社群媒體分析第七組第一次讀書會作業

職棒啦啦隊

第7組

N104020013 吳映儒

N104020005 李姿儀

N104020011 陳冠宏

N104020012 蔡京叡

N104020015 李宇婕

N104020016 楊世華

目錄

[一、 探討主題 1](#_Toc130150810)

[二、 分析流程 1](#_Toc130150811)

[(一) 標題 1](#_Toc130150812)

[1 爬蟲 1](#_Toc130150813)

[2 替換字串 2](#_Toc130150814)

[3 中文斷詞 2](#_Toc130150815)

[4 清除停用字 3](#_Toc130150816)

[5 詞頻計算 3](#_Toc130150817)

[6 資料分析 3](#_Toc130150818)

[(二) 內文 4](#_Toc130150819)

[1 爬蟲：設定與標題相同。 4](#_Toc130150820)

[2 第一層資料處理 4](#_Toc130150821)

[3 第二層資料處理 9](#_Toc130150822)

[4 第三層資料處理 11](#_Toc130150823)

[5 第四層資料處理 11](#_Toc130150824)

[6 第五層資料處理 16](#_Toc130150825)

[7 第六層資料處理 16](#_Toc130150826)

[(三) 留言 17](#_Toc130150827)

[1 爬蟲：設定與標題相同。 17](#_Toc130150828)

[2 留言萃取 17](#_Toc130150829)

[3 替換字串 18](#_Toc130150830)

[4 中文斷詞 18](#_Toc130150831)

[5 清除停用詞 18](#_Toc130150832)

[6 資料分析 18](#_Toc130150833)

[7 資料篩選 18](#_Toc130150834)

[三、 分析結果 18](#_Toc130150835)

[(一) 標題 18](#_Toc130150836)

[(二) 內文 19](#_Toc130150837)

[(三) 留言 26](#_Toc130150838)

1. 探討主題

國際棒壇最大的棒球盛事-第5屆世界棒球經典賽(WBC; World Baseball Classic)於2023年3月8日開打（2023/3/8-2023/3/21），共計20隊參賽；除了積極備賽，中華職棒也推派旗下啦啦隊女孩組成「台灣精品經典女孩」，在場外同步為臺灣隊吶喊助威，意外在國際爆紅，甚至引來韓國啦啦隊隊員加盟。本組此次希望透過文字探勘，了解臺灣最有影響力的網路社群之一-PTT的特定看板中，使用者在指定日期區間內對前開所述啦啦隊隊員的討論熱度及情感基調。

1. 分析流程

本組本次使用課程提供之文字探勘工作流程設計平台（Tarflow）進行文字探勘，並預計分別針對標題、內文及留言進行資料處理及相關分析（圖 1）。

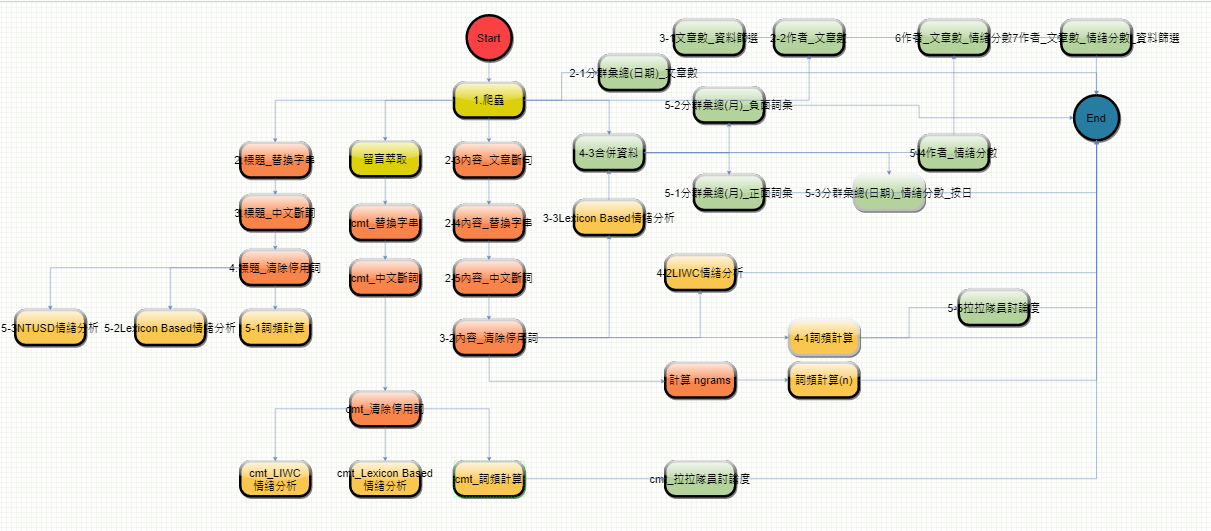


圖 1、總分析流程

本組本次分析包含標題、留言及內文（由左至右）。

1. 標題
2. 爬蟲
3. 目的：獲取分析標的。
4. 資料抓取標的：PTT八卦板。
5. 搜尋關鍵字：啦啦隊、經典女孩…等30個，主要為本次WBC中職啦啦隊隊員姓名。  
   【關鍵字選擇原則】

* 啦啦隊女孩成員眾多，並且擁有各自啦啦隊隊名，故關鍵字加入所有女孩名字以及其隊伍名稱。
* 有時討論版之議題不僅限於某一隊伍或是女孩，而是關心整個經典賽或是所有啦啦隊的各種相關主題，故關鍵字加入經典賽、啦啦隊、啦啦隊女孩等所有相關之統一稱呼。
* 在整理資料時發現，李多慧有不同選字組合，故關鍵字加入其變化形態。

1. 排除關鍵字：元旦、元旦升旗、丹丹漢堡…等57個，包含後續反覆檢查抓取結果後，回饋修正添加而成。

【關鍵字排除原則】

* 為了聚焦本次情緒分析主體，審視爬蟲結果後，將防疫宣導、與啦啦隊女孩一起被新聞報導過之人名、啦啦隊女孩們曾參與過之綜藝節目、啦啦隊女孩們於國外發展之相關內容等設定為排除關鍵字。
* 2023年台中市政府元旦升旗典禮邀請中信兄弟啦啦隊 Passion Sisters熱舞，與此次分析主題無關，故排除相關關鍵字。
* 啦啦隊女孩「丹丹」與知名速食店「丹丹漢堡」同名，為獨立出此次分析主題所需資料，故排除丹丹漢堡的相關關鍵字。
* 在爬蟲資料中包含許多啦啦隊隊伍、未聚焦有參與經典賽之隊伍，故將無參與經典賽的隊伍名稱設定為排除關鍵字。

