社群媒體分析第七組第二次讀書會作業

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](#_Toc131981046)

[二、 分析流程 1](#_Toc131981047)

[(一) 資料搜集 1](#_Toc131981048)

[1. 爬蟲 1](#_Toc131981049)

[2. 替換字串 2](#_Toc131981050)

[3. 中文斷詞 2](#_Toc131981051)

[4. 清除停用字 3](#_Toc131981052)

[(二) 資料分析 3](#_Toc131981053)

[1. 標題 3](#_Toc131981054)

[2. 留言 4](#_Toc131981055)

[三、 分析結果 9](#_Toc131981056)

[1. 標題 9](#_Toc131981057)

[2. 留言 9](#_Toc131981058)

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，第一次讀書會分析流程如圖 1，並簡要摘述2次讀書會之資料前處理流程如第一段落-資料蒐集。

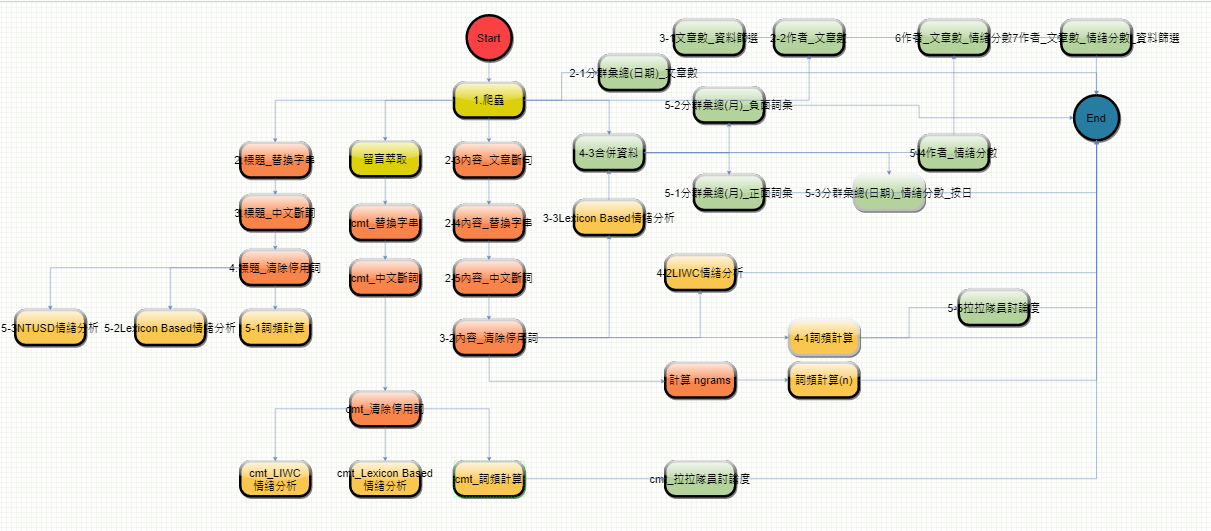


圖 1、總分析流程

本組本次分析包含標題、留言及內文（由左至右）。

1. 資料搜集
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，可參考上次作業。
6. 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。
7. 抓取資料區間：2022年9月14日～2023年3月19日，共804筆。

