**社群媒體分析第一次讀書會**

**議題分析**

指導老師：黃三益 教授

組別：第六組

組員：

N104020001 李采容、N104020002 廖英捷

N104020007 郭育雯、N104020008 游淑媛

N104020009 蔡雨臻、N104020010 盧貴聰

M106020015 林猷盛、B096060020 黃湘安

**目錄**

[一、分析議題說明 3](#_Toc130207455)

[二、工作流程設計 4](#_Toc130207456)

[三、資料清理 6](#_Toc130207457)

[四、情緒分析 9](#_Toc130207458)

[五、分群彙總與視覺化圖表 13](#_Toc130207459)

[**分群彙總設計 13**](#_Toc130207460)

[**視覺化圖表 20**](#_Toc130207461)

[六、心得與結論 24](#_Toc130207462)

# **一、分析議題說明**

* 主題：社會大眾對於政府普發六千元政策的網路情緒分析
* 議題發想：

去年全國稅收超徵4500億，行政院長蘇貞昌表示，預計將其中的1400億「還稅於民」，每人可以領取六千元現金。此政策造成民眾諸多關注，網路上也開始有了無排富政策、詐騙問題、未繳稅也可領取等等的話題討論。故本組將探討以下議題：

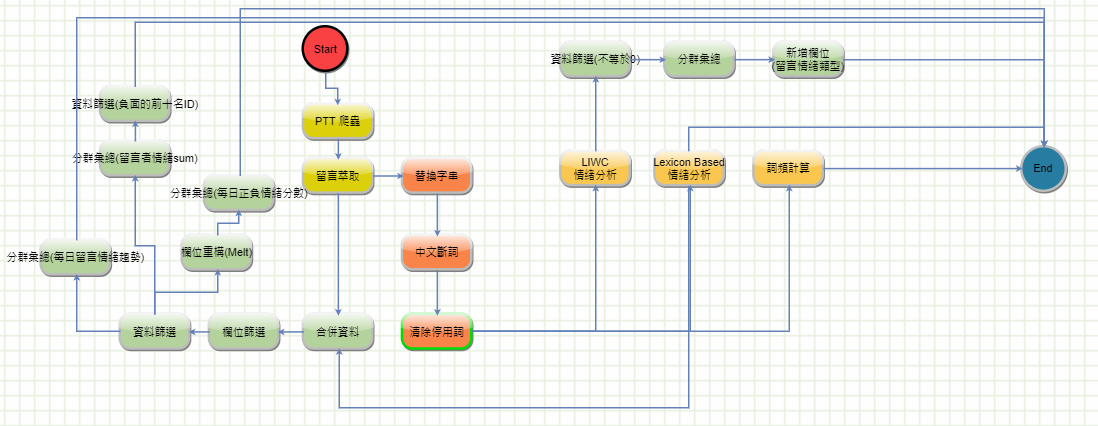
* 對於普發六千有哪些情緒詞彙？
* 網路上對於此政策是正面態度還是負面態度？
* 網路上的情緒趨勢走向為何？
* 使用平台：文字探勘工作流程設計平台

# **二、工作流程設計**

* 工作流程：Chai、testflow
* 視覺化儀表板：Chai、testflow
* 資料來源：PTT留言萃取（省錢、政黑、八卦版）
* 分析期間：112/3/1~112/3/18
* 流程概述：

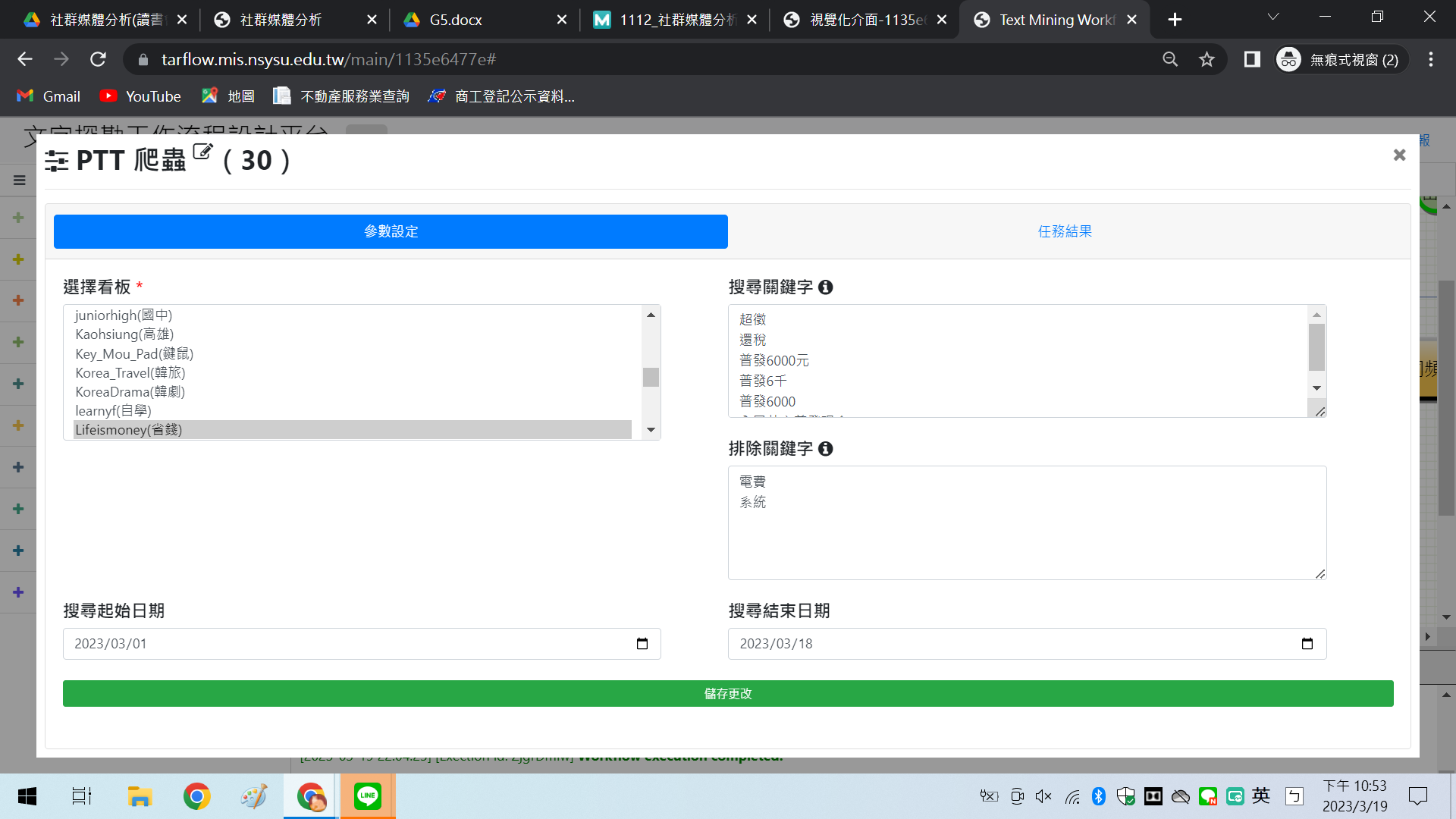
1. 指定六千元相關的關鍵字（超徵、還稅、普發6000元、普發6千、普發6000、全民共享普發現金），讀入PTT爬蟲資料並使用留言萃取讀入PTT底下留言。
2. 將網址、PTT平台發文文字和/n等字詞替換成空白。
3. 用中文斷詞加大詞彙權重。
4. 將不具分析效益的字詞進行停用。
5. 將清理之後的資料，進行Lexicon Based和LIWC兩個情緒分析字典，以瞭解情緒詞彙、正面與負面詞彙的情緒分數。
6. 用詞頻計算選取出現頻率最高的前50個詞組來產生文字雲。
7. LIWC情緒分析之後的資料，篩選出具有情緒的資料（正數與負數），並新增留言情緒類型（Max）的欄位。
8. 將Lexicon Based情緒分析的結果，用合併資料合併PTT留言萃取的爬蟲資料。
9. 用欄位篩選篩選出需要的欄位，並用資料篩選篩選出具有情緒值的資料（負數及正數）。
10. 進行分群彙總（留言情緒趨勢、留言者ID分群、每日留言的情緒分數）等。