1. 抓取資料區間：2022年9月14日-2023年3月19日。

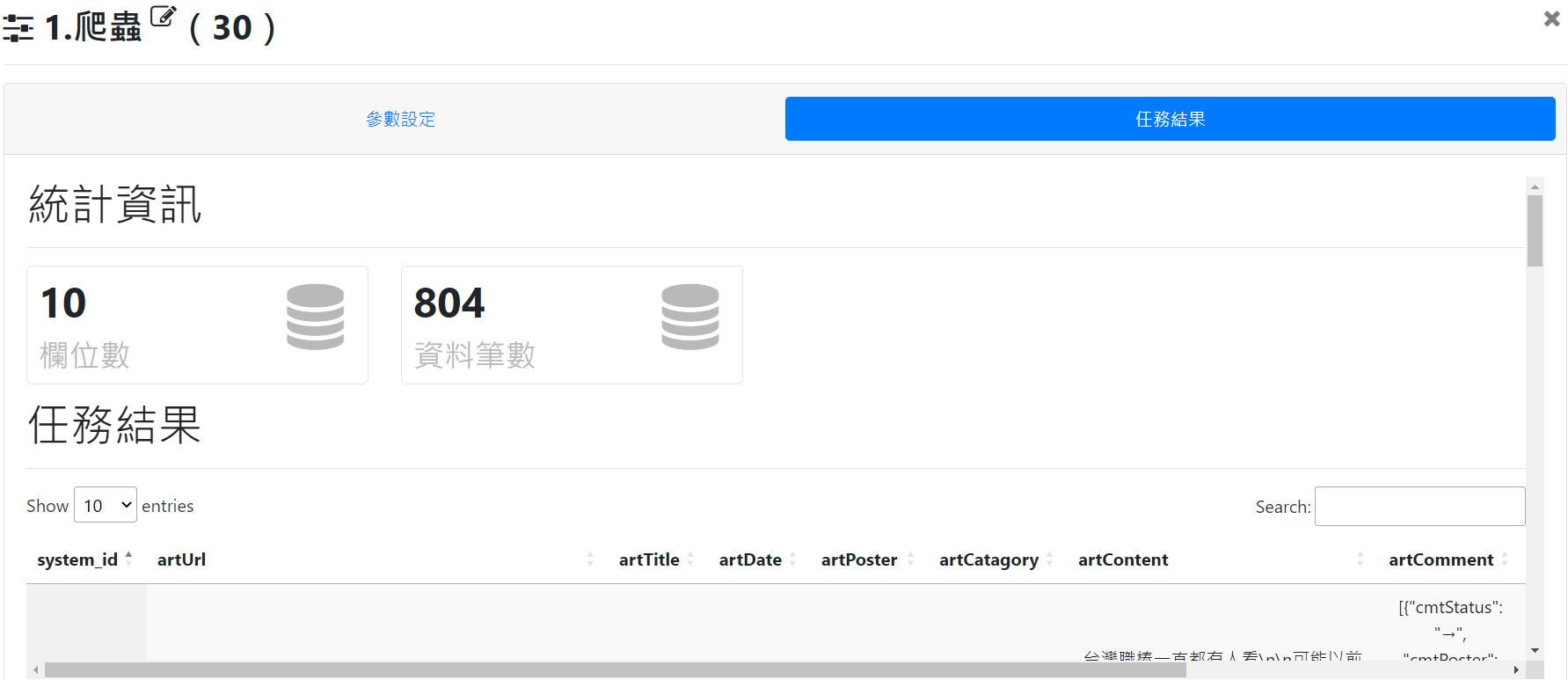


圖 2、爬蟲結果

1. 替換字串
2. 目的：將特殊符號或特定字串替換，以利後續分析。
3. 替換字串設定：去除標點符號及特殊字元，共32種符號變化需替換為空白。

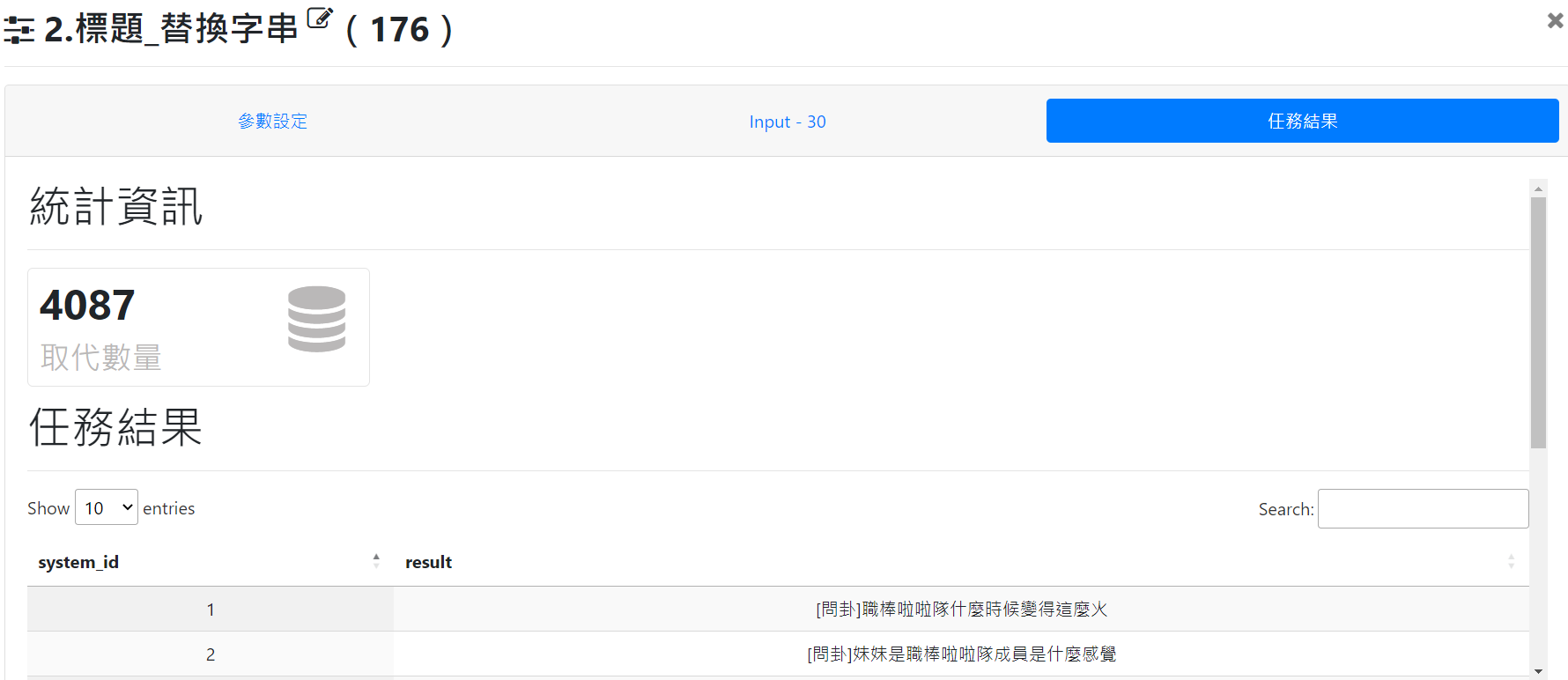


圖 3、替換字串結果

1. 中文斷詞
2. 目的：將句子進一步切割成適當詞語，以利後續分析。
3. 定義詞彙：自行定義啦啦隊、拉拉隊、啦啦隊女孩名字、隊伍名稱、女孩們的綽號等具相關性詞語進行權重設定…等67個詞彙。

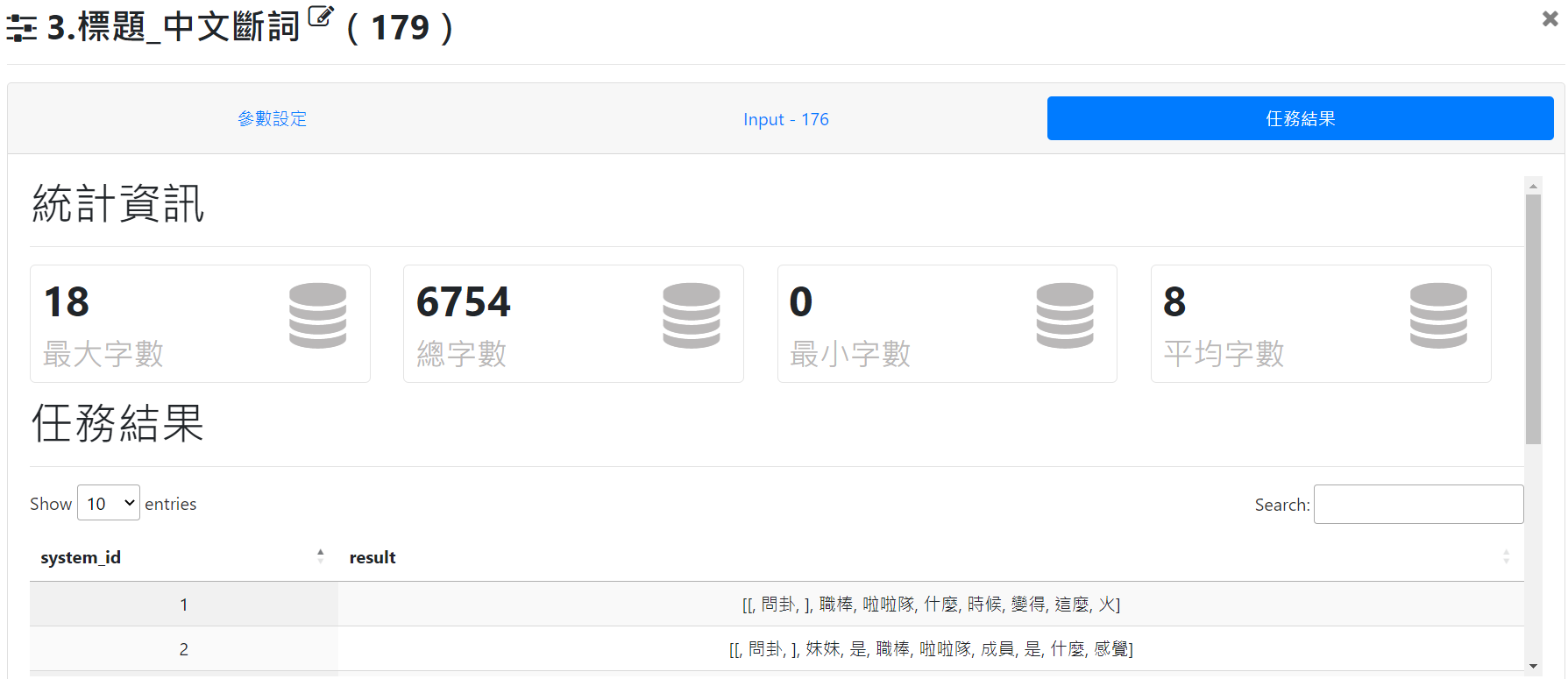


圖 4、中文斷詞結果

1. 清除停用字
2. 目的：過濾掉無實際含意或非分析主體等之詞彙，以利後續分析。
3. 定義詞彙：使用預設停止詞且轉換小寫英文，並清除換行符號、html tag、數字、特殊標點符號，除了系統設定外，亦加入自定義停止詞，內容包括較無意義的語助詞、轉貼新聞報導的詞彙，停用381個詞彙，清除5,039筆。



圖 5、清除停用字結果

1. 詞頻計算
2. 目的：進行標題出現詞彙之頻率計算。
3. 篩選詞彙數量：以預設100組，生成文字雲。
4. 資料分析
   1. NTUSD情緒分析
5. 目的：接續4處理，進行留言詞彙之情緒分析。
6. 定義正面詞彙：喜歡、開心、可愛、美女、人氣、漂亮…等27個。
7. 定義負面詞彙：討厭、不喜歡、噁心、可怕、做怪…等18個。
   1. Lexicon Based情緒分析
8. 目的：接續4處理，進行留言詞彙之情緒分析。
9. 定義正面詞彙：與6-1處理相同。
10. 定義負面詞彙：與6-1處理相同。
11. 內文
12. 爬蟲：設定與標題相同。
13. 第一層資料處理
    1. 分群彙總（依日期）（圖 6）：
14. 目的：以文章數為計算標的，依月份計算文章數。
15. 處理方式：使用「artDate」作為日期欄位，並依月進行分析；計算「system\_id」欄位，並保留「artDate」欄位。
16. 執行結果：以月份分群，計算各月文章數。



