萃取後欄位篩選



10-4 合併欄位

將兩篩選後的資料皆以 SYSTEM ID 做合併



10-6 **社會網路圖(Shiny**)

節點欄位(來源): artPoster

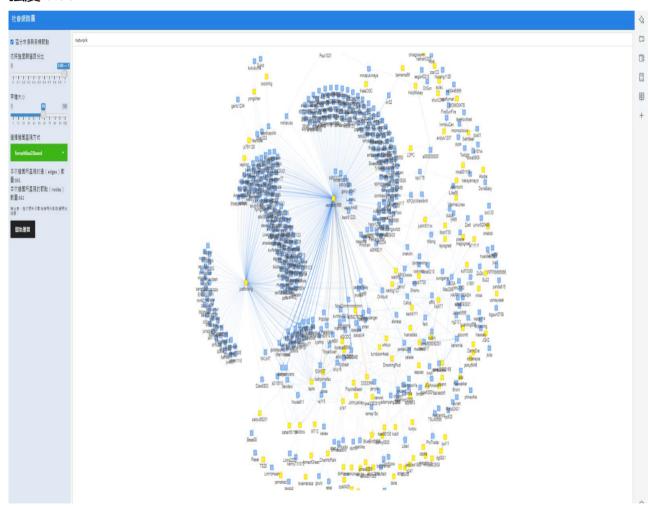
節點欄位(目標): cmtPoster

連結欄位: system_id@count

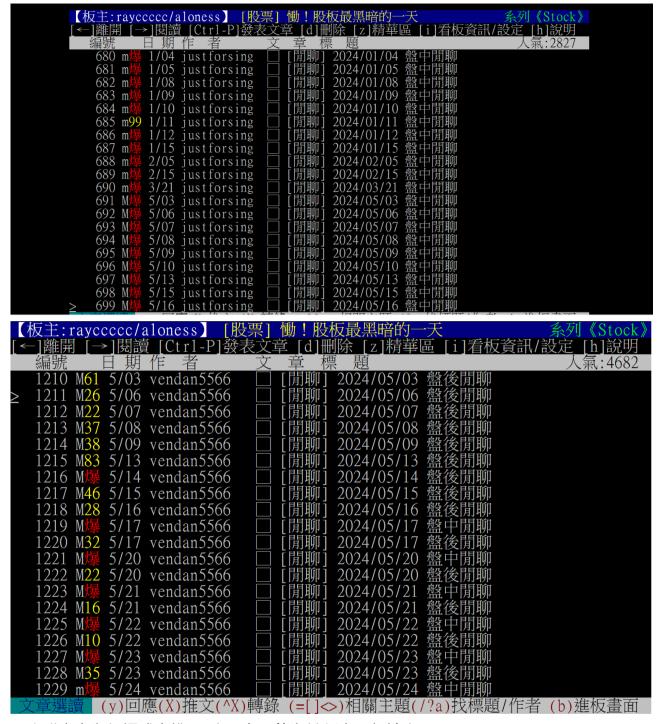


任務結果

強度 0.99-1



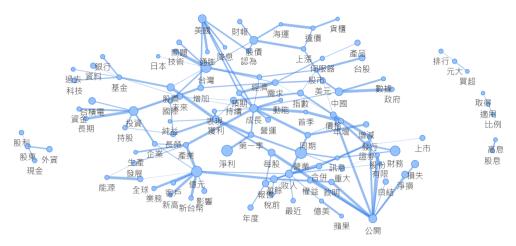
由畫出圖形看出有兩個人是很多人會在其文章底下留言,分別為 justforsing 跟 vendan5566, 後續於 PTT 檢驗



兩人發言多為紅爆或多推 M 文, 有可能為該版意見領袖之一。

10-2 字詞網路圖(Shiny)

幸字詞網路圖 (Shiny) [☑] (184) ቃ數有做更動・建議量新執行				
多數設定	Input - 182	任務結果		
	保存更改			



以 3 個較大的節點進行分析解讀:

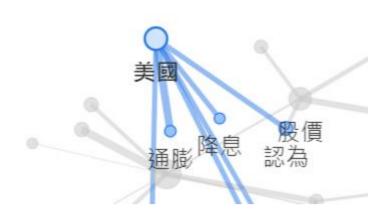
(1) 字詞"投資"-關聯字包括台積電、長期、資金、持股。顯示長期資金若選擇投資標的, 仍以台股資優生 台積電的討論度最高。



字詞"外資"-關聯字包括"股利""現金""股東",實證結果顯示外資持股與公司現金股利發放有關聯,與多篇論文研究"外資持股比例與股利發放政策的研究結果相呼應。(外資持股增加與公司現金股利發放有正向關係,外資持股增加與公司股票購回則為負向關係)



字詞"美國"-關聯字包括"通膨""降息",與美國近期深受通膨嚴重影響,且討論度極高的降息議題息息相關。



10-3 單中心網路圖(Shiny)

設定原始檔案元件 100

DTM: 141;

共線/相關性矩陣 186

臺 單中心網路圖 (Shiny) ^②(188) 参數有做更動 · 建議重新執行

多數設定	Input - 186		₹			
選擇關鍵字典檔案		原始檔案元件 *				
	~	100	•			
DTM *		共線/相關性矩陣 *				
(4) (100 (100 (100 (100 (100 (100 (100 (10						
	1201					

