**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

一張含有 螢幕擷取畫面, 文字, 圖表 的圖片

自動產生的描述

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

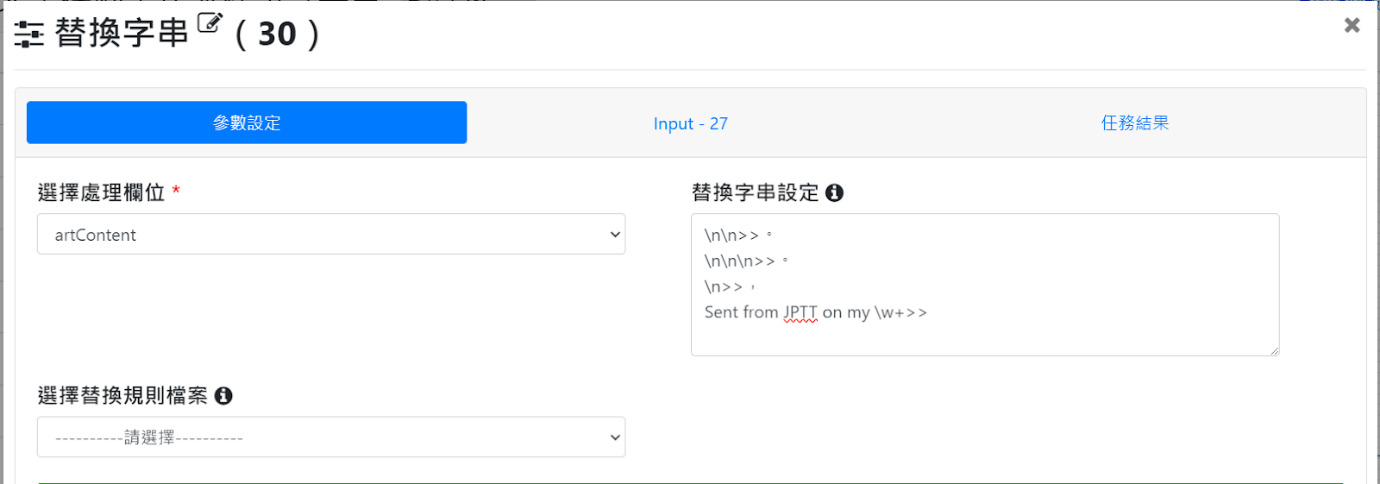
一張含有 文字, 螢幕擷取畫面, 軟體, 數字 的圖片

自動產生的描述

一張含有 文字, 軟體, 數字, 字型 的圖片

自動產生的描述

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



一張含有 文字, 螢幕擷取畫面, 軟體, 數字 的圖片

自動產生的描述

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)

一張含有 文字, 數字, 字型, 螢幕擷取畫面 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 軟體, 數字 的圖片