圖 6、依日期分群彙總文章數執行結果

* 1. 分群彙總（非數值）（圖 7）：

1. 目的：以文章數為計算標的，依作者計算文章數。
2. 處理方式：使用「artPoster」分群，以彙總函數「count」計算「system\_id」欄位。

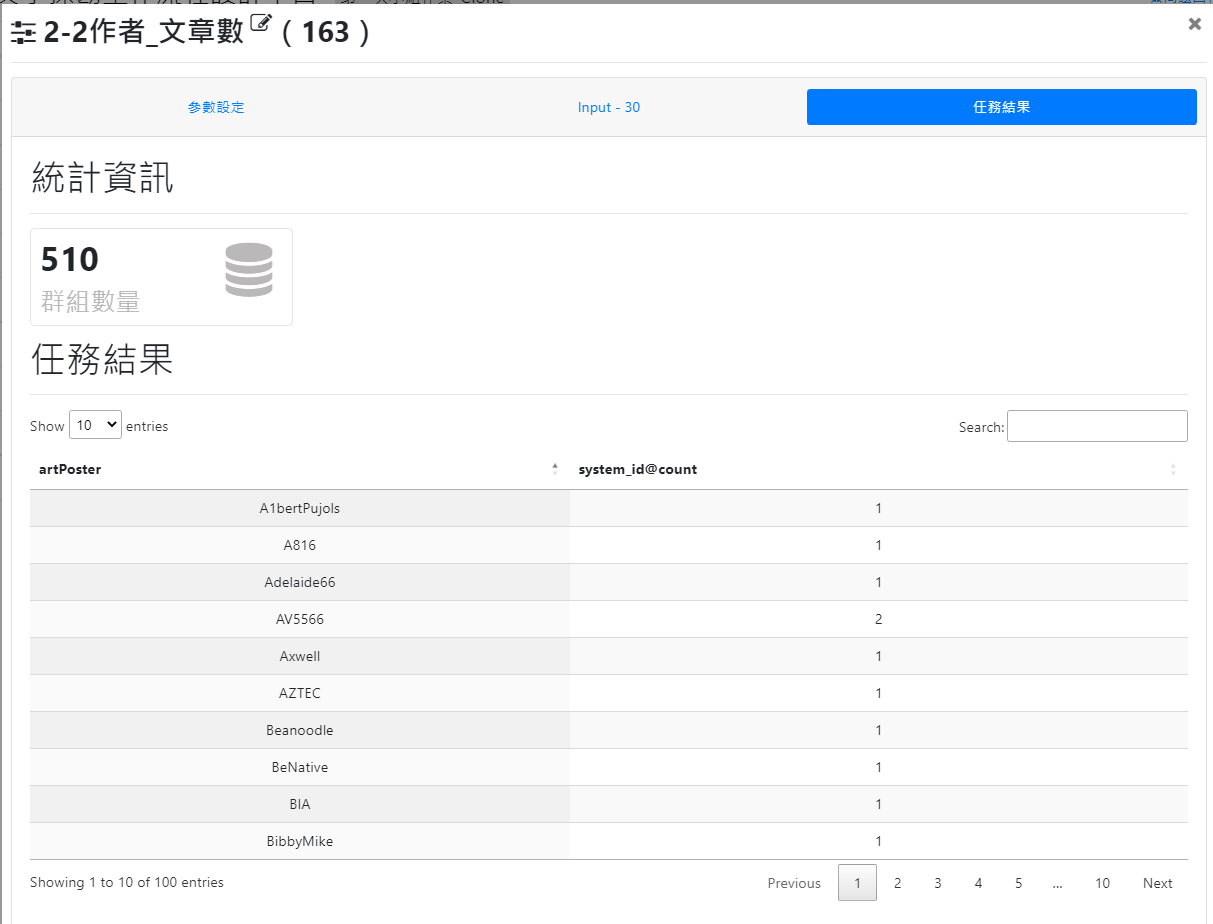


圖 7、依作者分群彙總文章數執行結果

* 1. 文章斷句（圖 8）：

1. 目的：將內容切割成適當的句子，以利後續分析。
2. 處理方式：處理「artContent」欄位，以默認符號進行斷句。



圖 8、將內文進行文章斷句處理結果

* 1. 替換字串（圖 9、圖 10）：

1. 目的：將特殊符號或特定字串替換，以利後續分析。
2. 替換字串設定：換行符號、sent from JPPT/BePTT/MoPTT、經n-gram計算及反覆修正，增加移除my iphone、on my、iphone、XD等，以及圖片格式結尾之網址、youtube網址；替換同音字如「林香」、「林鄉」等。



圖 9、替換字串處理前



圖 10、替換字串處理後

* 1. 中文斷詞（圖 11）：

1. 目的：將句子進一步切割成適當詞語，以利後續分析。
2. 定義詞彙：自行定義啦啦隊、拉拉隊…等60個詞彙，包含經n-gram分析後及檢視最終結果回饋修正納入之詞彙。



圖 11、中文斷詞處理結果

1. 第二層資料處理
   1. 資料篩選（圖 12）：
2. 目的：接續2-2處理，篩選出發表文章數大於9篇之作者。
3. 條件式：'$system\_id@count>=9'。

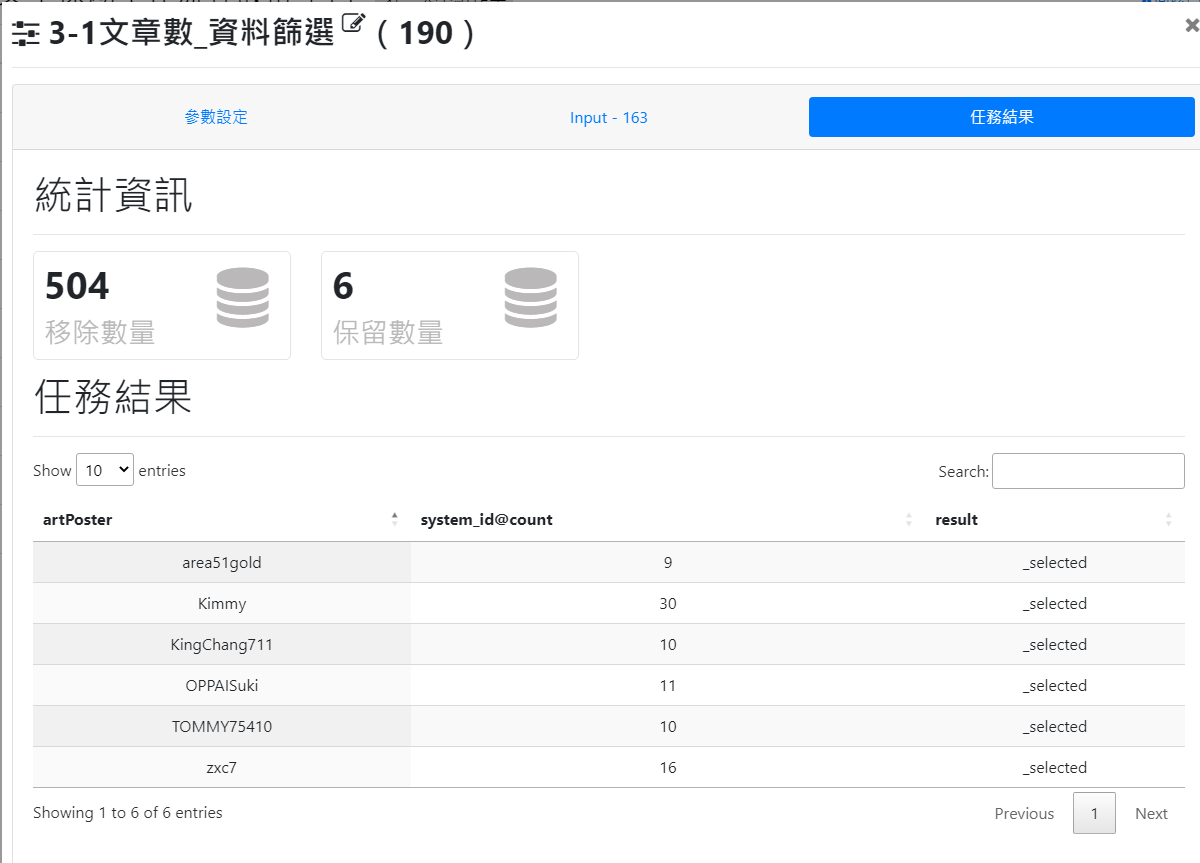


圖 12、2-2處理流程資料篩選結果

* 1. 清除停用詞（圖 13）：

1. 目的：接續2-5處理，過濾掉無實際含意或非分析主體等之詞彙，以利後續分析。
2. 定義詞彙：自行定義啦啦隊、棒球…等691個詞彙，包含經n-gram分析後及檢視最終結果回饋修正納入之詞彙。



圖 13、2-5處理流程清除停用詞處理結果

* 1. Lexicon Based情緒分析：

1. 目的：接續3-2處理，進行內文情緒分析。
2. 定義正面詞彙：喜歡、開心…等30個。
3. 定義負面詞彙：討厭、不喜歡…等18個。
4. 第三層資料處理
   1. 詞頻計算：
5. 目的：接續3-2處理，進行內文出現詞彙之頻率計算。
6. 篩選詞彙數量：100。
   1. LIWC情緒分析：
7. 目的：接續3-2處理，進行內文詞彙之頻率計算。
8. 定義正面詞彙：開心、可愛…等25個。
9. 定義負面詞彙：討厭、不喜歡…等20個。
   1. 合併資料：
10. 目的：接續3-3處理，將Lexicon Based情緒分析與原爬蟲資料合併，以進行其他欄位分析。
11. JOIN KEY：system\_id。
12. 第四層資料處理
    1. 分群彙總（依日期）（圖 14）：
13. 目的：接續4-3處理，以正面詞彙數為計算標的，依月份計算正面詞彙數。
14. 處理方式：使用「artDate」作為日期欄位，並依月進行分析；計算「positive\_count」欄位，並保留「system\_id」欄位。

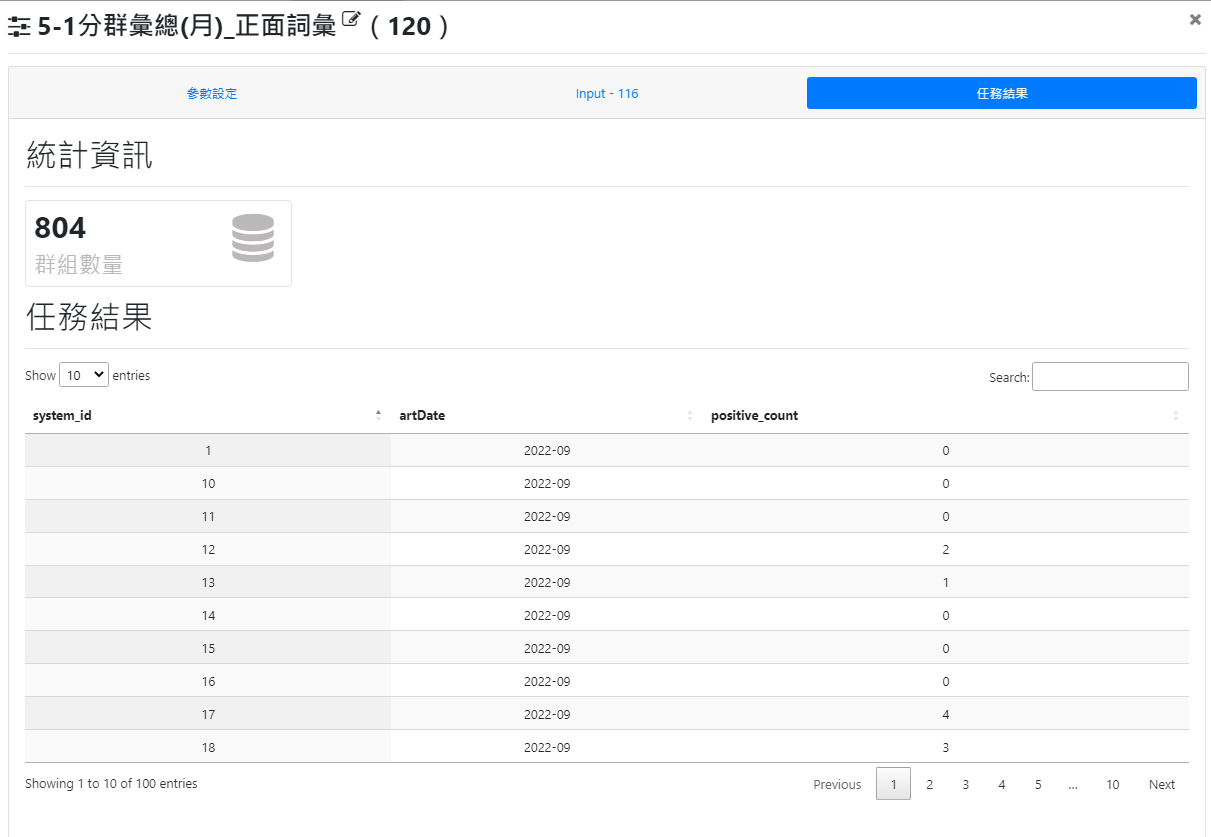


圖 14、4-3處理流程依日期按月分群之正面詞彙數彙總處理結果

* 1. 分群彙總（依日期）（圖 15）：

1. 目的：接續4-3處理，以負面詞彙數為計算標的，依月份計算負面詞彙數。
2. 處理方式：使用「artDate」作為日期欄位，並依月進行分析；計算「negative\_count」欄位，並保留「system\_id」欄位。

