# **社群媒體分析**

# **第二次讀書會報告\_第十二組**

## **一、分析主題**

探討東森新聞記者撰寫鄰近東亞國家(中國、韓國、日本、香港、新加坡、馬來西亞)相關新聞之情緒分析及新聞類別預測。

## **二、組員**

N114320002\_李湘怡

N114320008\_陳家卉

N114320009\_林育伈

N114320017\_黃敏喩

N114320020\_曾子瑋

N114320027\_吳政翰

N114320030\_吳郁文

N114320031\_洪宜綾

## **三、分析工具**

* 中山大學工作流程平台：Tarflow
* 工作流程名稱：第二次讀書會
* 情緒分析字典：LIWC、進階字典法、Lexicon Based
* 視覺化圖表：字詞網路圖、單中心網路圖
* 分類器：Decision tree、Logistic regression、SVC、KNN

## **四、動機目的**

小組從東森新聞撰寫中收集與中港日韓新馬地區相關的政治、社會、國際和旅遊方面的新聞文章。目標為進行情緒分析和新聞類別預測，想要獲取台灣鄰近且往來密切的國家新聞動態和情緒資訊。

利用資料庫爬蟲將從東森新聞撰寫中檢索新聞文章，了解文章中表達的情緒趨勢(正面、負面或中立)。同時使用機器學習分類器技術，便於預測新聞文章類別歸屬。

最終目的為幫助理解中港日韓新馬地區的新聞動態和情緒趨勢，從而獲得有價值的信息和洞察，關切台灣鄰近國家的政治、社會、旅遊等國際動態對我們的影響。

## **五、資料來源**

* 資料來源：東森新聞資料庫之政治、社會、旅遊、國際看板。
* 關鍵字：中國、韓國、日本、香港、新加坡、馬來西亞
* 搜尋日期：2023/04/06~2024/04/06
* 資料筆數：4387筆

****

## 

## 

## **六、系統流程圖**

6-1.資料前處理：以東森新聞爬蟲資料庫，進行資料觀察和前處理(斷詞、停用詞)。

6-2.情緒分析：以LIWC、進階字典法、Lexicon Based三種方式分析新聞內容情緒，再合併RAW DATA進行欄位篩選與分群，產生出分析結果。

6-3.將資料轉為DTM進行相關性計算，以產生視覺化圖表(字詞網路圖、單中心網路圖)。

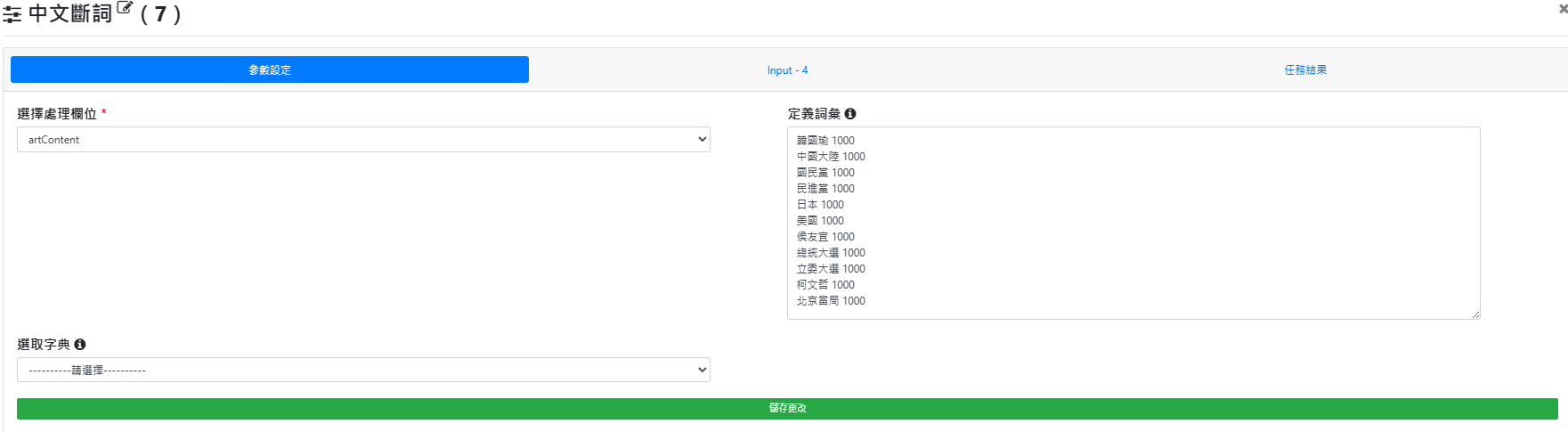
6-4.將資料進Tf-idf詞頻計算後做資料訓練與測試，最後使用機器學習(Decision tree、Logistic regression、SVC、KNN)建立分類器預測模型。