自動產生的描述

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

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

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

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

使用功能<詞性篩選>，依照預設選項

一張含有 文字, 軟體, 數字, 字型 的圖片

自動產生的描述

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

一張含有 文字, 螢幕擷取畫面, 軟體, 網頁 的圖片

自動產生的描述

一張含有 文字, 字型, 數字, 螢幕擷取畫面 的圖片

自動產生的描述

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。

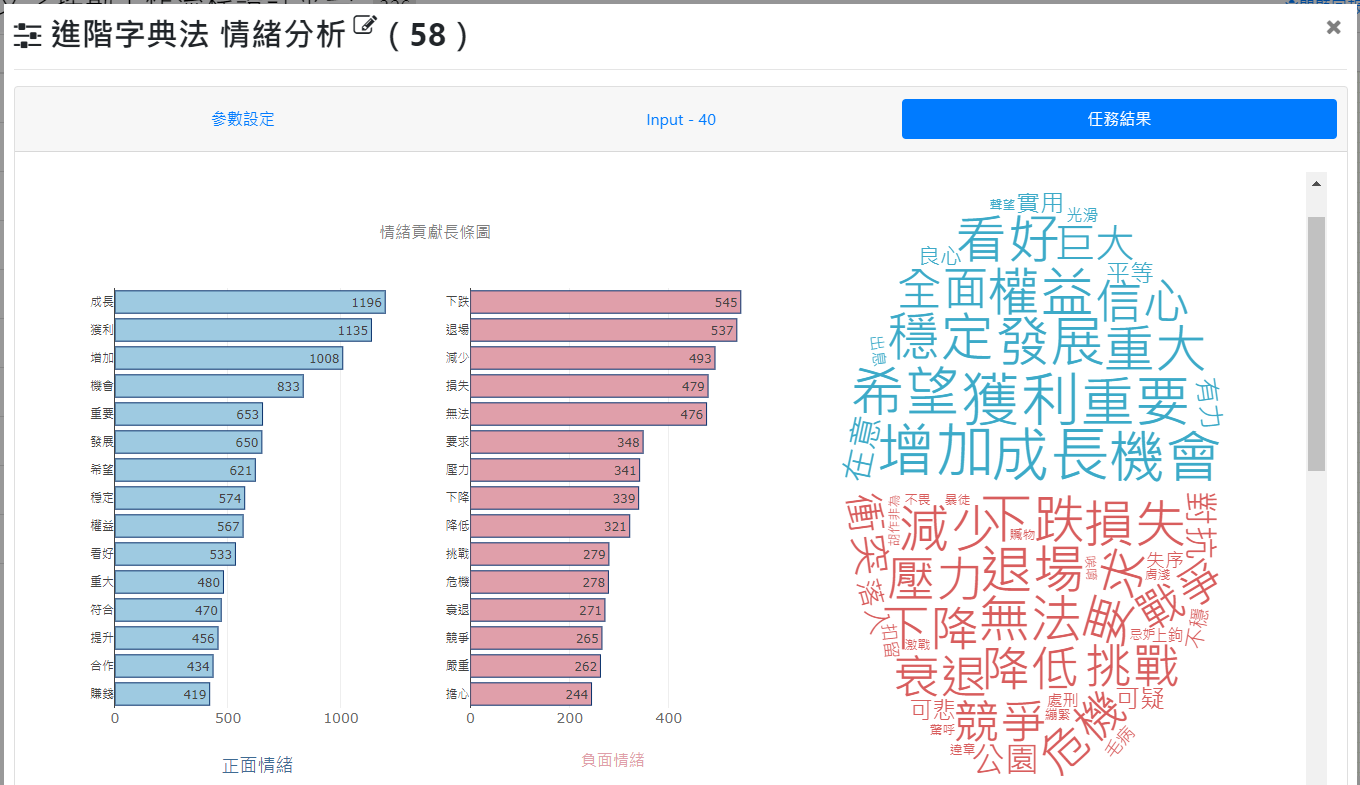
依股票版中習慣用語定義正面詞彙，包含有「反彈、樂觀、漲、看好、回升、賺、牛、歐印