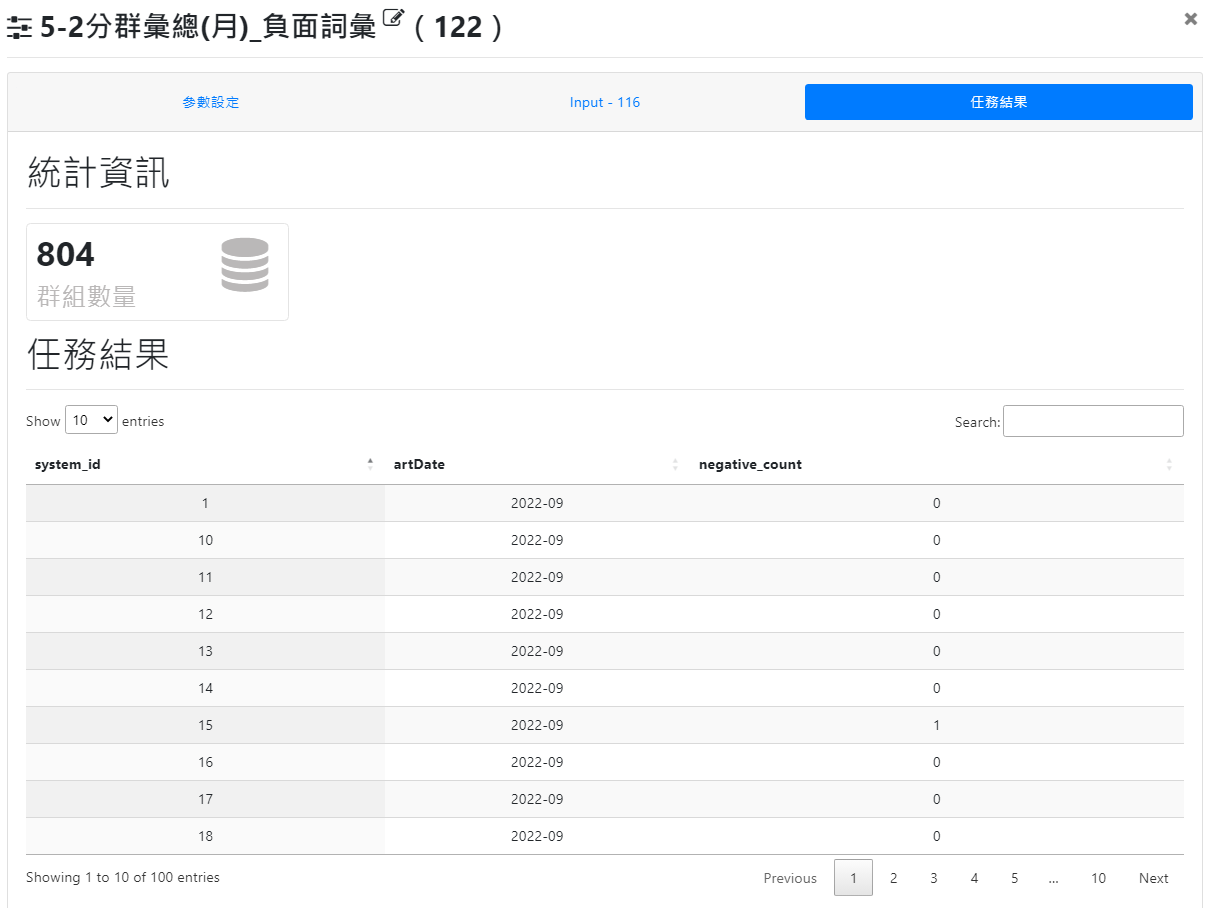


圖 15、4-3處理流程依日期按月分群之負面詞彙數彙總處理結果

* 1. 分群彙總（依日期）（圖 16）：

1. 目的：接續4-3處理，以情緒分數為計算標的，依日計算情緒分數總和。
2. 處理方式：使用「artDate」作為日期欄位，並依日進行分析；計算「sentiment\_value」欄位，並保留「system\_id」欄位。

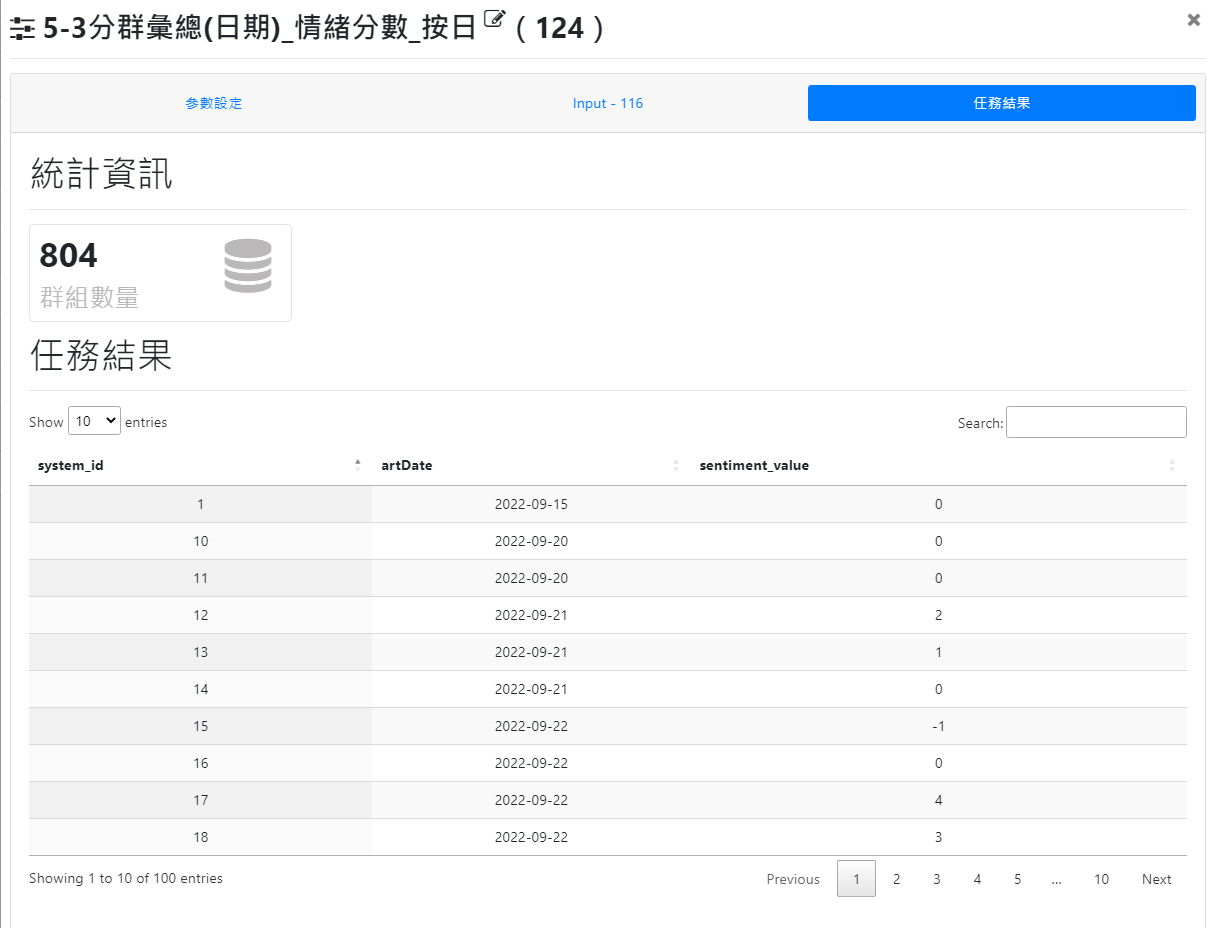


圖 16、4-3處理流程依日期按日分群之情緒分數彙總處理結果

* 1. 分群彙總（非數值）（圖 17）：

1. 目的：接續4-3處理，以情緒分數為計算標的，依作者計算情緒分數總和。
2. 處理方式：使用「artPoster」分群，以彙總函數「sum」計算「sentiment\_value」欄位。

