

第一次讀書會報告

組別：11

組員：

邱昱榕 N114320004

朱怡樺 N114320005

林威呈 N114320011

董力慈 N114320019

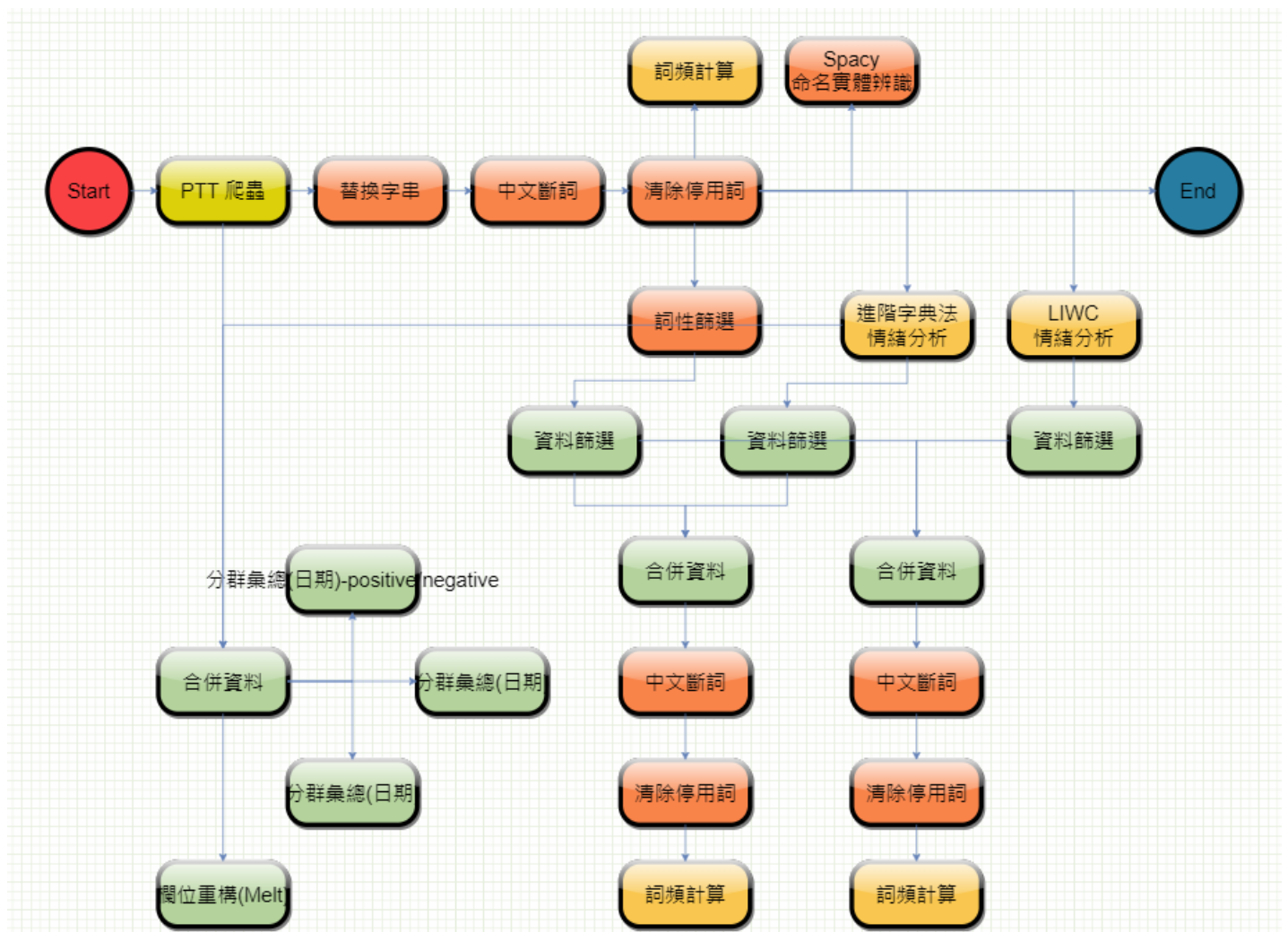
花秀旻 N114320021

趙丞德 N114320023

陳昱維 N114320026

李晉安 N114320028

下圖為我們這次整個情緒分析的流程圖



本次報告主題標的：取 Tarflow 中，PTT 股票版

標的資料期間：2023/09/01-2024/02/29

資料筆數：9131 筆資料

參數設定

任務結果

選擇看板 *

Stock(股票)
studyteacher(實習教師)
TaichungBun(台中)
Tainan(台南)
TaiwanDrama(台劇)
Teacher(教師)
Tech_Job(科技工作)

搜尋關鍵字 ①

以換行區隔 · e.g.
國立中山大學
西子灣
...