* 工作流程設計圖一覽：

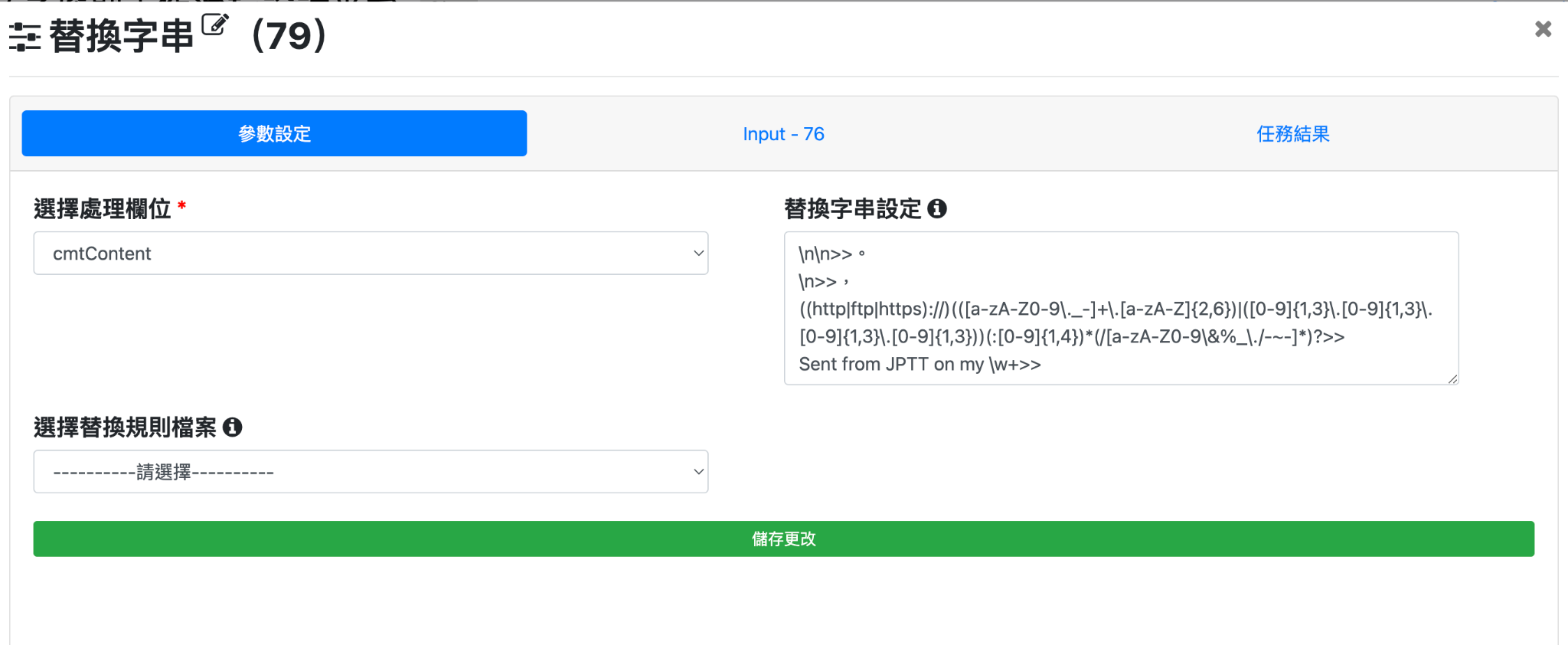


# **三、資料清理**

1. 以「PTT爬蟲」元件截取所需資料，選取「八卦」、「政黑」、「省錢」三個看板的文章，時間區段為2023/01/01~2023/03/18。搜尋的關鍵字包含：超徵、還稅、普發6000元、普發6千、普發6000、全民共享普發現金，並排除關鍵字電費與系統。



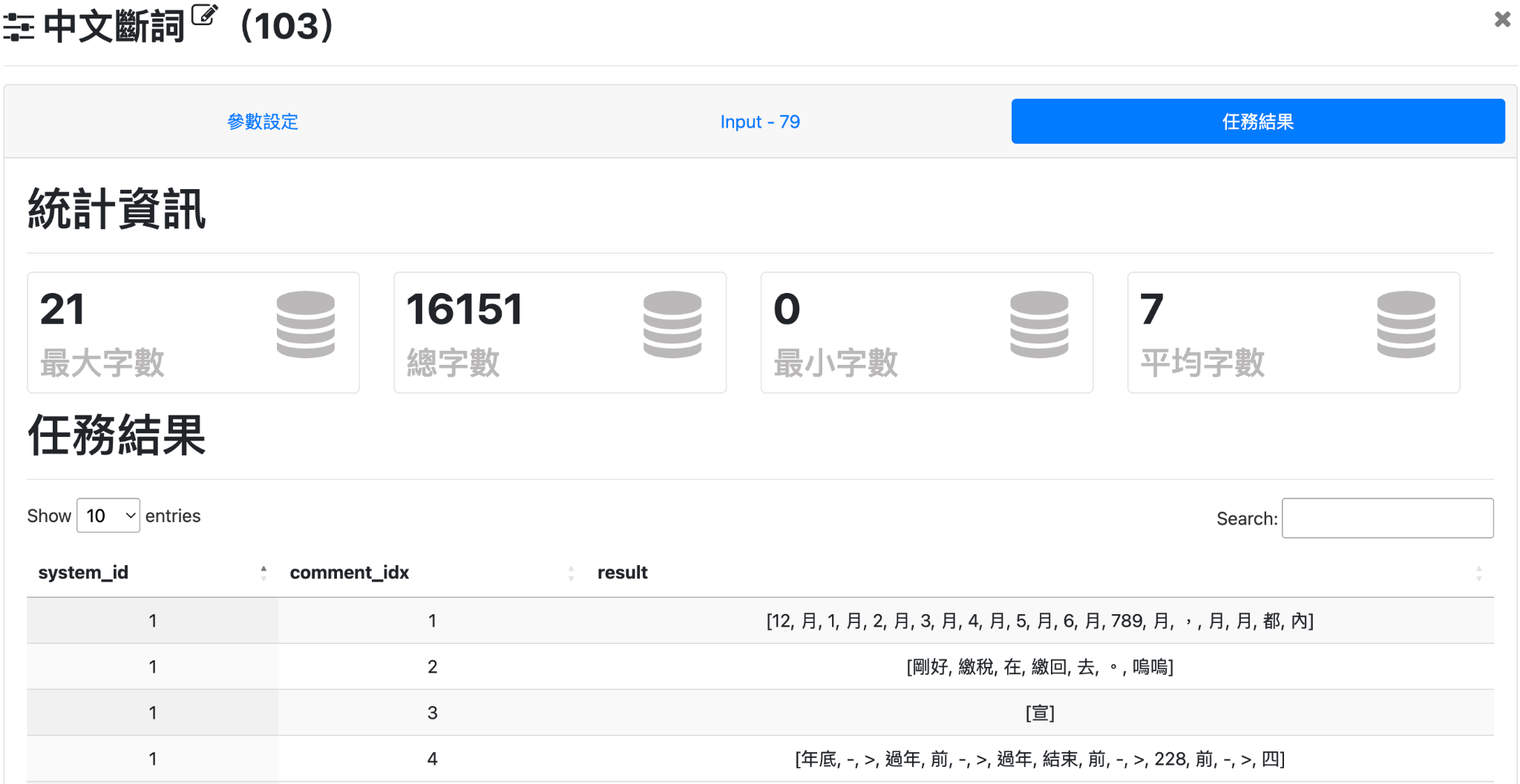
1. 以「替換字串」元件將非文字符號取代



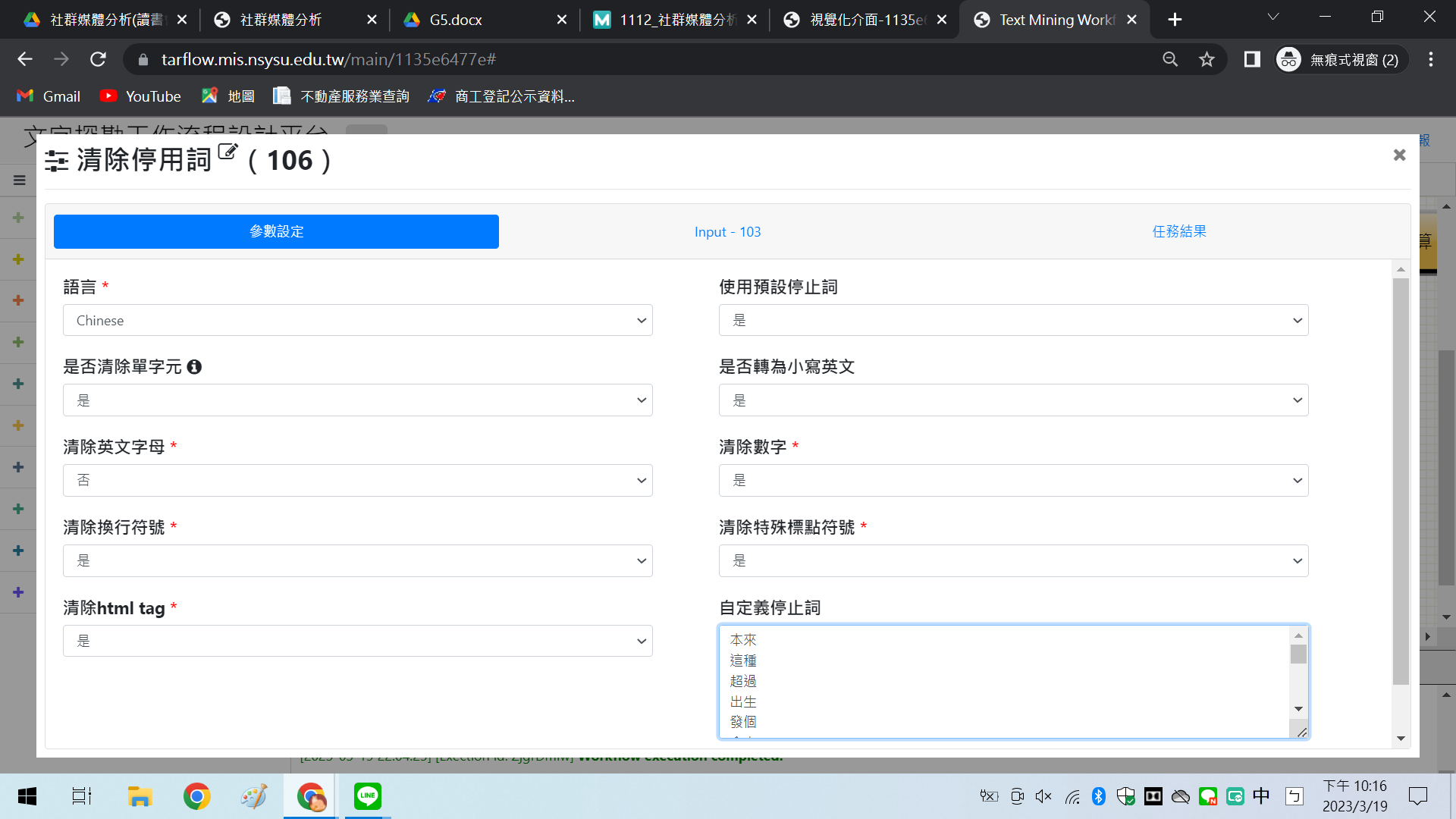
1. 以「中文斷詞」元件將文句分段，並在檢視輸出結果後增加部分字詞權重，使其能夠被篩選出來。例如：「一次性」一詞被分割成「一次」與「性」兩部分，因此手動輸入100的權重，以確保結果符合語意。