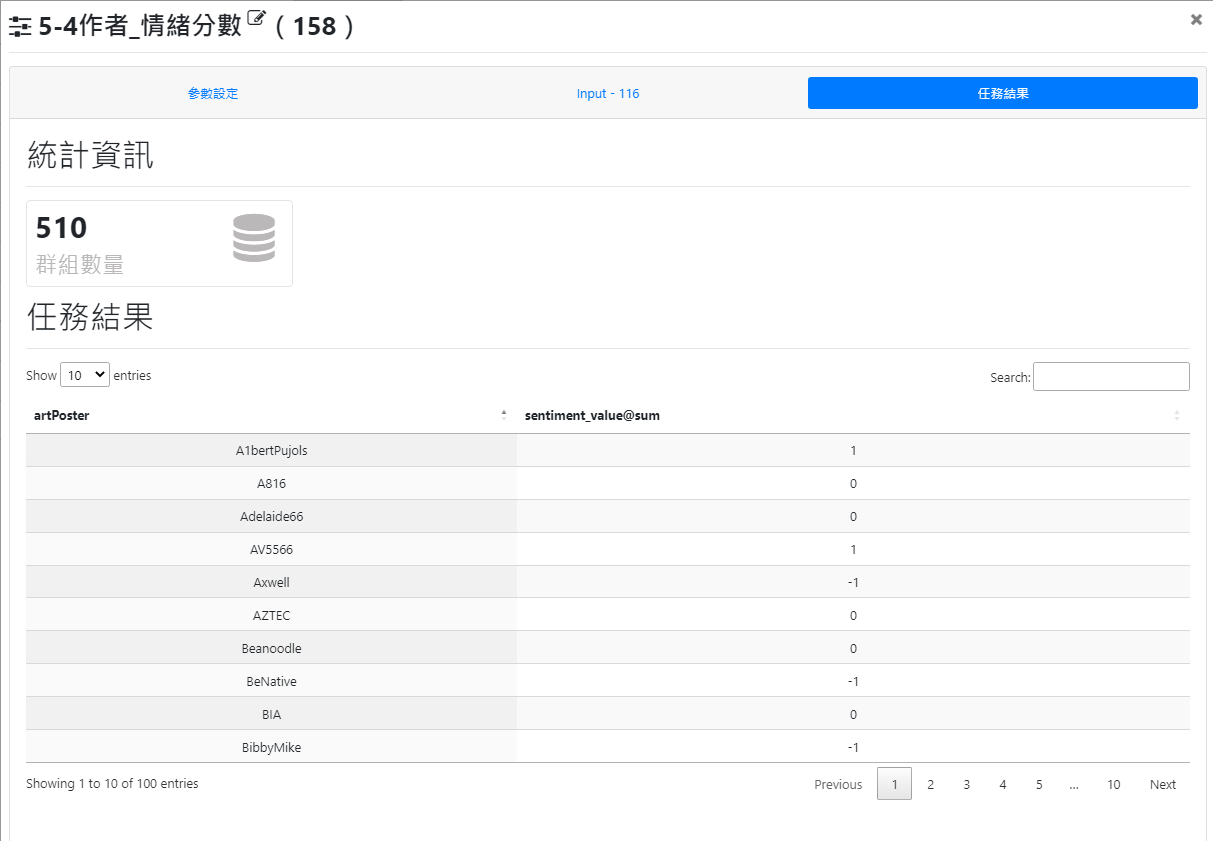


圖 17、4-3處理流程依作者分群彙總情緒分數總和執行結果

* 1. 資料篩選（圖 18）：

1. 目的：接續4-1處理，篩選出指定詞彙之詞頻計算結果。
2. 條件式：'$Term=="林襄"'或'$Term=="峮峮"'等21個條件。

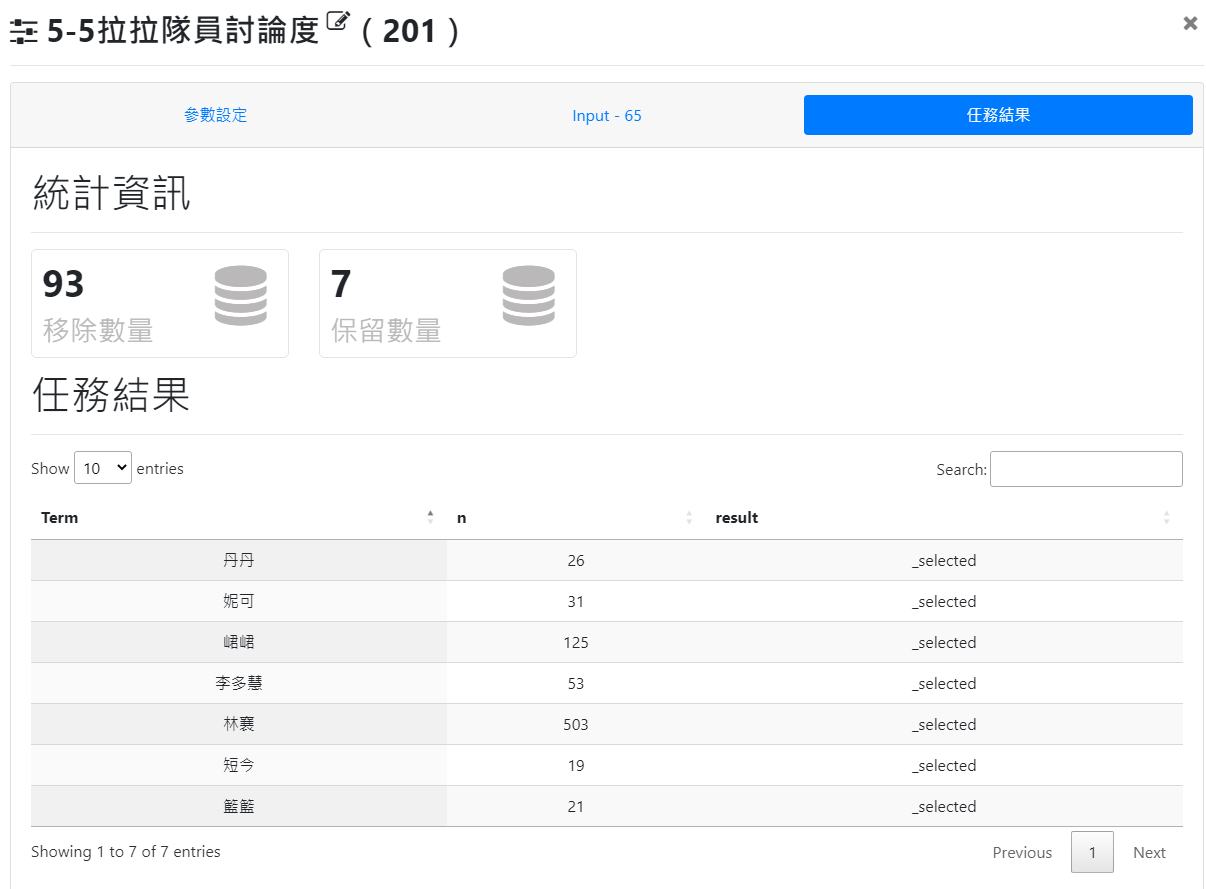


圖 18、4-1處理流程資料篩選結果

1. 第五層資料處理
   1. 合併資料（圖 19）：
2. 目的：使用2-2及5-4資料，以作者分群，將各作者之文章數與情緒分數合併。
3. JOIN KEY：artPoster。

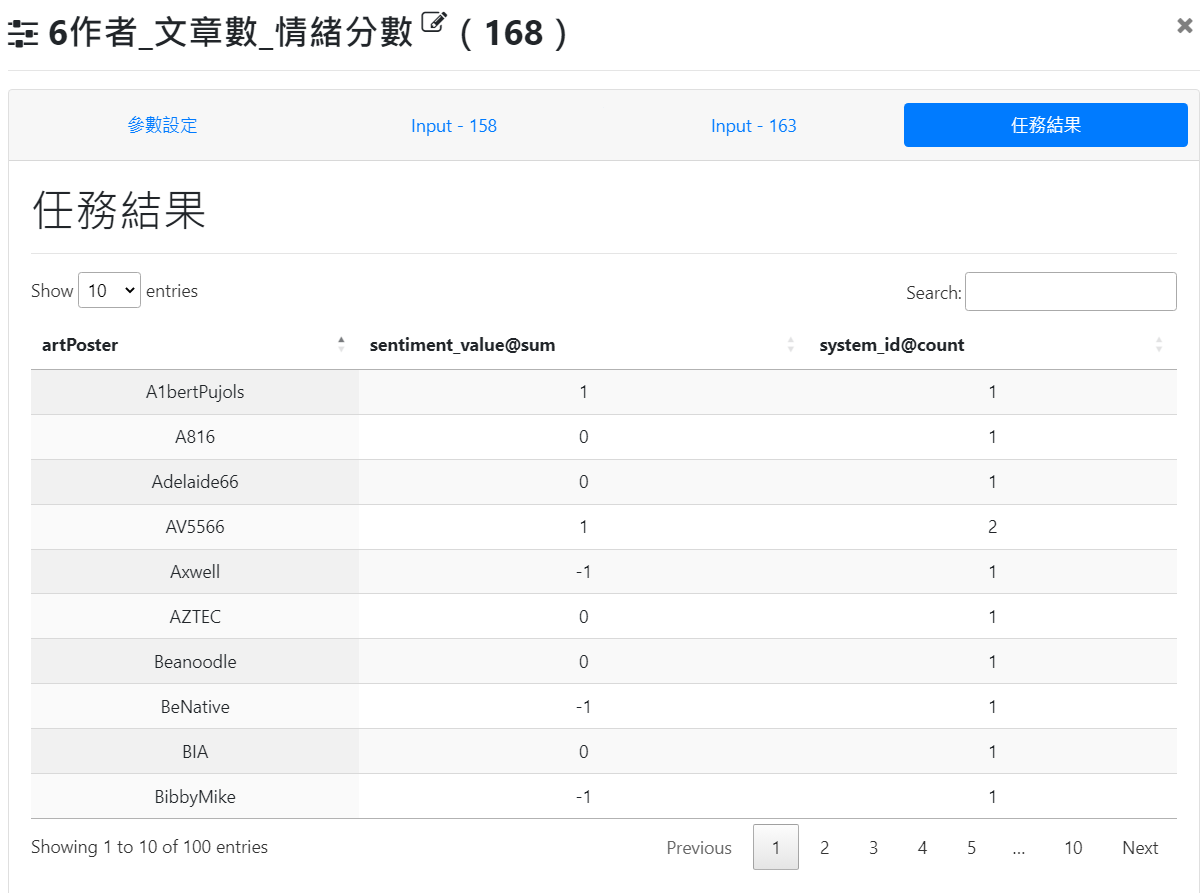


圖 19、使用2-2及5-4處理流程，依作者合併資料處理結果

1. 第六層資料處理
   1. 資料篩選：
2. 目的：接續6-1處理，篩選出發表文章數大於9篇之作者。
3. 條件式：'$system\_id@count>=9'。

