**第一次讀書會**

**議題分析**

**指導老師：黃三益 教授**

|  |
| --- |
|  |

**第五組**

組員

M094050003 周映君

M094050008 蔡旻蓉

M094111006 呂奕憲

N094220003 林賢育

N094220008 龔敖摩

N094220017 林慶儒

N094220028 王淵司

**目錄**

[一、分析議題簡介 1](#_Toc86356998)

[二、工作流程設計 2](#_Toc86356999)

[三、資料清理說明 3](#_Toc86357000)

[四、情緒分析探討 11](#_Toc86357001)

[五、分群彙總設計 14](#_Toc86357002)

[六、視覺化圖表呈現 19](#_Toc86357003)

[七、結論 25](#_Toc86357004)

# 一、分析議題簡介

**議題發想：**iPhone 13於今年9月發表及上市，立即成為當月份一個熱門的話題。考量蘋果發表會為台灣時間9/15零時，台灣正式上市的時間為9/24。因此擬從最多訊息傳布的PTT探索從發表會前三天到上市日之後約一週期間有關iPhone 13的討論。並進行下列分析與確認：

* + 對iPhone 13有哪些情緒詞彙？
  + 討論高峰是在哪幾天？
  + 是正面詞彙多，還是負面詞彙多？
  + 負面詞彙與正面詞彙在這段期間的趨勢為何？
  + 分析的過程是否有額外的發現？

**分析期間：**2021/9/13~2021/9/29

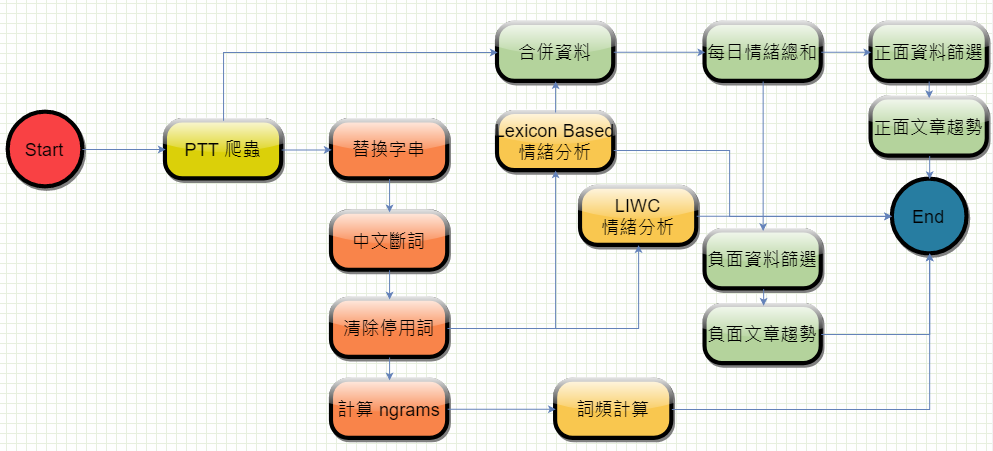
**資料來源：**PTT八卦板、手機通訊板及iPhone板

**使用工具：**工作流程系統

# 二、工作流程設計

工作流程平台建立的工作流程為Team5\_HW1，設計執行的步驟依序為：

1. 指定iPhone相關關鍵字與排除字，讀入”PTT爬蟲”的資料。
2. 將網址開頭、PTT平台發文字串用”替換字串”替換為空白。
3. 用”中文斷詞”加大特定詞彙的權重。
4. 對不具分析效益的詞彙用清除停用詞設定停用。
5. 設定Lexicon Based和LIWC兩個情緒分析字典，以瞭解情緒詞彙、正面與負面詞彙的發生情形。
6. 用”計算ngrams”設定2來彙總討論文中的詞組(兩個詞彙組成一組)，並用”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. 用合併資料合併PTT爬蟲與Lexicon Based情緒分析結果，並使用  
   ”分群彙總(日期)”依日期來彙總每日的情緒值(即每日情緒總和)。
8. 用”資料篩選”分別篩選情緒值為負數及正數的資料，然後再用  
   ”分群彙總(日期)”，彙總這段期間每天負面文章與正面文章趨勢。



# 三、資料清理說明

* **PTT爬蟲部分**

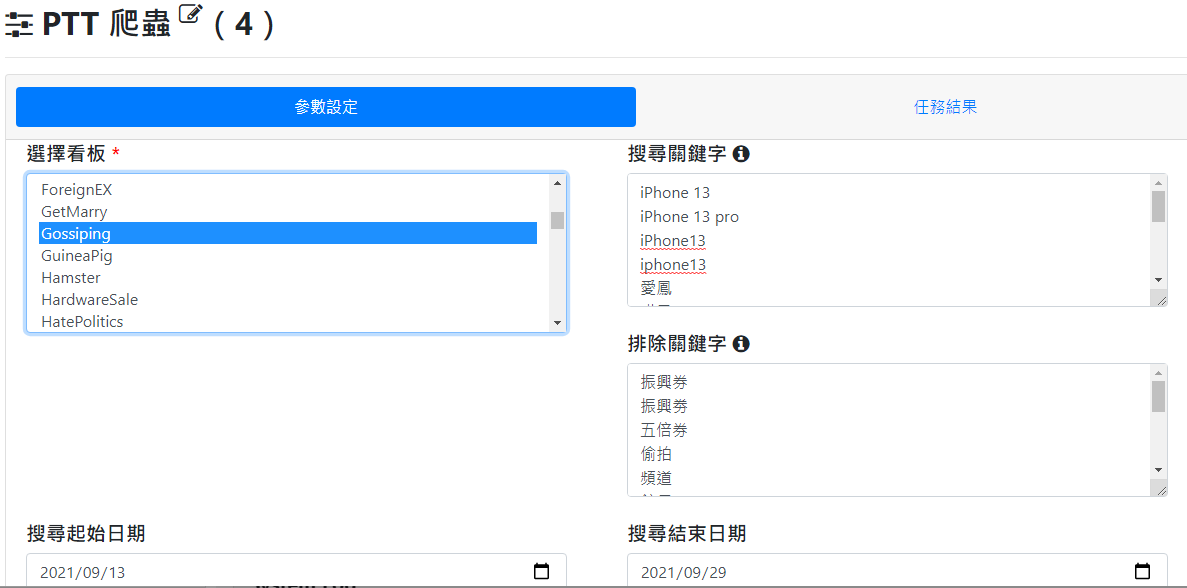
選擇的板為Gossiping(八卦板)、MobileComm(手機通訊)和iPhone版。

搜尋關鍵字為：

iPhone 13; iPhone 13 pro; iPhone13; iphone13; 愛鳳; 哎鳳; 哀鳳; 蘋果新手機; 蘋果發表會; iPhone新機; iPhone新色; iPhone發表會; 果粉; 天峰藍。

設定上述關鍵字的理由：

因為考慮iPhone 13有不同連字和大小寫組合，故做此設定，另外考慮年輕人對iPhone的中文說明，而加入哀鳳等說明。此外亦加入發表會與討論度高的蘋果手機新色天峰藍來聚焦蒐集與iPhone 13相關資訊。至於為什麼不用只用iPhone當關鍵字，因為有試過只用iPhone會找到太多與iPhone 13的文章。