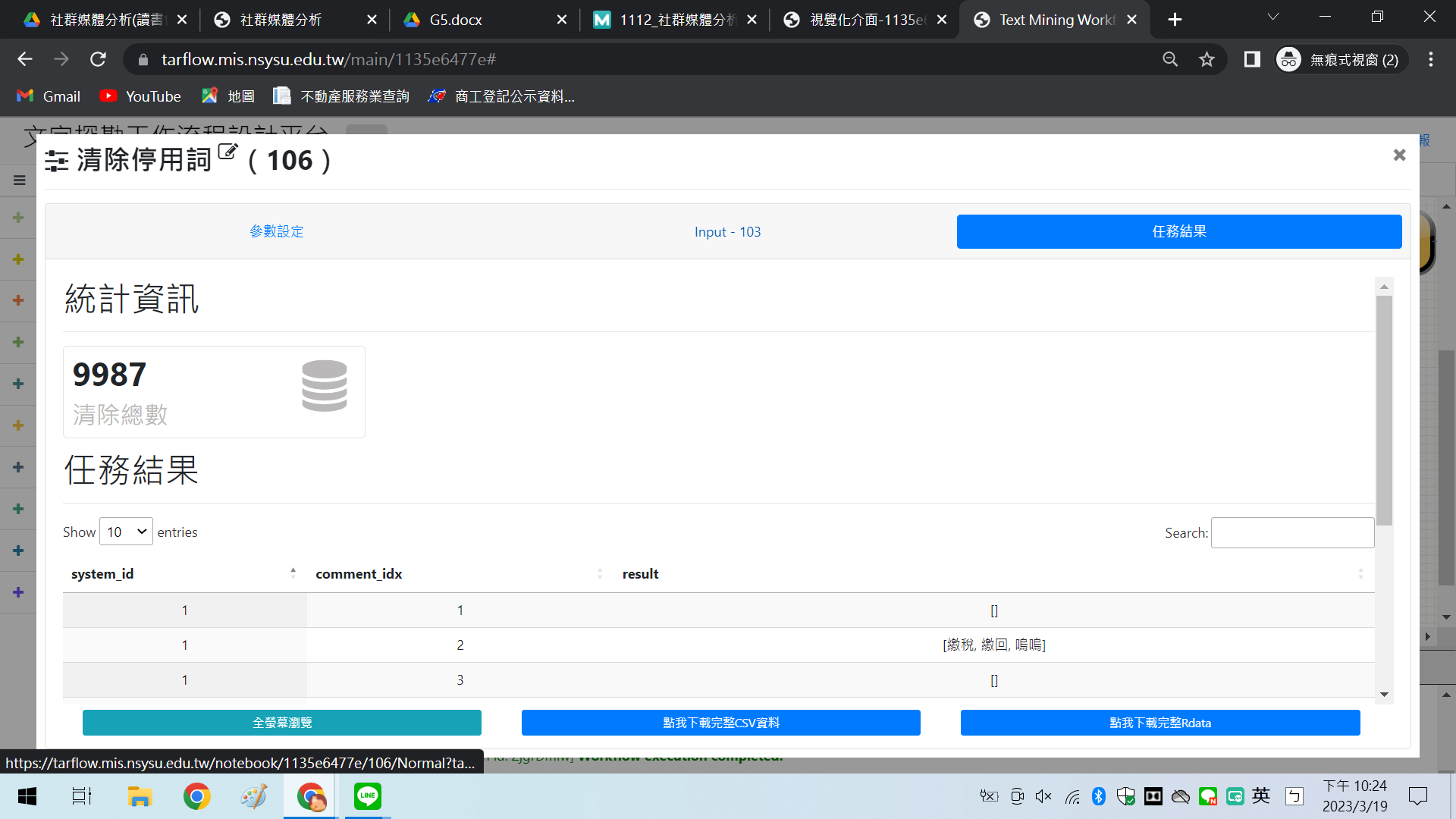


斷詞後獲得的總字數為16,151字



1. 清除停用字：本來、這種、超過、出生、發個、拿去、網站、發個、一年、ooo、東西、反正、這幾年、準備好、好像等等，停用總數為9987個字詞。





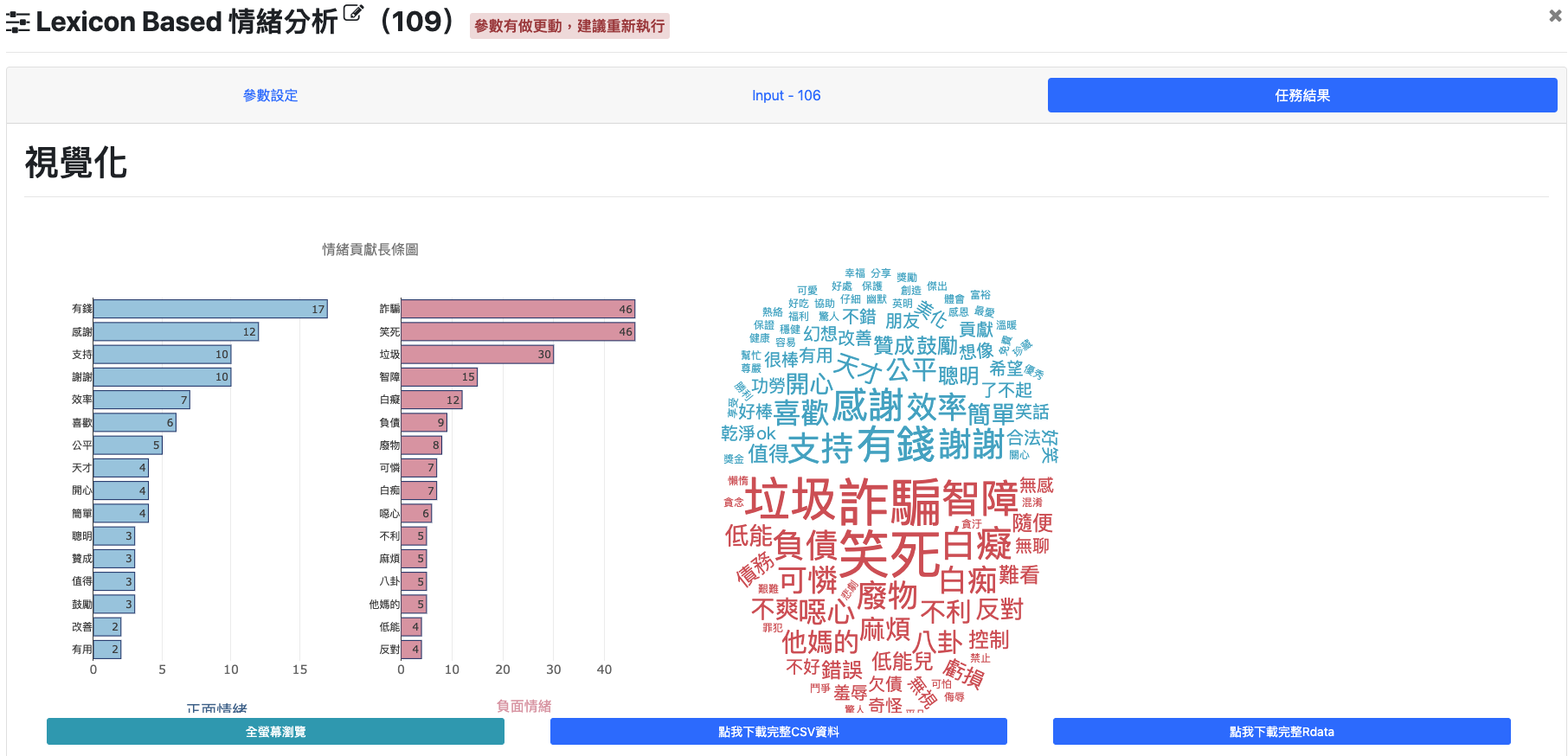
# **四、情緒分析**

採用Lexicon Based與LIWC這兩種情緒字典，目的用來分析PTT鄉民發文內容的情緒文字。文字以Lexicon Based為主要辨析目標。

1. 資料載入預設情緒字典，其設定如下，特別將現今常用鄉民回覆語「笑死」定義為負面詞彙進行排除。



1. Lexicon Based情緒分析結果如下圖，可看到負面詞類型多於正面詞。



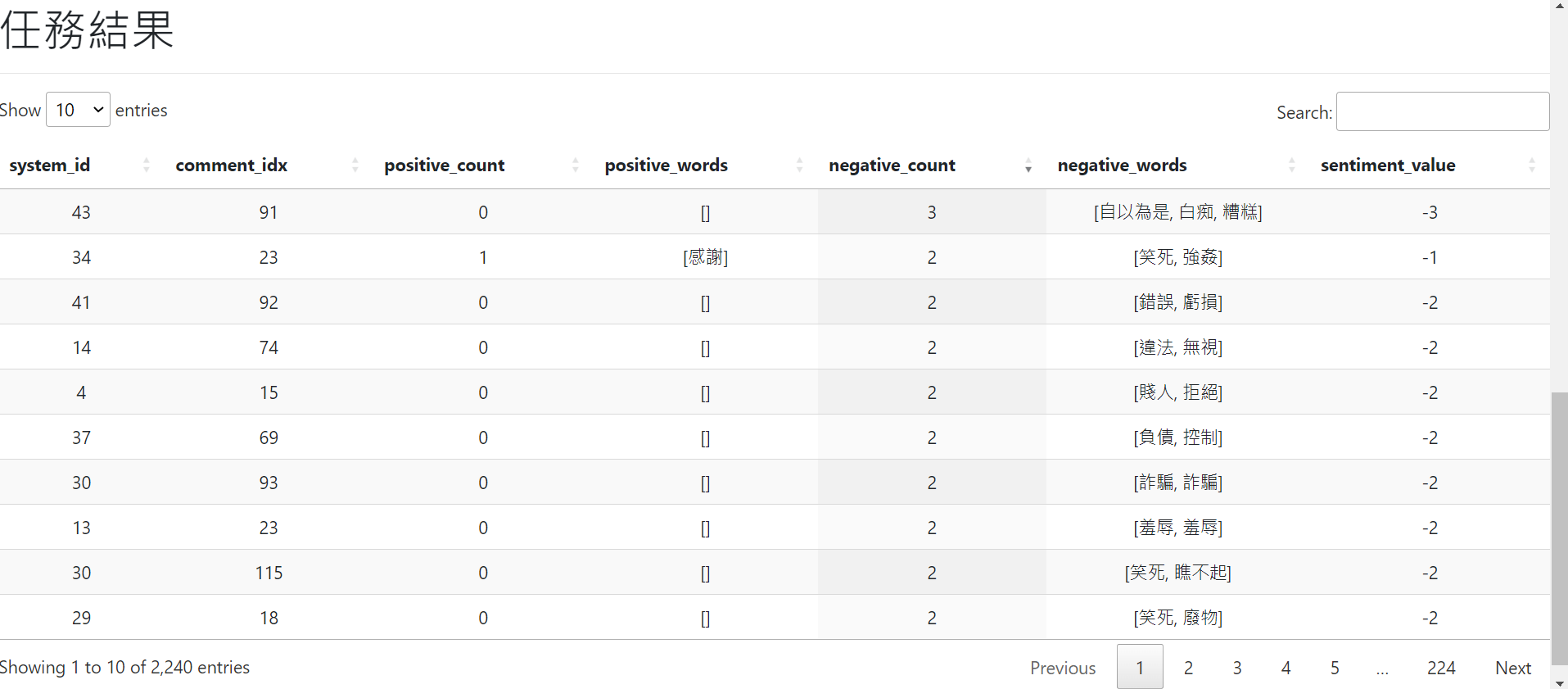
正面字詞，出現高頻率有「有錢、感謝、支持、謝謝、效率、喜歡」。

負面字詞，出現高頻率有「詐騙、笑死、垃圾、智障、白癡、負債」。



由上圖可知，但高頻的負面情緒略高於正面情緒。