排除關鍵字 ①

以換行區隔 · e.g.
壽山動物園
猴子
...

搜尋起始日期

2023/09/01



搜尋結束日期

2024/02/29



參數設定

任務結果

統計資訊

10

欄位數



9131

資料筆數



任務結果

Show 10 entries

Search:

使用功能<替換字串>-將 PTT 文章中，換行、簽名檔等字串替換成斷句用的標點符號或是空白

參數設定

Input - 27

任務結果

選擇處理欄位 *

artContent

替換字串設定 ①

\n\n>> "
\n\n\n>> "
\n>> "
Sent from JPTT on my \w+>>

選擇替換規則檔案 ①

-----請選擇-----

≡ 替換字串 (30)

參數設定

Input - 27

任務結果

統計資訊

87549

取代數量

任務結果

Show 10 entries

Search:

system_id	result
1	外資・排行 股票名稱 百萬 收盤價 漲跌・1 2454聯發科 1328 718 +7・2 4763材料-KY 1062 1090 +93・3 3034聯詠 625 422 +11.5・4 3008大立光 616 2165 +75・5 2204中華 530 112.5 +4・6 3324雙鴻 497 320 +14・7 2357華碩 492 394 -6・8 1477聚陽 433 342.5 +13.5・9 3035智原 427 357 +12・10 3044健鼎 425 195 +9・資料來源：・https://tinyurl.com/57bdwj9z・投信・排行 股票名稱 百萬 收盤價 漲跌・1 3324雙鴻 492 320 +14・2 6274台權 318 141.5 +12.5・3 3044健鼎 216 195 +9・4 3034聯詠 212 422 +11.5・5 2379 瑞昱 188 449 +14・6 6285啟碁 179 132 +2・7 2303聯電 173 46.55 +0.55・8 8358金居 158 70.6 +2.5・9 2603長榮 156 108.5 +1・10 8299群聯 156 427.5 0・資料來源：・https://tinyurl.com/2a4ur4ap・以上 謝謝

接下來針對 result 欄位做中文斷詞 (這部分我們沒有特別定義詞彙與權重)

≡ 中文斷詞 (36)

參數設定

Input - 30

任務結果

選擇處理欄位 *

result

定義詞彙

以換行符號區隔, e.g.
詞彙 權重
國立中山大學 1000
西子灣 500
...

選取字典

-----請選擇-----

≡ 中文斷詞 (36)

參數設定

Input - 30

任務結果

統計資訊

8253

最大字數

3021401

總字數

0

最小字數

330

平均字數

任務結果

Show 10 entries

Search:

system_id	result
1	[外資, , 排行, 股票, 名稱, 百萬, 收盤價, 漲跌, , , 1, 2454, 聯發科, 1328, 718, +, 7, , , 2, 4763, 材料, -, KY, 1062, 1090, +, 93, , , 3, 3034, 聯詠, 625, 422, +, 11.5, , , 4, 3008, 大立光, 616, 2165, +, 75, , , 5, 2204, 中華, 530, 112.5, +, 4, , , 6, 3324, 雙鴻, 497, 320, +, 14, , , 7, 2357, 華碩, 492, 394, -, 6, , , 8, 1477, 聚陽, 433, 342.5, +, 13.5, , , 9, 3035, 智原, 427, 357, +, 12, , , 10, 3044, 健鼎, 425, 195, +, 9, , , 資料, 來源, :, , , https, :, /, /, tinyurl, ...]