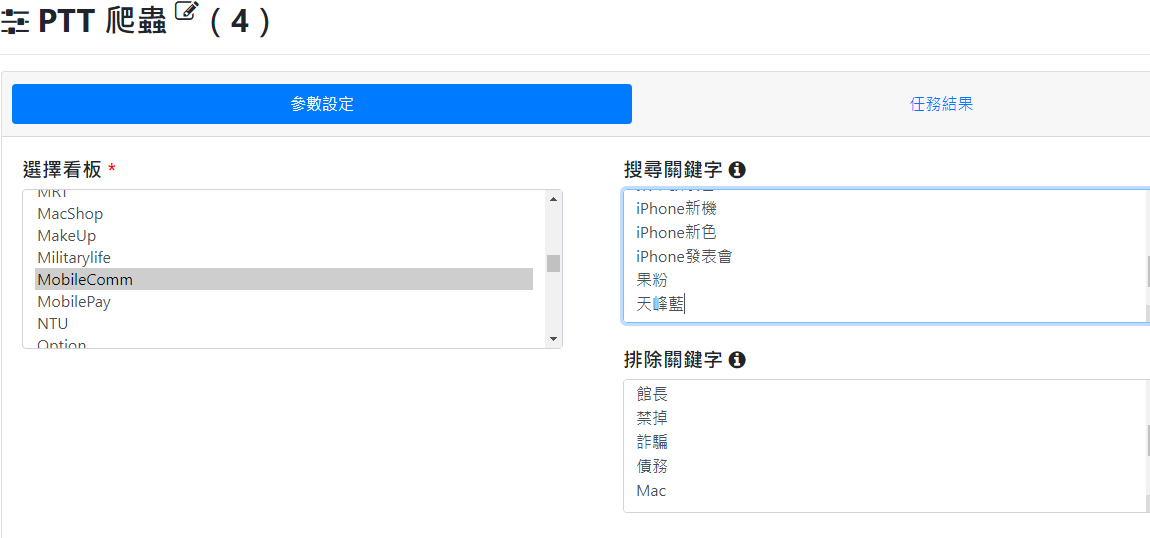


排除關鍵字為：

振興券; 振興劵; 五倍券; 偷拍; 頻道; 館長; 禁掉; 詐騙; 債務; Mac; mac; MacBook; iPad; ipad。

設定上述關鍵字的理由：

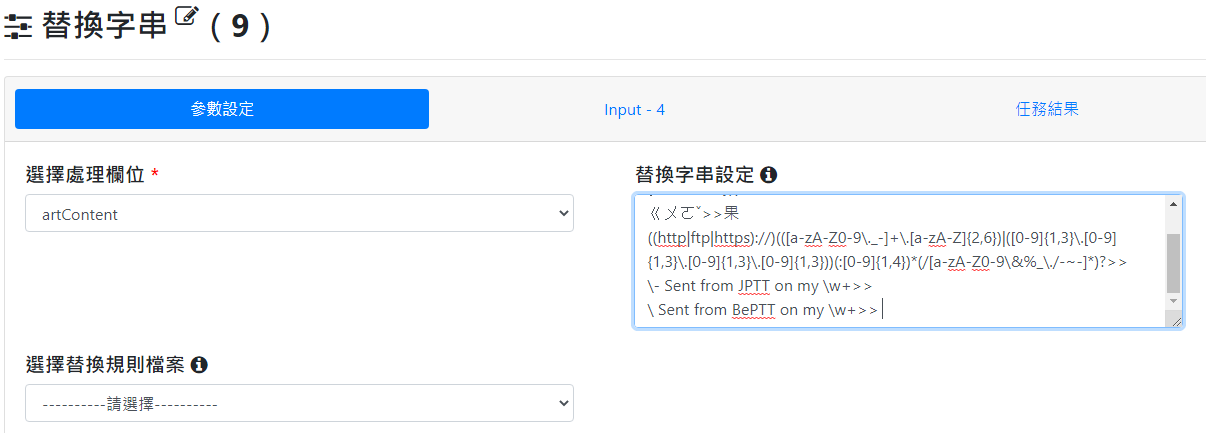
因為考慮聚焦在iPhone推出後情緒分析，因此在粗略審視第一次PTT爬蟲結果後，對於振興券怎麼用、蘋果手機偷拍、蘋果頻道、網紅館長、中國禁掉蘋果、手機詐騙及債務等進行排除。同時考慮蘋果發表會聚焦的項目是新iPhone，因此排除Mac和iPad的關鍵字。



另外，經檢視爬蟲後的資料，僅Gossiping(八卦板)、MobileComm  
(手機通訊)有資料，無iPhone板的資料。

* **替換字串**

除了參考助教範例，也加入另一個常見的發文字串Sent from BePTT，用下面正則表示式替換為空白。  
\ Sent from BePTT on my \w+>>



* **中文斷詞**

因為發現斷詞工具會把iphone和13，以及天峰和藍拆開，所以對iphone13、天峰藍和其他藍色設定較高的權重。



iphone13 20000

iPhone13 20000

天峰藍 10000

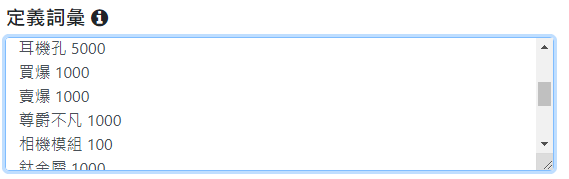
天鋒藍 10000

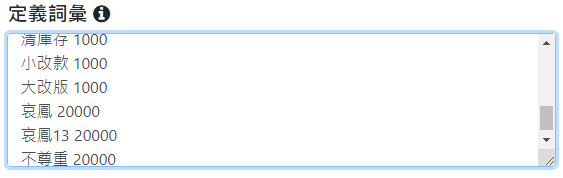
太平洋藍 10000

天空藍 10000

此外，看到耳機和孔、買和爆、賣和爆、清和庫存、小和改款、大和改版、相機和模組、尊爵和不凡、鈦和金屬、哀和鳳等字在第一次斷詞結果時被斷開，因此對它們設定較高的權重。

另外，在後面情緒分析發現”尊重”有被抓出來為情緒詞，但實際檢視內文為”不尊重”，因此對”不尊重”也設定較高的權重。





耳機孔 5000

買爆 1000

賣爆 1000

尊爵不凡 1000

相機模組 100

鈦金屬 1000

清庫存 1000

小改款 1000

大改版 1000

哀鳳 20000

哀鳳13 20000

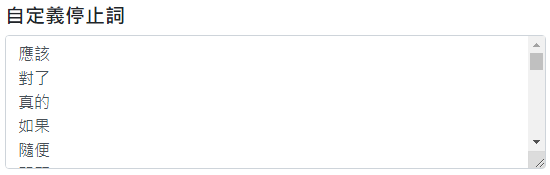
不尊重 20000

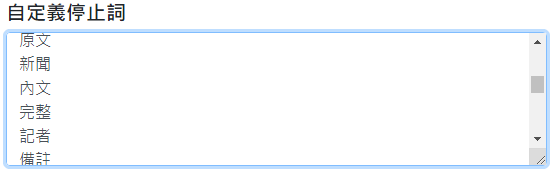
回不去 16000

* **清除停用詞**

除了使用使用預設停止詞，也加入不少自定義停止詞。停止詞的內容包括較無意義的語助詞，例如應該、隨便、還好，並加上轉貼新聞報導的詞彙，如媒體、來源、原文、記者、標題等。由於在第一次產生文字雲時有看到這些新聞報導詞組，因此排除掉這些字，讓後續的文字雲視覺分析可以更精確反映高頻率的詞組。