1. 下圖為Lexicon Based的情緒輸出結果，可看到「自以為是、白痴、糟糕、錯誤、虧損、違法、無視、賤人、拒絕、負債、控制、詐騙、羞辱、笑死、瞧不起、廢物、垃圾」等負面字詞。大多數鄉民都是以負面思考面向討論政府公告普發6000元的資訊。

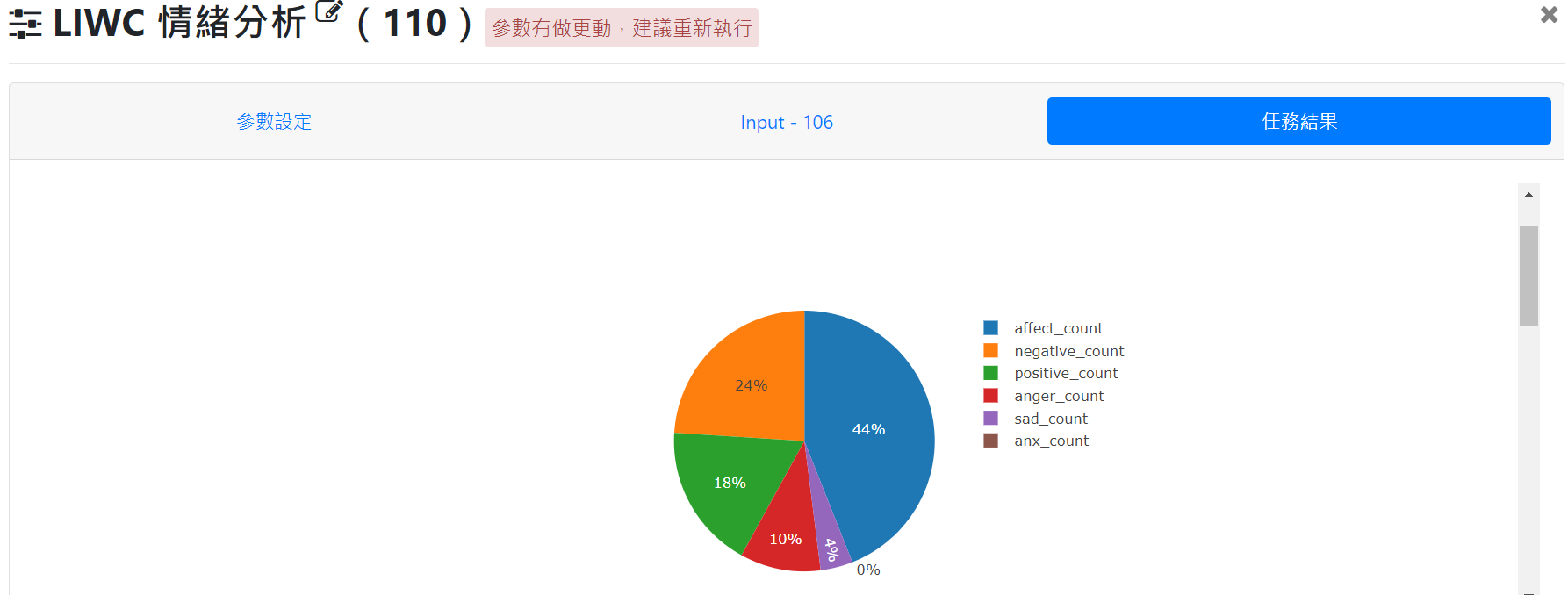


然而，LIWC情緒字詞的分析亦同Lexicon Based的設定「笑死」為負面字詞。



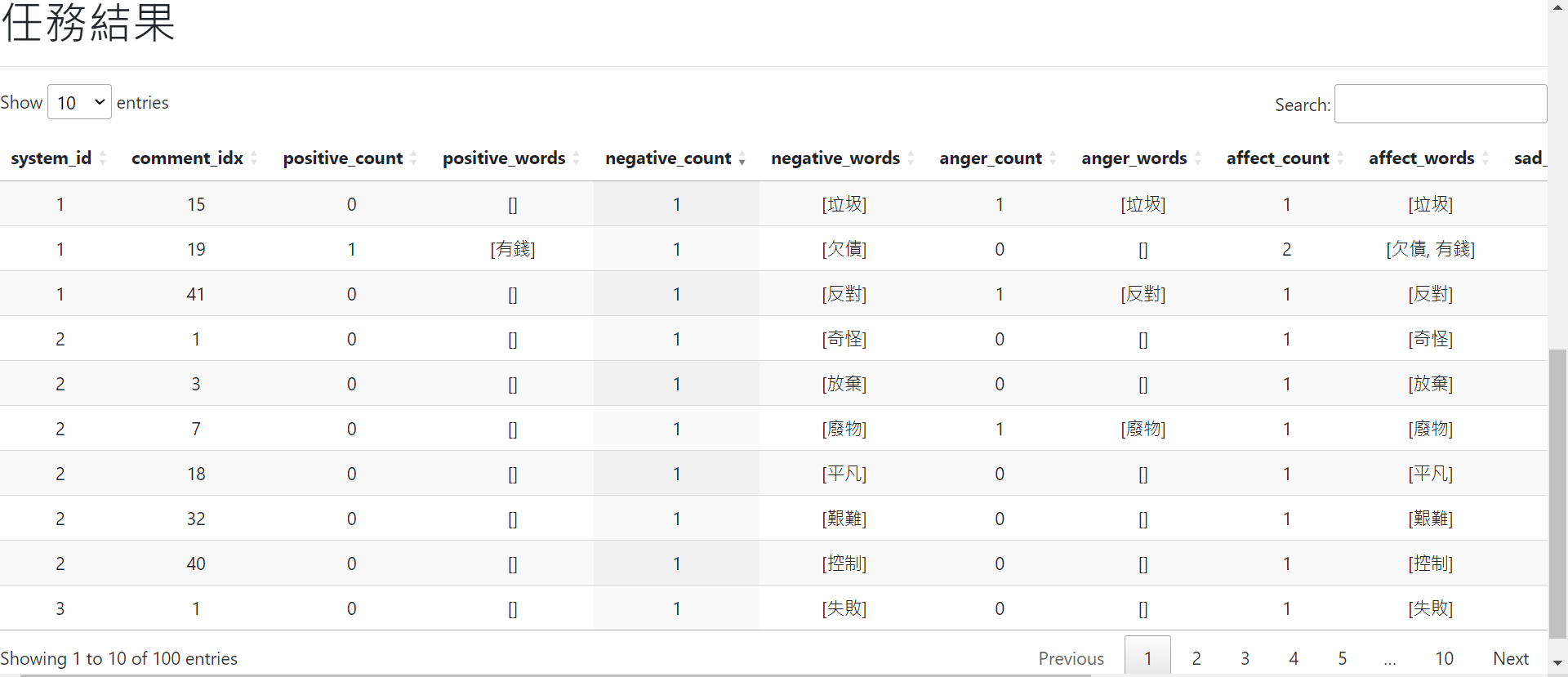
1. LIWC情緒分析雖有較多情緒類型，以分析結果顯示，可看到為負面情緒大於正面情緒，此段區間的討論版負向詞（橘色）高於正向詞（綠色），

* affect\_count 44%
* negative\_count 24%
* positive\_count 18%
* anger\_count 10%
* sad\_count 4%



LIWC較Lexicon Based不同之處在於，情緒字詞於文章回文有重複出現的話，LIWC只會計算一次。經組員們討論評估後，考量到文章以鄉民評論為主軸，其情緒字詞出現次數是可以做為分析正負向的參考，故決定以Lexicon Based的輸出結果，進行後續的情緒分群彙總分析。