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。

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

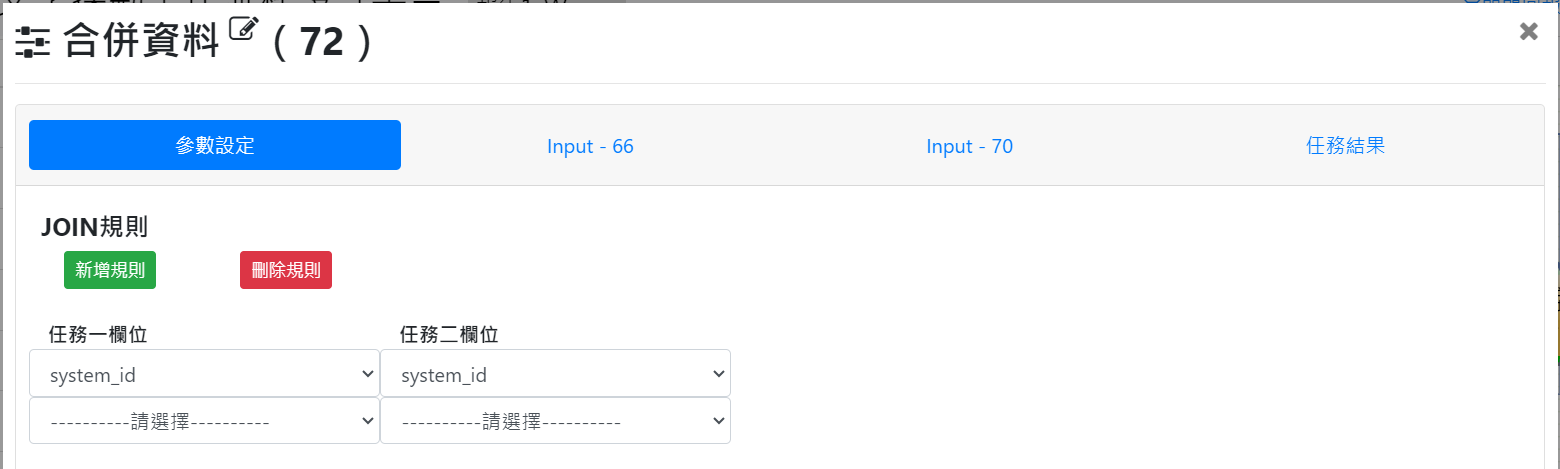


情緒分析後我們篩選整體情緒為負面的( value < 0 )

