**社群媒體分析第九組第二次讀書會作業**

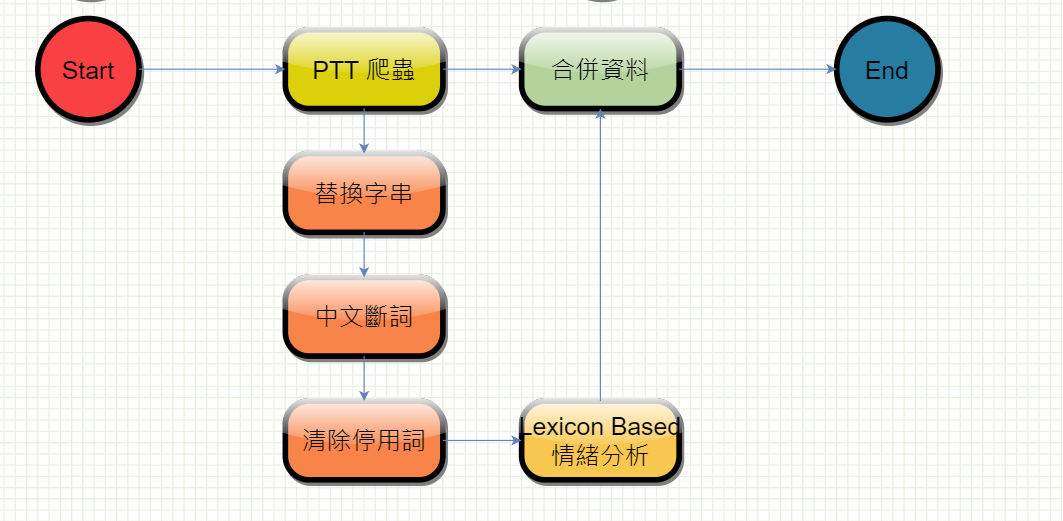
1. 分析熱門主題:家庭暴力
2. 使用平台:中山大學工作流程平台
3. 所使用的工作流程名稱:violence、violence\_inference
4. 建立時間: 2021/12/03 14:33:00