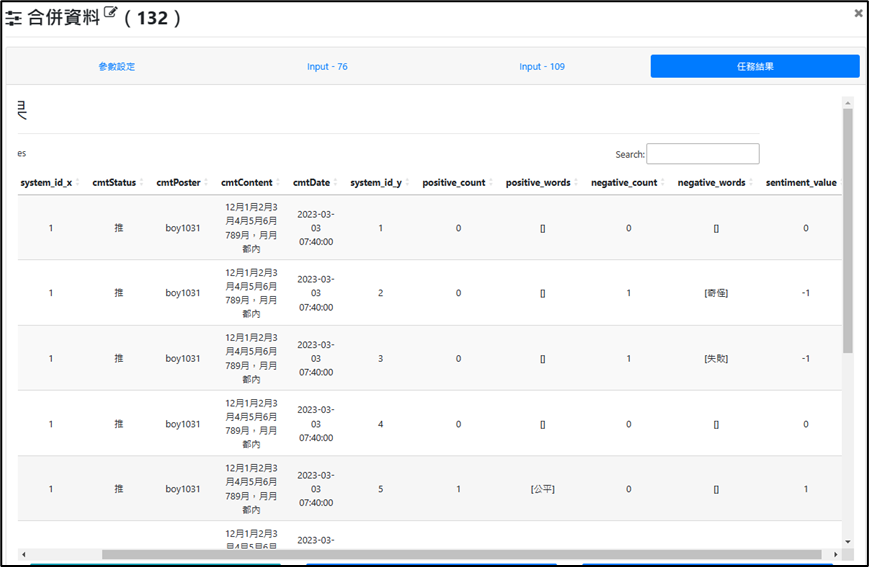
# **五、分群彙總與視覺化圖表**

**分群彙總設計**

1. 合併資料[132]

於分群匯總前先進行「資料合併」，將爬蟲後「留言萃取」與「Lexicon Based情緒分析」資料依comment\_idx進行合併，以利後續依日期來匯總每日情緒值（情緒分析資料無日期欄位，故與爬蟲後資料進行合併）。





1. 欄位篩選[165]

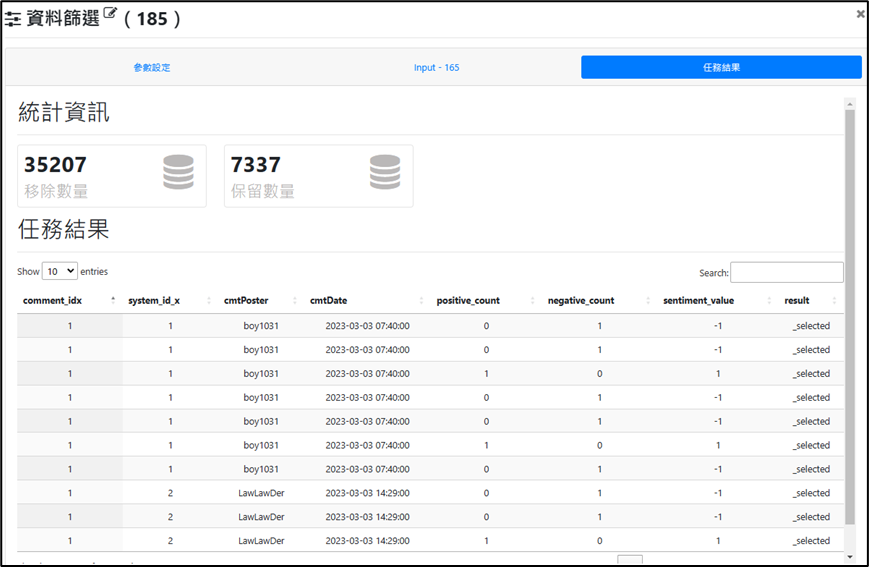
整理一下合併後的資料，移除無用欄位，保存後續可能使用的欄位即可。設定如下：



1. 資料篩選[185]

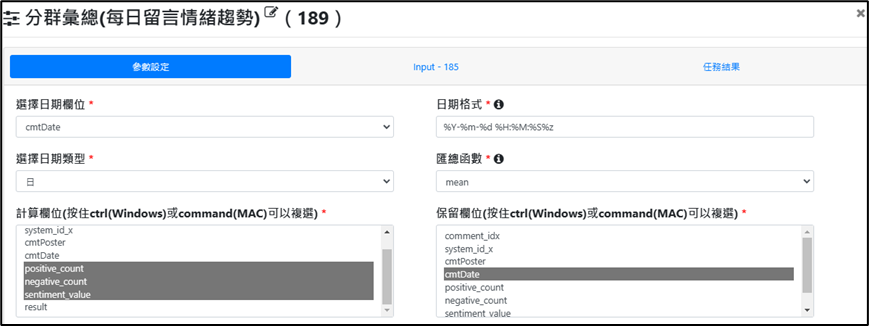
再用「資料篩選」模組保留（情緒值> 0 或 情緒值< 0）包含情緒字眼之資料，此階段保留了7337筆資料。設定與結果如下：

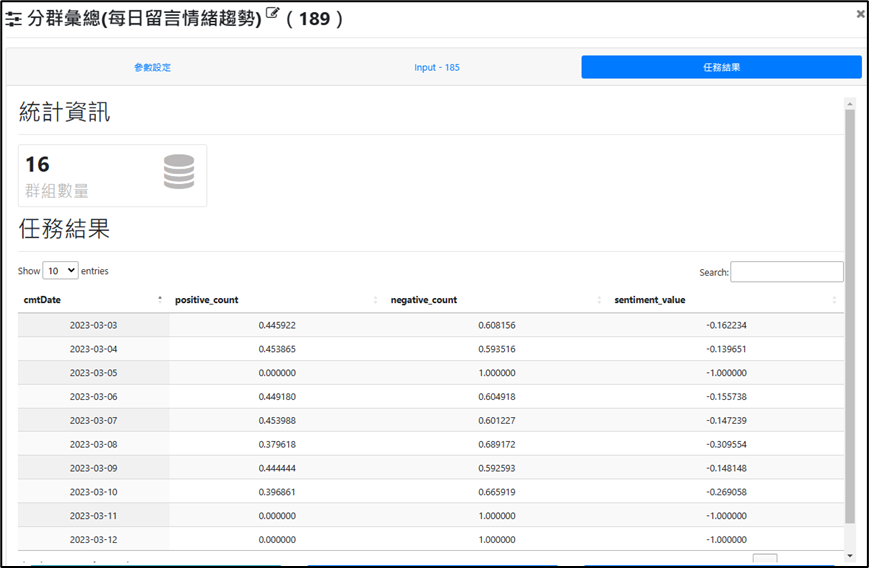




1. 分群匯總（每日留言情緒趨勢）[189]

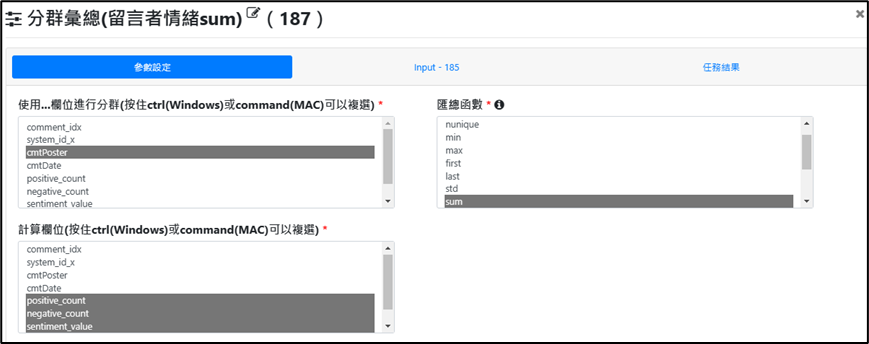
將篩選後之資料使用「分群彙總（日期）」模組，依日期作分群並將正向情緒、負向情緒、情緒值欄位計算平均值，以此看出每日留言之情緒趨勢。設定與結果如下：

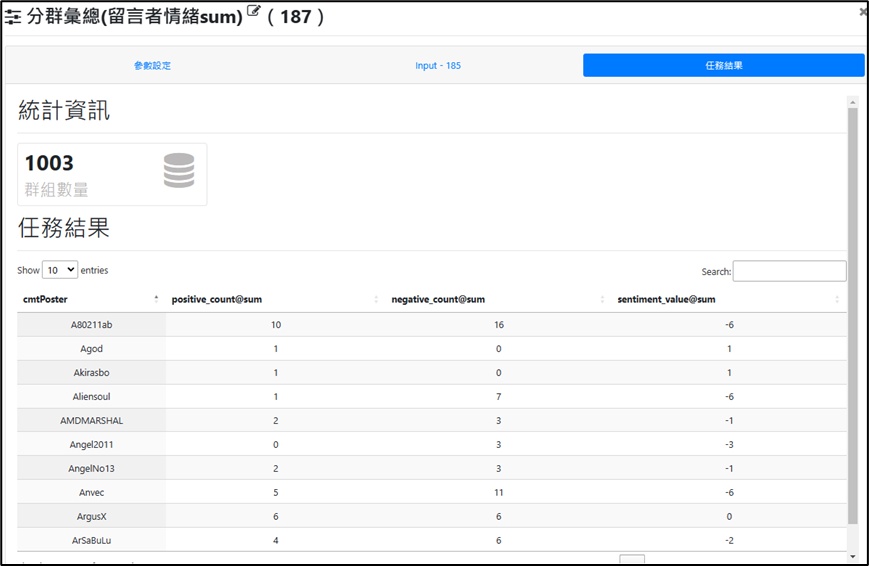




1. 分群彙總（留言者情緒總和）[187]/資料篩選（負面的前十名ID）[212]