<mark>任務結果</mark>

選擇詞類: others

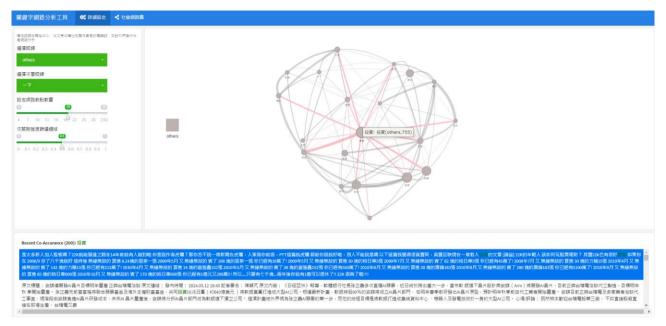
選擇次層詞類:一下

設定網路節點數量: 20

依關聯度篩選鏈結: 0.5

圖中顯示最大的節點為 "others" ,表示這個詞在文本中出現頻率最高,並且與多個詞語有密切的關聯性。

"投資"與"股票"之間的連線較粗且為紅色,表示這兩個詞語之間的共現頻率較高,代表文本中常常同時提到這兩個概念,反映了文本的主題可能涉及金融或投資相關內容。

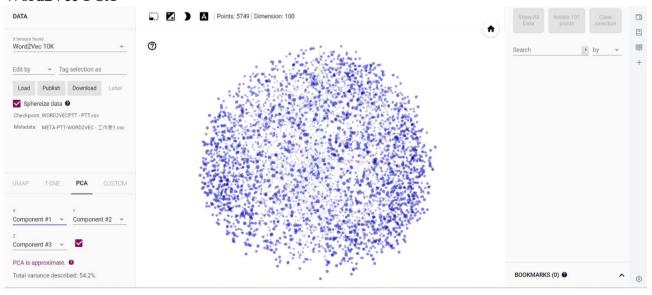


十一、分群視覺圖

11-1 Word2Vec-PCA法

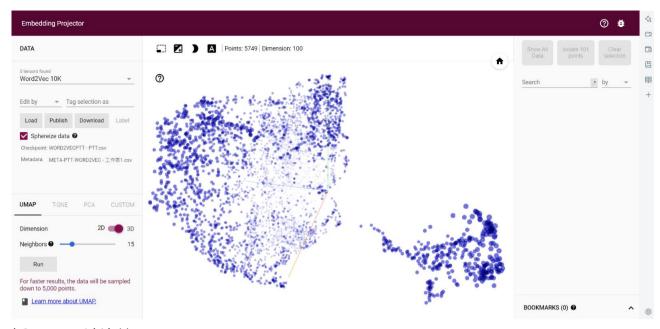
任務結果

Word2Vec-PCA



單一專版分類較不明顯

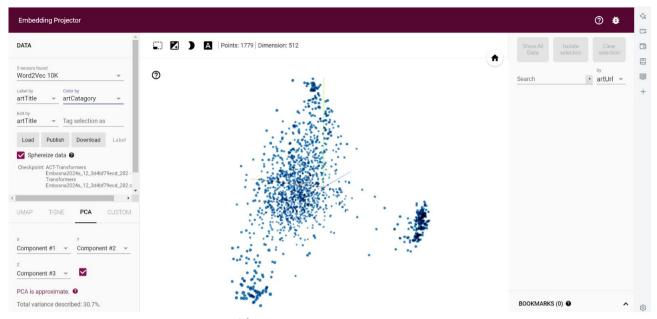
11-2 Word2Vec-UMAP法



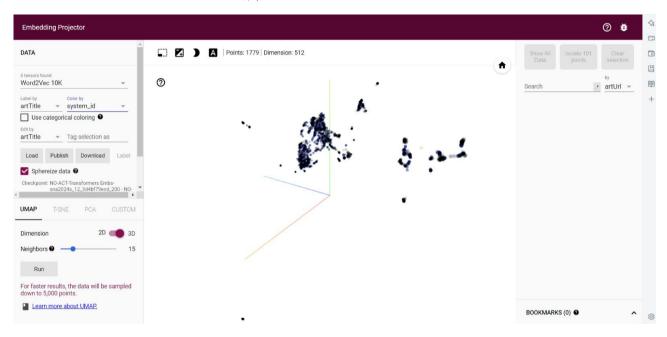
採 UMAP 法比較明顯

11-3 -PCA法

<mark>任務結果</mark>



11-4 Transformers Embs -UMAP法



十二、小結

本次讀書會報告聚焦於分析 2024 年總統就職期間 (2024/05/02~2024/05/22) PTT 股票看板的新聞內容,目的是識別該時期內的討論主題,本次分析共收集了 1779 筆新聞資料,分析顯示,在總統就職期間,PTT 股票看板的討論主要集中在三個主題:政治事件、股票市場波動和投資策略,另外高頻詞如 "公司"、"市場"、"股東"等顯示出文本的核心關鍵詞;網絡圖中,"投資"與"股票"之間的連線顯示了這兩個詞語經常一同出現,反映出投資者在討論股票市場時,對投資策略的關注度。

透過本次分析,可以了解總統就職期間市場參與者的情緒和關注點,進一步預測政治事件對股票市場的影響。