一張含有 桌 的圖片

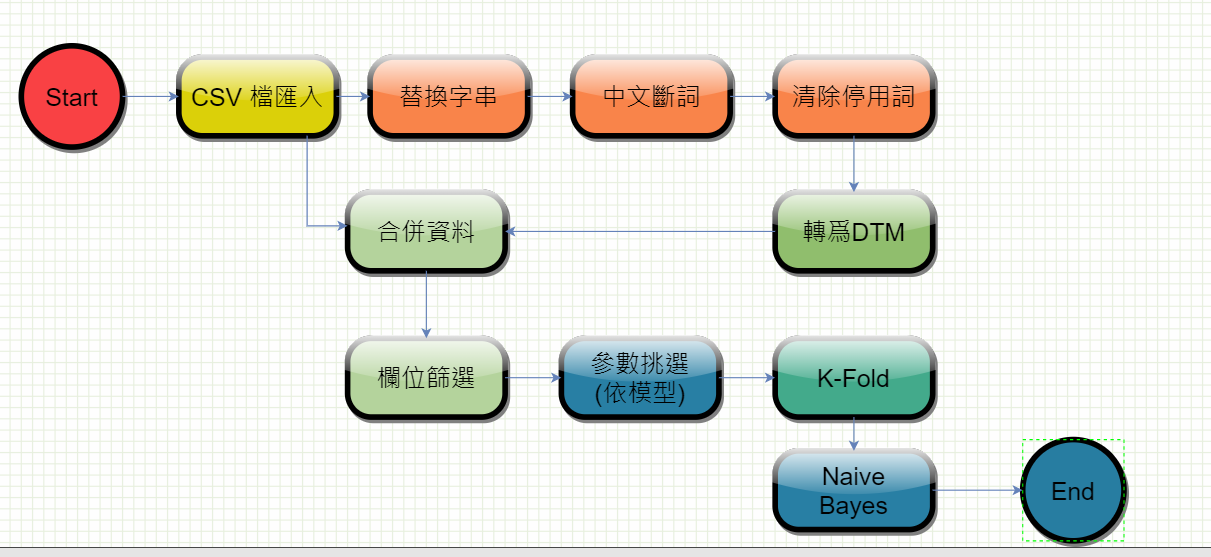
自動產生的描述

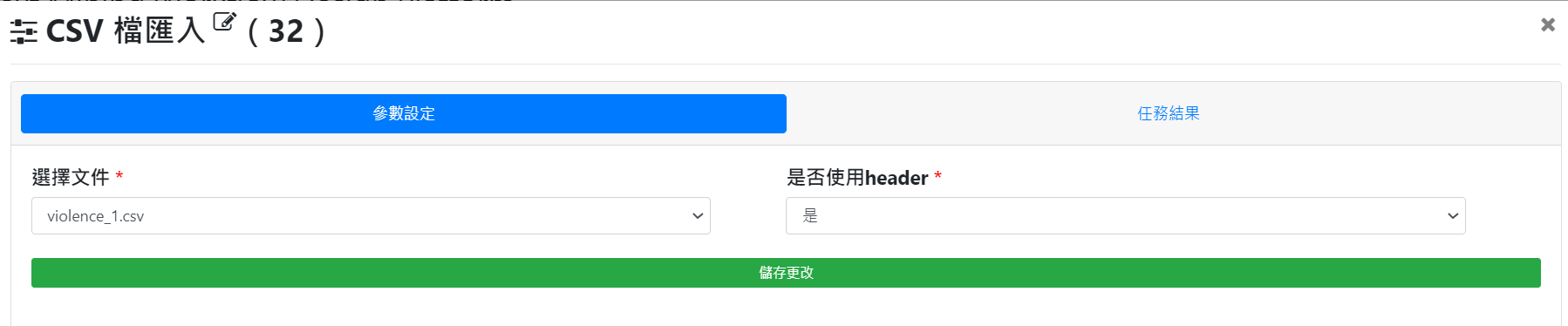
使用此工作流程

1. 家庭暴力工作流程圖:
2. 使用工作流程平台PTT元件爬取2021/11/30~2021/12/03、八卦版、關鍵字為: 高嘉瑜、林秉樞、家暴、家庭暴力、家庭暴力防治法、身體暴力、精神暴力、言語暴力、性暴力、緊急救援、緊急安置、家暴型態、家暴防治、社會局、保護令及家暴法等字詞，共計**1128**篇文章，並使用Lexicon Based情緒分析元件計算每篇文章的情緒值(sentiment\_value)，將PTT爬蟲的文章和每篇文章的情緒值合併之後，輸出CSV檔案，在excel使用統計四分位數進取小於第一四分數及第三四分位數，小於第一四分位數標記為negative情緒、大於第三四分位數為positive情緒，其餘為neutral情緒，因為本組擬訓練一個區分正負面情緒的分類器，所以刪除neutral文章。



1. 將標記好正負面情緒的CSV檔匯入工作流程平台進行模型訓練，共計**438**篇。





將標記好正負面情緒的CSV檔匯入CSV檔匯入



共計438篇

模型的精確度、精確率、召回率及F1皆大於80%

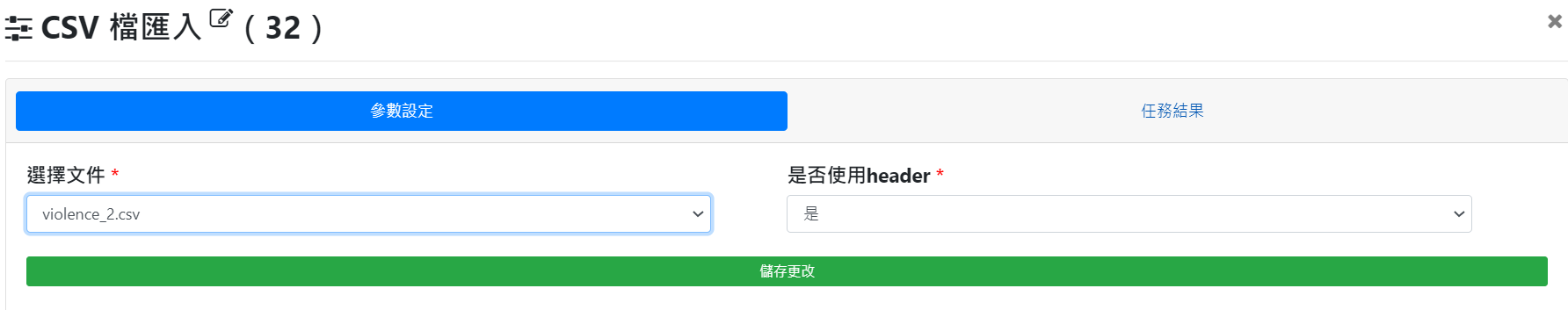


1. 接著使用工作流程平台的蘋果日報爬蟲元件爬取2021/11/30~2021/12/03、關鍵字為: 高嘉瑜、林秉樞的字詞，共計277篇文章



1. 將從蘋果日報爬蟲元件爬取的277篇文章輸入訓練好的情緒分類器進行正負面情緒分析:

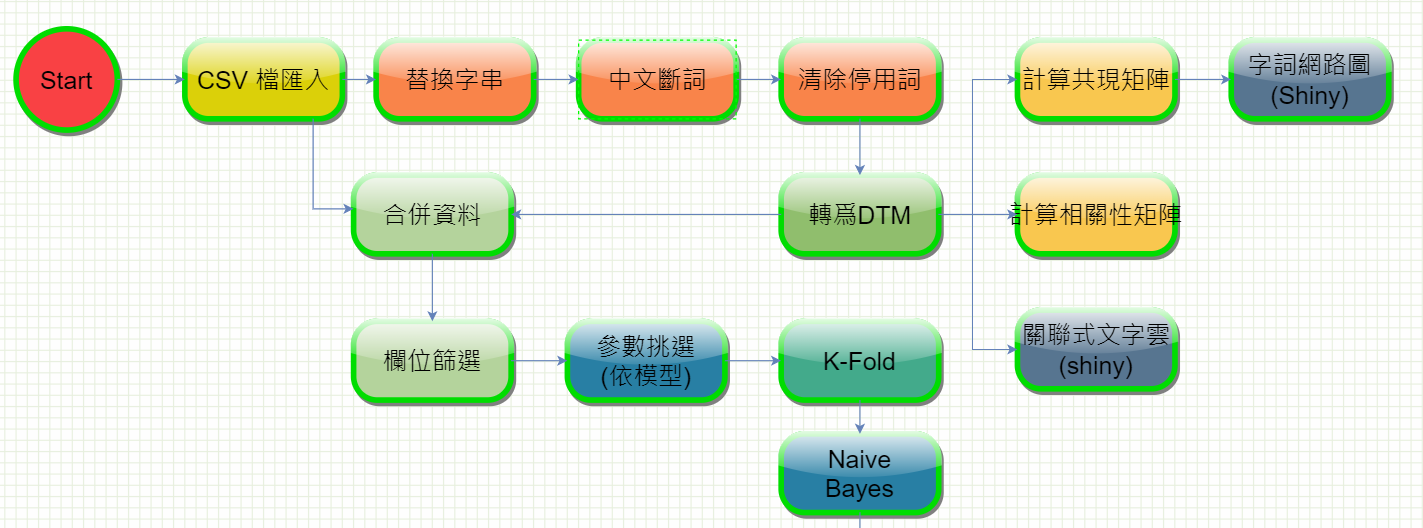


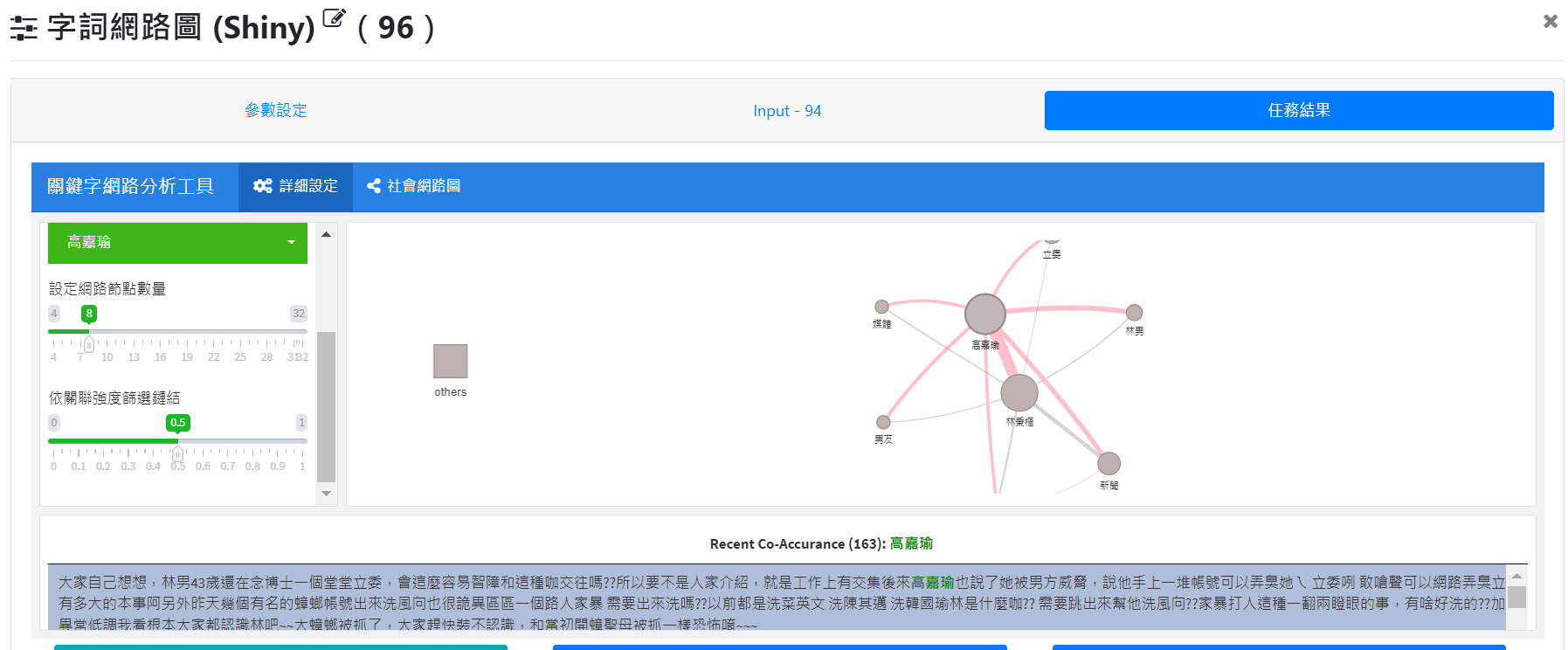


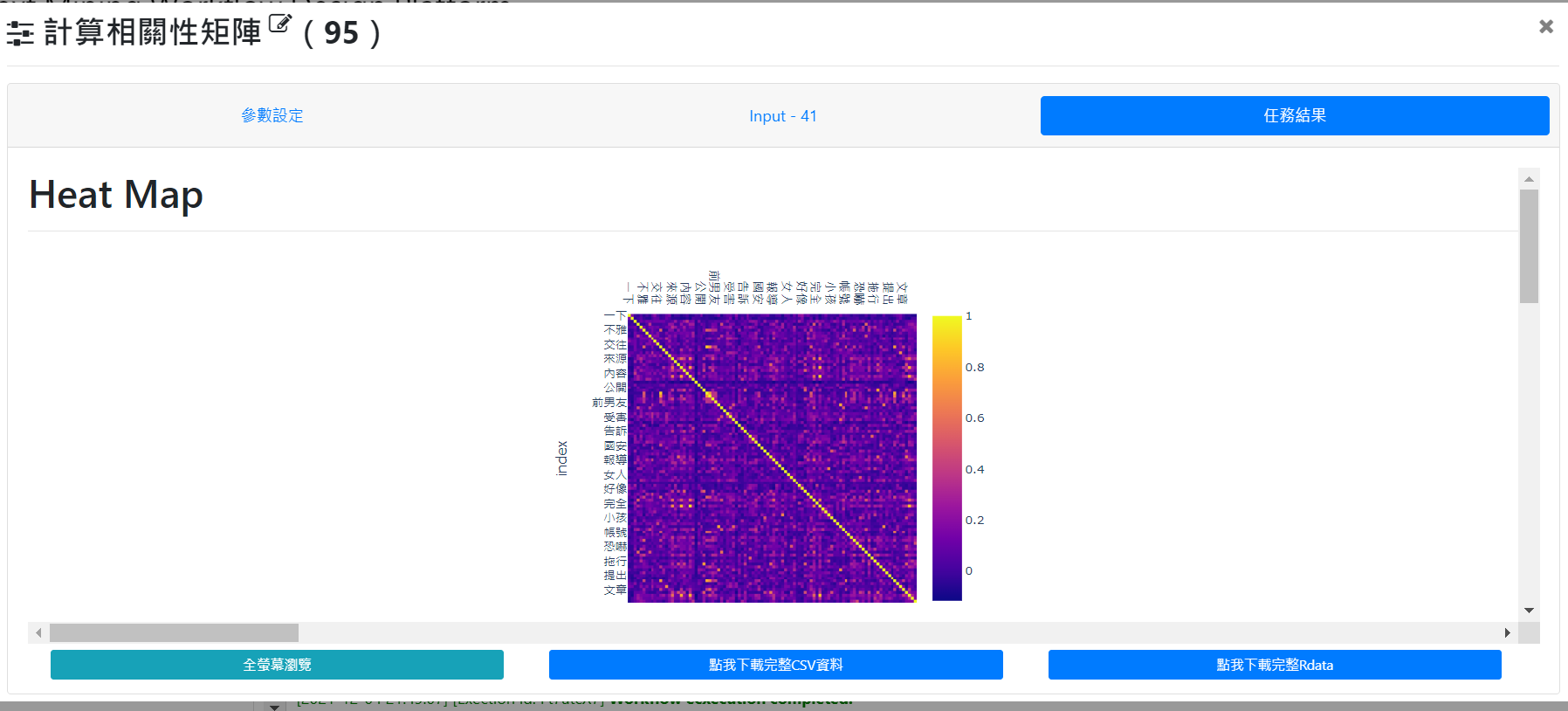