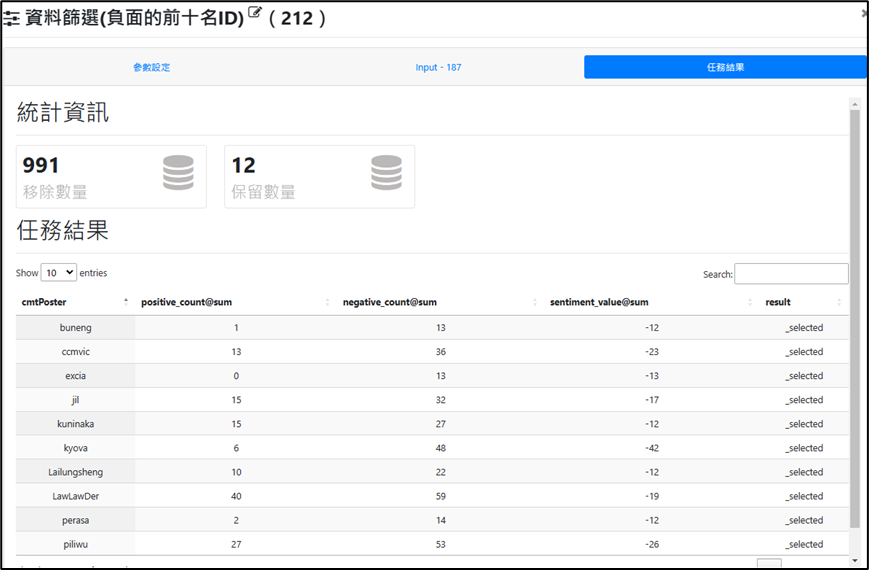
將篩選後之資料使用「分群彙總」模組，依張貼者/留言者（以下簡稱發文者）作分群，並將正向情緒、負向情緒、情緒值欄位計算總合，從而看出各發文者各言論情緒程度。設定與結果如下：





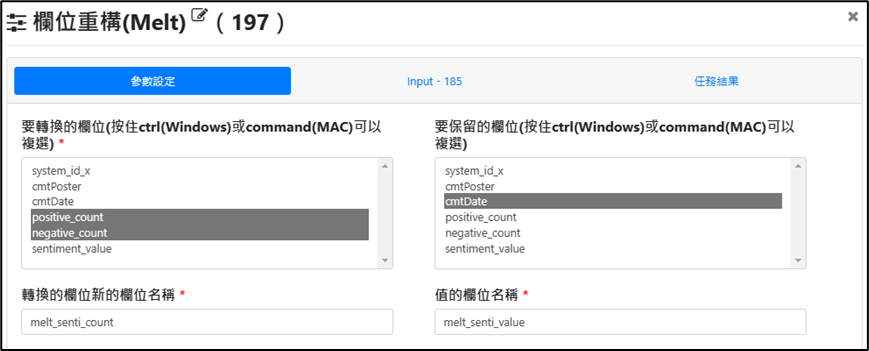
1. 根據各留言者之情緒值總和，「資料篩選」出負面情緒前10名的ID。設定與結果如下：

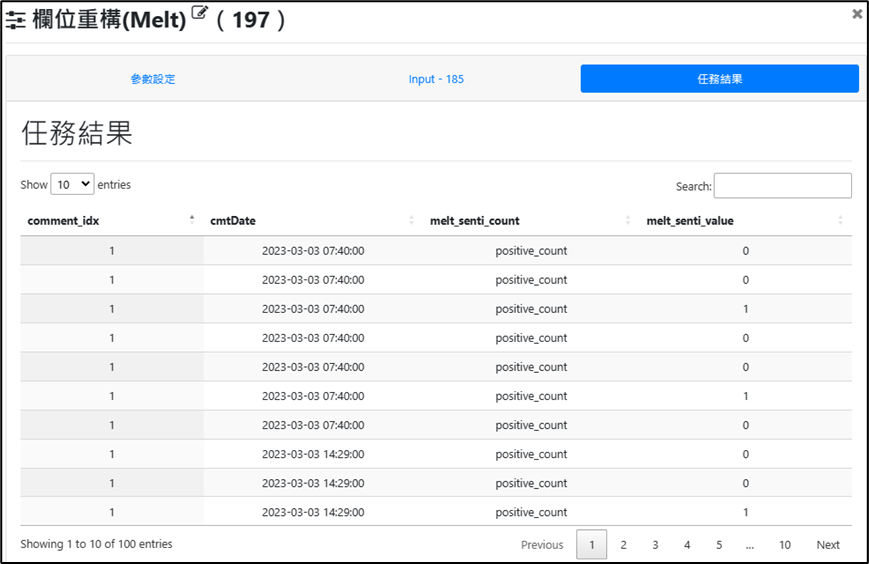


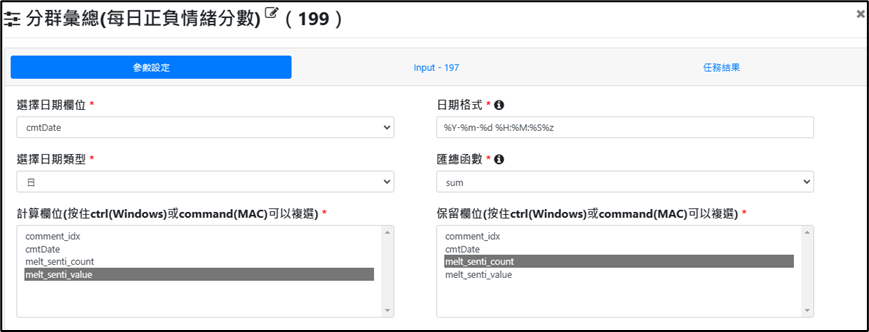


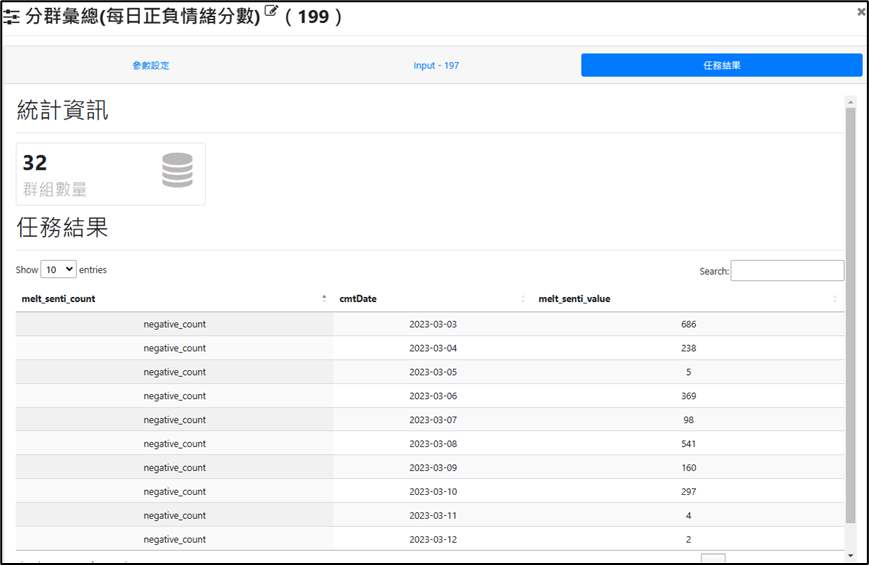
1. 欄位重構（Melt）[197]/ 分群彙總（每日正負情緒分數）[199]