## 

## **七、資料預處理、清除停用字**

賦予專有名詞權重提升中文斷詞的精準度

"韓國瑜"、"中國大陸" 、"國民黨"、"民進黨"、 "日本"、 "美國"、 "侯友宜"、 "總統大選"、 "立委大選" 、"柯文哲"、 "北京當局"。



參數設定：自定義停用字45個

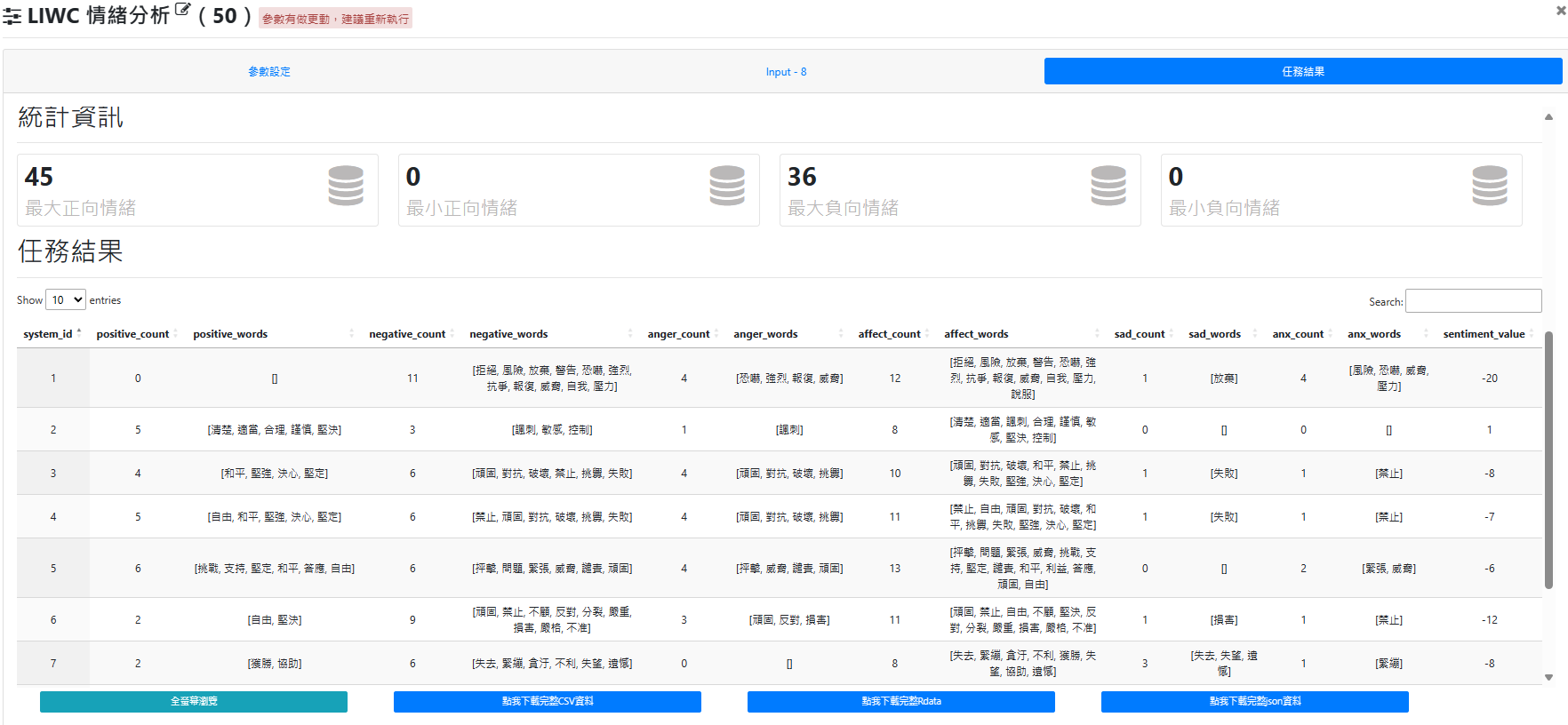
## 

將以下文字設為停用字："一名"、"男子"、"女子”、”關係”、”當個”、”朋友”、”報導"、"網友"、"紛紛"、"文章"、"來源"、"表示"、"一下"、"立即"、"立即一下"、"命定"、"有限公司"、"過去"、"重要"、"行為"、"代表"、"無法"、"對此"、"之後"、"目前"、"現在"、"指出"、"成為"、"知道"、"最後"、"今天"、"發生"、"導致"、"包括"、"決定"、"造成"、"最後"、"進行"、"網路"、"傳出"、"超過"、”下午”、”早上”、”中午”、”晚上”