設定的全部停用詞清單如下：

應該

對了

真的

如果

隨便

問題

八卦

還好

看看

欸

來源

媒體

綜合

報導

原文

新聞

內文

完整

記者

備註

署名

連結

標題

作者

網址

轉載

摘錄

如下

詳細

附註

摘要

全文

照貼

範圍

心得

評論

文章

嘻嘻

為啥

如題

聽說

感覺

樓上

有人

請問

不用

只能

視頻

朋友

謝謝

感謝

保護

免費

遊戲

優惠

正義

信義

希望

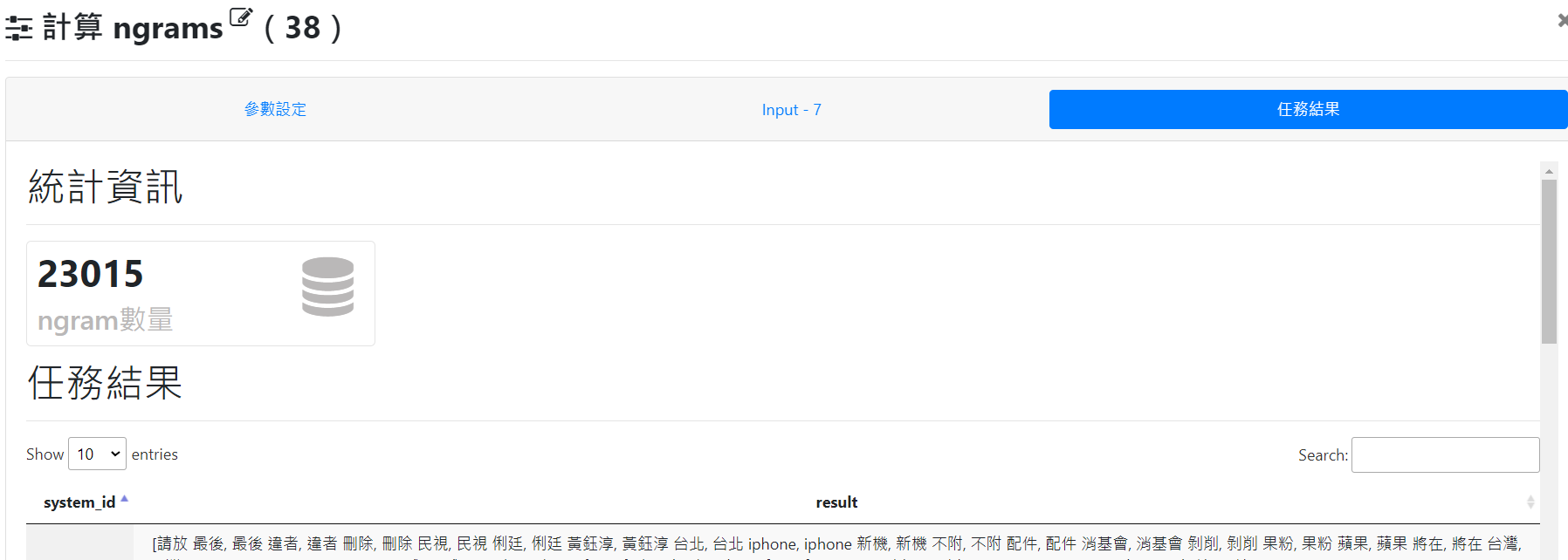
想像

放鬆

mm

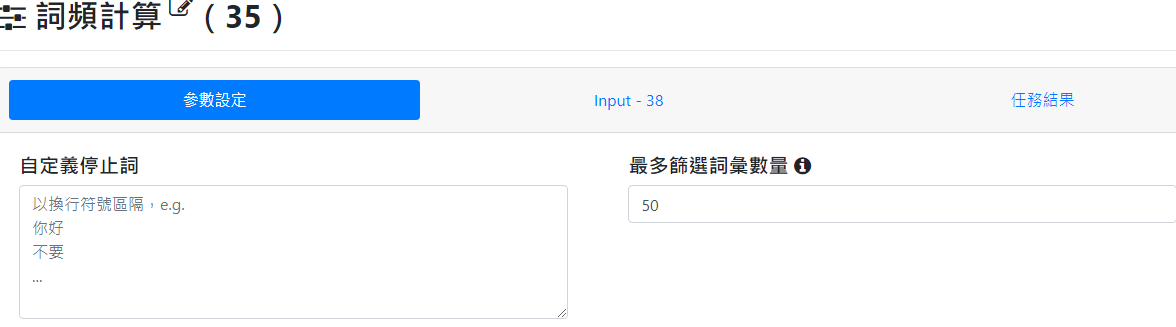
* 使用ngrams另外彙總兩個詞的組合



****

* **使用詞頻計算來產生前面ngrams詞組彙總的前50個高頻詞組的文字雲。**

關於這個文字雲的分析，將在後面視覺化圖表之章節做說明。





# 四、情緒分析探討

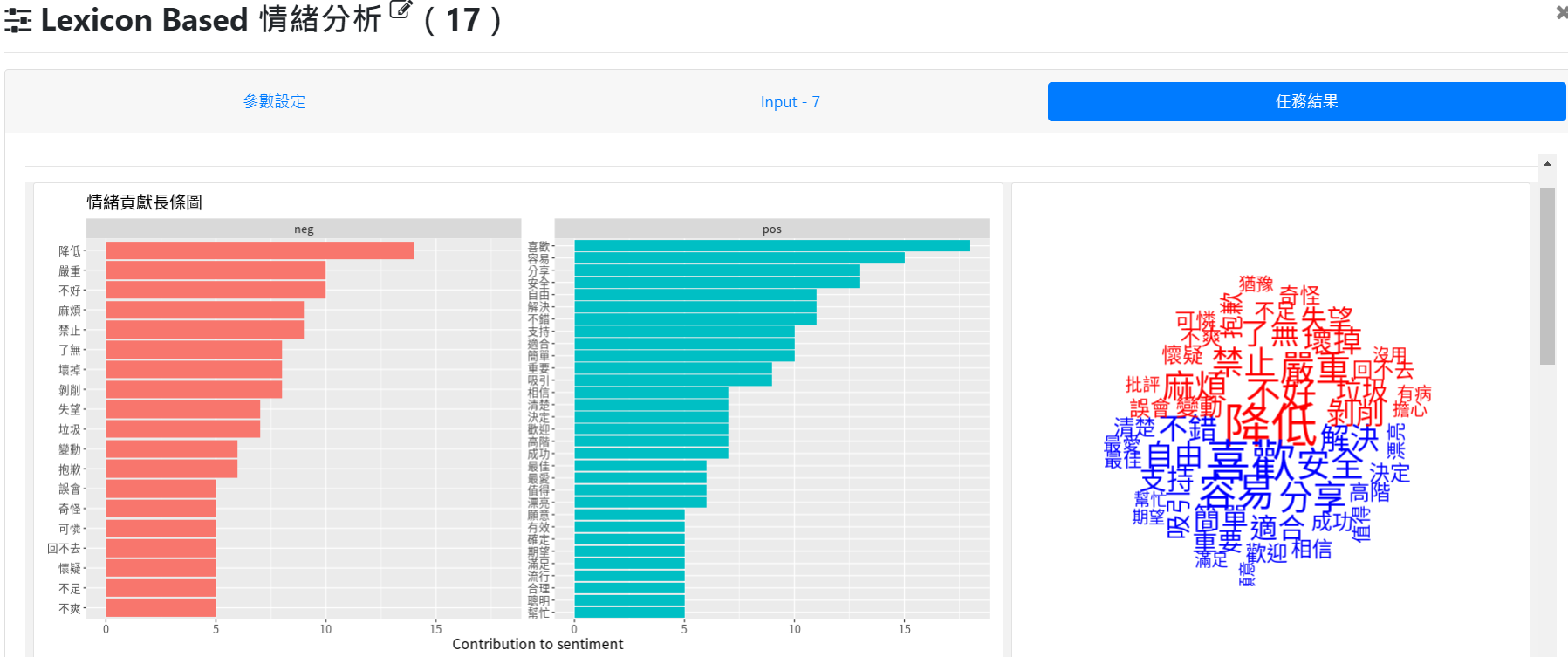
分別使用Lexicon Based與LIWC這兩個情緒字典來分析發文內容的情緒文字。但每日情緒彙總的趨勢則選擇以Lexicon Based的情緒分析結果為主。



經檢視情緒分析第一次輸出結果發現，大部分正面詞彙均有抓到，因此正面詞彙的部分僅新增「贏了」這個詞。此外，「尊重」雖然是正向詞彙，但實際檢視內文發現該斷詞應為「不尊重」，故調整「不尊重」在斷詞設定的權重，並將「不尊重」放入新增的負面詞彙。此外，經檢視，有許多負面詞彙被找出，因此除了「不尊重」外，也新增下面負面詞彙：

崩潰、剝削、有病、投訴、壞掉、了無、笑死、好笑、回不去。

Lexicon Based情緒分析結果如下圖，可看到正面詞類型多於負面詞。



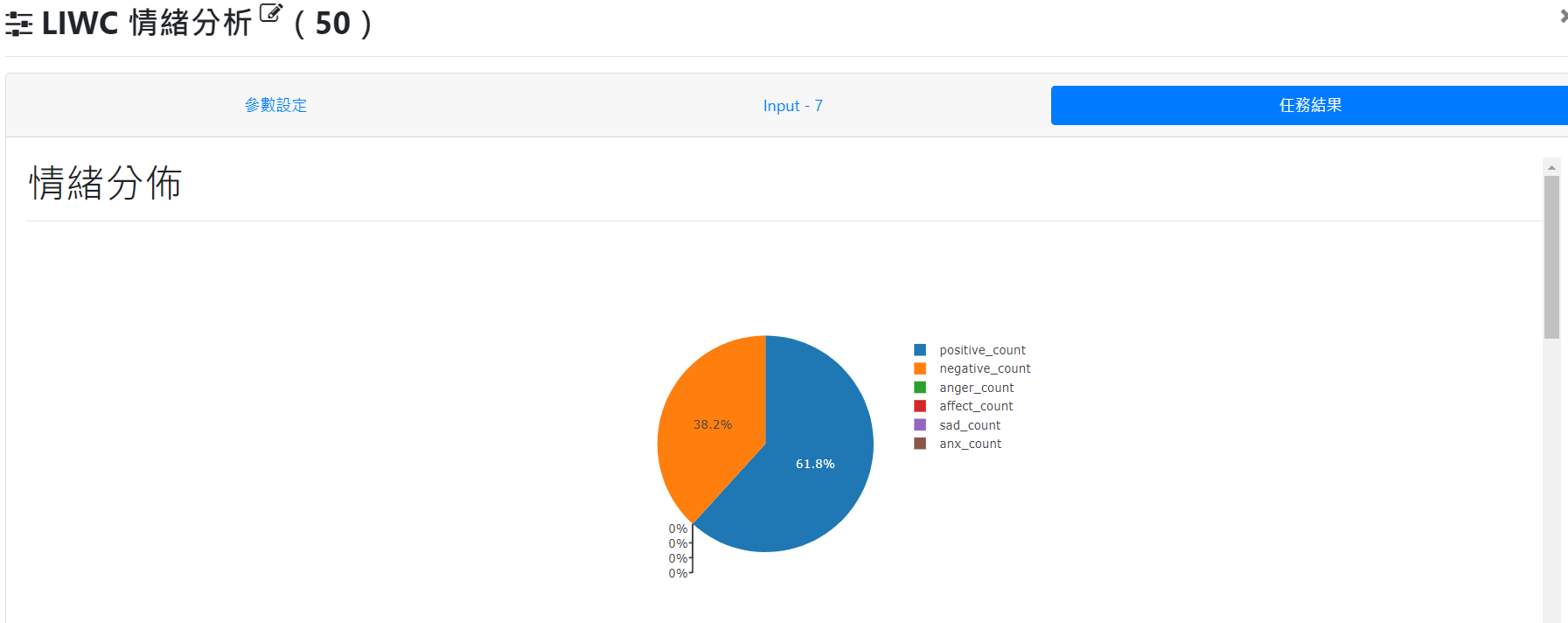
但高頻的正向情緒詞類型數則與負向情緒相當接近。(如下圖)



下面為Lexicon Based情緒分析的輸出結果，可看到負向詞在文章中若重複出現，則將一併列出來計算情緒值。如下圖可看到剝削即警告重複出現。



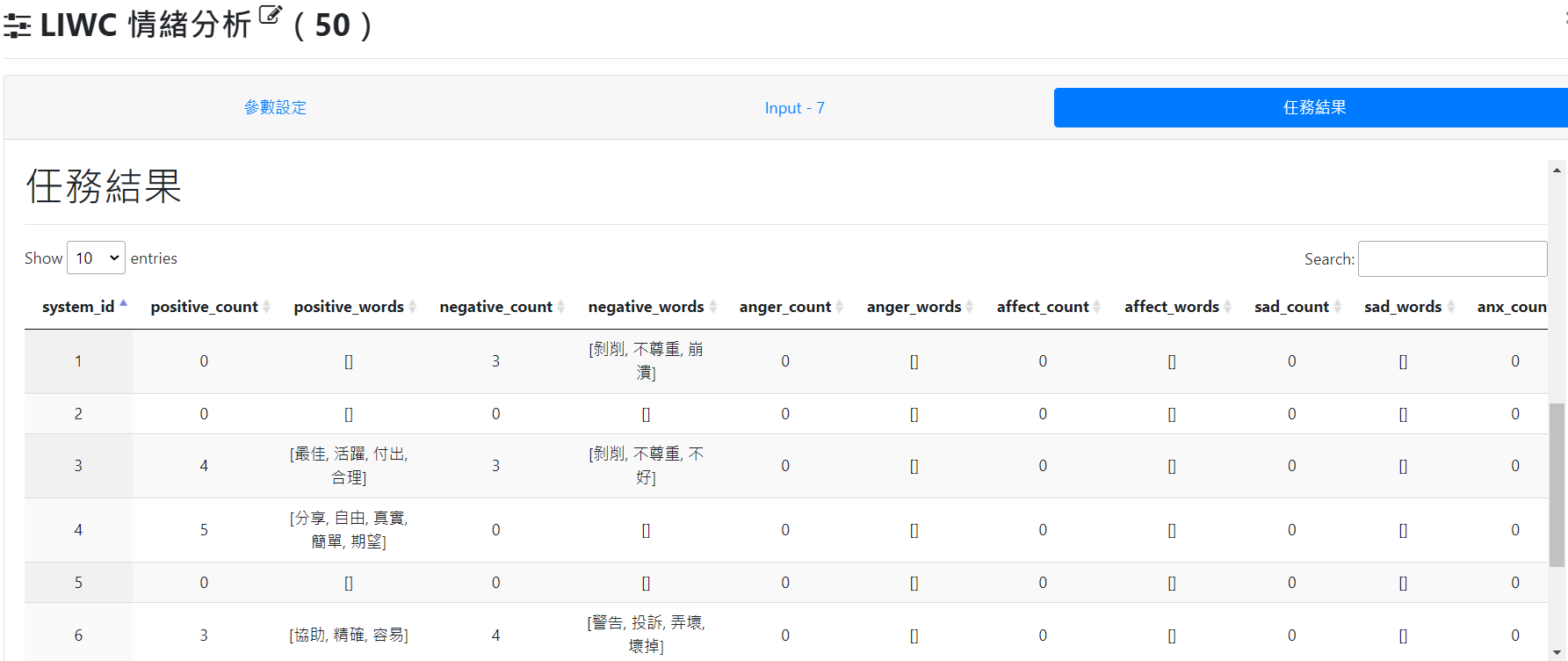
而LIWC情緒分析雖然內建較多情緒類型，但就輸出結果可看到僅有正向詞彙與負向詞彙的比例彙總，其他類型為0。此外，可看到這段期間正向詞的比例(藍色)高於負向詞(橘色)。(如下圖)