將「欄位重構」後之資料使用「分群彙總（日期）」模組，依日期作分群計算melt\_senti\_value總和，藉以看出每日正/負兩種情緒各總量。設定與結果如下：



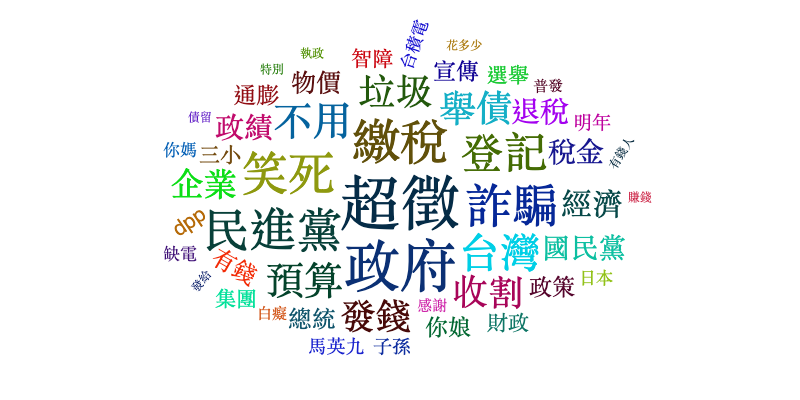






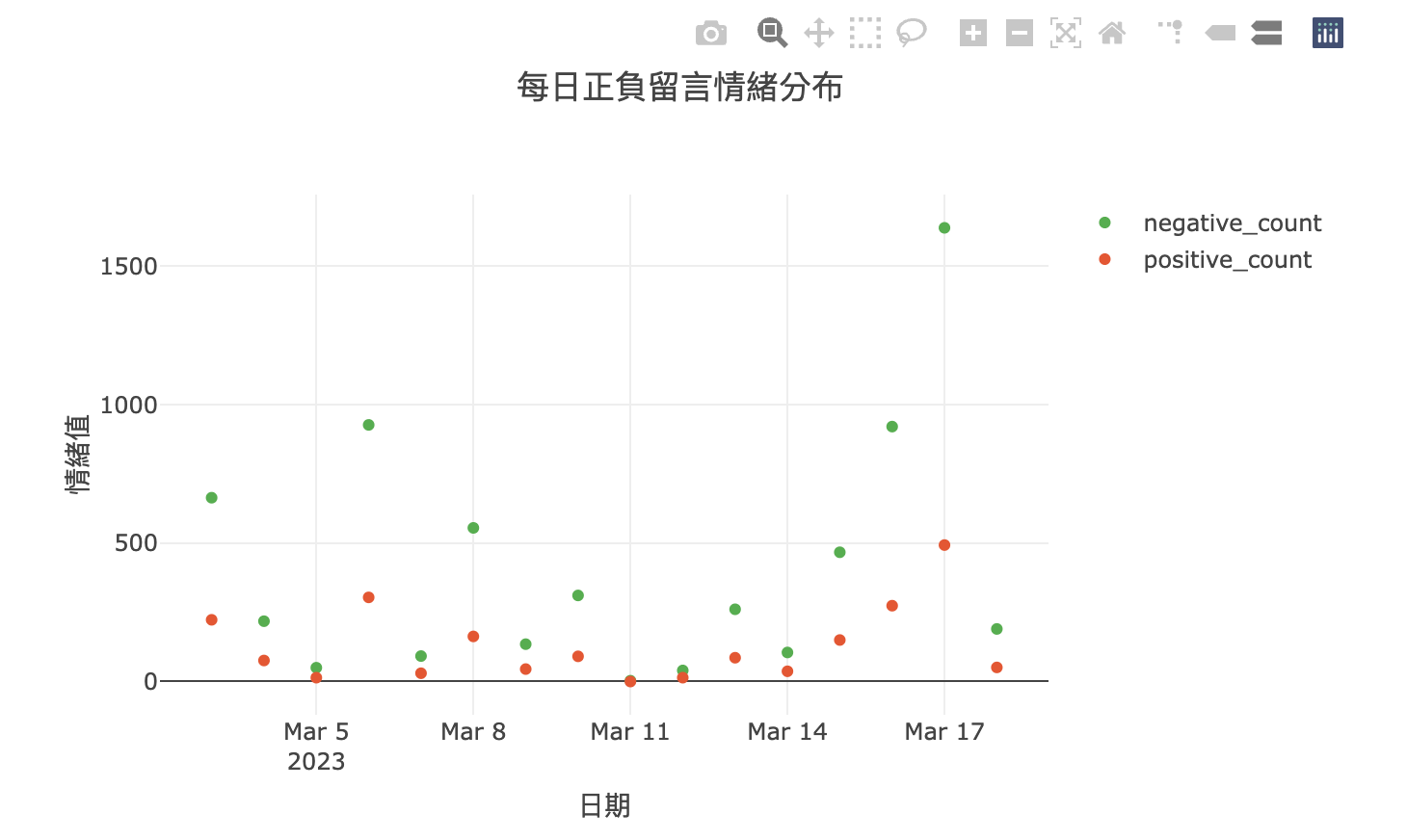
**視覺化圖表**

文字雲：使用詞頻計算篩選出前50名的字詞呈現文字雲，其中可看出「超徵」、「政府」、「繳稅」、「民進黨」、「笑死」等字出現，由此可發現，網友對於該項議題為政府超徵稅收引起。