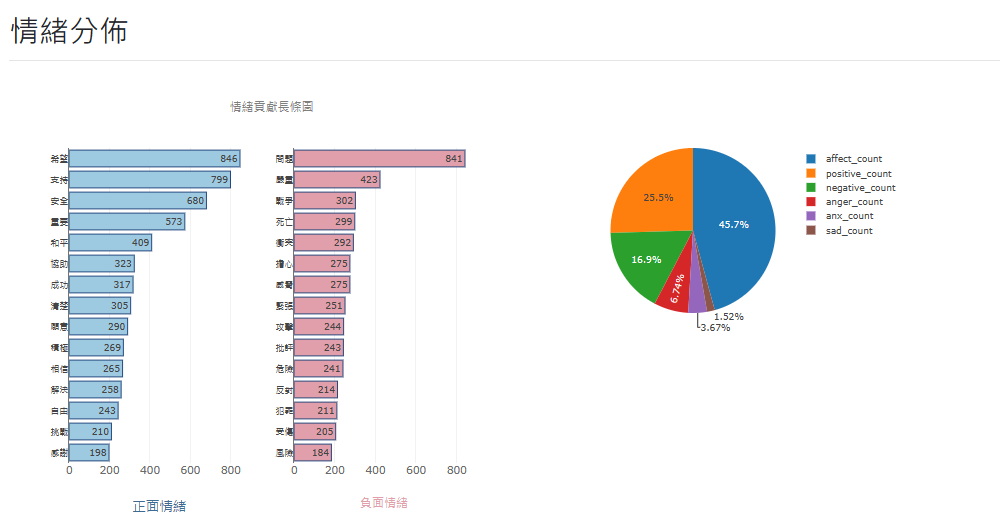
## **八、情緒分析**

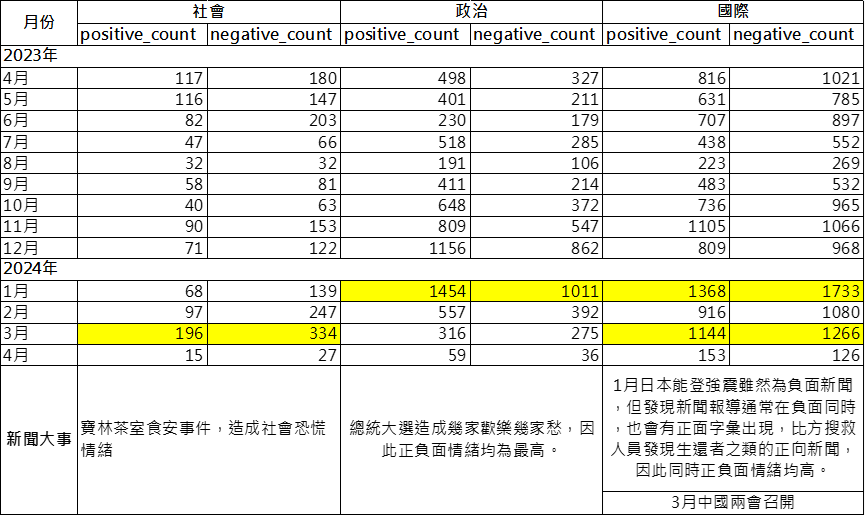
### 8-1 LIWC

透過LIWC字典將新聞情緒分析如下，單篇文章內含有45個最大正向情緒，36個最大負向情緒。



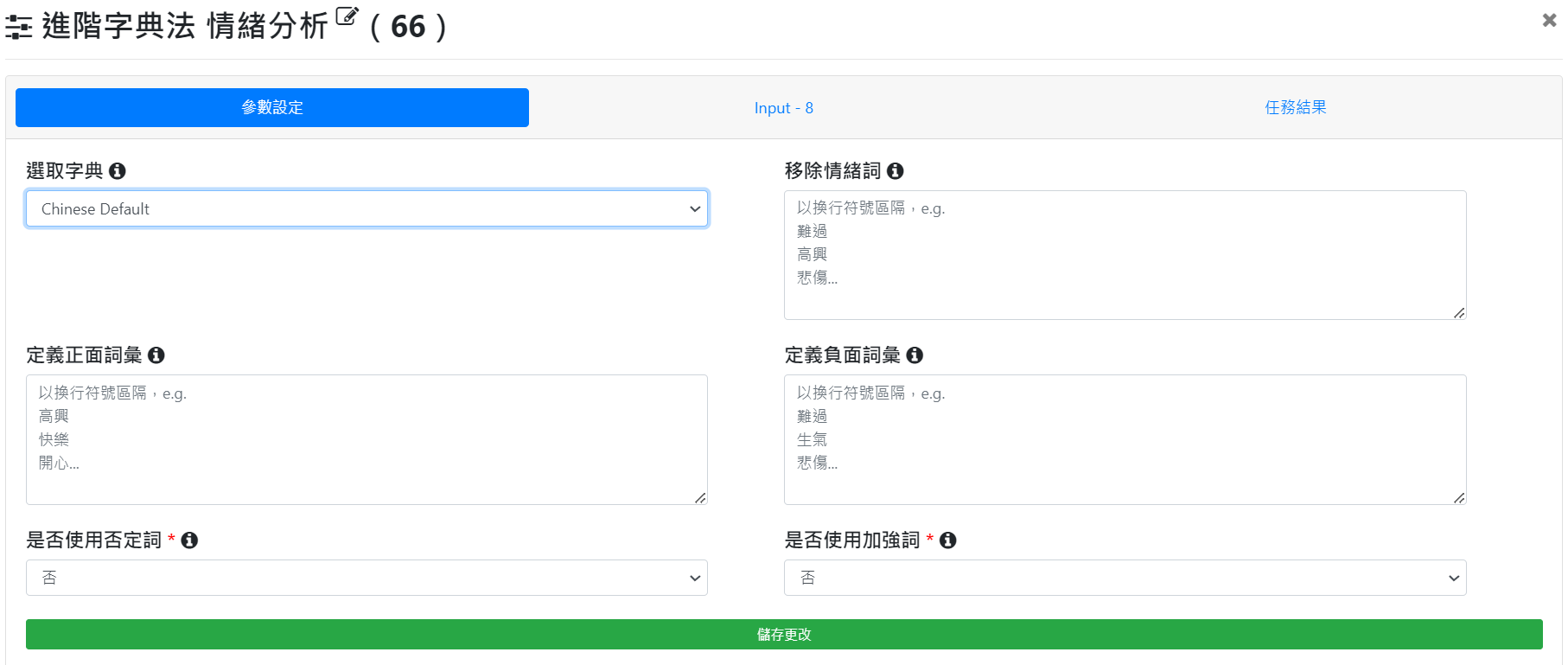
將情緒分布透過圖表顯示如下，正向情緒當中以"希望"、"支持"、"安全" 為前三多的情緒字，負面情緒當中則以"問題"、”嚴重”、"戰爭"為前三多的情緒字。