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

再跟我們以長榮為關鍵字篩選的資料合併



就是把紅線產出的資料與綠線產出的資料以ID合併

一張含有 文字, 螢幕擷取畫面, 圖表, 行 的圖片

自動產生的描述

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

一張含有 文字, 字型, 數字, 螢幕擷取畫面 的圖片

自動產生的描述

再次清除停用詞

一張含有 文字, 螢幕擷取畫面, 數字, 軟體 的圖片

自動產生的描述

最後詞頻計算產出文字雲，了解那些關鍵字常出現在偏向負面情緒的文字內容裡

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

我們另外用 LIWC 情緒分析進行比較，一樣將反彈等詞彙設為正面詞彙，公園等詞彙設為負面詞彙

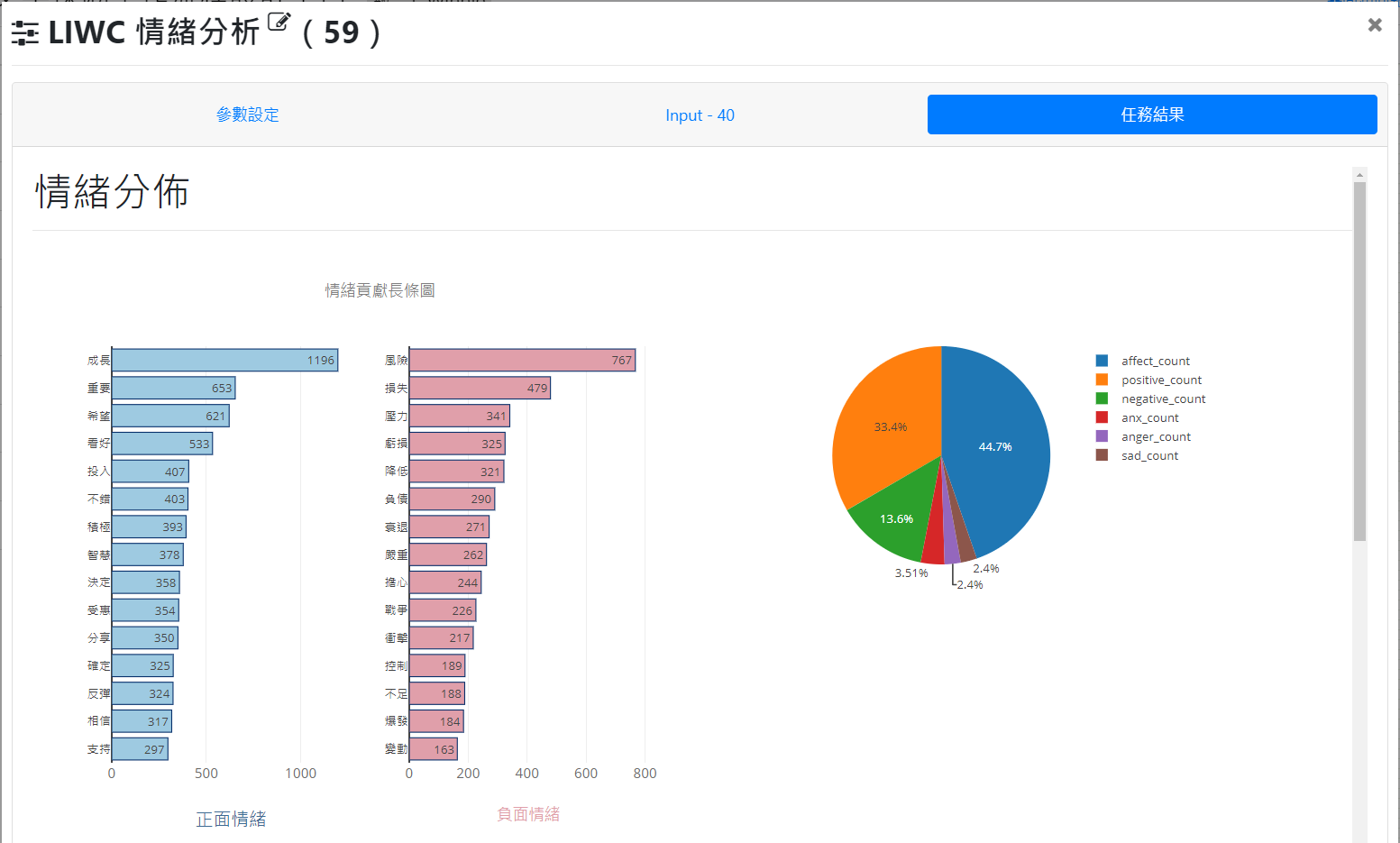
依股票版中習慣用語定義正面詞彙，包含有「反彈、樂觀、漲、看好、回升、賺、牛、歐印

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。

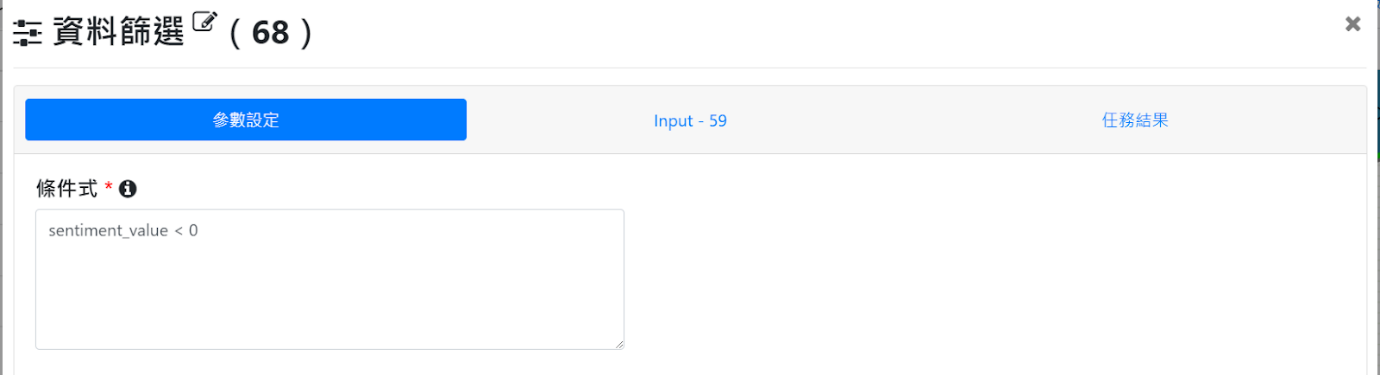
依股票版中習慣用語定義反面詞彙，包含有「恐慌、熊、賠、公園、綠、利空、衰退」，有趣的詞彙例如「公園」，在PTT的討論風氣中，當面臨股票市場走跌，網友常會表達要睡公園的說法，來表達對股票市場負面情緒。