使用功能<清除停用詞>，詳細設定可以參閱下圖，我們透過多次情緒分析結果的文字雲，來回觀察並增訂所需停用之字詞

選擇清除英文字母，因文章內容常用有網址、無效英文縮寫等

選擇清除數字，例如股票代碼等與情緒分析較無關之資料，以及技術性專有名詞等無情緒表達的字詞。
定義停止詞設定，我們特別設定 88 個 PTT 版上包含有，

1. 常有文本的固定格式字詞，例如有依版規、內文、標題、資料來源等詞彙
2. 感謝詞包含謝謝、感謝等非表正面情緒的禮貌性用詞
3. 文章內表達時間序，例如過去、最近、現在、目前、未來、今天、今日、今年、去年、明年等詞彙
4. 常出現但與情緒分析較無相關之詞彙，例如成交量、股名、股票代碼、政府、產業等詞彙

清除停用詞 (40)

參數設定

Input - 36

任務結果

語言 *

Chinese

是否清除單字元

是

清除英文字母 *

是

清除換行符號 *

是

清除html tag *

是

使用預設停止詞

是

是否轉為小寫英文

是

清除數字 *

是

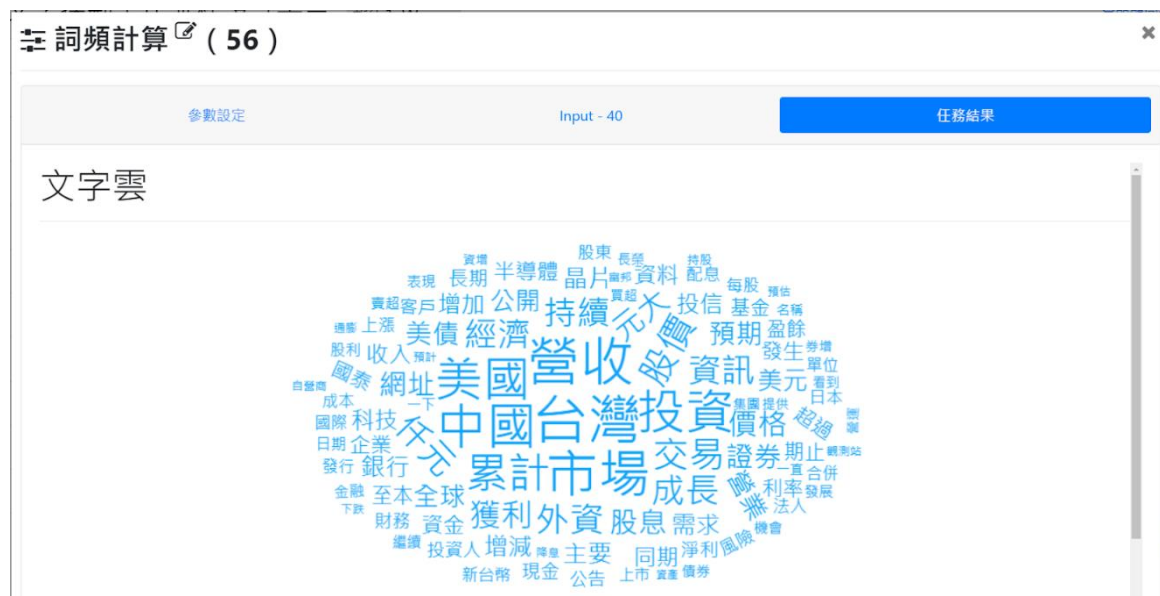
清除特殊標點符號 *

是

自定義停止詞

imgur
JPG
HTTP
資料來源
內文

經由來回設定停用字，透過<詞頻計算>產出文字雲，取得較為整潔、且可初步了解概況的文字雲



使用功能<詞性篩選>，依照預設選項

詞性篩選 (64)

參數設定

Input - 40

任務結果

語言 *
Chinese

選擇保留詞性 *
Noun
Proper Noun
Verb
Adjective
Adverb

儲存更改

使用功能<資料篩選>，我們想針對「長榮」作為分析的主要範圍，選出 666 筆資料

資料篩選 (70)

參數設定

Input - 64

任務結果

條件式 * ⓘ
result.str.join(' ').str.contains('長榮')

儲存更改

資料篩選 (70)

參數設定

Input - 64

任務結果

統計資訊

8428
移除數量

666
保留數量

使用功能<進階字典法-情緒分析>，我們選擇< NTUSD 字典>，進階字典法情緒分析是一種將文字轉換為情緒分數的方法，用來評估文本的情感是正面或負面。

為了能提高文本中隱含情感，能對詞彙分析更加準確，我們適當清除停用字、及自定義正負面詞彙。

依股票版中習慣用語定義正面詞彙，包含有「反彈、樂觀、漲、看好、回升、賺、牛、歐印紅、利多、成長」，「反彈」此詞彙在平常日常用語中較屬於表負面，但在股票討論時是表示正面情緒的。