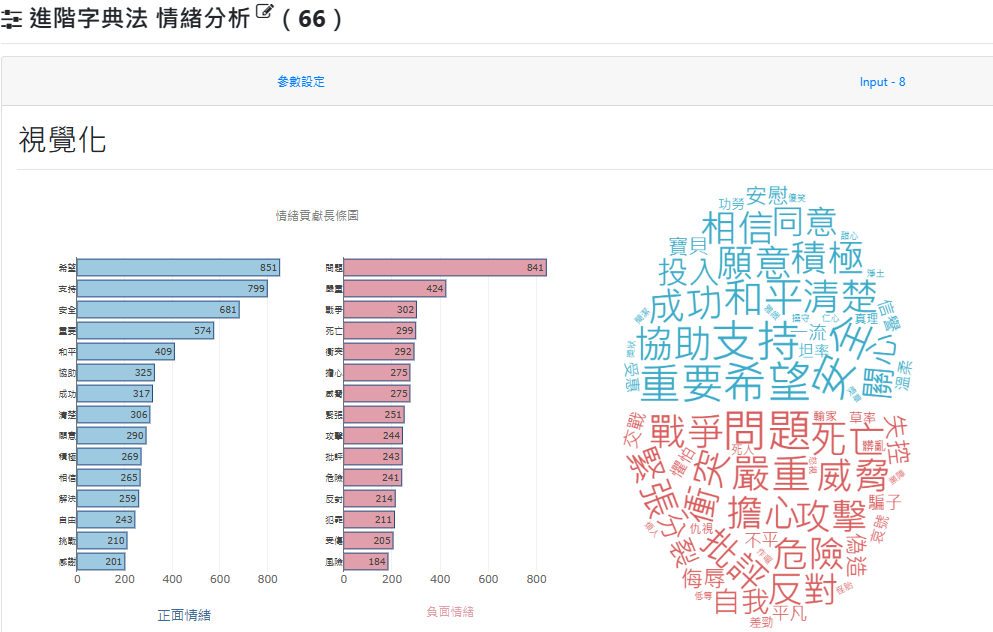
### 8-2進階字典法

**8-2-1 參數設定：**選擇Chinese Default 作為情緒字典。

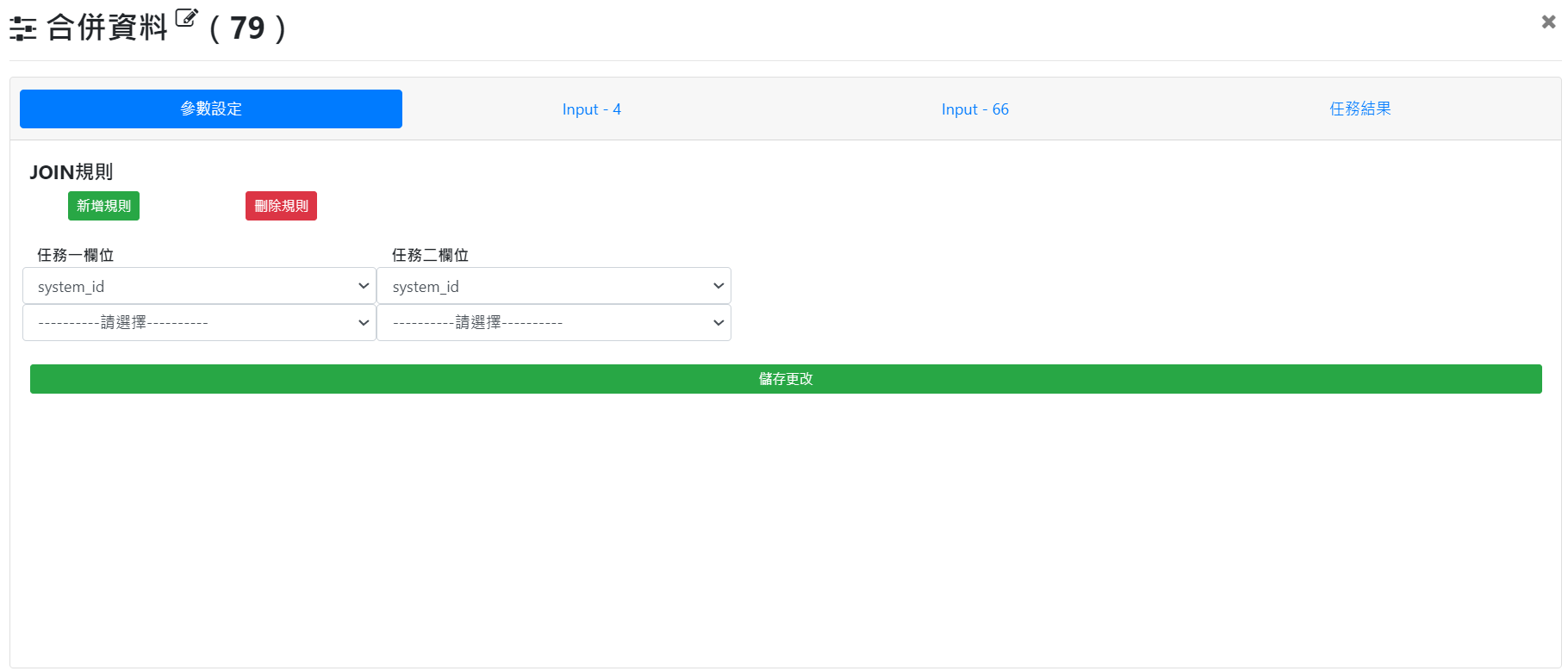


**8-2-2 任務結果 ：**最大正向情緒有46個，最大負向情緒有36個。

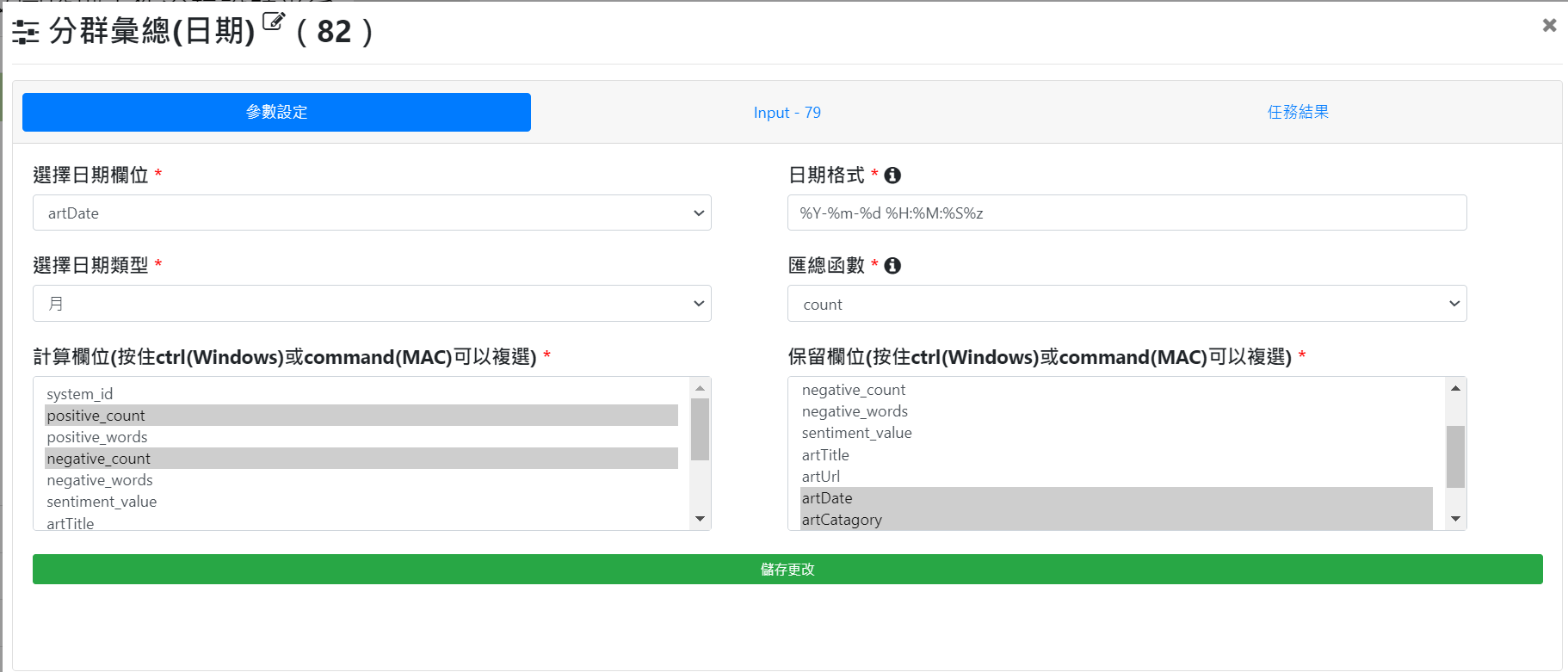




**8-2-3合併資料：**因要進行正負面值情緒分析,利用已分析出來的進階情緒字典分析與 東森新聞爬蟲 來源，以 system\_id 欄位做為 key 值，進行 join。



**8-2-4分群彙總(日期)：**因要進行分群分析,利用合併欄位資料之positive\_count和negative\_count進行計算。