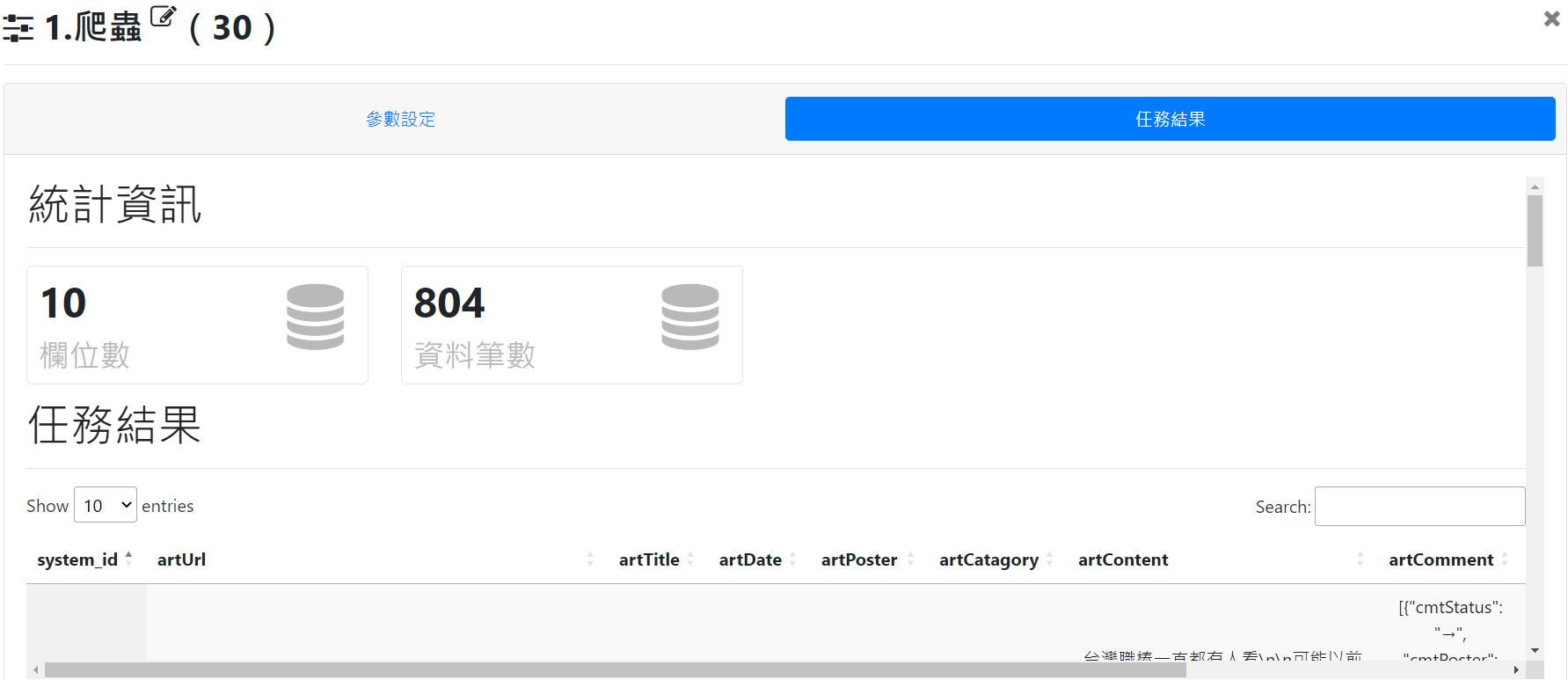


圖 2、爬蟲結果

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。

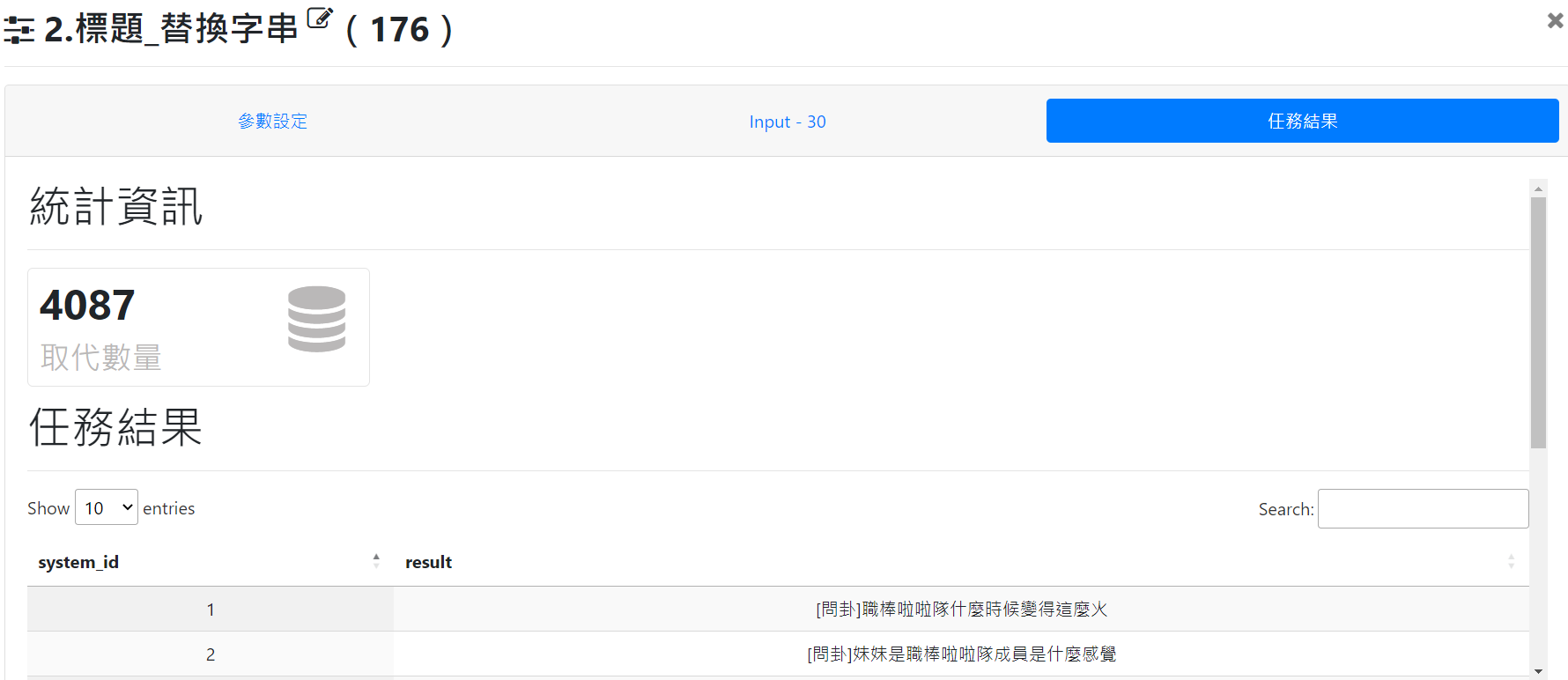


圖 3、替換字串結果

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。

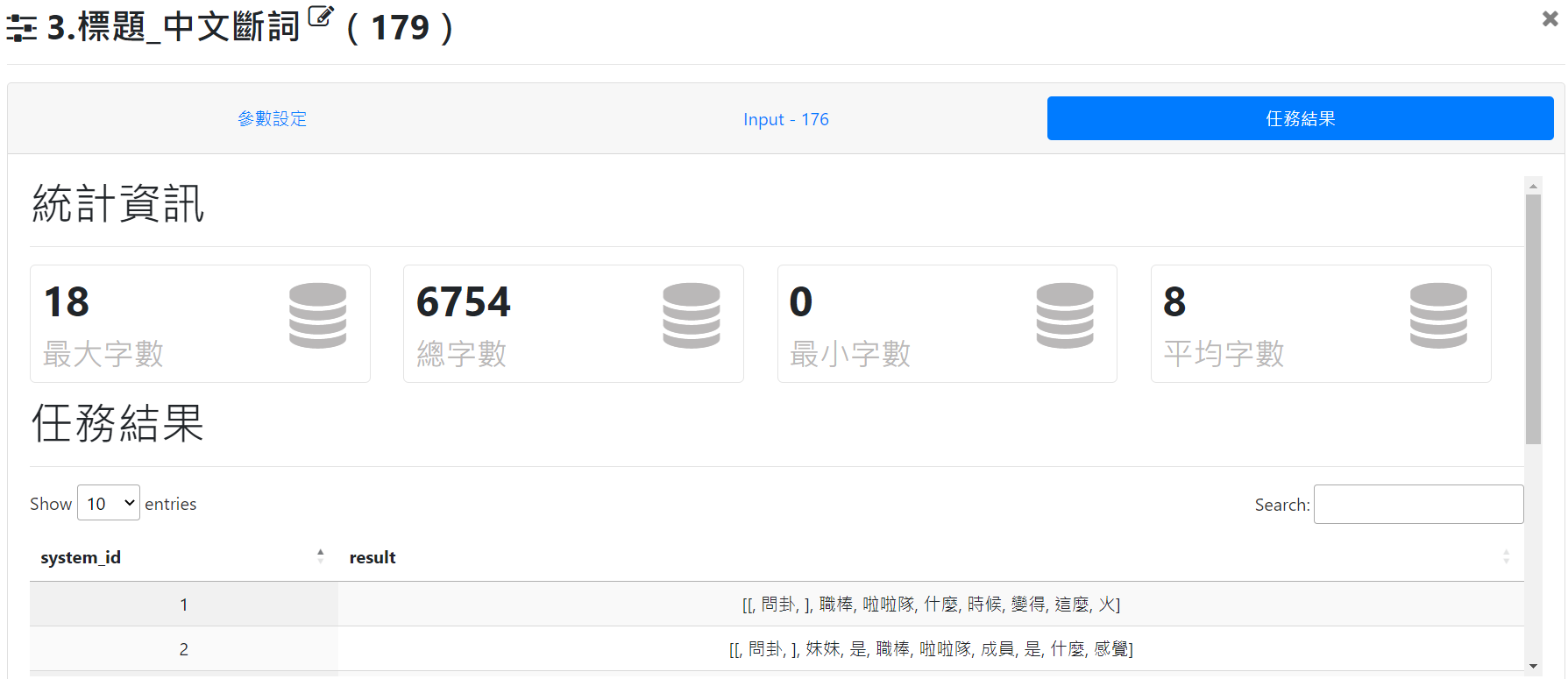


圖 4、中文斷詞結果

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. 資料分析**(第2次讀書會新增內容)**
2. 標題
   1. 詞性篩選+詞頻計算：
      1. 目的：分別篩選標題中「專有名詞」與「形容詞」，並進行詞頻計算，以檢視熱門關鍵字。
      2. 處理方式：
3. 保留詞性：專有名詞、形容詞。
4. 進行詞頻分析。