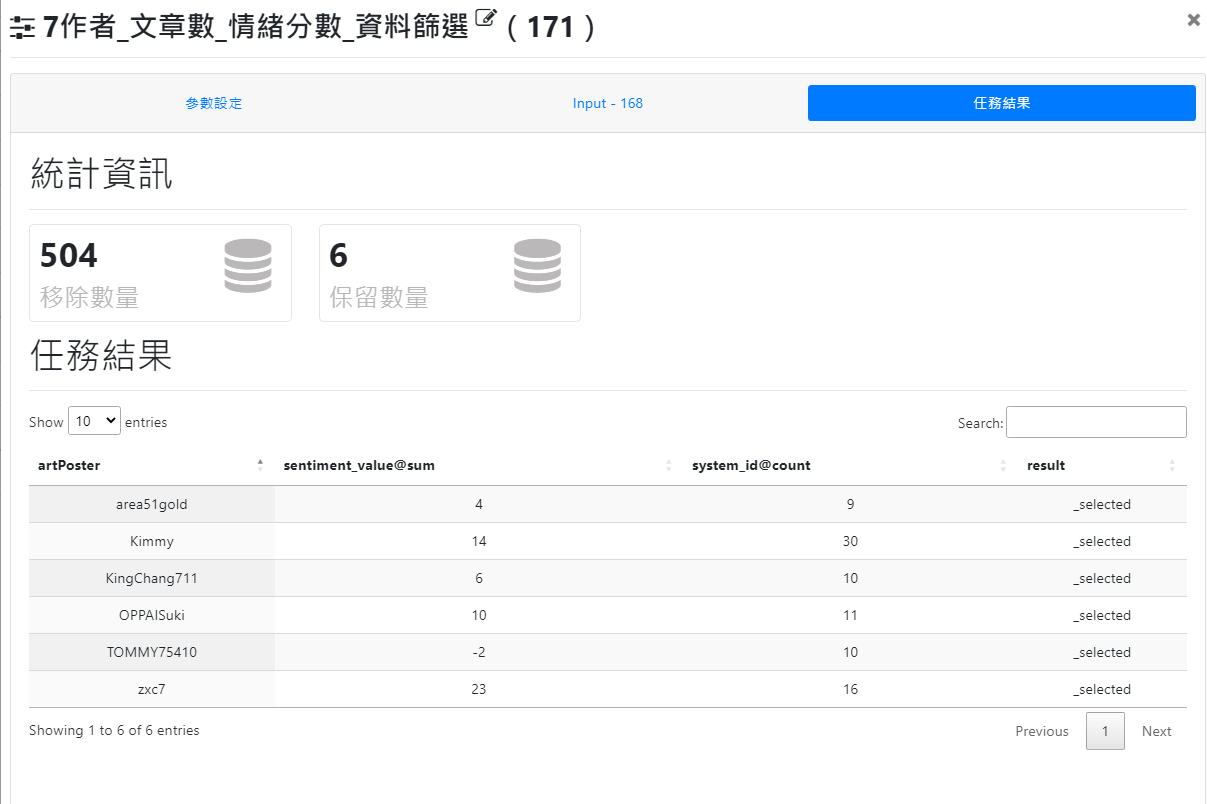


圖 20、6-1處理流程資料篩選結果

1. 留言
2. 爬蟲：設定與標題相同。
3. 留言萃取
4. 目的：提取各文章留言資訊。

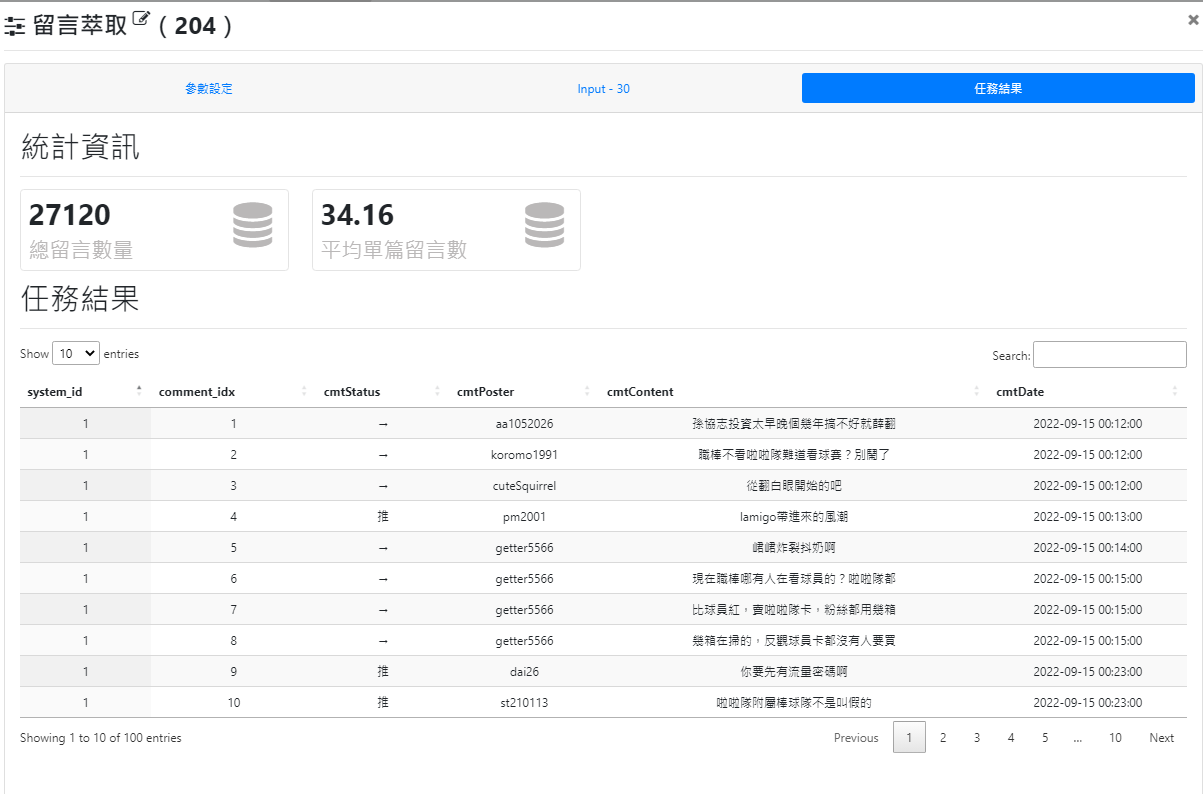


圖 21、留言萃取結果

1. 替換字串
2. 目的：將特殊符號或特定字串替換，以利後續分析。
3. 替換字串設定：設定與內文相同。
4. 中文斷詞
5. 目的：將句子進一步切割成適當詞語，以利後續分析。
6. 定義詞彙：自行定義啦啦隊、拉拉隊…等61個詞彙，包含內文分析之部分詞彙。
7. 清除停用詞
8. 目的：過濾掉無實際含意或非分析主體等之詞彙，以利後續分析。
9. 定義詞彙：自行定義啦啦隊、棒球…等513個詞彙，包含內文分析之部分詞彙。
10. 資料分析
    1. LIWC情緒分析：
11. 目的：接續5處理，進行留言詞彙之情緒分析。
12. 定義正面詞彙：喜歡、開心…等32個。
13. 定義負面詞彙：討厭、不喜歡…等21個。
    1. Lexicon Based情緒分析：
14. 目的：接續5處理，進行留言詞彙之情緒分析。
15. 定義正面詞彙：與6-1處理相同。
16. 定義負面詞彙：與6-1處理相同。
    1. 詞頻計算：
17. 目的：接續5處理，進行留言出現詞彙之頻率計算。
18. 篩選詞彙數量：100。
19. 資料篩選
20. 目的：接續6處理，篩選出發表文章數大於9篇之作者。
21. 條件式：'$Term=="林襄"'或'$Term=="峮峮"'等21個條件。
22. 分析結果
23. 標題

依據Lexicon Based情緒分析結果，以正面情緒較多，如喜歡、美女、人氣等，且多針對外表評論，如可愛、好看、甜心等；負面情緒較少，標題相較內文或留言比較簡約，故出現正負面之詞彙也相對簡單 （圖 22）。

再觀察NTUSD情緒分析結果，仍以正面情緒較多，且與Lexicon Based分析大致相似，較不同的正面詞會有期待、有名、優質…等，但負面情緒詞彙數也大幅增加；其中負面情緒字彙中出現台女、秘密、驚人，該詞彙雖被分類於負面，但在不同語境中可能出現正、負面意義，如加入此一考量，則較難判斷留言整體屬正面或負面（圖 23）。