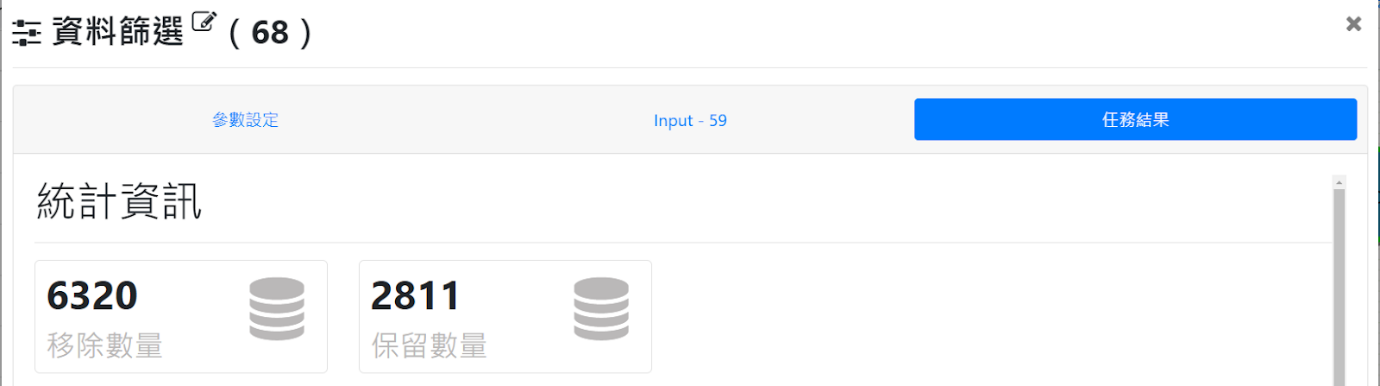
一張含有 文字, 行, 數字, 軟體 的圖片

自動產生的描述



篩選整體情緒偏向負面的資料





一樣與詞性篩選後的資料做合併

一張含有 文字, 軟體, 電腦圖示, 數字 的圖片

自動產生的描述

設定斷詞

一張含有 文字, 字型, 數字, 螢幕擷取畫面 的圖片

自動產生的描述

清除停用詞

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

再次詞頻計算產出文字雲，了解那些關鍵字常出現在偏向負面情緒的文字內容裡

一張含有 文字, 螢幕擷取畫面, 軟體, 數字 的圖片

自動產生的描述

使用功能<分群彙總>

將從 PPT 爬蟲完的文檔和用 "進階字典法情緒分析" 完的結果透過合併資料功能合併起來，接著透過分群彙總計算出我們要觀察的數值，選定 artDate 當作分群的日期欄位，計算 postive\_count 和 negative\_count 的 sum。

可以分日查詢

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 收據, 數字 的圖片

自動產生的描述

也可以分月查詢

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

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

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

欄位重構

一張含有 文字, 字型, 數字, 軟體 的圖片

自動產生的描述

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

一張含有 文字, 螢幕擷取畫面, 字型, 圖表 的圖片

自動產生的描述

一張含有 文字, 筆跡, 字型, 行 的圖片

自動產生的描述

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

一張含有 螢幕擷取畫面, 文字, 圖表 的圖片

自動產生的描述

一張含有 文字, 字型, 設計 的圖片

自動產生的描述

3.文字雲圖

選擇<LIWC情緒分析>的分析結果

一張含有 文字, 軟體, 螢幕擷取畫面, 電腦圖示 的圖片

自動產生的描述

一張含有 文字, 字型, 筆跡, 設計 的圖片

自動產生的描述

4.長條圖

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

一張含有 文字, 螢幕擷取畫面, 軟體, 設計 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 鮮豔, 繪圖 的圖片

自動產生的描述

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

一張含有 文字, 螢幕擷取畫面, 字型, 圖表 的圖片

自動產生的描述