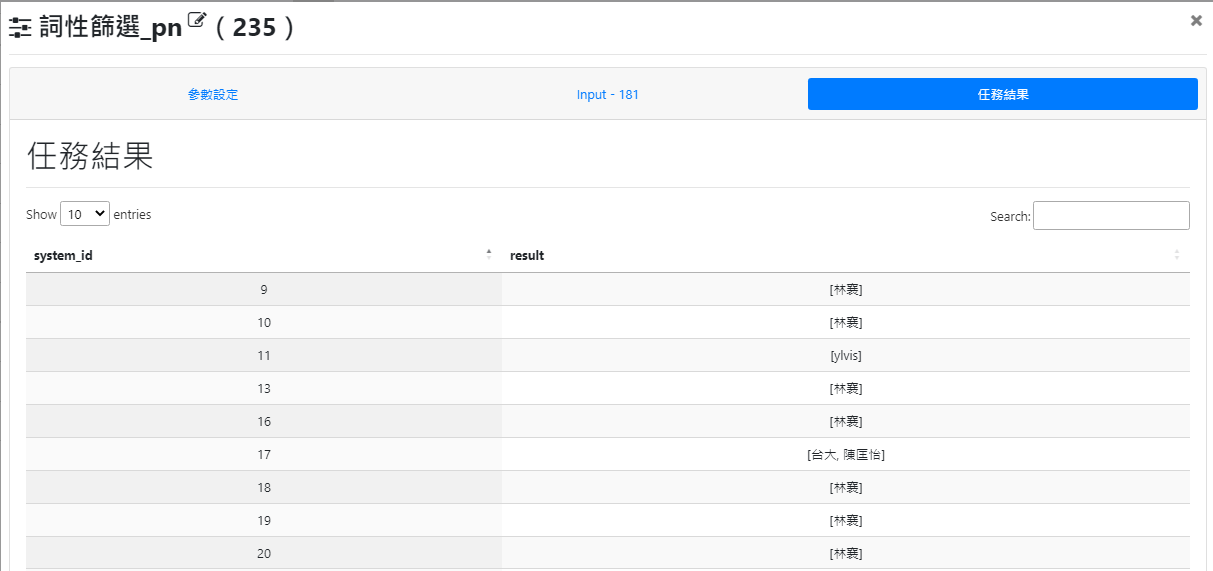


圖 6、標題專有名詞詞性篩選結果

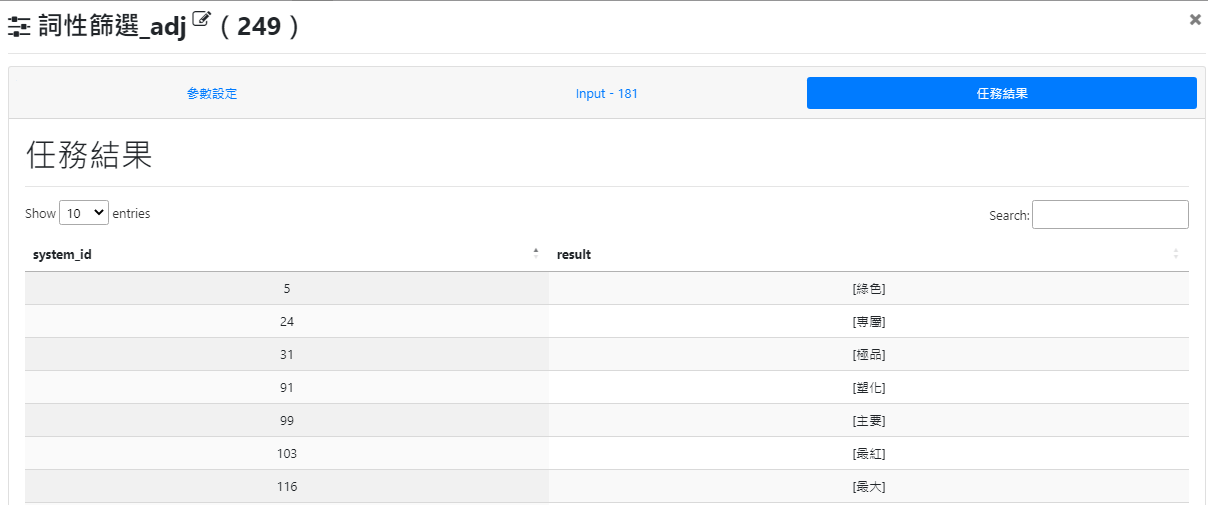


圖 7、標題形容詞詞性篩選結果

* 1. 合併資料：

1. 目的：將篩選結果與爬蟲資料合併以取回原始資訊，以利進行後續分析。
2. 處理方式：執行資料合併，並將CSV檔下載另作分析。
3. 留言
   1. 進階字典情緒分析：
4. 目的：接續清除停用詞，針對留言詞彙進行進階字典法之情緒分析。
5. 處理方式：設定部分因留言可能出現否定詞，於是在清除停用詞部分選擇不清除單字詞，反覆將輸出結果透過NTUSD字典法進行正負面情緒分析(圖 8)。

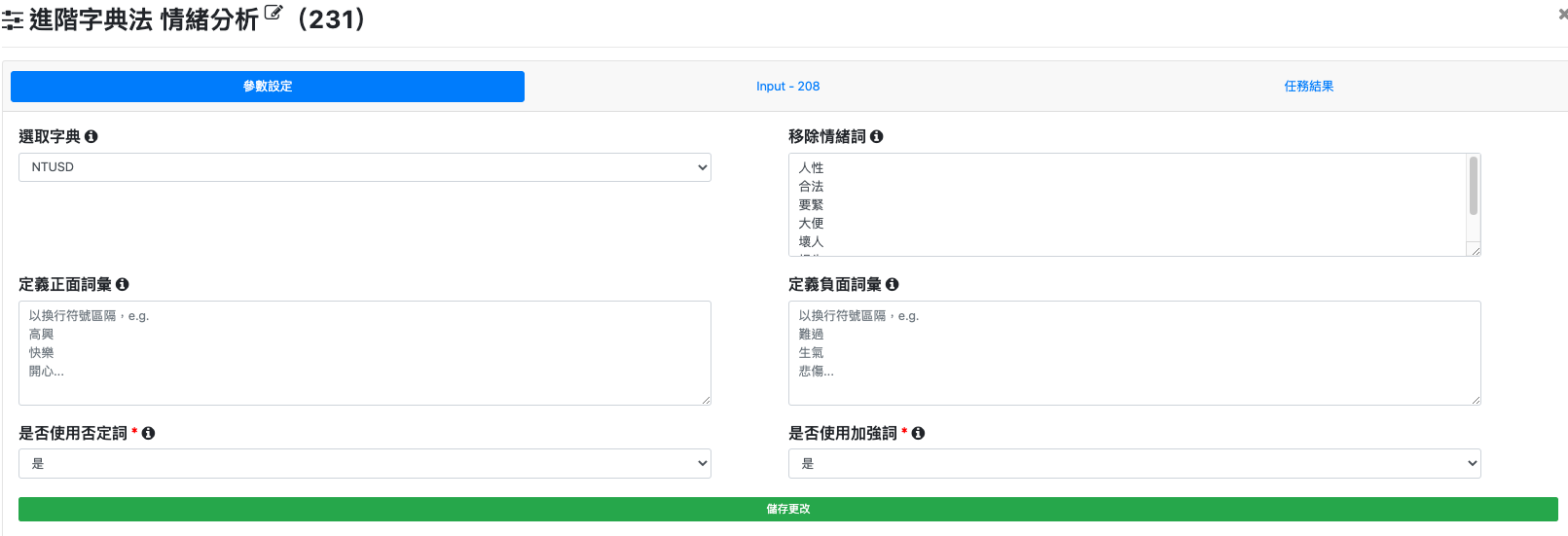


圖 8、進階字典法分析設定

* 1. 詞性篩選+詞頻計算：

1. 目的：篩選出專有名詞和動詞，進行出現詞彙之頻率計算。
2. 篩選詞彙數量：50。
   1. Corpus Base分析：
3. 目的：接續清除停用詞，針對留言詞彙進行Corpus Base分析取得情緒分數。
4. 輸出結果：得到數據以0.5為基準，超過為正面情緒，反之為負面情緒。
   1. 合併資料：
5. 目的：接續2-3處理，將Corpus Base分析與原留言萃取資料合併，以進行其他欄位分析。
6. JOIN KEY：comment\_idx及system\_id。
   1. 分群匯總\_每日平均情緒分數(圖 9)
7. 目的：接續4-3處理，計算每日平均情緒分數。
8. 處理方式：使用「cmtDate」作為日期欄位，並依日進行分析；計算「sentiment\_value」欄位平均數，並保留「cmtDate」欄位。
9. 執行結果：以日期分群，計算每日平均情緒分數。