依股票版中習慣用語定義反面詞彙，包含有「恐慌、熊、賠、公園、綠、利空、衰退」，有趣的詞彙例如「公園」，在 PTT 的討論風氣中，當面臨股票市場走跌，網友常會表達要睡公園的說法，來表達對股票市場負面情緒。

進階字典法 情緒分析 (58)

參數設定Input - 40任務結果

選取字典 ⓘ
NTUSD

移除情緒詞 ⓘ
以換行符號區隔，e.g.
難過
高興
悲傷...

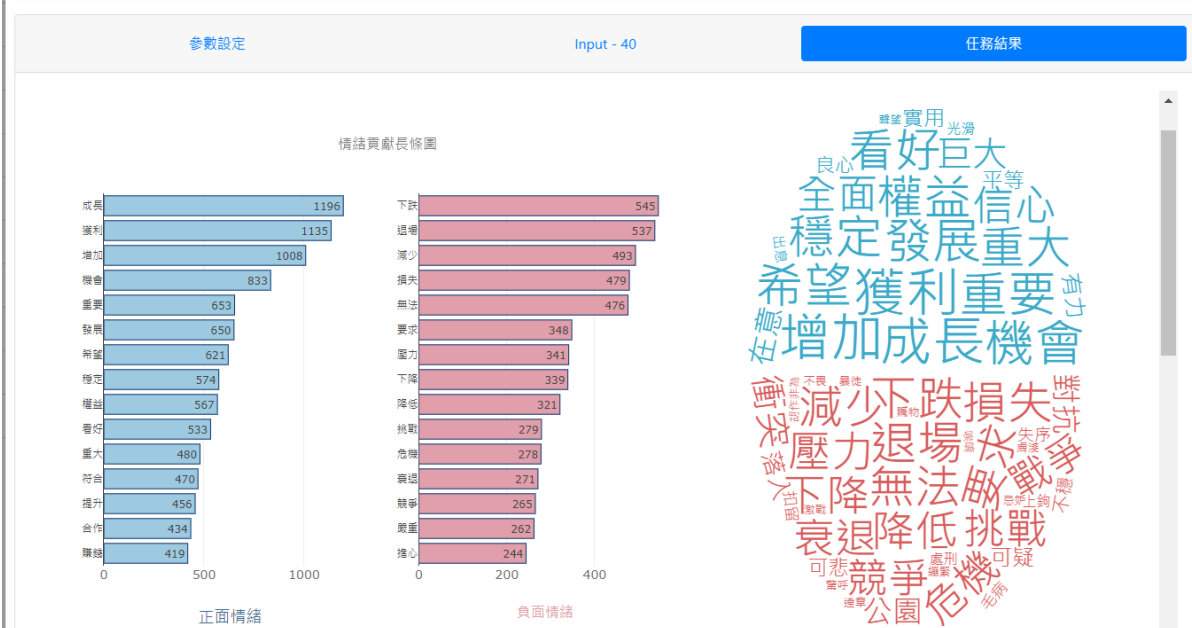
定義正面詞彙 ⓘ
反彈
樂觀
漲
看好
回升

定義負面詞彙 ⓘ
跌
恐慌
熊
賠
公園

是否使用否定詞 * ⓘ
否

是否使用加強詞 * ⓘ
否

進階字典法 情緒分析 (58)



情緒分析後我們篩選整體情緒為負面的 (value < 0)

資料篩選 (66)

參數設定Input - 58任務結果

條件式 * ⓘ
sentiment_value < 0

再跟我們以長榮為關鍵字篩選的資料合併

合併資料 (72)