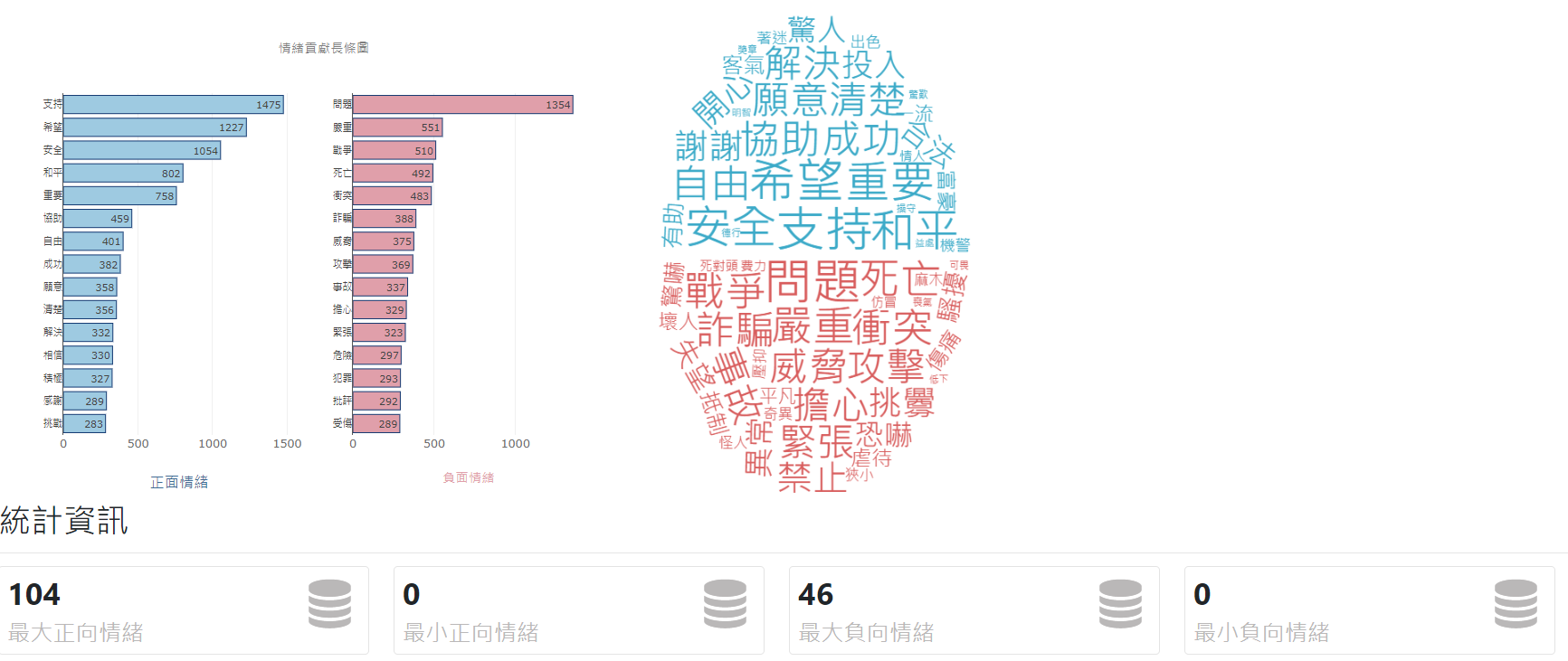


**8-2-5任務結果：**將計算出來的正負情緒值之月group\_by後得到正負值的sum值。



### 8-3Lexicon Based

**8-3-1參數設定：**選擇 chinese，並載入預設情緒字典





**8-3-2任務結果：**最大正向情緒有104個，最大負向情緒有46個。



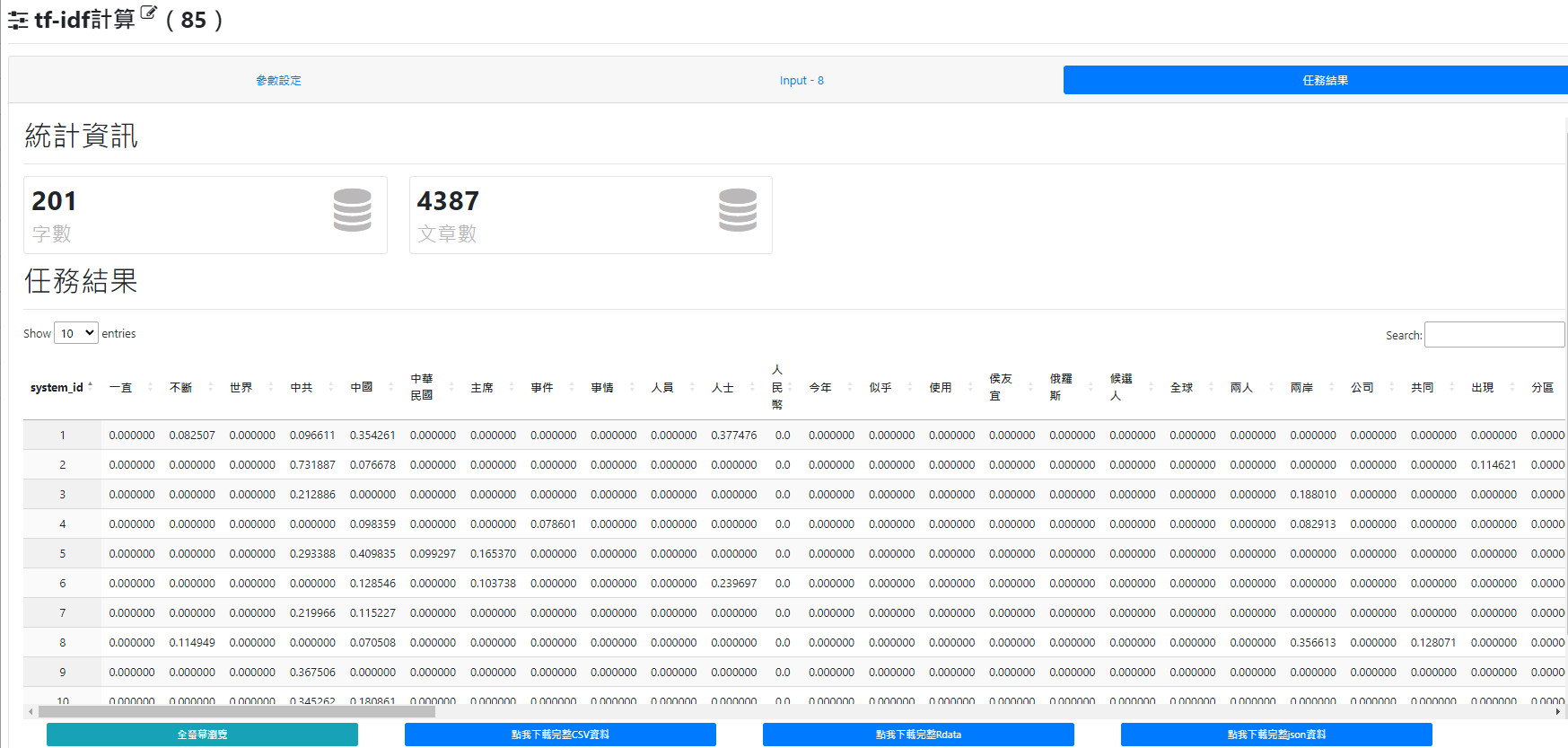
**8-3-3合併資料：**因要進行正負面值情緒分析,利用已分析出來的Lexicon Based 情緒分析與 東森新聞爬蟲 來源，以 system\_id 欄位做為 key 值，進行 join。