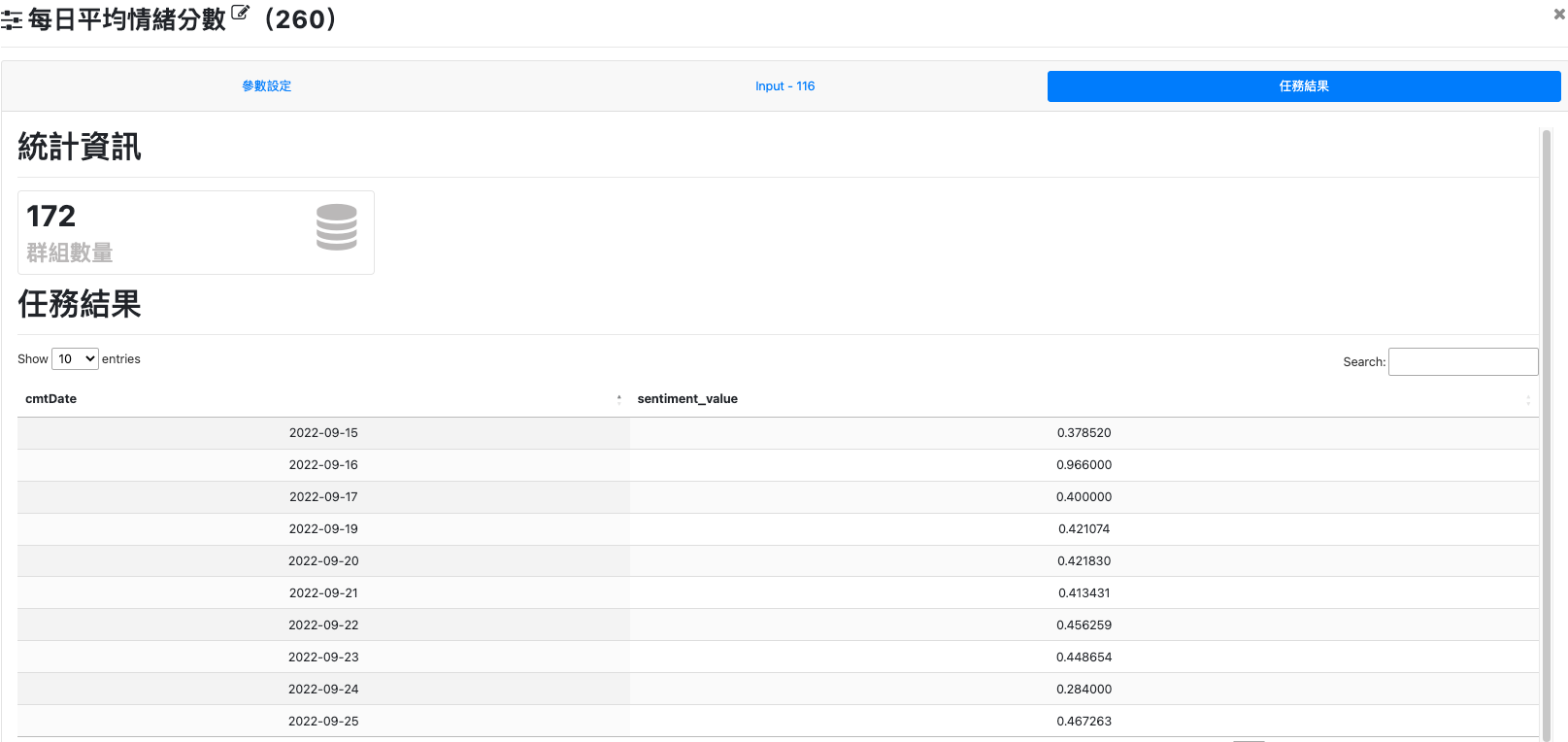


圖 9、每日平均情緒分數

* 1. 資料篩選＿正面情緒(圖 10)

1. 目的：接續4-3處理，篩選出情緒分數大於及等於0.5，即呈現出偏正向之留言。
2. 條件式：'$sentiment\_value>=0.5'。

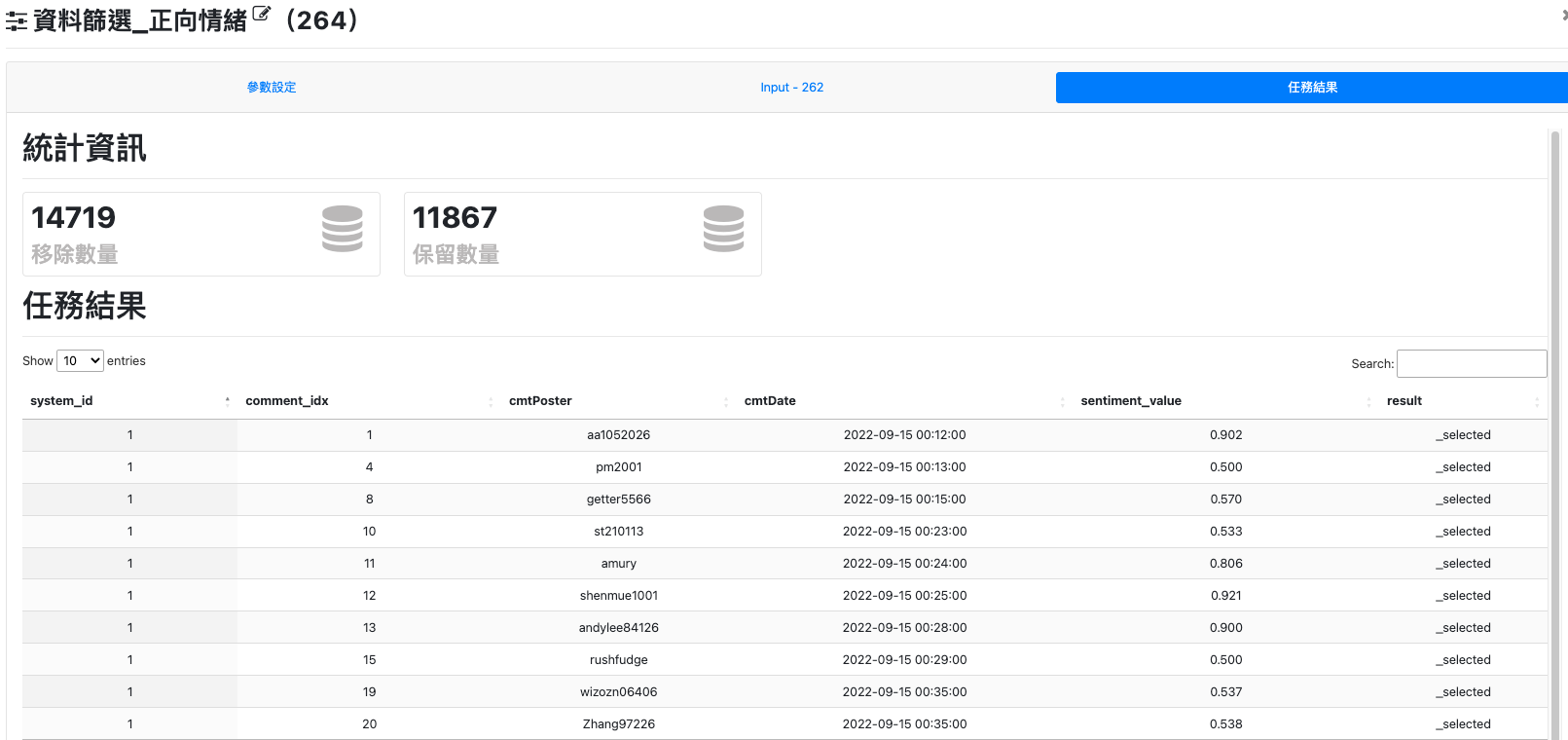
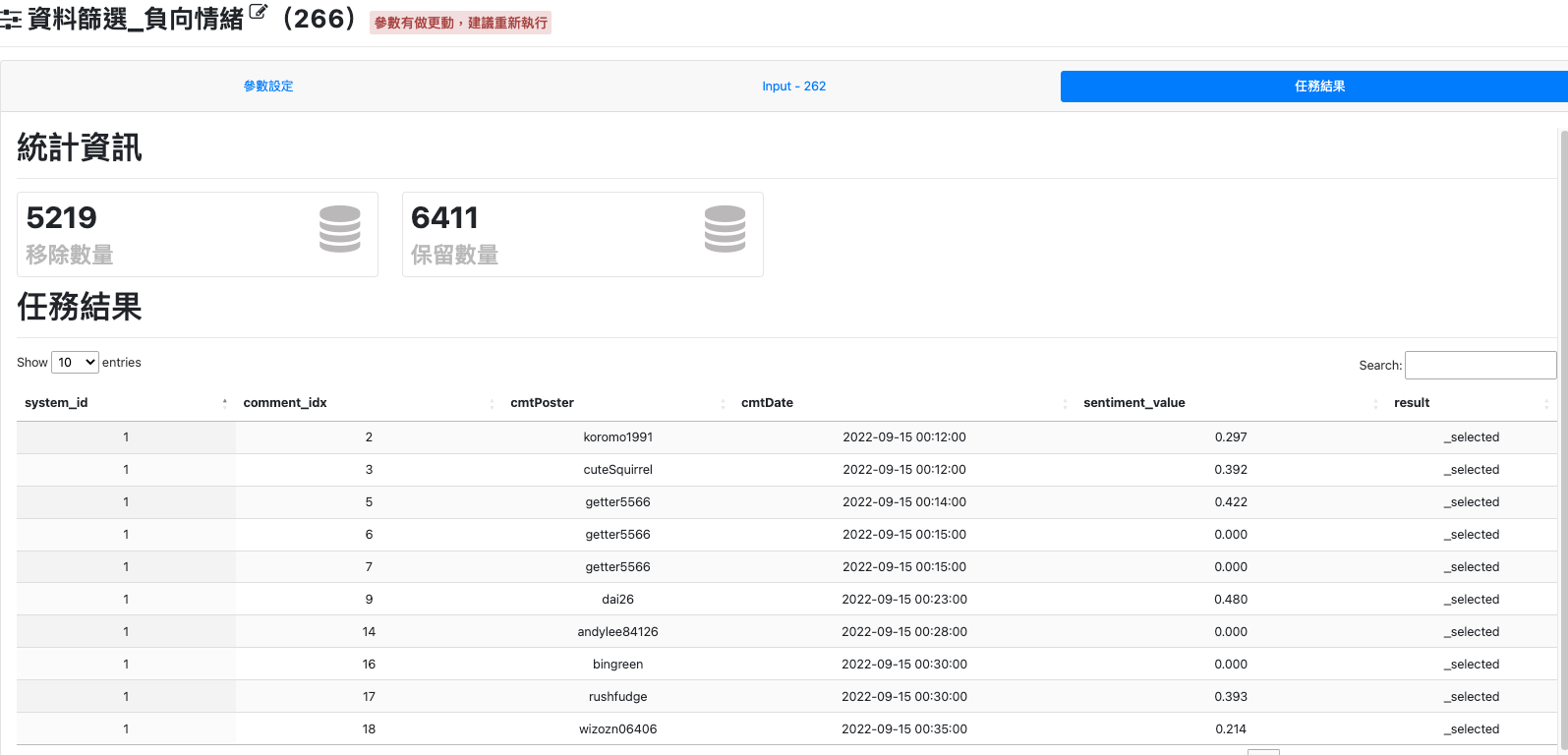