而以詞頻分析結果而言，留言中出現最高頻率之詞為「林襄」，再來是峮峮及李多慧，顯見使用者對於這些啦啦隊成員有較高的討論度（圖 24）。

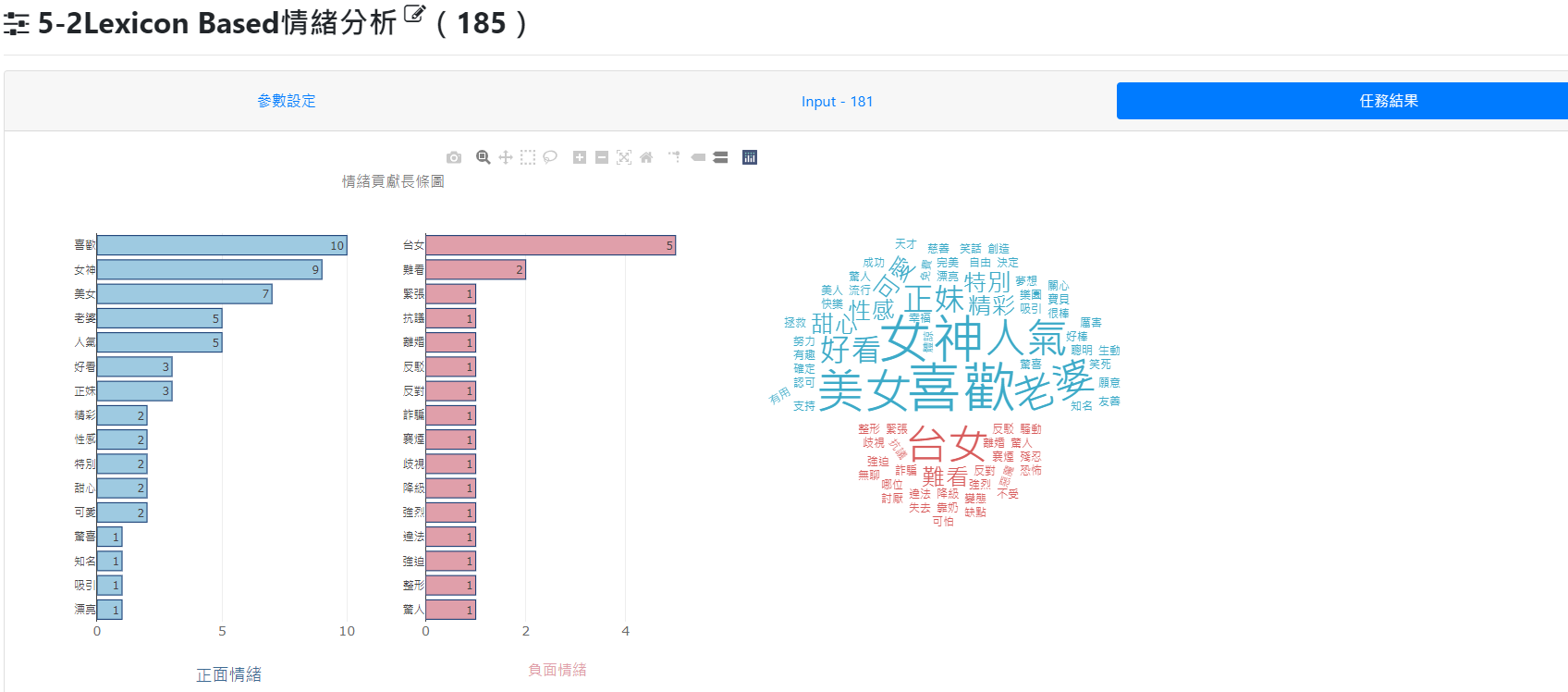


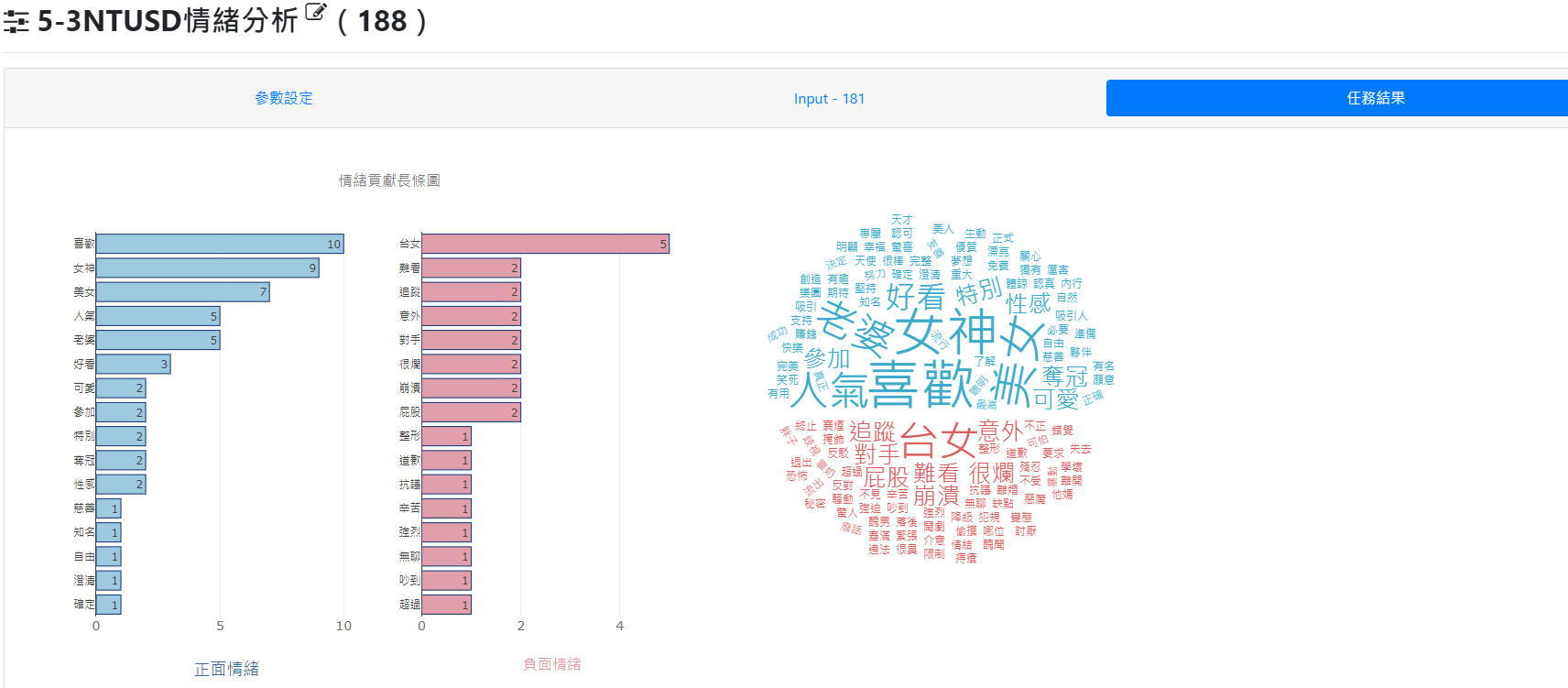
圖 22、指定資料區間內標題Lexicon baesd情緒分析結果

圖 23、指定資料區間內標題NTUSD情緒分析結果



圖 24、指定資料區間內標題詞頻計算結果

1. 內文

自2022年9月起即有關鍵字相關文章討論，表示本組探討主題並非新興議題，但可見在2023年2月起，文章數增多，至3月明顯衝高，表示在近期討論較熱烈（圖 25）。

在指定資料區間中，發文數最多之ID為Kimmy，共發30則貼文，表示該作者可能相當關注這個議題（圖 26），但又可見發文數前六名的作者中，5位均對討論議題抱持正面態度，僅1名作者偏向中立（圖 27）。

將內文進行詞頻分析，以「林襄」最多、「峮峮」次之，及日本、身材等，也提到其他的啦啦隊成員，如yuri、短今、妮可等，表示針對本組探討主題，使用者可能較多關注於外表，且以林襄、峮峮討論度最高（圖 28）；依據LIWC情緒分析結果，內文中Affect word約佔45.7%，正面詞彙43.5%、負面詞彙6.52%，整體而言正面詞彙比例明顯大於負面（圖 29），再觀察Lexicon Based情緒分析結果，仍以正面情緒明顯較多，且以外型為主，如可愛、美女、甜美、正妹等，與詞頻分析結果結論相同（圖 30）。

依據Lexicon Based分析結果觀察指定資料區間中整體情緒分數的變化，可見約2/17整體文章情緒分數極低，表示該日對討論議題抱持負面態度，可能於當日有負面新聞見報；而約3/8開始，整體文章情緒分數均較高，表示該日以後，使用者對於討論議題之文章均抱持正面態度（圖 31）。

承上，接著探討正、負面詞彙出現數量與按月分布情形，可見資料期間正面詞彙數量中位數均約1個，但在2022年12、2023年1、3月都有較多的離群值，但最大異常值則出現在2022年的9月，表示以上月份可能有使用者發表非常正面的文章（狂粉）（圖 32）；負面詞彙數量中位數均約0個，整體數量也較正面少，但在2023年2月有最大異常值出現，表示這個月可能有使用者發表非常負面的文章（黑粉）（圖 33）。

接著，本組選擇7位新聞能見度較高之啦啦隊員分析指定資料區間內之討論度，可見以「林襄」遠高於其他隊員，但可注意的是，「李多慧」於3/17宣布加盟樂天桃猿旗下啦啦隊，不過約2天多的時候，其討論度已較其他原臺灣職棒啦啦隊隊員高，顯示使用者對其可能有高度興趣（圖 34、圖 35）。