參數設定 Input - 66 Input - 70 任務結果

JOIN規則

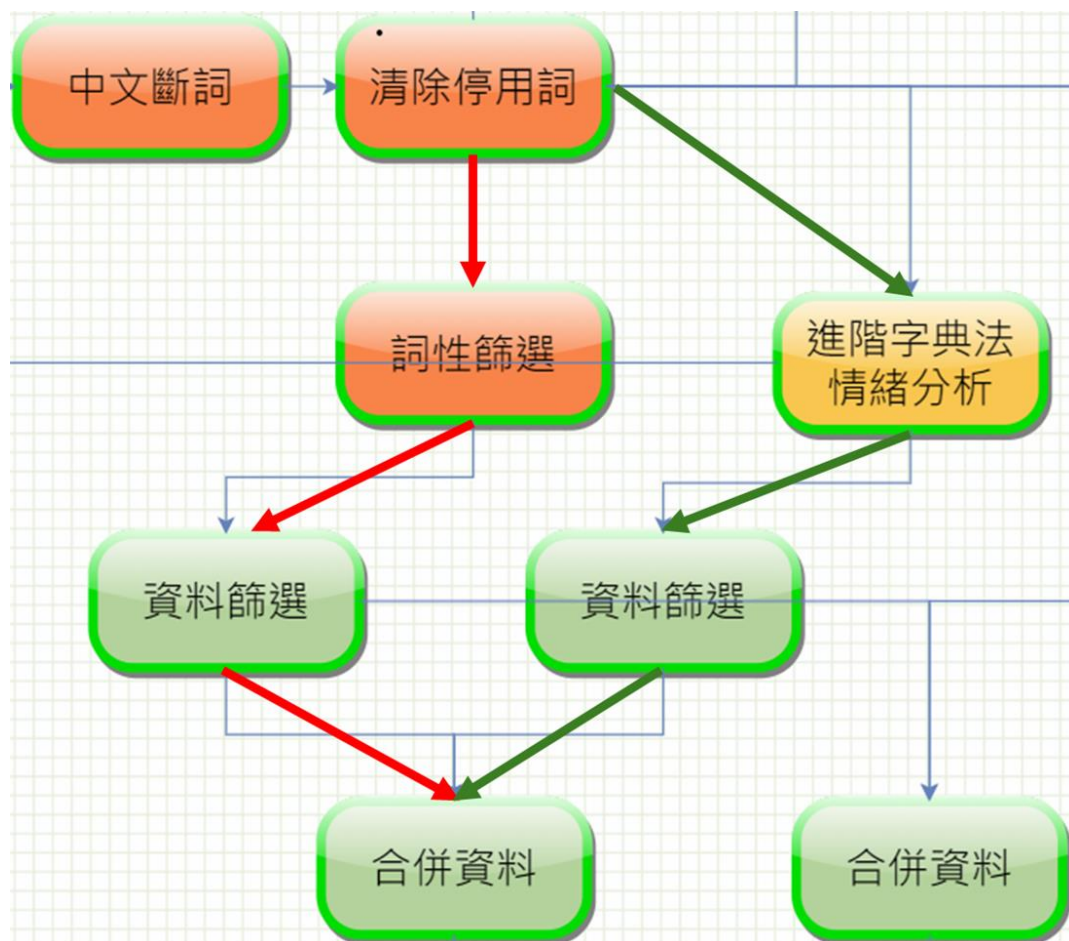
新增規則 刪除規則

任務一欄位 任務二欄位

system_id system_id

請選擇 請選擇

就是把紅線產出的資料與綠線產出的資料以 ID 合併



並再對合併後的 result_Filter 欄位做斷詞，依幾次詞頻計算下來找到出現頻率較高的專有名詞(Ex：脫硫塔、未依板規等)，並以此設定斷詞

中文斷詞 (79)

參數設定 Input - 72 任務結果

選擇處理欄位 *

result_Filter

定義詞彙

脫硫塔 900
未依板規 500
刪文 500
刪文處分 500

選取字典

請選擇

再次清除停用詞

參數設定

Input - 40

任務結果

語言 *

Chinese

正面詞彙 ①

反彈
樂觀
漲
看好
回升

負面詞彙 ①

跌
恐慌
熊
賠
公園

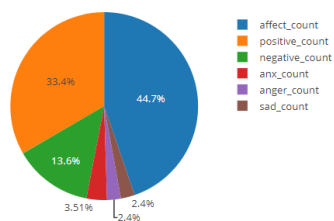
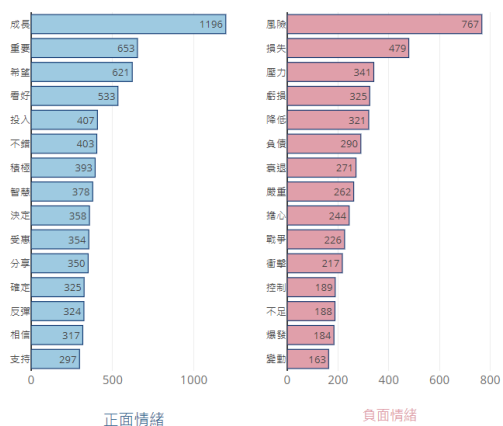
參數設定