另外，下圖可看到LIWC所抓到的最大正向情緒類型數及最大負向情緒類型數均小於Lexicon Based，且Lexicon Based預設字典比LIWC能夠定義出文章中較多的負面情緒字。此外，若情緒詞在發文中有重複出現，則LIWC只會計算一次。

經過考量，擬使用Lexicon Based情緒分析的輸出結果來進行後續的每日情緒總和分析。



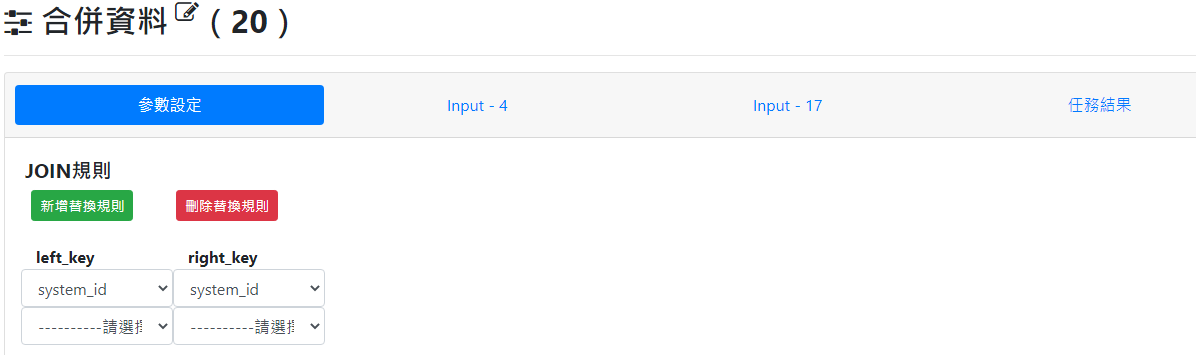


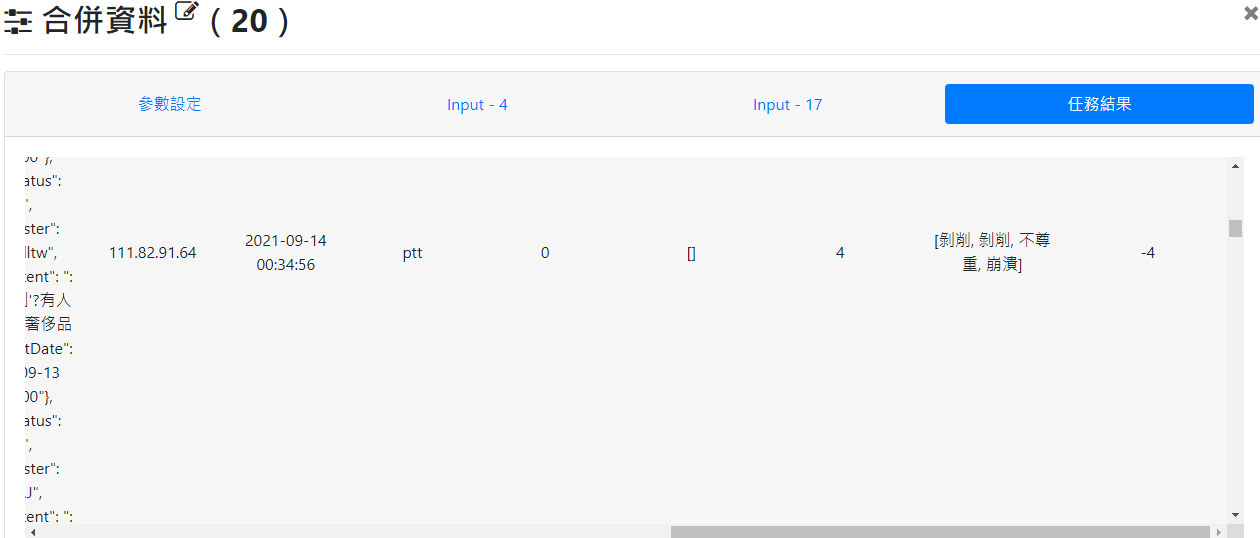
# 五、分群彙總設計

於分群彙總前進行資料合併，將爬蟲資料與Lexicon Based情緒分析資料依system\_id進行合併，以利後續依日期來彙總每日情緒值。

(因為每日爬蟲資料無情緒值資訊，但情緒分析資訊亦無日期欄位，故需進行合併)

* 合併資料

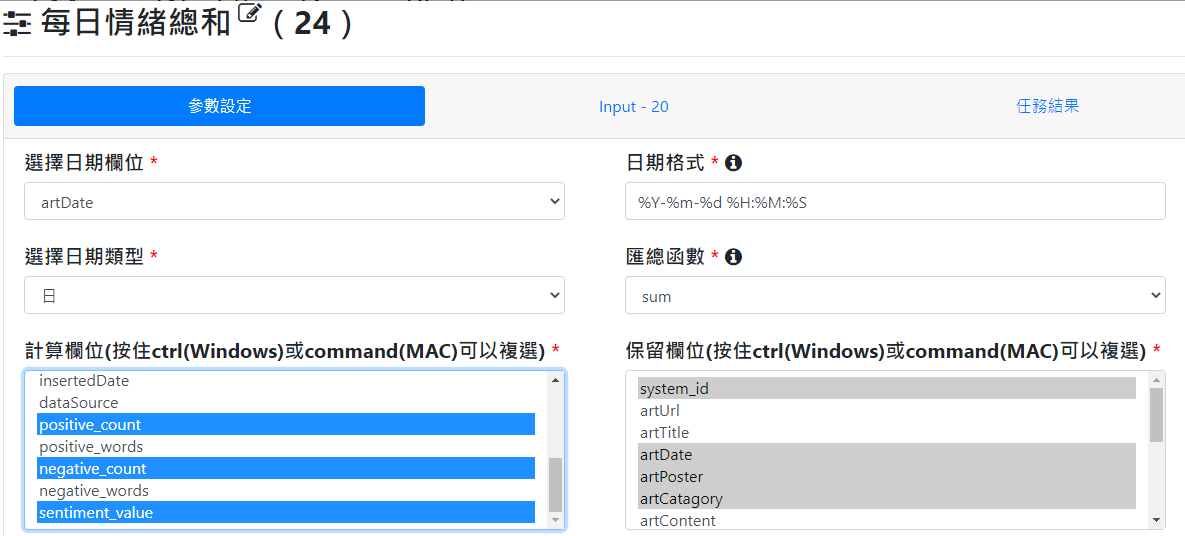




* **每日情緒總和**

使用分群彙總(日期)模組，依照日期選擇正向文字數、負向文字數與情緒值等計算欄位來彙總，並加入PPT板類別、發文人及system\_id等額外參考之保留欄位來產生每日情緒總和的資料。