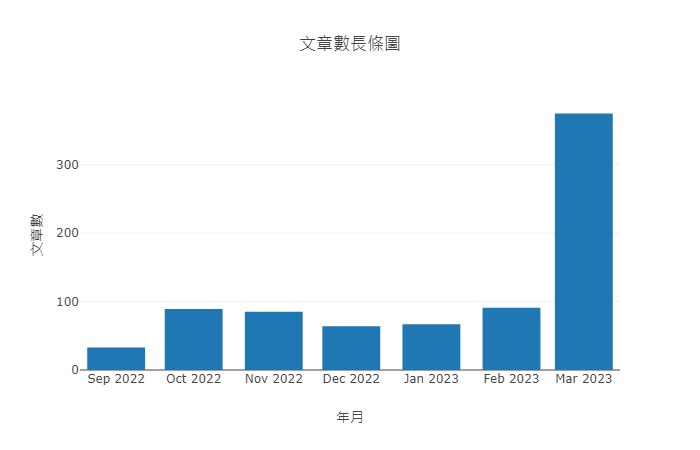


圖 25、指定資料區間內相關文章數分布

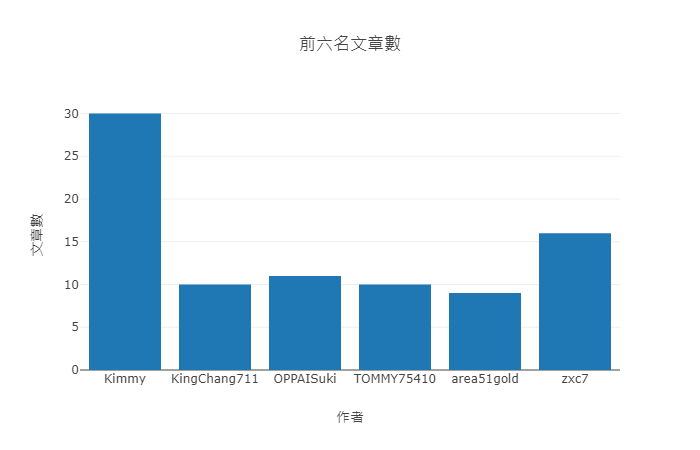


圖 26、指定資料區間內發文數前六名作者

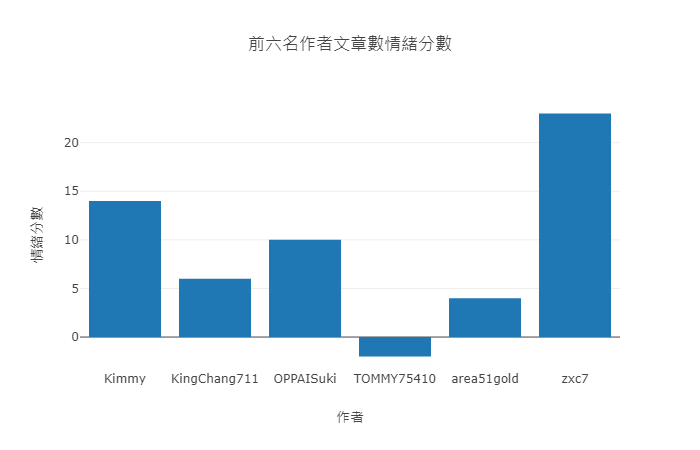


圖 27、指定資料區間內發文數前六名作者之情緒分數概況

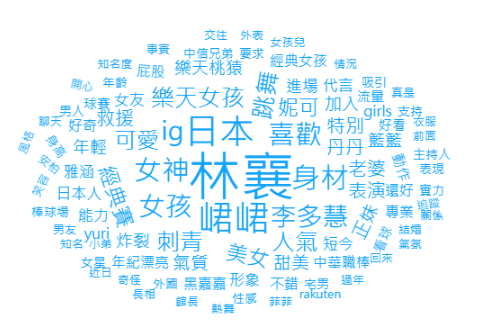


圖 28、指定資料區間內內文詞頻計算結果

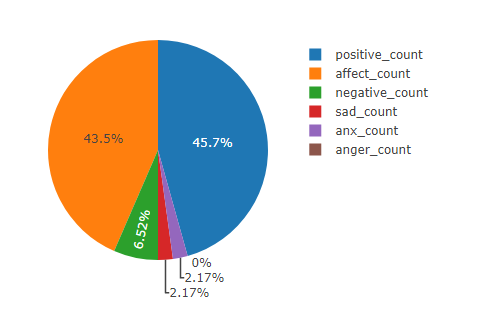


圖 29、指定資料區間內內文LIWC情緒分析結果

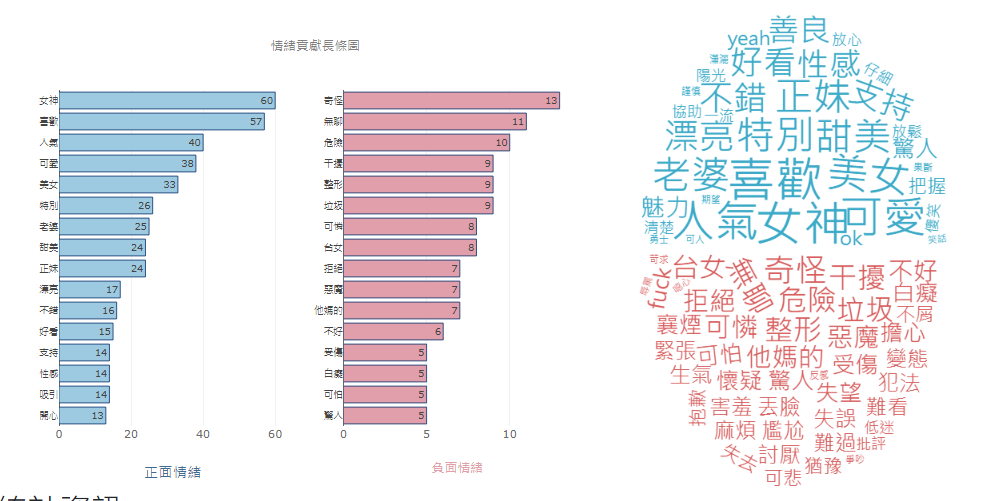


圖 30、指定資料區間內內文Lexicon baesd情緒分析結果

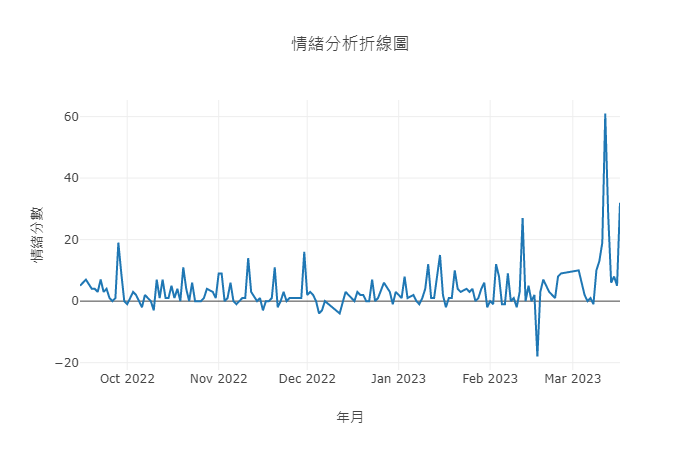


圖 31、指定資料區間內內文情緒分數變化

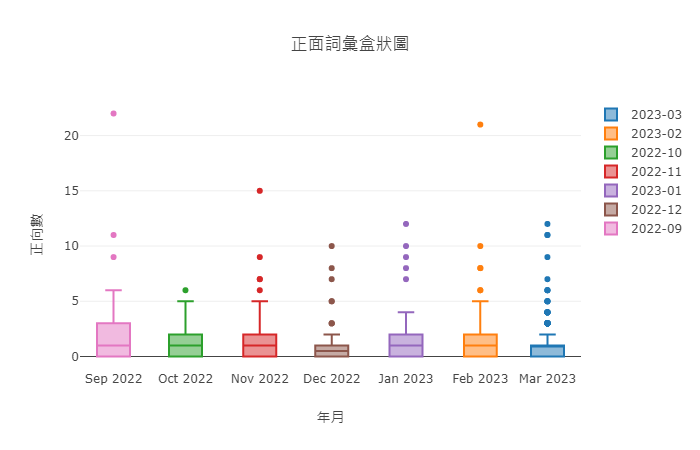


圖 32、指定資料區間內內文正面詞彙數分布

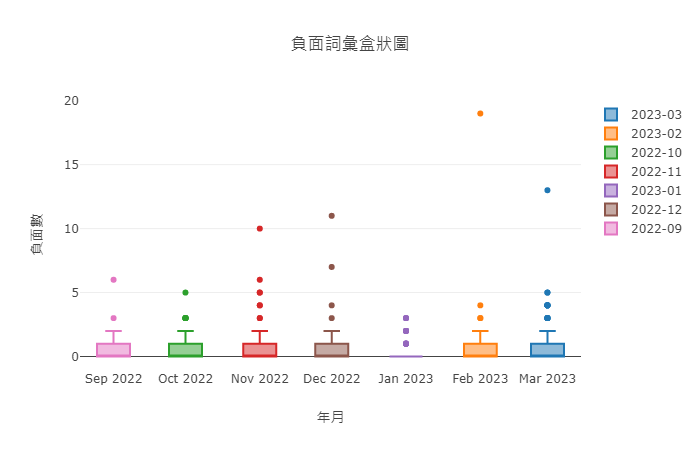


圖 33、指定資料區間內內文負面詞彙數分布

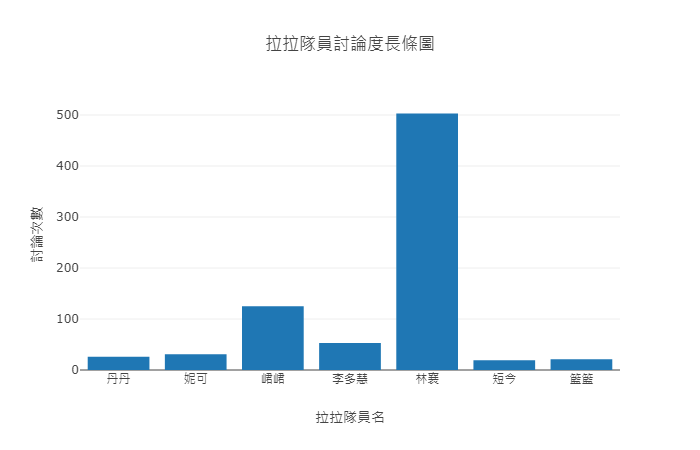


圖 34、指定資料區間內，內文特定詞彙討論度

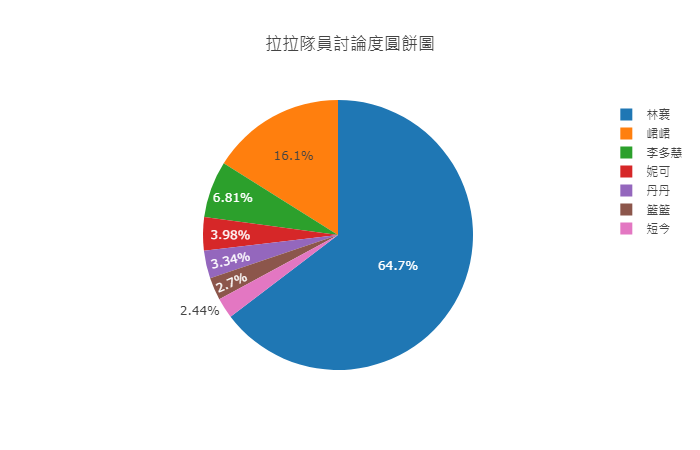


圖 35、指定資料區間內，內文特定詞彙討論度占比

1. 留言

依據LIWC情緒分析結果，內文中Affect word約佔42.1%，正面詞彙36.8%、負面詞彙15.8%，整體而言正面詞彙比例大於負面；但與內文相比（正面43.5%、負面6.52%、悲傷2.17%），正面詞彙比例下降、負面詞彙比例上升，且悲傷詞彙也上升，顯示留言評價結果可能與內文不同（圖 36）。

再觀察Lexicon Based情緒分析結果，仍以正面情緒較多，且與內文相同以外型為主，如可愛、美女、甜美、正妹等，但負面情緒詞彙數也大幅增加；其中正面情緒字彙數最多者為「笑死」，該詞彙雖被分類於正面，但在不同語境中可能出現正、負面意義，如加入此一考量，則較難判斷留言整體屬正面或負面（圖 37）。

而以詞頻分析結果而言，留言中出現最高頻率之詞仍為「林襄」，顯見使用者確實對其有較高的討論或興趣，接續為「笑死」、「喜歡」、「垃圾」等詞彙，依據前述「笑死」暫難全數分類為正面或負面之考量，留言整體對探討主題之評價可能相當兩極。

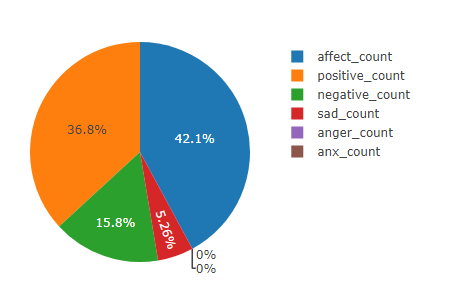


圖 36、指定資料區間內留言LIWC情緒分析結果

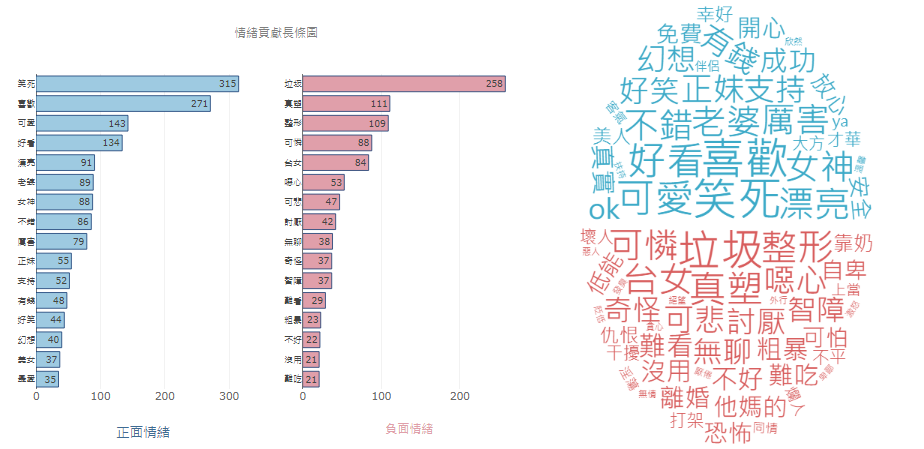


圖 37、指定資料區間內內文Lexicon Based情緒分析結果



圖 38、指定資料區間內留言詞頻計算結果