Input - 40

任務結果

情緒分佈

情緒貢獻長條圖



篩選整體情緒偏向負面的資料

資料篩選 (68)

參數設定

Input - 59

任務結果

條件式 * 1

sentiment_value < 0

資料篩選 (68)

參數設定

Input - 59

任務結果

統計資訊

6320
移除數量

2811
保留數量

一樣與詞性篩選後的資料做合併

合併資料 (75)

參數設定

Input - 68

Input - 70

任務結果

JOIN規則

新增規則

刪除規則

任務一欄位

system_id

-----請選擇-----

任務二欄位

system_id

-----請選擇-----

儲存更改

設定斷詞

中文斷詞 (82)

參數設定

Input - 75

任務結果

選擇處理欄位 *

result_Filter

定義詞彙 1

脫硫塔 900
未依板規 500
刪文 500
刪文處分 500

選取字典 1

-----請選擇-----

清除停用詞

分群彙總(日期)-positive/negative (112)

參數設定

Input - 95

任務結果

選擇日期欄位 *

artDate

選擇日期類型 *

日

計算欄位(按住ctrl(Windows)或command(MAC)可以複選) *

artComment
e_ip
insertedDate
dataSource
positive_count
positive_words
negative_count

日期格式 * ⓘ

%Y-%m-%d %H:%M:%S%z

匯總函數 * ⓘ

sum

保留欄位(按住ctrl(Windows)或command(MAC)可以複選) *

system_id
artUrl
artTitle
artDate
artPoster
artCategory
artContent

任務結果

Show 10 entries

Search:

artDate	positive_count	negative_count
2023-09-05	46	9
2023-09-06	381	226
2023-09-07	372	182
2023-09-08	446	275
2023-09-09	280	249
2023-09-10	233	178

也可以分月查詢

分群彙總(日期) (116)

參數設定

Input - 95

任務結果

選擇日期欄位 *

artDate

選擇日期類型 *

月

計算欄位(按住ctrl(Windows)或command(MAC)可以複選) *

artComment
e_ip
insertedDate
dataSource
positive_count
positive_words
negative_count

日期格式 * ⓘ

%Y-%m-%d %H:%M:%S%z

匯總函數 * ⓘ

sum

保留欄位(按住ctrl(Windows)或command(MAC)可以複選) *

system_id
artUrl
artTitle
artDate
artPoster
artCategory
artContent

任務結果

Show 10 entries

Search:

artDate	positive_count	negative_count
2023-09	7352	4997
2023-10	7683	5059
2023-11	7387	4134
2023-12	8720	5749
2024-01	8608	5899
2024-02	7901	4819

同時還想要觀察每天的平均的情緒分數，所以計算 sentiment_value 的 mean

≡ 分群彙總(日期) (92)

參數設定

Input - 95

任務結果

選擇日期欄位 *


artDate

選擇日期類型 *


日

計算欄位(按住ctrl(Windows)或command(MAC)可以複選) *

insertedDate
dataSource
positive_count
positive_words
negative_count
negative_words
sentiment_value

日期格式 * 

%Y-%m-%d %H:%M:%S%z

匯總函數 * 

mean

保留欄位(按住ctrl(Windows)或command(MAC)可以複選) *

system_id
artUrl
artTitle
artDate
artPoster
artCategory
artContent

欄位重構

≡ 欄位重構(Melt) (118)

參數設定

Input - 95

任務結果

要轉換的欄位(按住ctrl(Windows)或command(MAC)可以複選) *

artComment
e_ip
insertedDate
dataSource
positive_count
positive_words
negative_count