## 

## **九、Tf-idf**

## 

篩選詞彙數量為200。



在4387篇文章中挑出201詞進行Tf-idf比重計算。

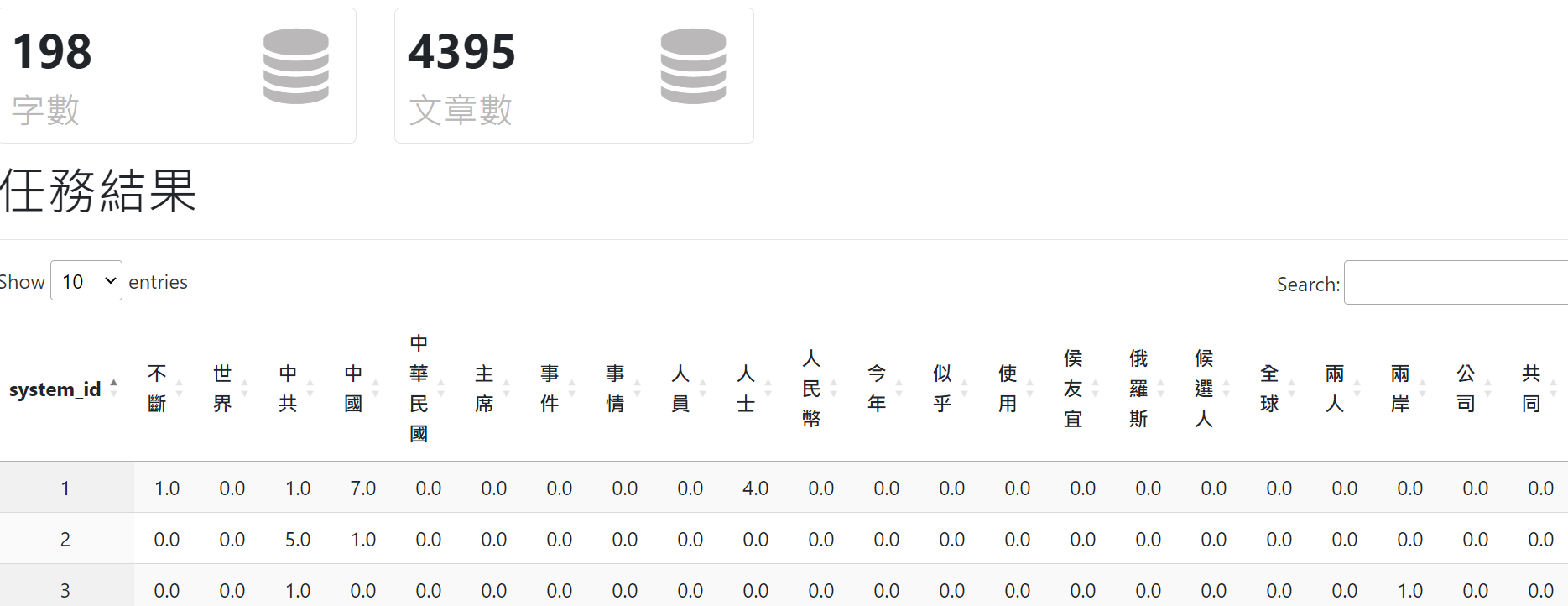
## **十、計算相關性矩陣及字詞網路圖(Shiny)**

**12.1 轉為DTM**

* 參數設定：保留200個詞彙數量

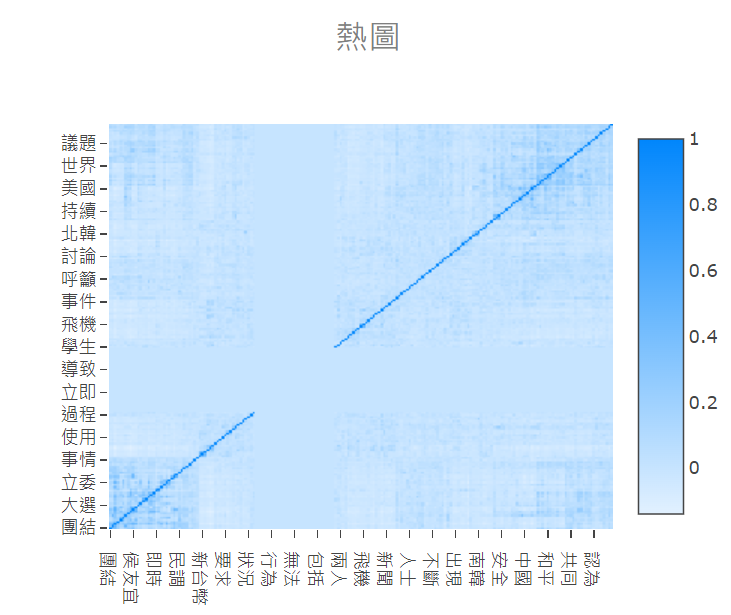
****

* 任務結果：4395文章數



**12.2 計算相關性矩陣**

* 任務結果

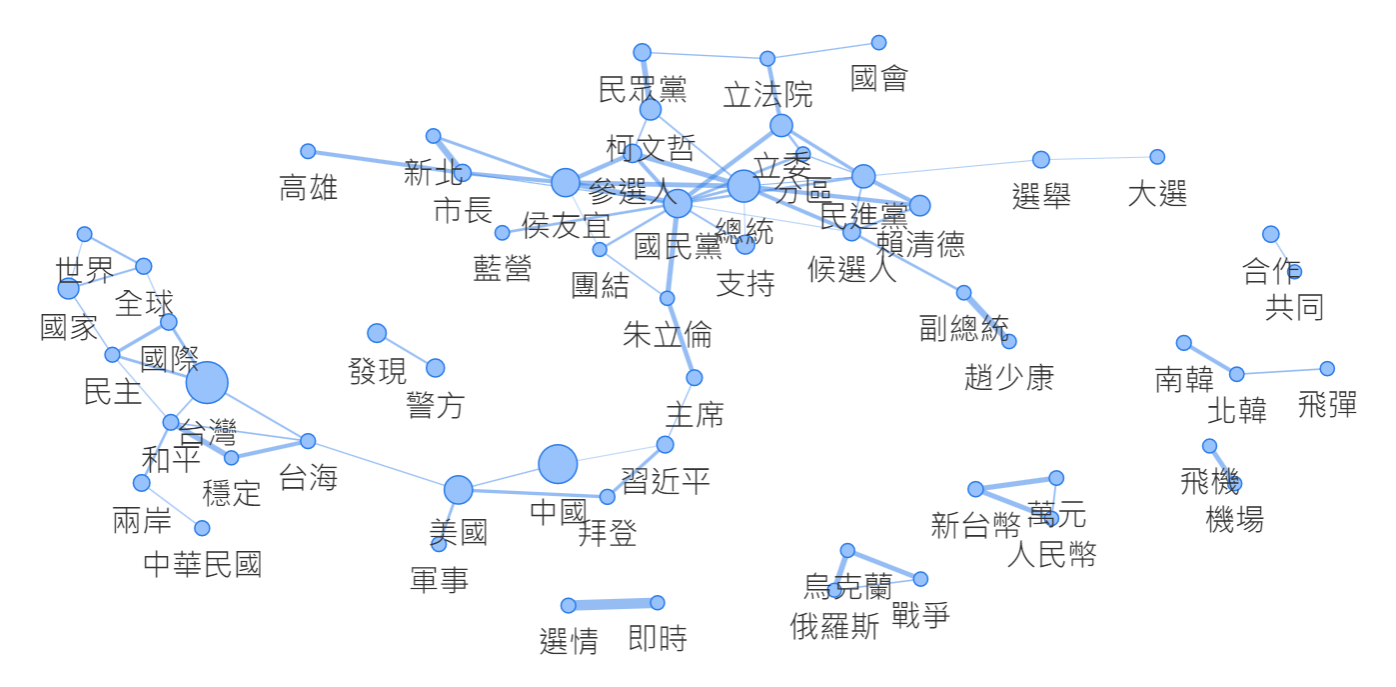




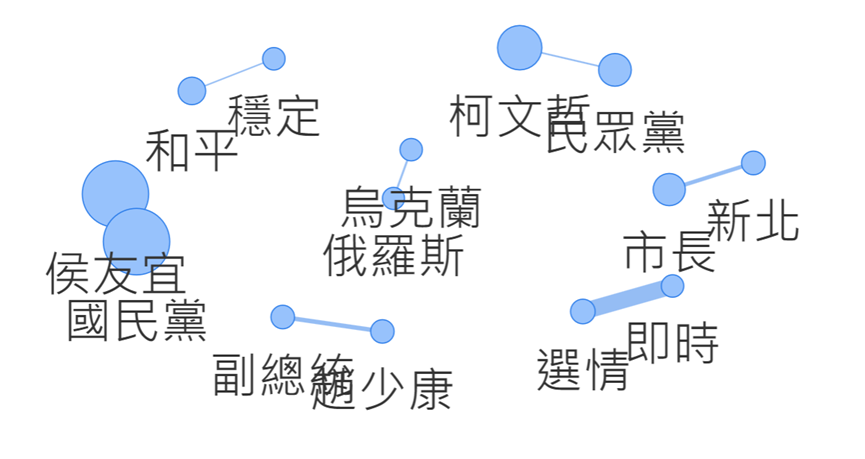
**12.3 字詞網路圖(Shiny)**

* 任務結果：

詞彙關聯度範圍0.3~1



詞彙關聯度範圍0.5~1



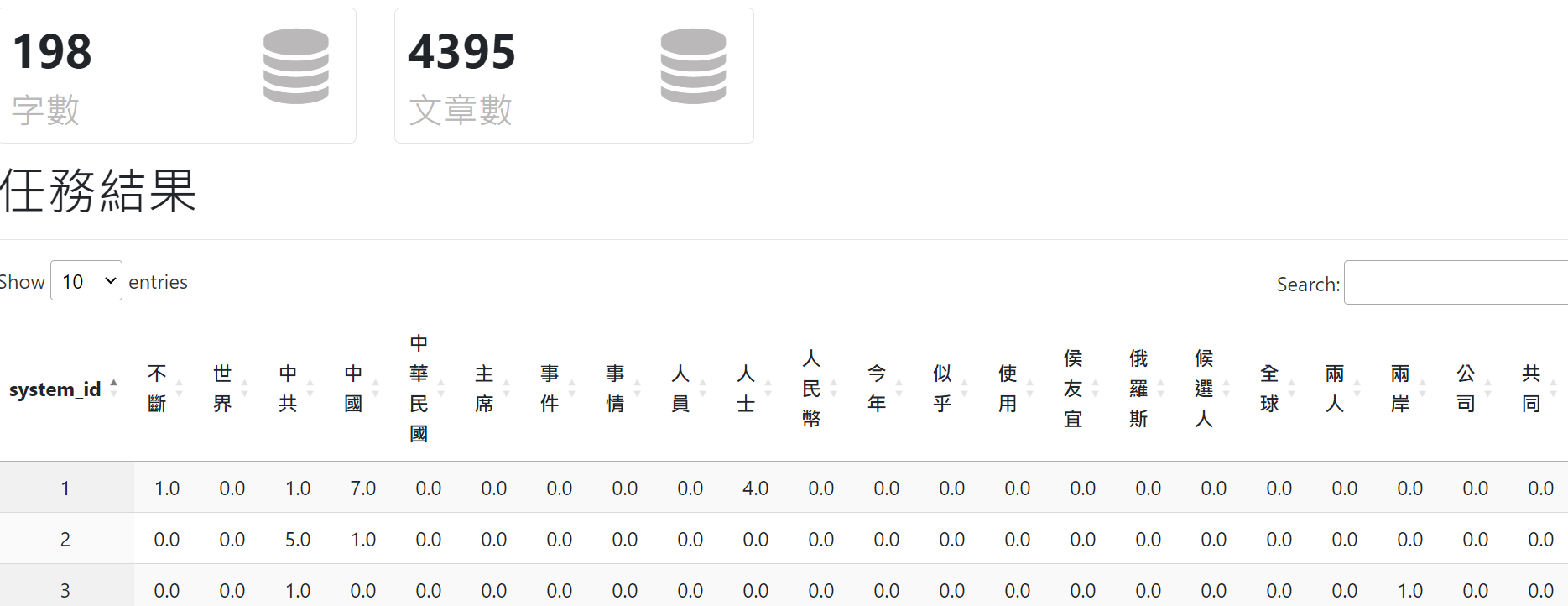
## **十一、計算共現矩陣及單中心網路圖(Shiny)**

**13.1 轉為DTM**

* 參數設定：保留200個詞彙數量