圖 10、正向留言結果

* 1. 資料篩選＿負面情緒(
  2. 圖 11)

1. 目的：接續4-3處理，篩選出情緒分數小於0.5，但不為0，即呈現出偏負向之留言。
2. 條件式：{'AND':['$sentiment\_value<0.5','$sentiment\_value>0']}

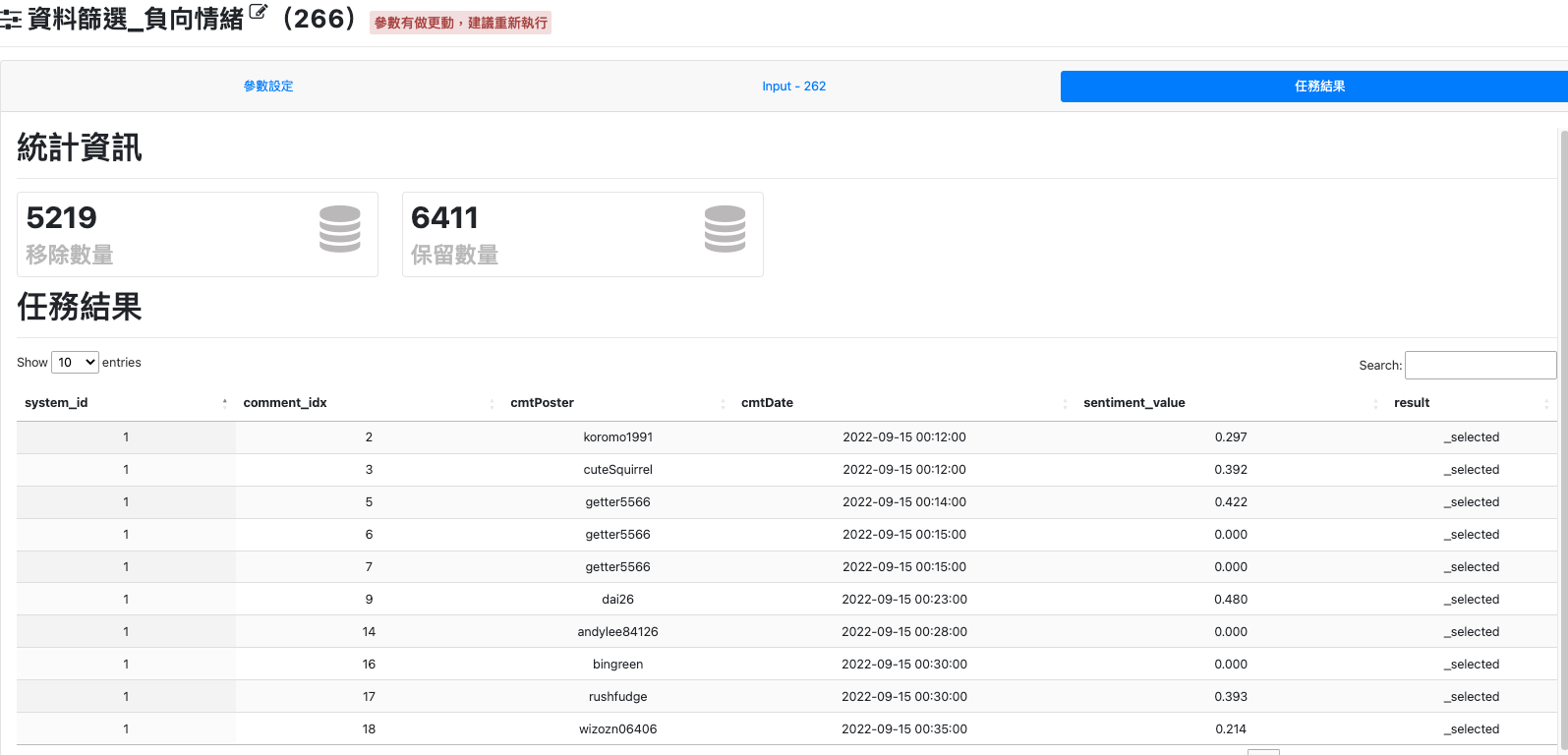


圖 11、負向留言結果

* 1. 分群匯總＿留言者之正面留言數(圖 12)

1. 目的：接續2-6處理，以留言者與同篇文章為計算標的，計算留言數。
2. 處理方式：使用「cmtPoster、system\_id」作為分群依據，並計算「sentiment\_value」欄位之數量。