要保留的欄位(按住ctrl(Windows)或command(MAC)可以複選)

artUrl
artTitle
artDate
artPoster
artCategory
artContent
artComment

轉換的欄位新的欄位名稱 *

variable

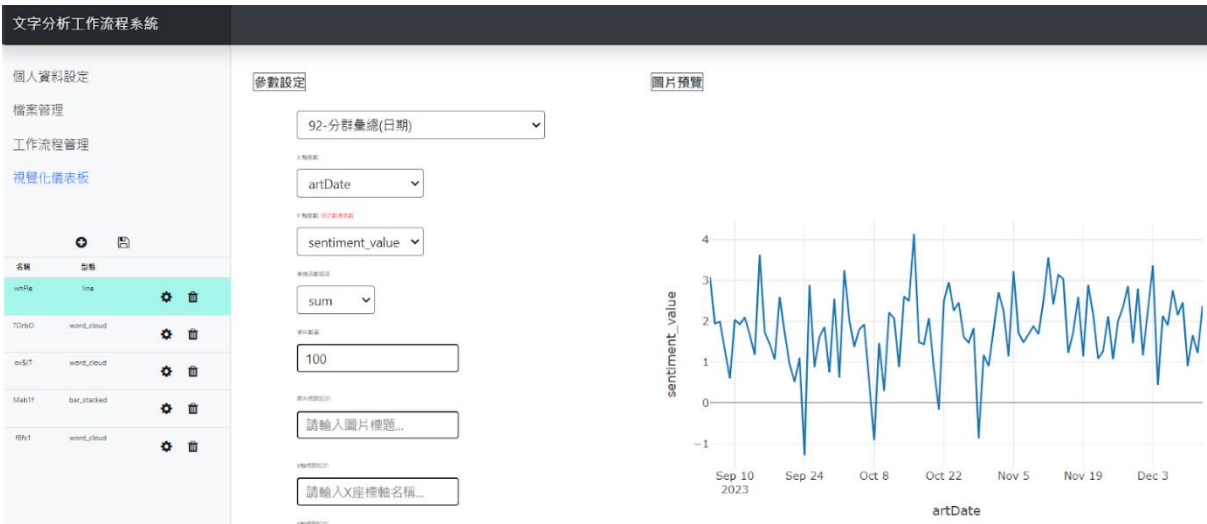
值的欄位名稱 *

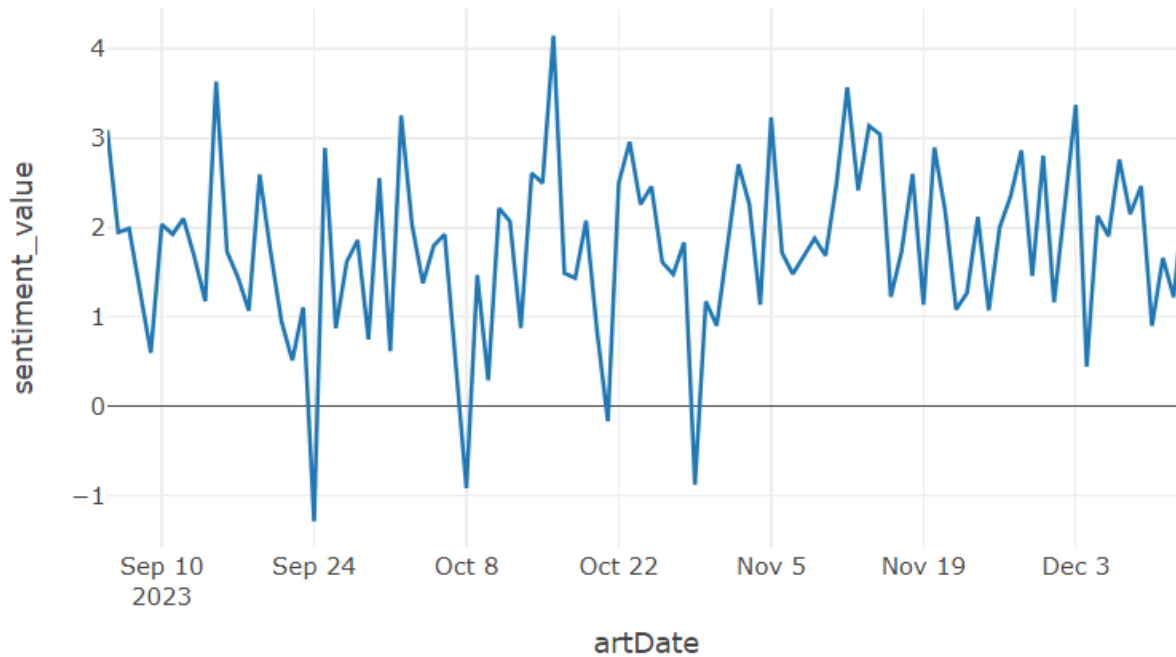
value

視覺化儀表板

1.折線圖

使用折線圖來表達去年 9 月-12 月 · ptt 股票版中使用情緒分析所得知的每日平均情緒分數





2.文字雲圖