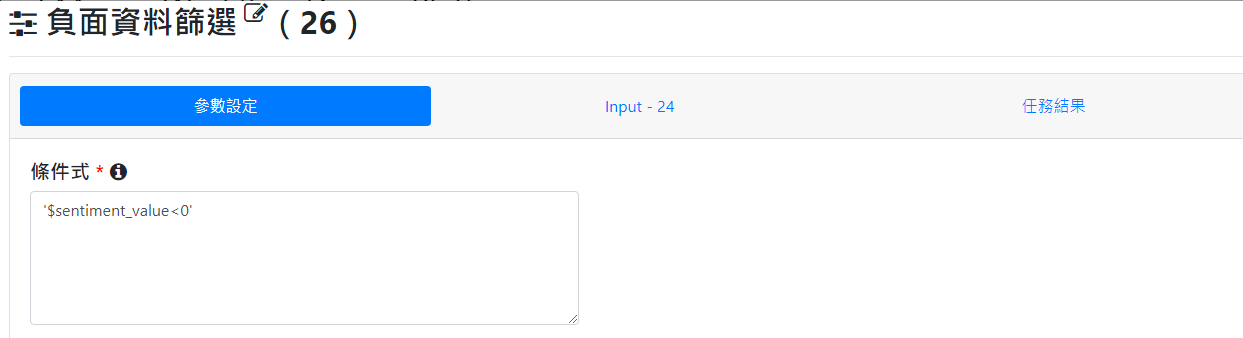
設定如下：

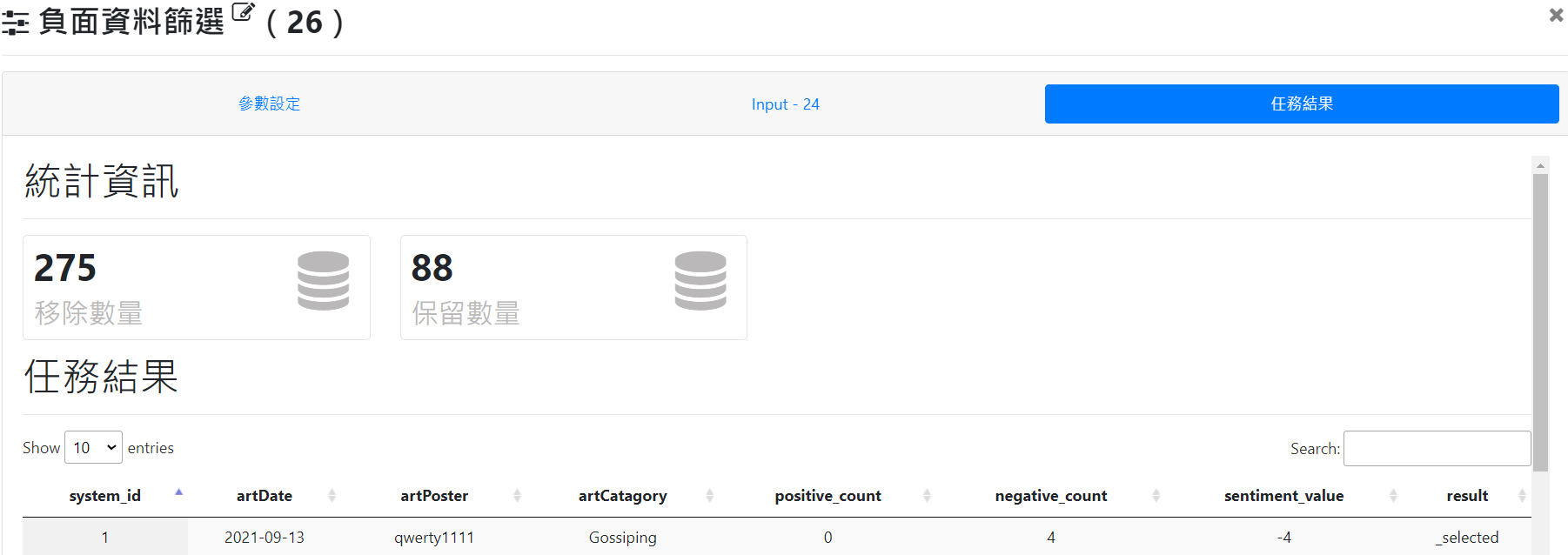




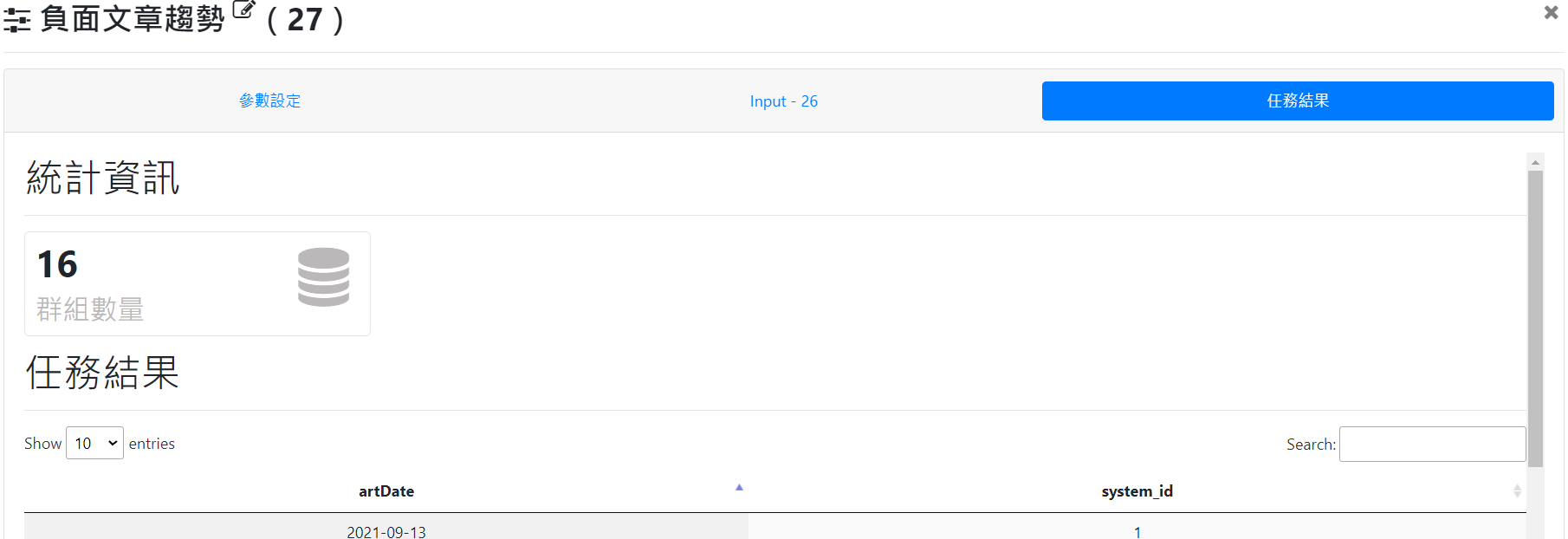
* **彙總負向文章趨勢**

先用資料篩選模組來篩選情緒值小於0的資料。然後用分群彙總(日期)模組，依照日期來計算system\_id筆數，以取得每日負向的文章數。設定如下：





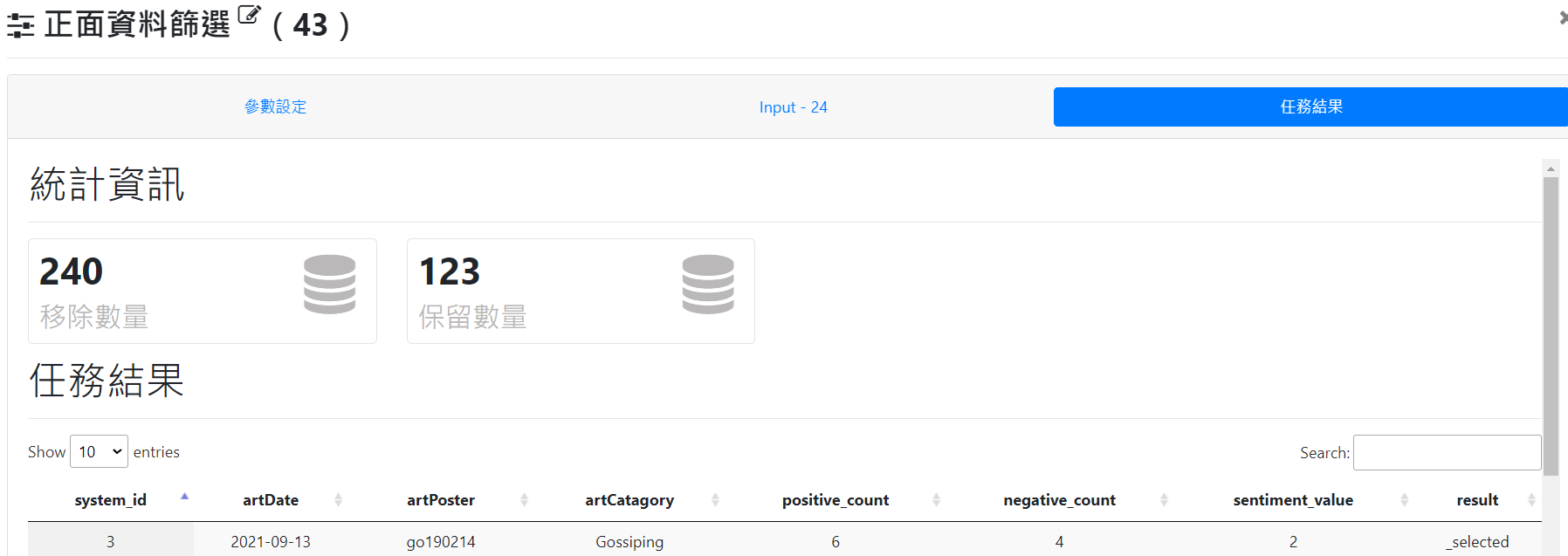




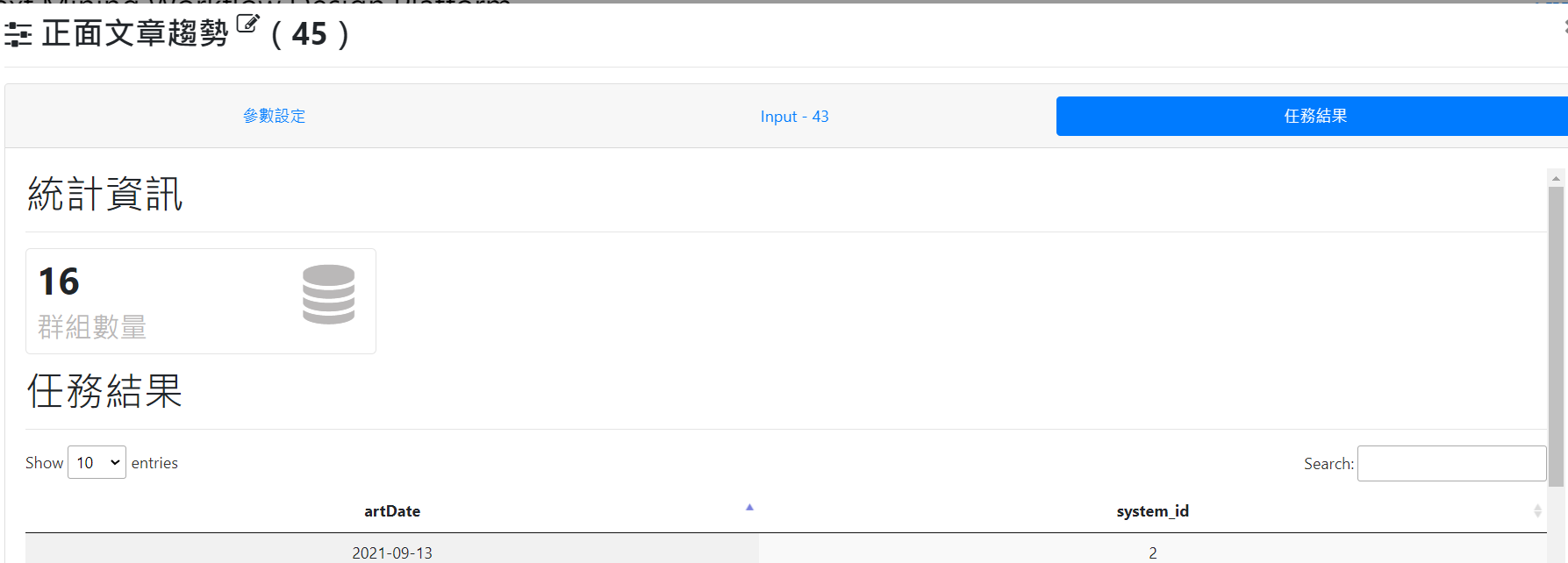
* **彙總正向文章趨勢**

先用資料篩選模組來篩選情緒值大於0的資料。然後用分群彙總(日期)模組，依照日期來計算system\_id筆數，以取得每日負向的文章數。設定如下：