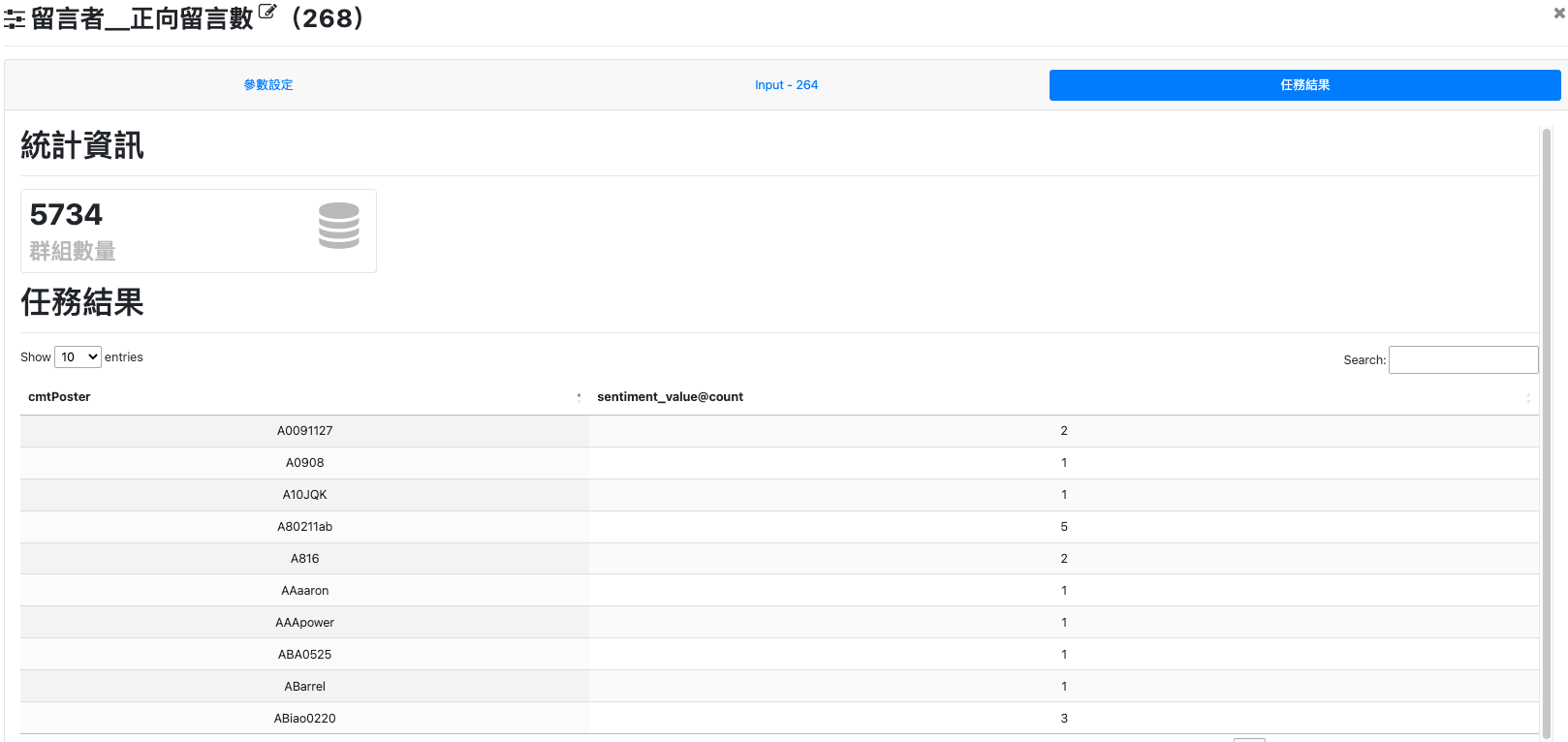


圖 12、留言者正向留言數結果

* 1. 分群匯總＿留言者之負向留言數(圖 13)

1. 目的：接續2-7處理，以留言者與同篇文章為計算標的，計算留言數。
2. 處理方式：使用「cmtPoster、system\_id」作為分群依據，並計算「sentiment\_value」欄位之數量。

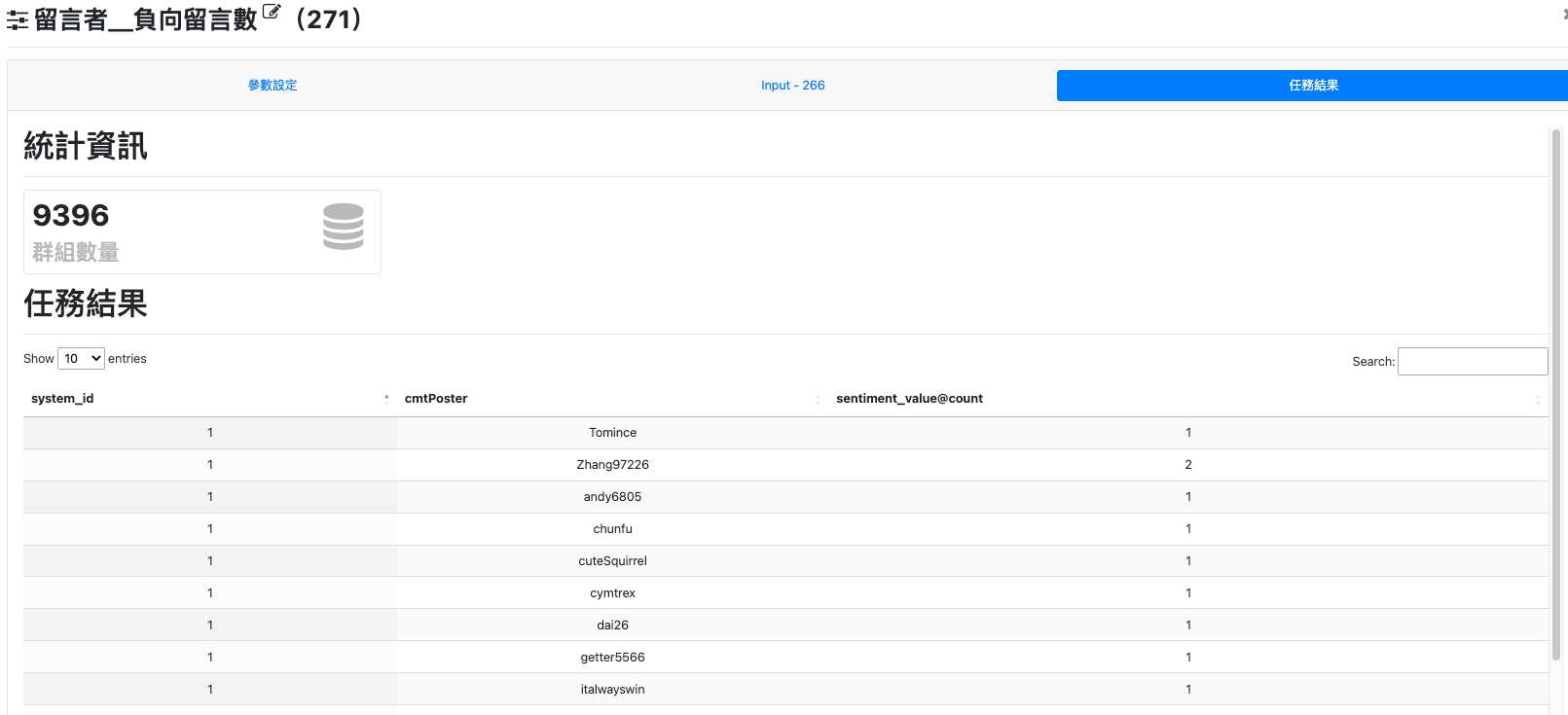


圖 13、留言者負向留言數結果

* 1. 資料篩選＿正面情緒留言者排名前六名(圖 14)

1. 目的：接續2-8處理，篩選出發表正向留言數大於8篇之留言者。
2. 條件式：'$sentiment\_value@count>8。

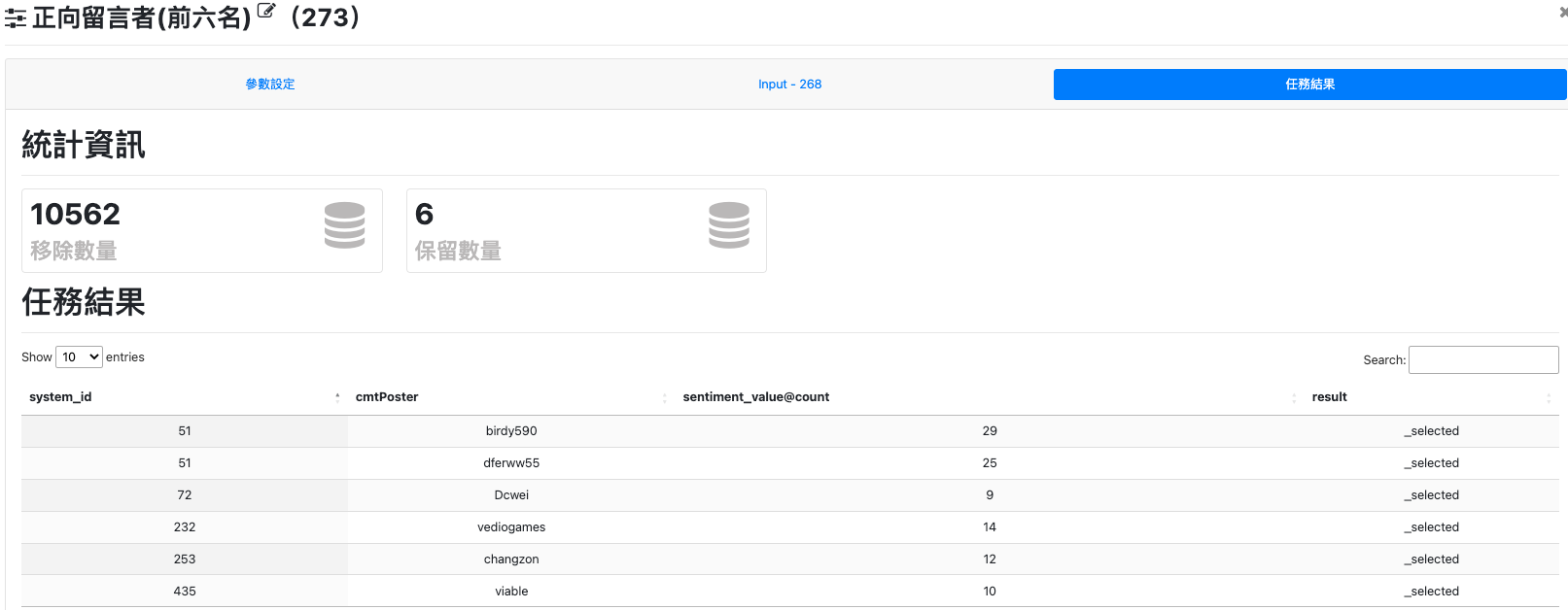


圖 14、排名前六名正面情緒留言者

* 1. 資料篩選＿負面情緒留言者排名前六名(圖 15)