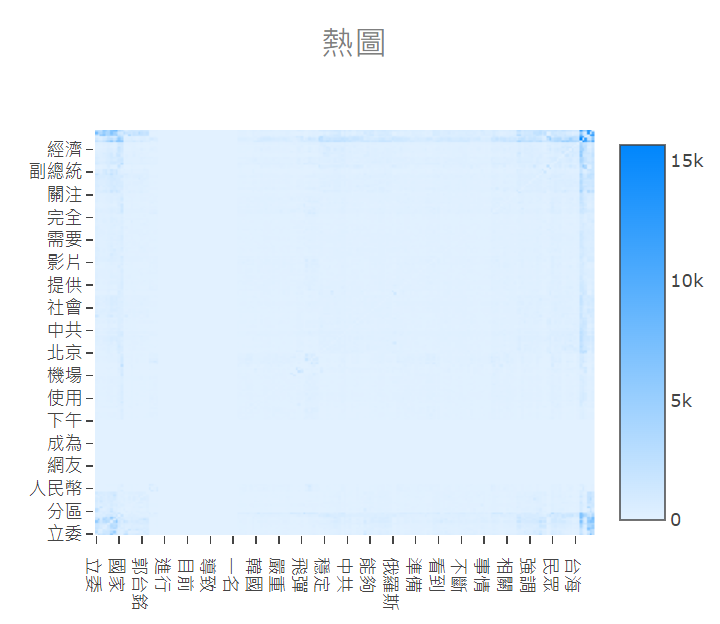
****

* 任務結果：4395文章數



**13.2 計算共現矩陣**

* 任務結果：





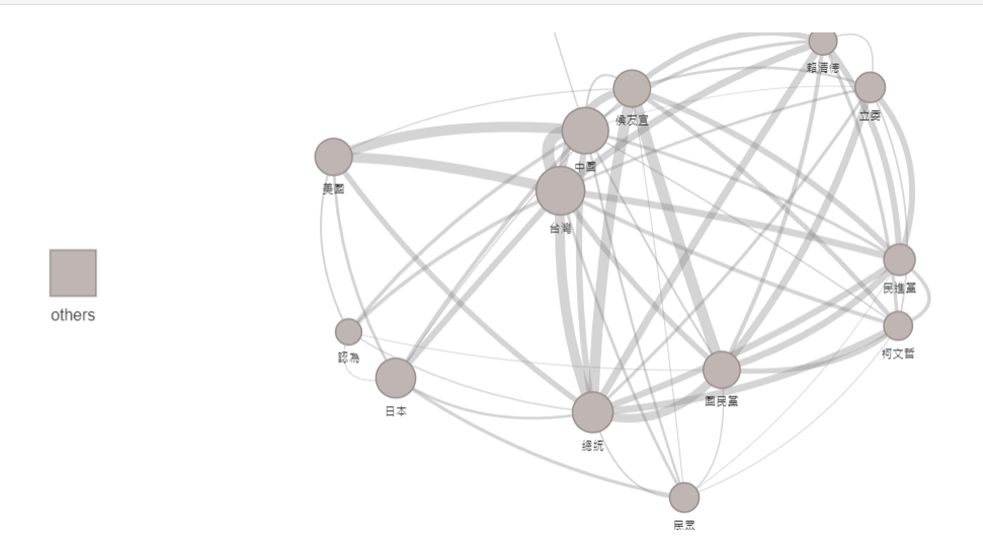
**13.3 單中心網路圖(Shiny)**

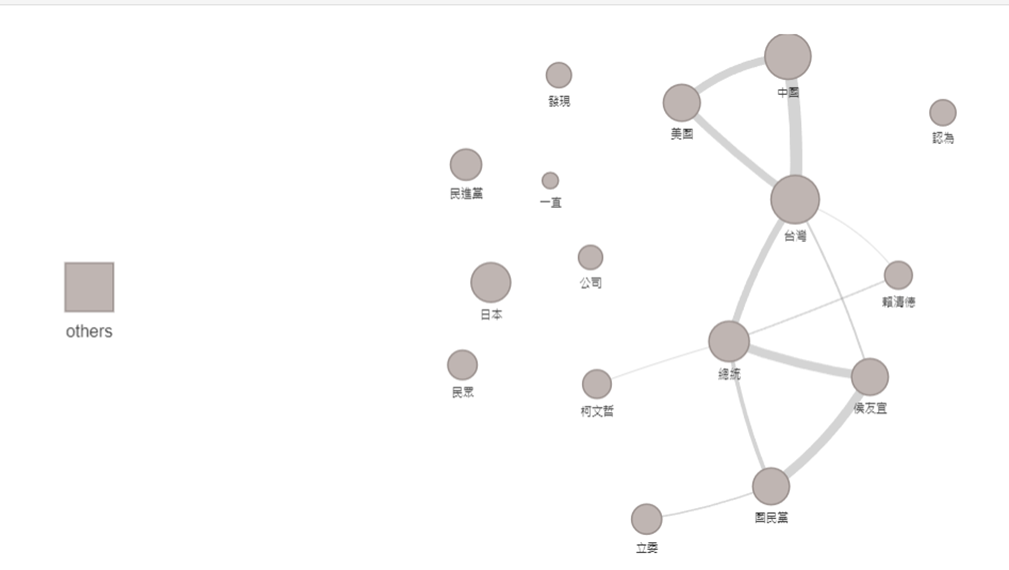
* 參數設定：原始檔案元件4、DTM35、共線/相關性矩陣38



* 任務結果：

網路節點數量15、關聯強度篩選鏈結0~0.5

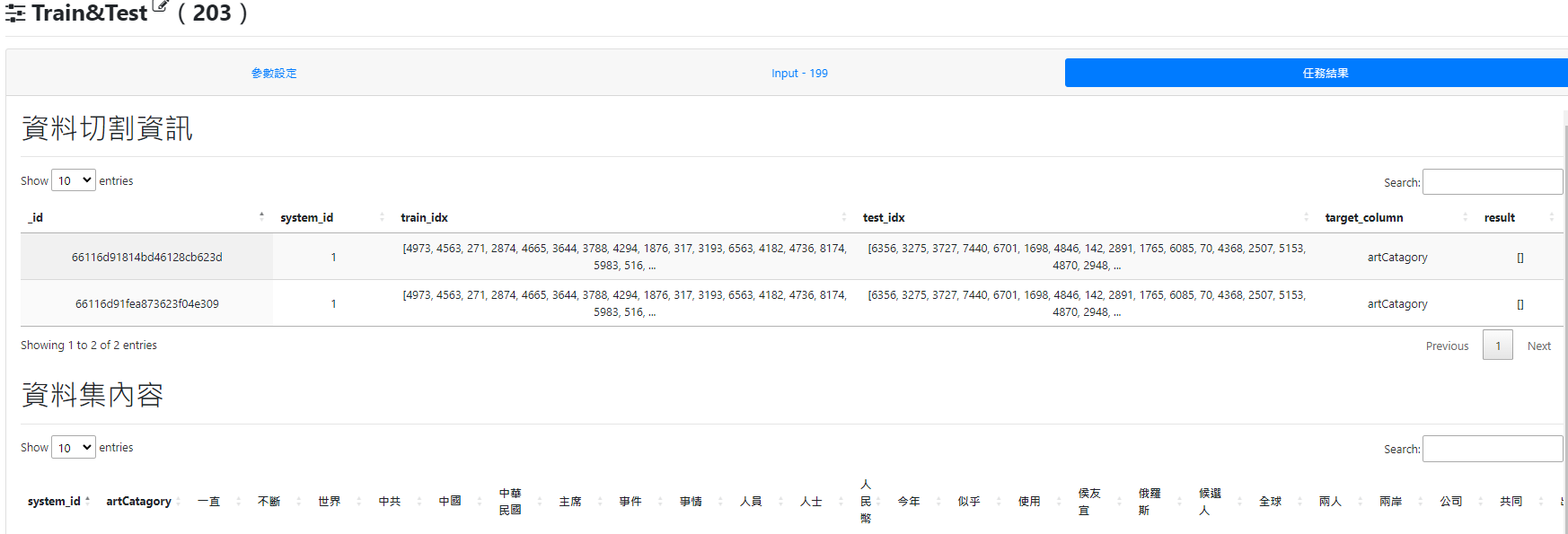


網路節點數量15、關聯強度篩選鏈結0~0.9

## **十二、test& training**



切割訓練與測試資料集，抽出20%做測試資料，結果如下:



## 十三、Classifer

將資料分割為訓練集與測試集後，分類器選擇Decision tree（決策樹）、Logistic Regression（邏輯迴歸）、KNN、SVC，參數設定如圖所示。









任務結果：

各分類模型衡量指標彙整如下表所示。在對比不同分類模型的預測結果，可以觀察到所有模型在給定數據集上都展現了相當不錯的表現。其中，在衡量指標上以邏輯回歸（Logistic Regression）模型的整體表現最為突出。

| 衡量指標 | Decision tree | Logistic  Regression | SVC | KNN |
| --- | --- | --- | --- | --- |
| 測試資料準確度 | 0.807 | 0.893 | 0.874 | 0.833 |
| 測試資料加權F1 | 0.810 | 0.883 | 0.851 | 0.798 |
| 測試資料加權精確率 | 0.814 | 0.886 | 0.878 | 0.835 |
| 測試資料加權召回率 | 0.807 | 0.893 | 0.874 | 0.833 |

### 13-1. Decision tree



### 13-2. Logistic Regression



### 13-3. SVC



### 13-4. KNN

****

## 

## 

## 

## 

## 