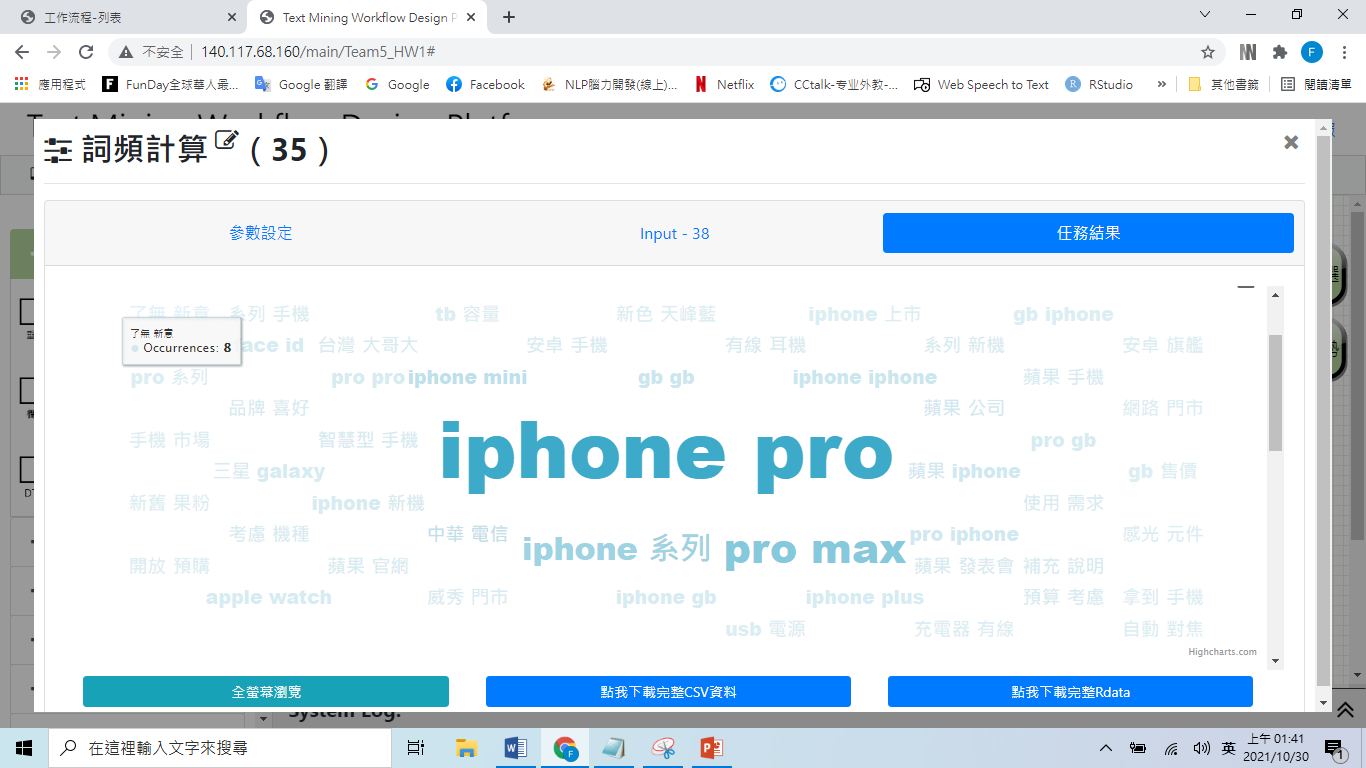






# 六、視覺化圖表呈現

除了前述透過情緒分析模組對情緒詞所產生的視覺化呈現外，另外使用詞頻計算篩選前50名高頻詞組所做的文字雲，不含預期出現的高頻iphone系列詞組，可看到「了無 新意 」和 「新色 天峰藍」均出現在前50名。是本次分析瞭解客戶對iPhone新機主要態度的有利參考資訊。





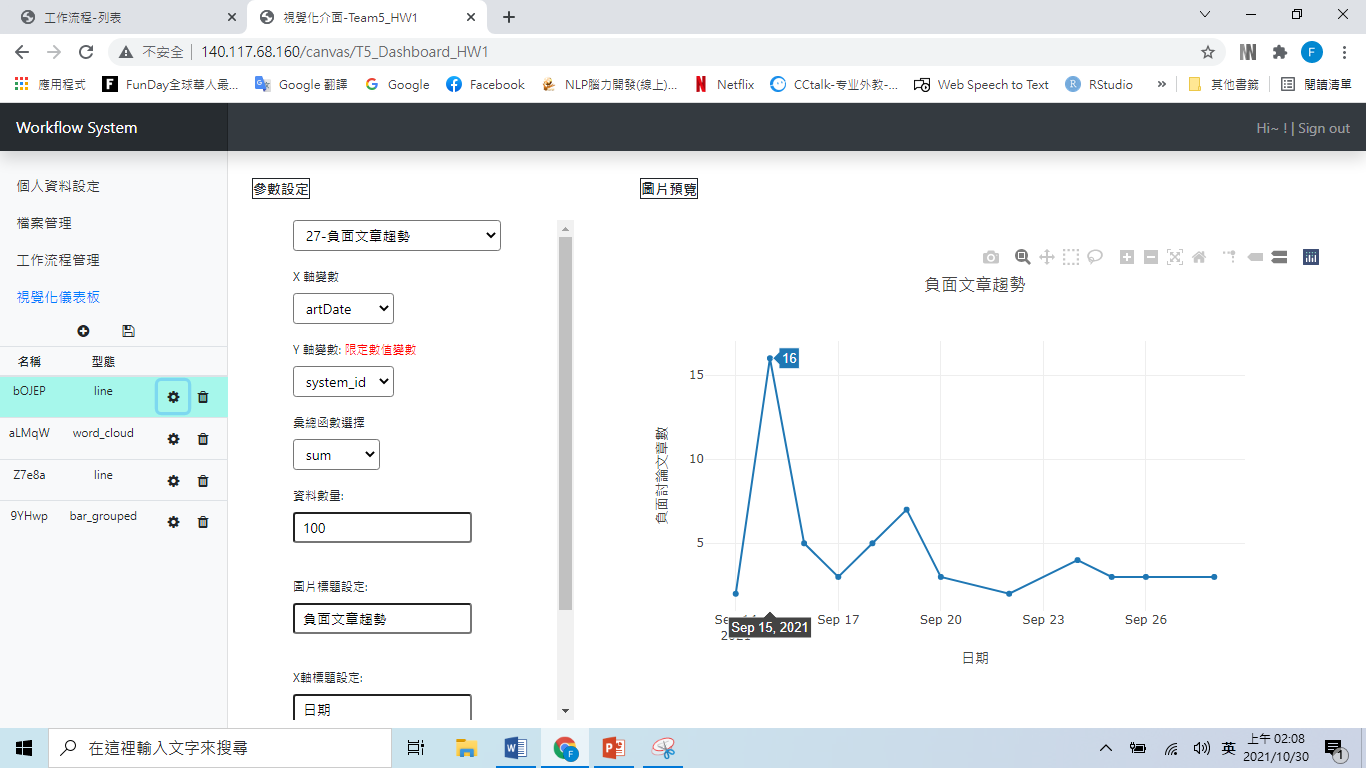
此外，本組亦透過工作流程平台的視覺化儀表板，建立其他視覺化分析圖表，儀表板的名稱為T5\_Dashboard\_HW1。



依序說明建立的儀表板如下：  
(因為未找到修改圖表名稱的功能，故沿用系統建立時的名稱)