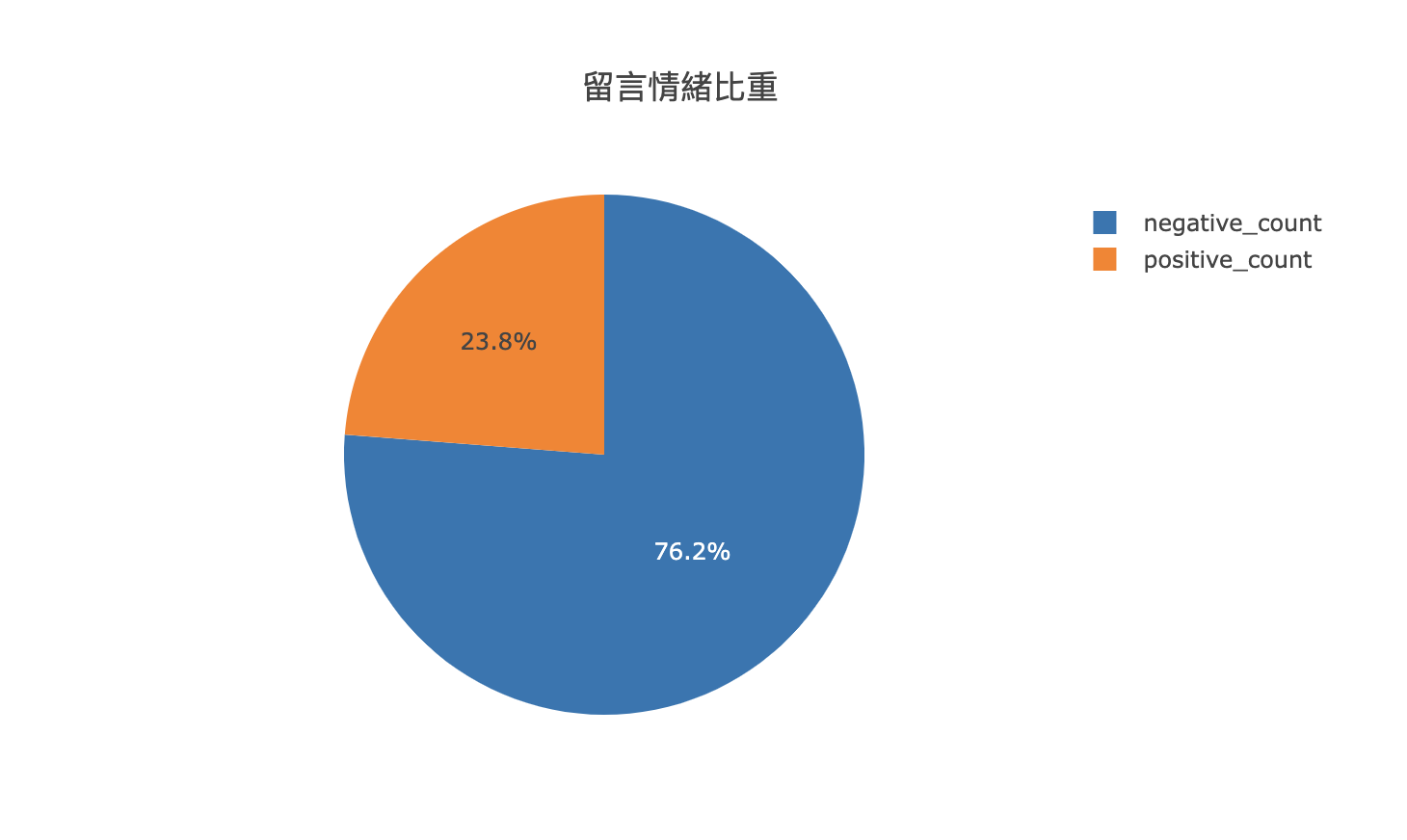
1. 每日正負情緒值

根據「日期」進行匯總，統計每日留言的正向與負向數量，透過點陣圖呈現，發現PTT留言負向情緒皆大於正向，其中3/17為負向最高，共有1639則。



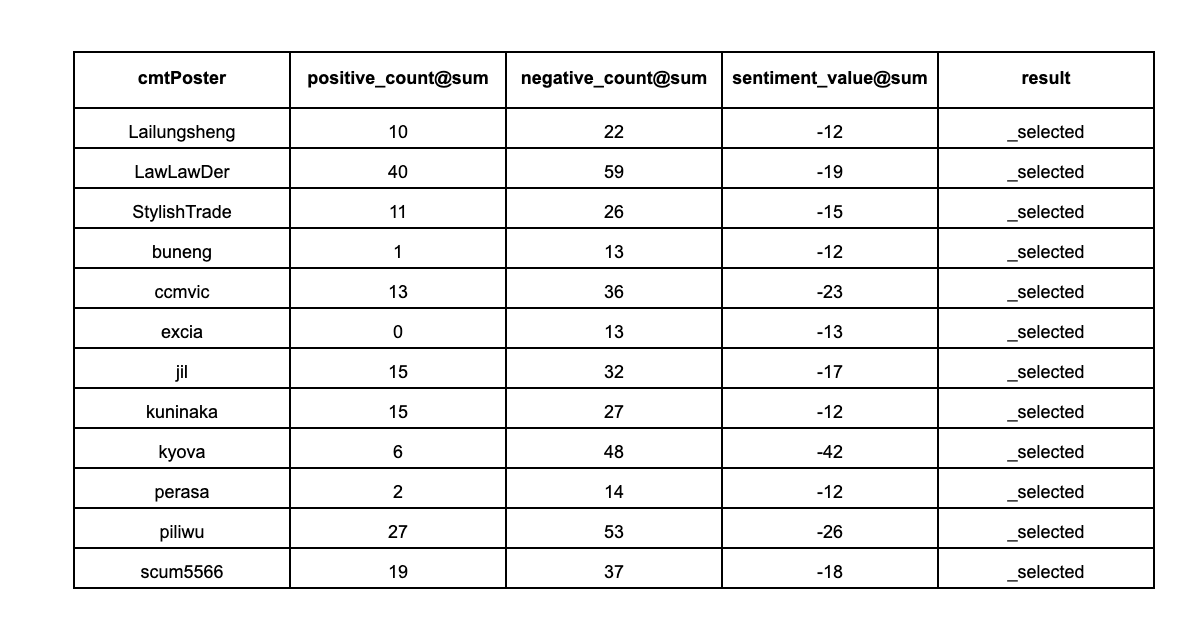
1. 留言情緒比重

根據每日正負情緒進行平均計算，透過圓餅圖呈現，平均每日負向情緒為76.2%，正向情緒為23.8%。



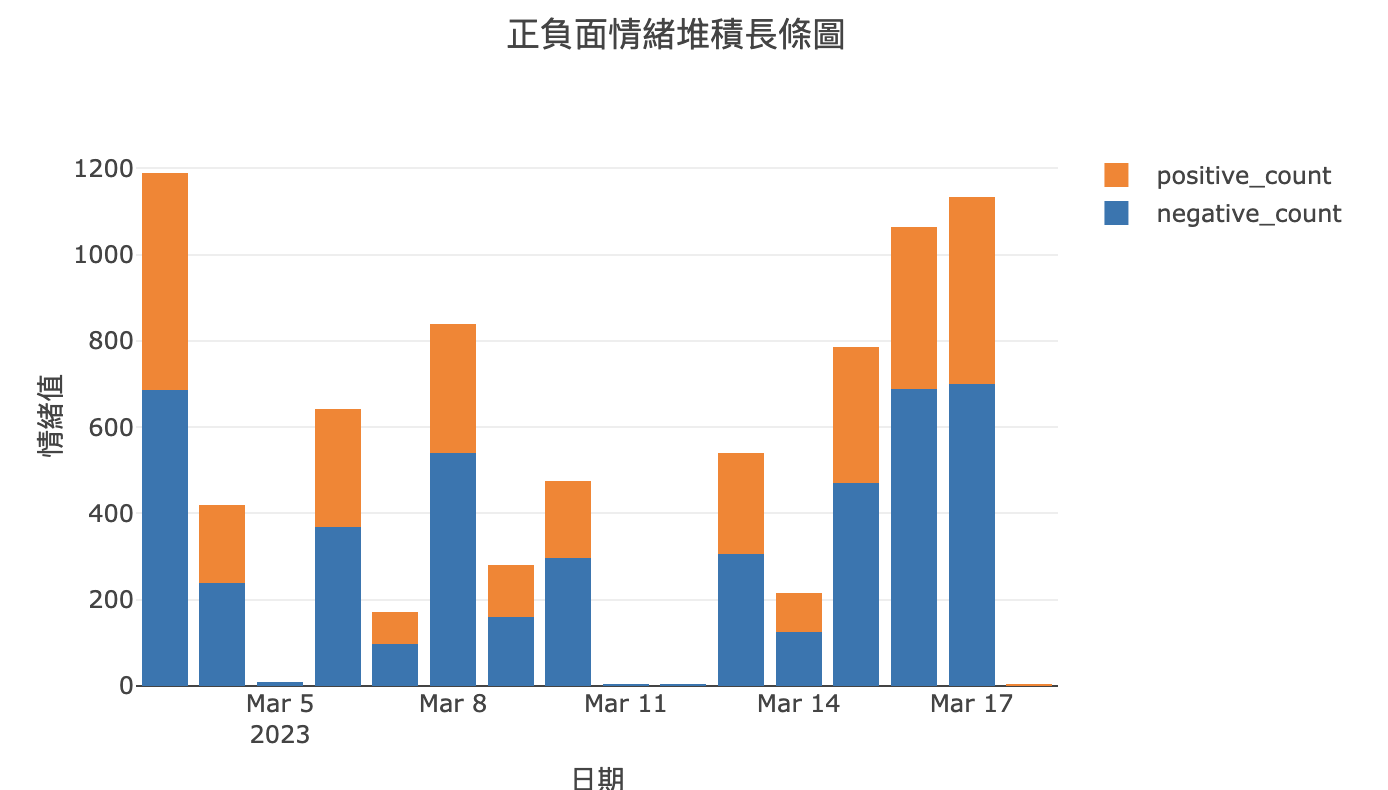
1. 前十名負向留言ID

根據分群匯總和資料塞選的結果，透過表格呈現，發現LawLawDer、pillwu、kyova負向次數最高，分別為59、53、48次數。



1. 正負向情緒堆疊長條圖

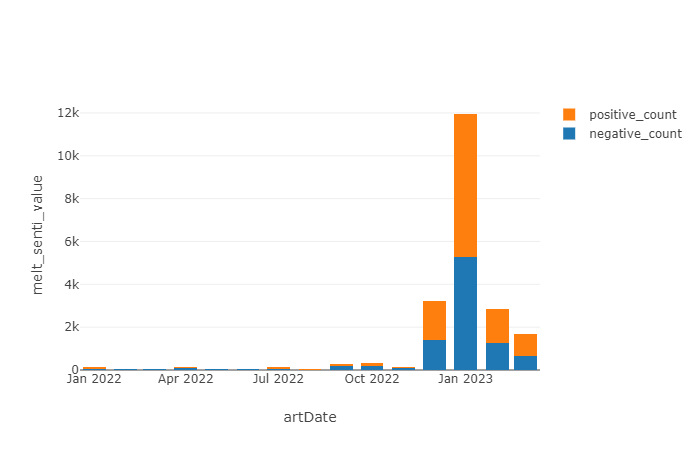
根據日期匯總每日正負向情緒，透過長條圖呈現，發現以3/1加總情緒值為最多，其次為3/17、3/16，其中3/5、3/11、3/12數量較少，因為較少留言數的日期都是六日（有可能是上班族六日較少上PTT），另外3/11為棒球經典賽開打日期，也有可能造成較少留言的現象。



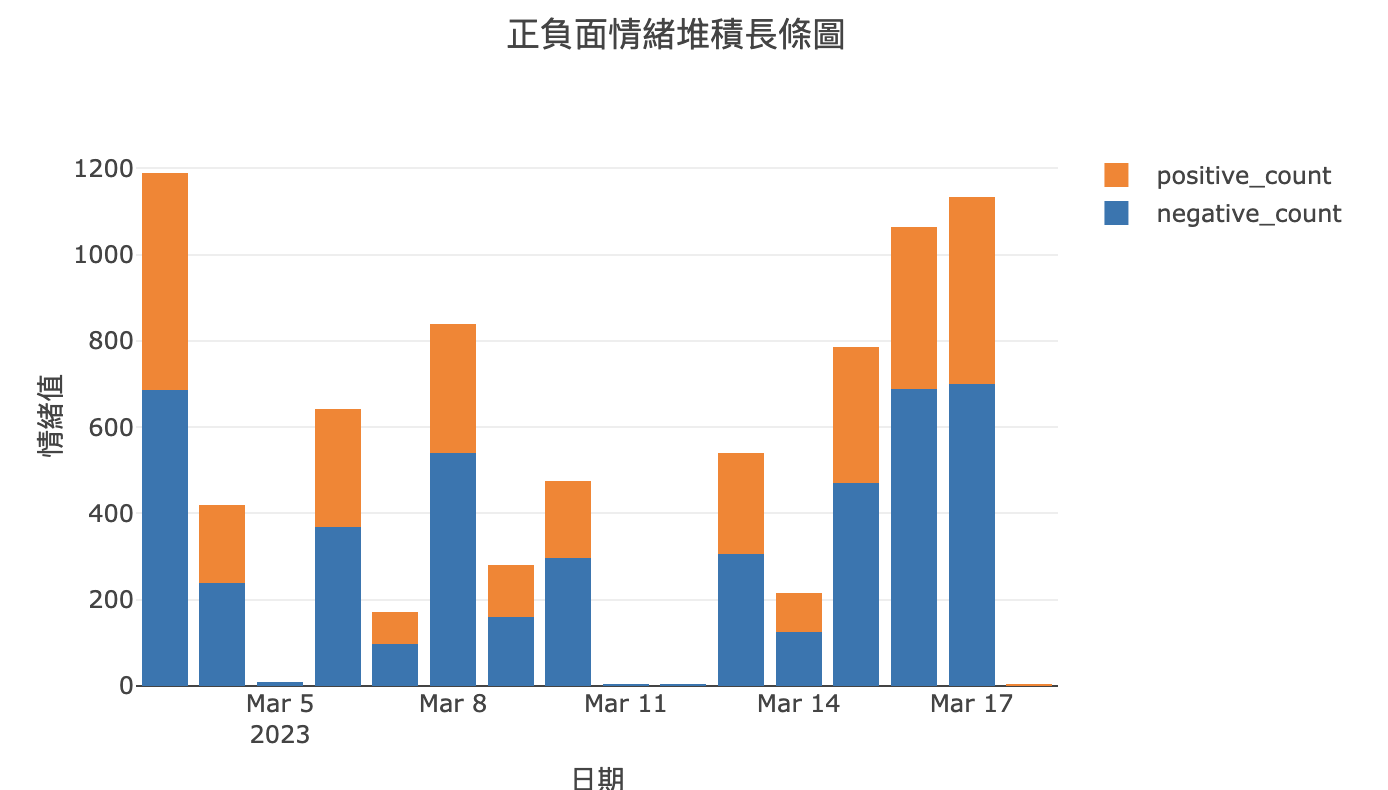
# **六、心得與結論**

隨著社群媒體的普及，愈來愈多的人們在社交媒體上表達情緒、討論事件和分享經驗。因此，常用字典法來分析社群媒體文本中的情緒取向。

在資料來源的選擇上，PTT的看板眾多，挑選「八卦版」、「政黑版」、「省錢版」有其目的，由於實質多了6,000元，在「省錢板」上應該會掀起一波討論，例如如何使用最划算等，「八卦版」則是不論大小事都能在「八卦版」上找到，尤其是本議題有高度政治意涵，因此「政黑版」和「八卦版」是必選的看板。



但另有同學加入「中時新聞」、「聯合新聞」分析結果，則是以正向情緒較多，本組認為，可能是因為新聞版需經媒體審核結果所致。



本組正負情緒堆積長條圖，顯示的日期皆有對應到相關的事件，如3/3為「行政院長赴立法院報告普發現金（疫後特別預算案）」的時間，3/5晚間財政部也公佈普發6000元所編列之經費，3/6立法院財政委員會審查疫後特別預算案，3/15公佈領取管道16號公開登記網址。可以發現PTT上的情緒堆疊皆是有相對應事情發生時才會在當天或幾天內堆疊起來，也可以看做在PTT上若無相關事件發生，中間的討論就不會被延續而冷下來。

根據文字雲詞頻計算可以發現，大部分的議題是集中在政府、超徵、稅收、舉債、預算等，隨著時間推移，登記、詐騙、宣傳等也慢慢出現。

本組研究心得，在PTT上討論風向是由政府在超徵議題上的問題開始，其中可以看到兩黨和政治人物的名稱出現，可以推知在PTT上，討論以黨和政客來比較兩黨的政績或狀況。討論議題上也可以發現子孫等，為討論初期時常會提到的債留子孫，其中不乏有許多罵人的詞彙，在結合情緒分析的正面負面比例，可以得知，在超徵或普發6000等議題上，PTT是負面情緒大於正面情緒的。若拼湊這些詞彙並深入了解可以發現，反對此議題者多以舉債、債留子孫、為了選舉等詞語留言。

最後，謝謝學校提供文字分析工作流程系統，讓無程式背景的同學，可以很輕易的進行資料蒐集、文字探勘及相關進階分析，十分方便。