選擇<進階字典法情緒分析>的分析結果，相較 LIWC 有更多負面字詞

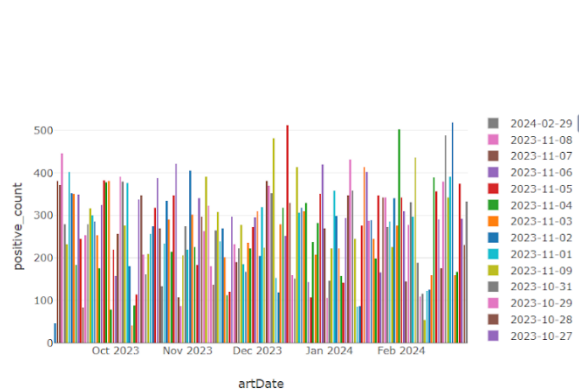


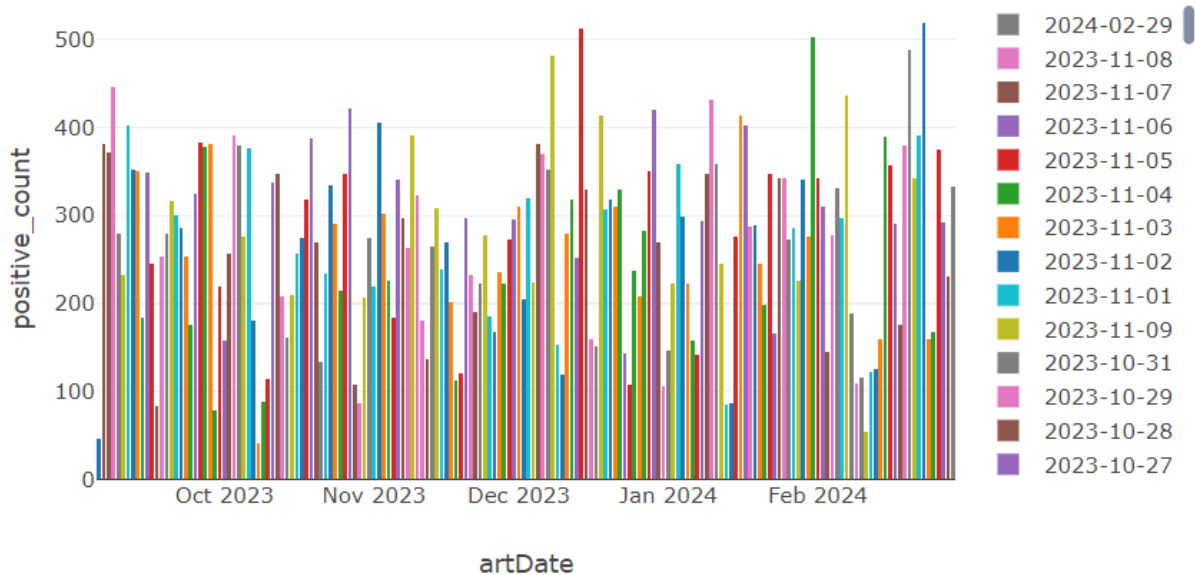
3.文字雲圖

選擇<LIWC 情緒分析>的分析結果



以長條圖表達自去年 9 月-12 月，ptt 股票版中使用情緒分析取得每日正面詞彙的數量





5. 文字雲

選擇<進階字典法情緒分析>的分析結果

資料數字設為更大的 500，可以觀察到文字雲的改變

參數設定

圖片預覽

76-詞頻計算

文字類別

Term

詞頻類別: 詞頻計算結果

n

詞頻數量

500

預覽