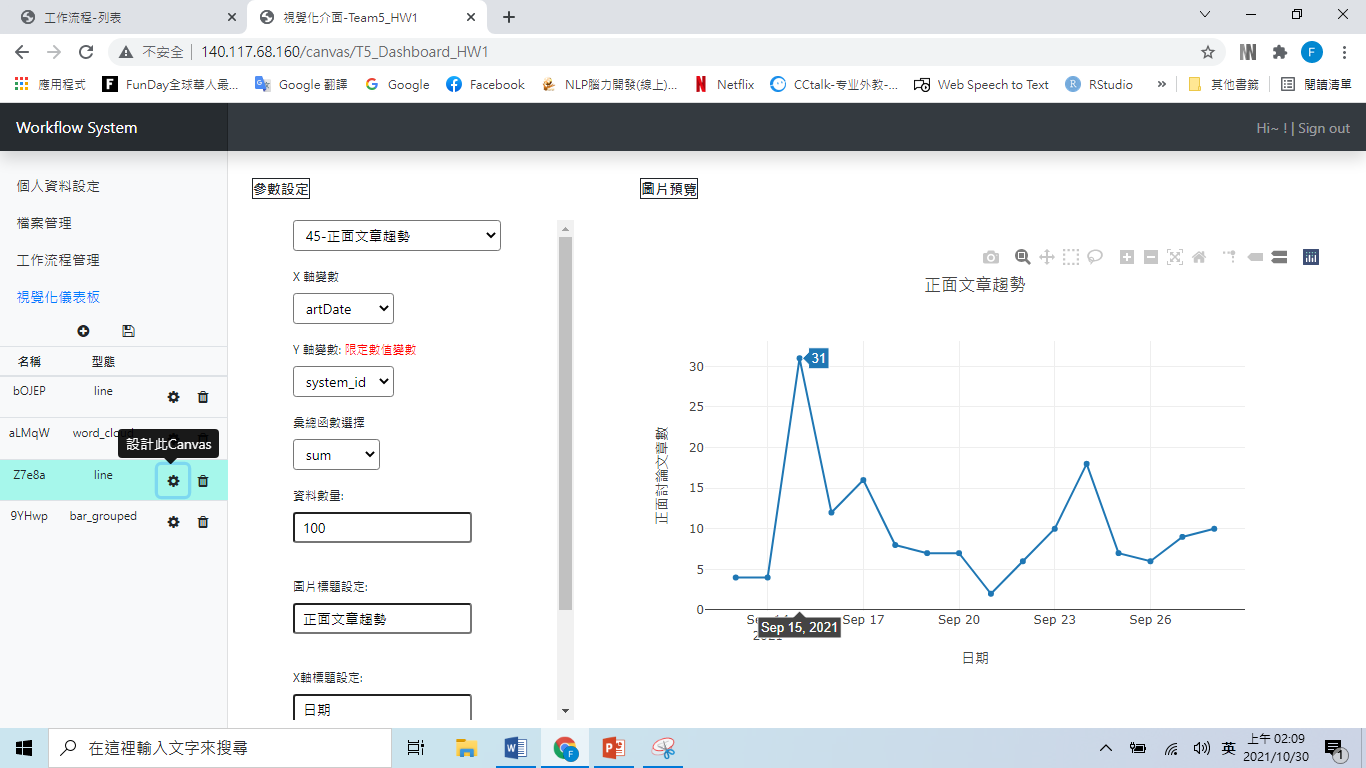
1. **負面文章趨勢的折線圖**

使用資料為負面文章趨勢的彙總資料，可看到高峰是蘋果發表會9/15  
當天，顯示新iPhone發表後，失望人數頗多，之後負面討論開始趨緩。



1. **正面文章趨勢的折線圖**

正面文章同樣在iPhone發表會9/15那天來到高峰，且正面文章數多於負面文章數，顯示同時也有不少人對新iPhone的推出買單。討論熱度之後下降，一直到9/24上市日又再小幅升起。



1. **詞頻計算的文字雲**

同樣使用前面詞頻計算資料產生文字雲，雖然數位儀表板的文字雲可提供較多的顏色變化，但可惜解析度不佳。原本詞頻計算模組內建的文字雲較易於分析辨識高頻詞組資訊。

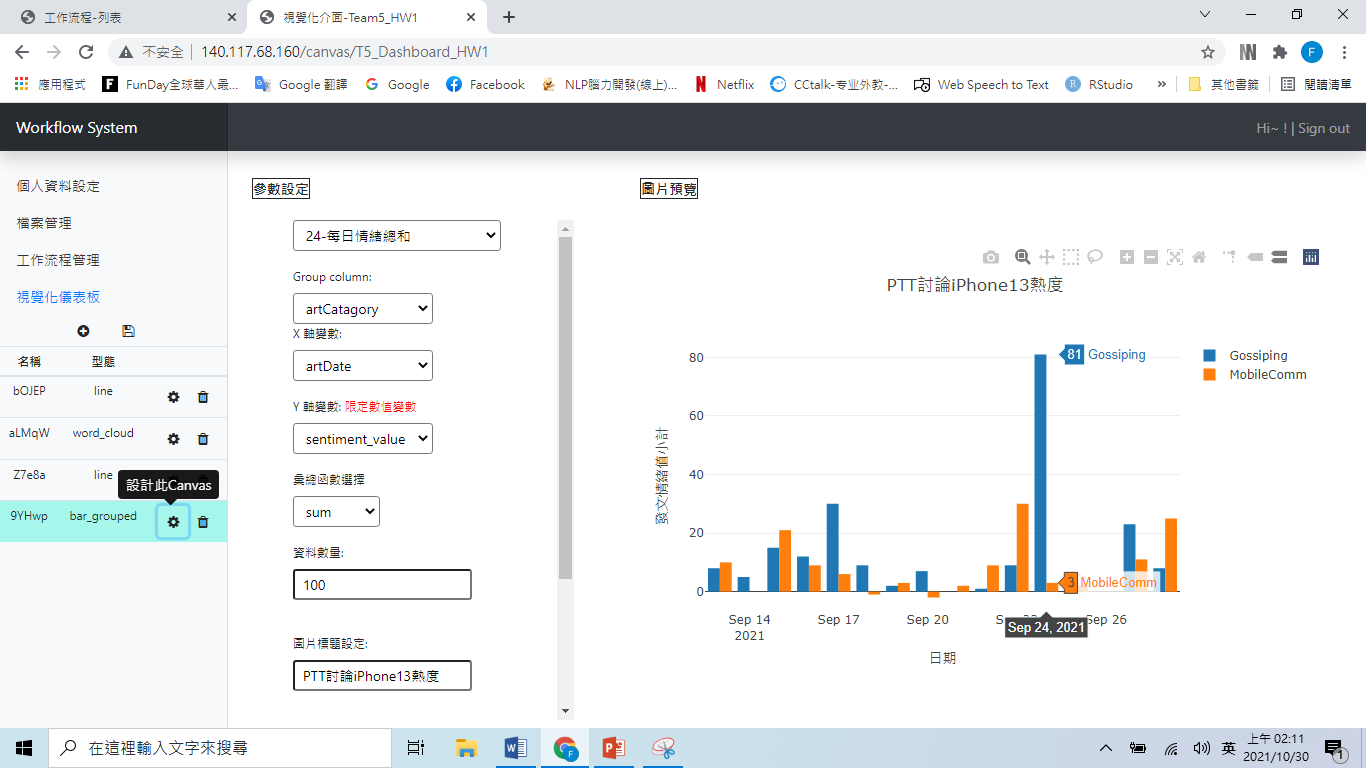


1. **每日情緒值彙總之群組長條圖**

使用每日情緒總和的資料，以PTT討論板類型為群組，以群組長條圖來分別彙總八卦板和手機通訊板的每日情緒值趨勢。

可以發現9/15發表會那天雖然有討論熱度，但卻在實際上市日9/24那天才達到高峰(如下圖二)。此外，iPhone屬於手機，理論上討論應集中在手機通訊板，但可看到八卦板依然是新訊息風向的首要指標。在iPhone上市的9/24當天，情緒討論熱度遠高於手機通訊板。

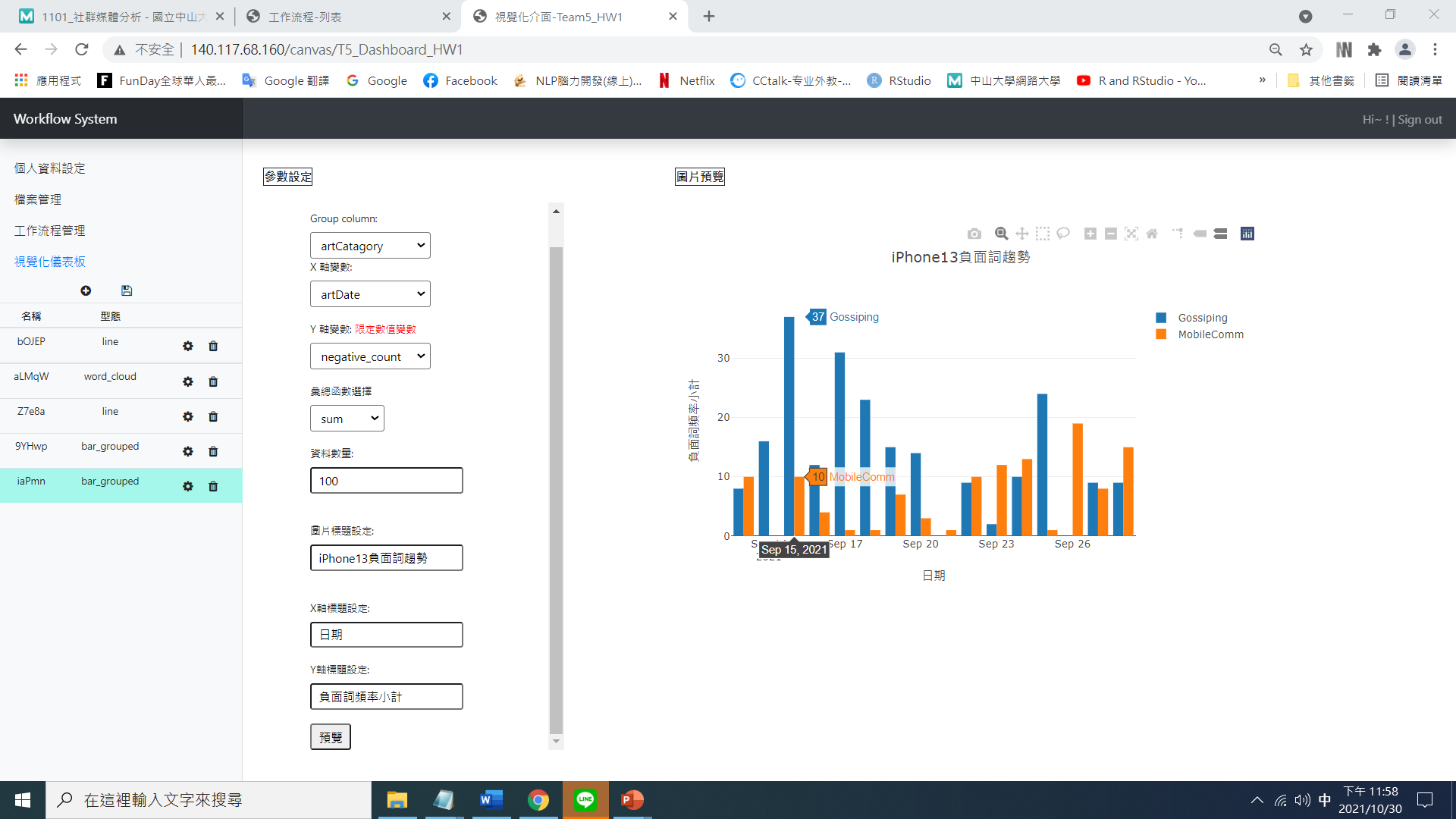




1. **每日負面文字數彙總的群組長條圖**

同樣使用每日情緒總和的資料，以PTT討論板類型為群組，以群組長條圖來分別彙總八卦板和手機通訊板的每日負向文字數的趨勢。

可以看到9/15發表會那天的八卦板文章有關iPhone13的負面詞數目為高峰，之後逐漸下降，直到上市日隔天(9/25)又上升(如下圖二)，然後討論熱度轉移到手機通訊板。