1. 目的：接續2-8處理，篩選出發表負向留言數大於8篇之留言者。
2. 條件式：'$sentiment\_value@count>8。

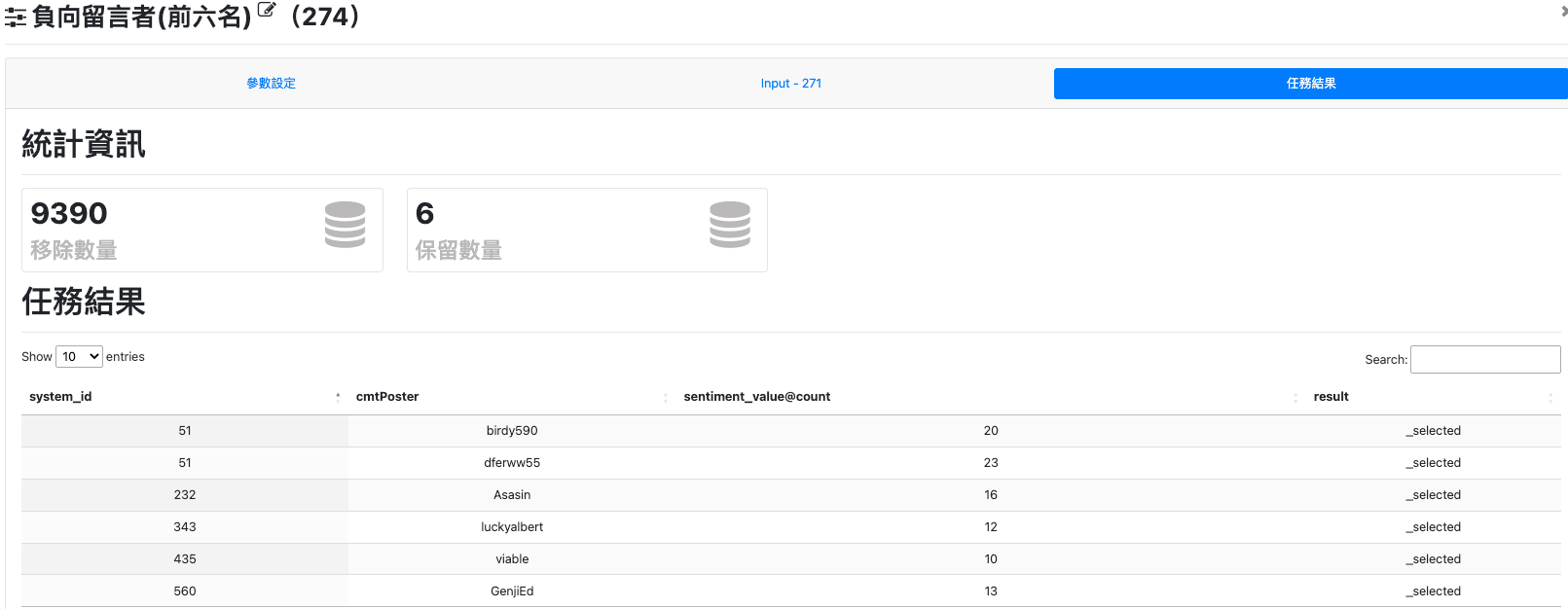


圖 15、排名前六名負面情緒留言者

1. 分析結果
2. 標題

依據篩選專有名詞之詞頻計算結果，以「林襄」最多，佔180次；「李多慧」(含類似專有名詞)44次，其他名詞則均少於10次，表示標題中以林襄、李多慧討論最為熱門。檢視分析結果，以專有名詞篩選詞性的結果算是相當精準，沒有出現明顯分類錯誤的詞語（圖 16）。

形容詞的篩選結果則相當多樣化，進行詞頻分析後，最常出現的詞語也只佔2次，分別為「最紅」、「年輕」、「最大」等。檢視分析結果，以形容詞篩選詞型的結果出現比較大的誤差，包含「科學」、「平均」、「應援團」、「短今」、「滴妹」等應該不屬於形容詞的詞語（圖 17）。

再將標題標記詞性並進行詞頻計算後，以NN 758次佔最多、NR 401次次之、VV 384次第三；總體而言，仍以名詞居多（圖 18）。

將詞性篩選結果與爬蟲資料合併，取得資料原始資訊後，本組將標題出現「林襄」的文章篩選出執行進一步分析，首先，於本組分析的資料區間內，於2022年9月至2023年2月，標題提到林襄的文章數每日平均約15篇，但2023年3月卻急遽衝高到86篇，推測是受經典賽影響，使得「林襄」在八卦版的討論度水漲船高（圖 19）。

再分析於文章標題提到「林襄」的作者發文數，前六名分別為kimmy、area51gold等人，其中又以kimmy以19篇顯著多於其他人（圖 20）；進一步分析前六名作者資料區間內之發文概況，可見kimmy於資料區間中每個月均會發表以林襄為標題之文章，可推測為林襄的忠實粉絲；而第二名的area51gold則僅於2023/3有發文紀錄，但單月發8篇文，推測可能跟autokey都是在此月份起開始關注林襄相關話題（圖 21）。

1. 留言

依據NTUSD進階字典法情緒分析結果分析，正面情緒以「笑死」出現最高，與Lexicon Based情緒分析一樣，但因較難去判斷笑死為正面或負面，暫不討論此詞彙，再來為喜歡、可愛、漂亮…等辭彙，而負面情緒留言詞彙出現最多的是垃圾，其次為可憐、小心、羨慕、噁心，透過字典法分析留言詞彙可見大家對棒球啦啦隊的評價相當多也相當兩極化(圖 22)。

而針對專有名詞和動詞進行篩選的詞頻分析結果而言，專有名詞部分最多為林襄，其次是峮峮，還有許多啦啦隊成員的名字(圖 23)；動詞部分出現最多為喜歡，其次為不行、好看、可愛，因許多形容詞的詞彙出現，也難以判斷留言者使用這些詞彙當下的用意，較難從字典法判斷詞性正確性與否(圖 24)。

依據Corpus Base情緒分析結果，可看出每則留言的情緒分數，透過合併原資料，觀察指定資料區間中整體平均情緒分數的變化，可見約去年9/16為整體文章情緒分數最高，表示該日對討論議題為正向留言較多，也可能是留言數較少，但皆是正向詞彙留言；而最低今年1/14，當天平均情劇分數為零，表示該日無任何相關留言出現，整體而言，留言之平均分數算是在正負向平均範圍中(圖 25)。

在指定資料區間中，依據情緒分數本組也將正面及負面分開討論，也篩選出各自留言者是偏正面(圖 26)或負面(圖 27)，找出最多的前六名，表示該發文者可能相當關注這個啦啦隊這個議題，其中較特殊的為birdy590及dfeww55兩位留言者，留言中出現正向及負向詞彙比例都相當高，可以判斷假設兩位皆是相當中立的留言者，也是啦啦隊的忠實觀眾，留言相對大量也很直接。