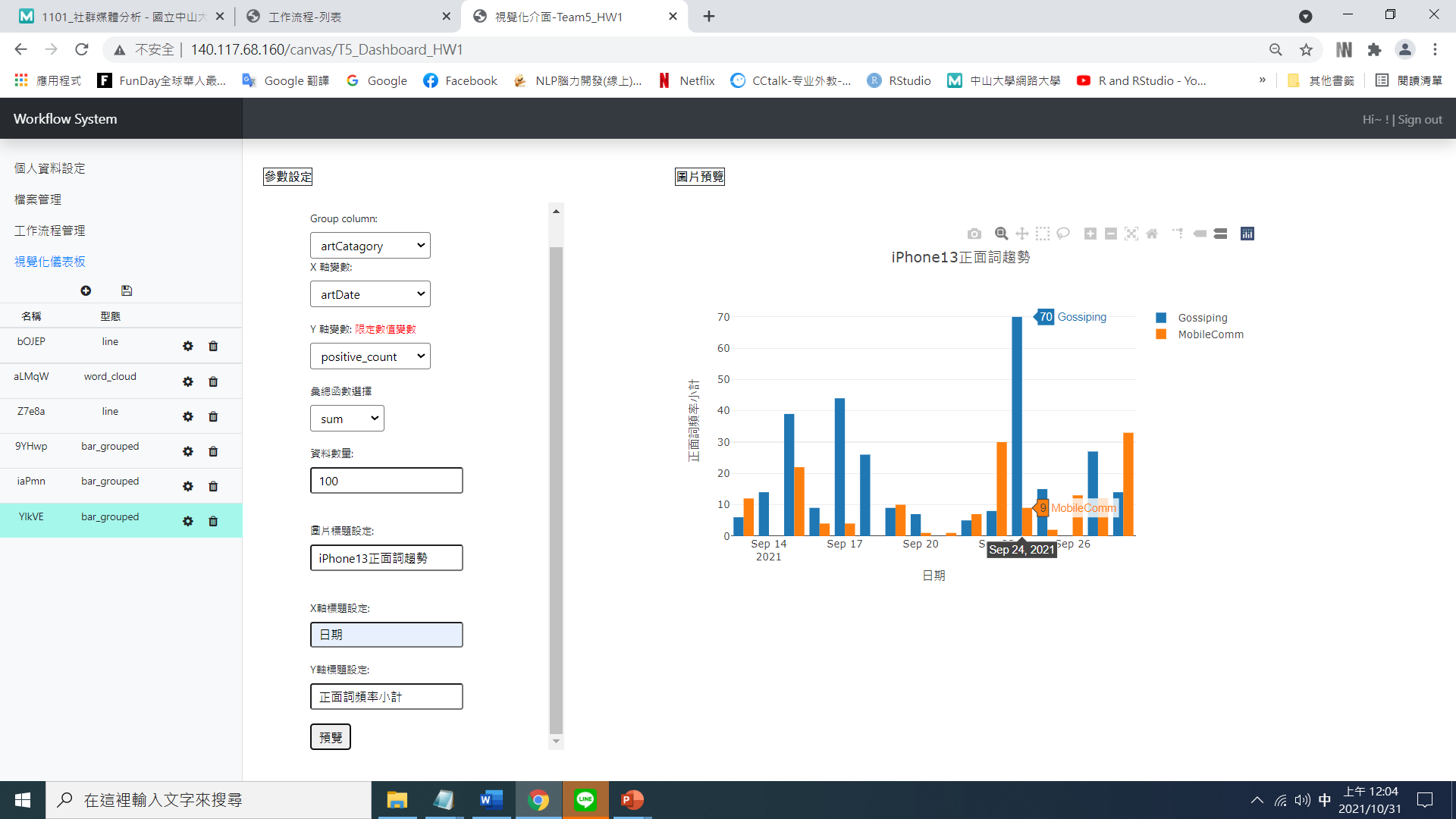
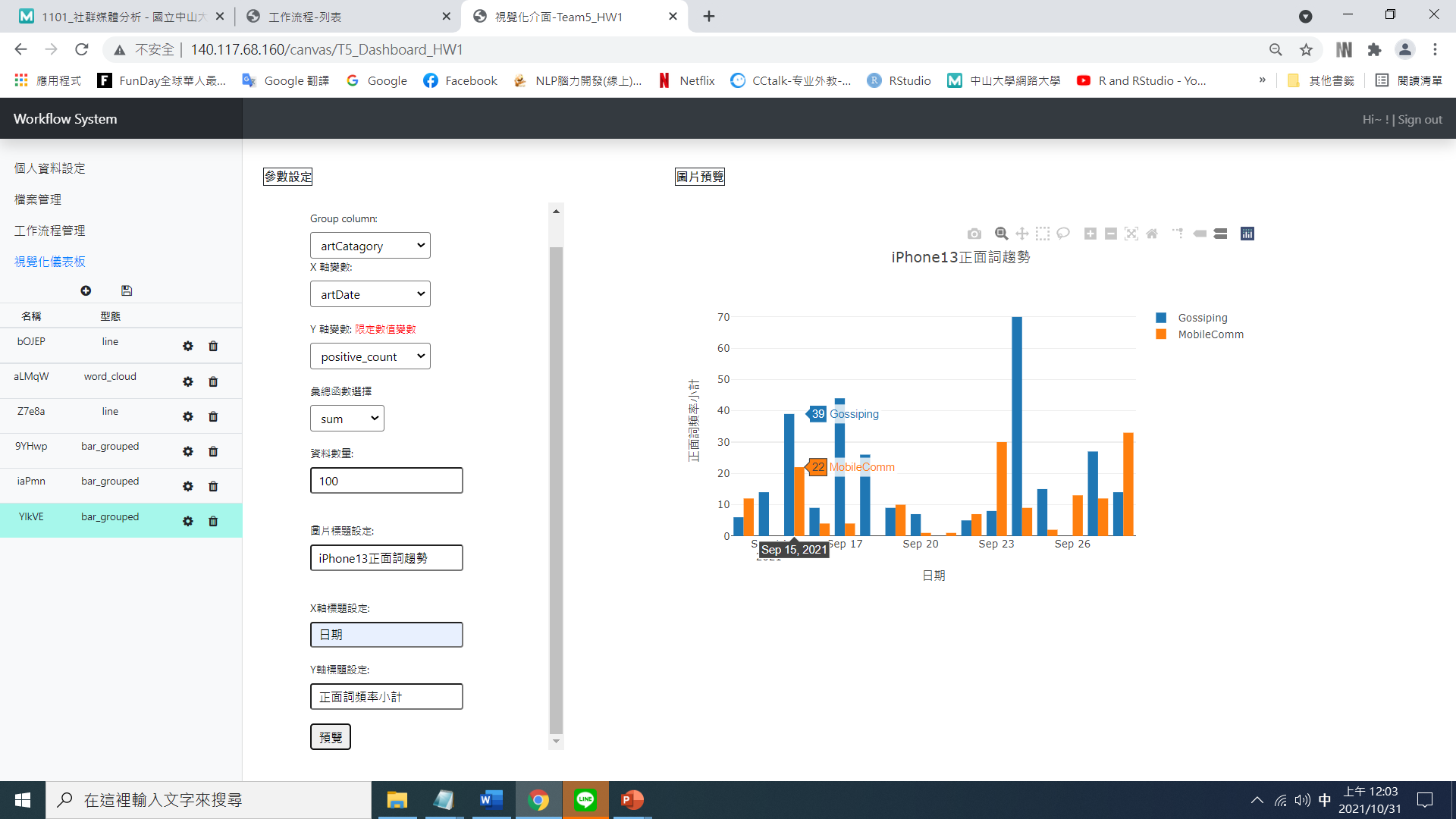




1. **每日正面文字數彙總的群組長條圖**

同樣使用每日情緒總和的資料，以PTT討論板類型為群組，以群組長條圖來分別彙總八卦板和手機通訊板的每日正向文字數的趨勢。

可以看到9/15發表會那天八卦板正向文字數目不是這段期間的高峰，雖然非高峰，但正向文字數仍高於發表會當天的負向文字數。而一直到上市日9/24那天才達到正向文字數的高峰。推測可能有不少人在上市那天拿到新手機後在八卦板發文炫耀。但隔天即討論度下降，討論熱度轉移到手機通訊板。



# 七、結論

* 情緒詞頻率出現最高的時間如同預想是出現在發表日及上市日。
* 八卦板依舊是風向首要指標，在iPhone13上市首日討論熱度達到高峰。然後討論熱度逐漸轉向手機通訊板，不再於八卦板討論。
* 整體來說，正向情緒詞的比例多於負向情緒詞。
* 負面詞在發表日達到高峰，但正面詞卻在上市日達到高峰。
* 文字雲的詞組頻率中，「了無 新意」 進入前50名，顯示很多人認為iPhone13沒有亮點。
* 「新色 天峰藍」 也進入文字雲前50名，顯示有許多人對iPhone這個新色有興趣。