圖 16、指定資料區間內標題專有名詞之詞頻分析結果



圖 17、指定資料區間內標題形容詞之詞頻分析結果



圖 18、指定資料區間內標題詞性之詞頻分析結果

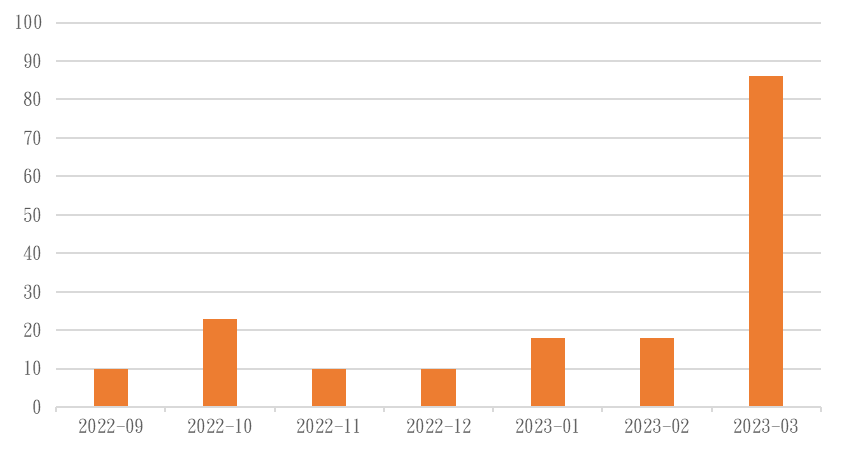


圖 19、指定資料區間內，各月份標題含「林襄」關鍵字之文章數

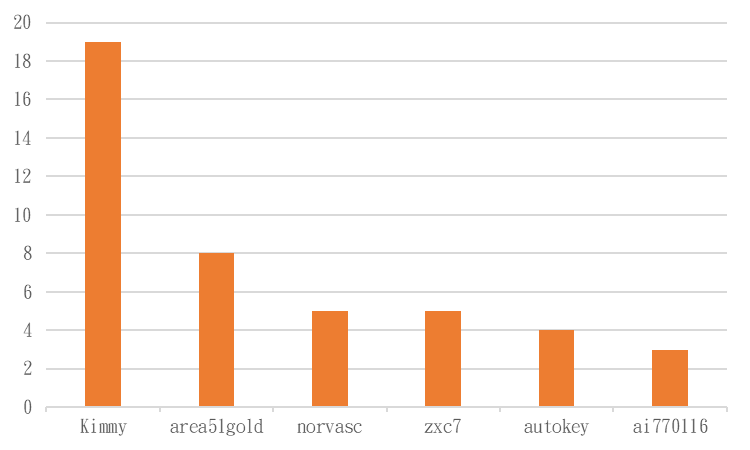


圖 20、指定資料區間內，發表標題含「林襄」關鍵字文章數之前六名作者

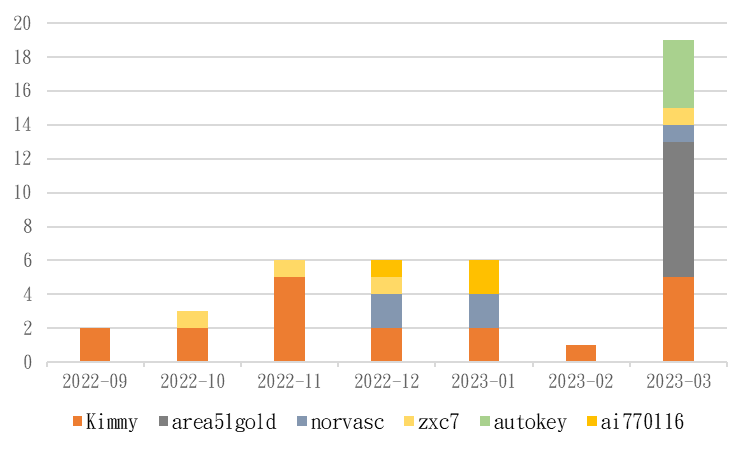


圖 21、指定資料區間內，圖20之前六名作者發文概況

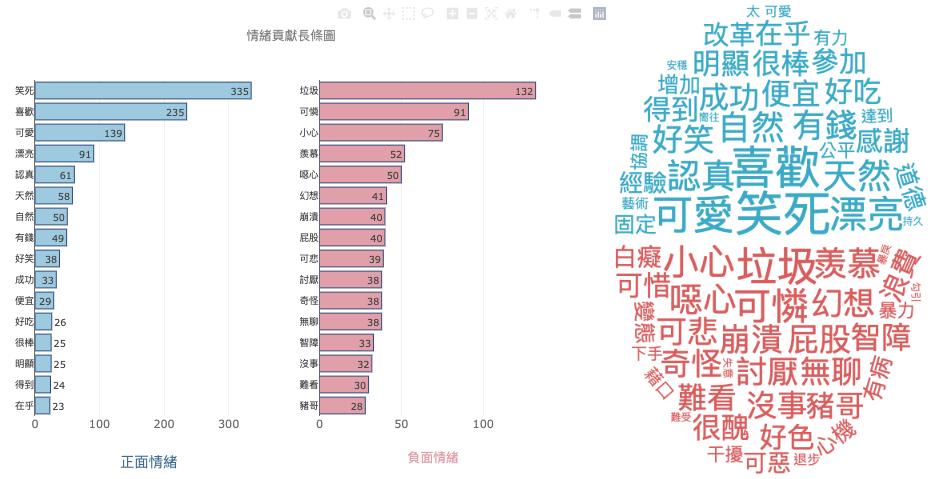


圖 22、指定資料區間內留言進階字典法情緒分析結果



圖 23、指定資料區間內留言專有名詞之詞頻計算結果

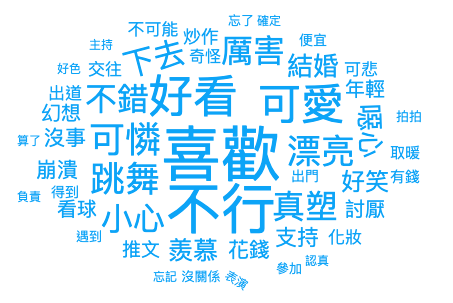


圖 24、指定資料區間內留言動詞之詞頻計算結果

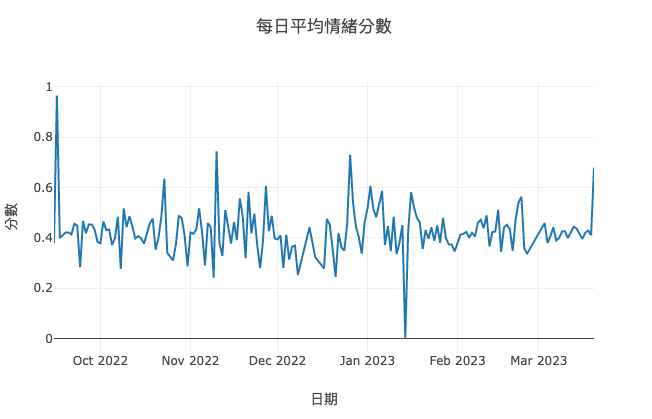


圖 25、指定資料區間內每日平均情緒分數結果

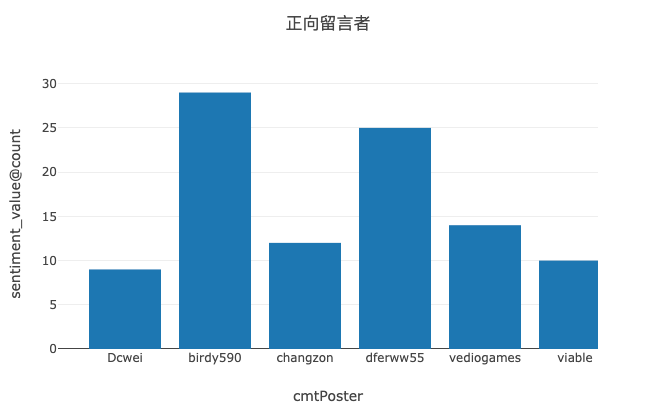


圖 26、指定資料區間內正向情緒前六名留言者結果

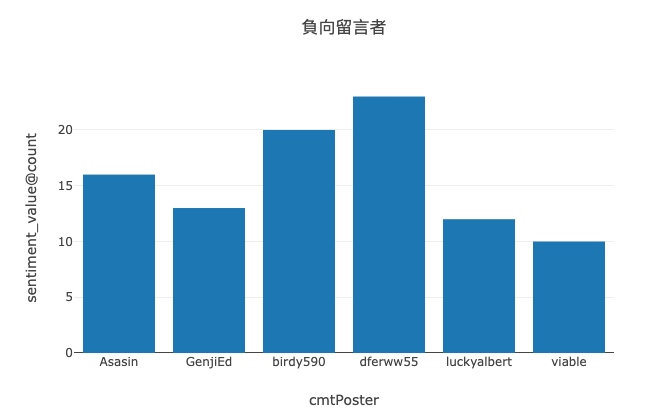


圖 27、指定資料區間內負向情緒前六名留言者結果