將從蘋果日報爬取的277篇文章輸入

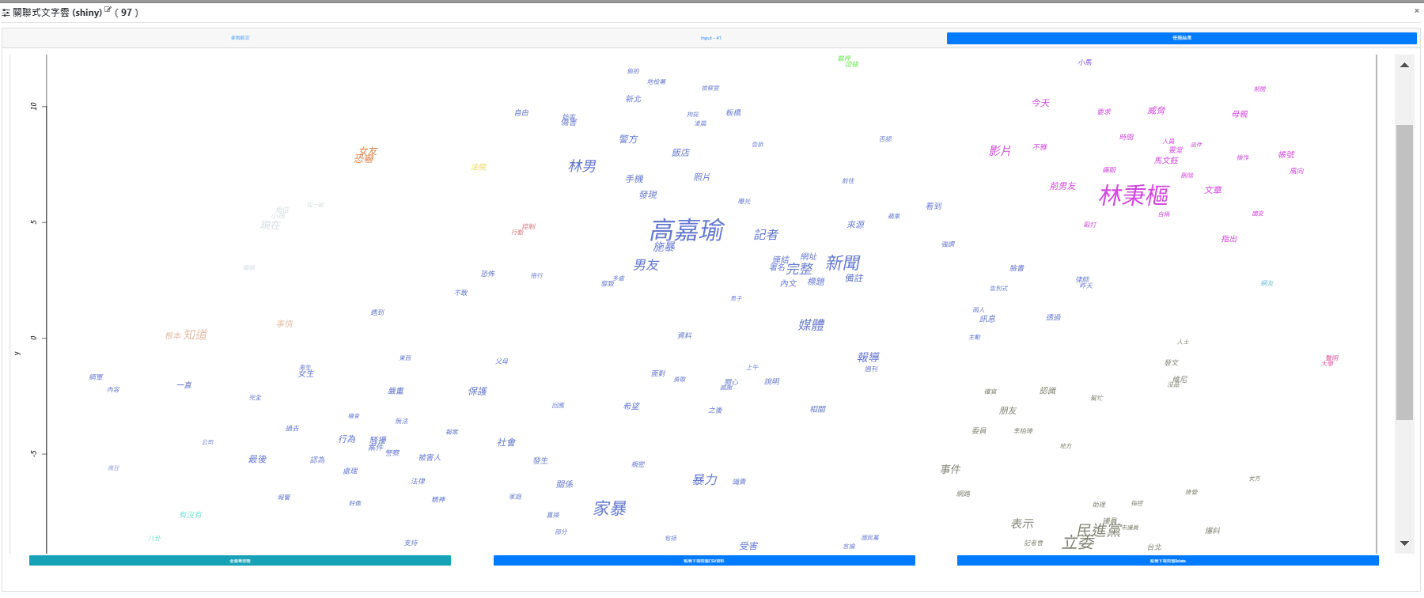


共計277篇

1. 使用標記好正負面情緒的CSV檔匯入工作流程平台進行資料清理(替換字串、中文斷詞及清除停用字)，其中「馬文鈺」斷詞會斷為「馬文」、「鈺」，所以提高馬文鈺的權重為1000。轉為DTM之後，計算共現矩陣並產生字詞網路圖(Shiny)、計算相關性矩陣及產生關聯式文字雲(shiny)  
   



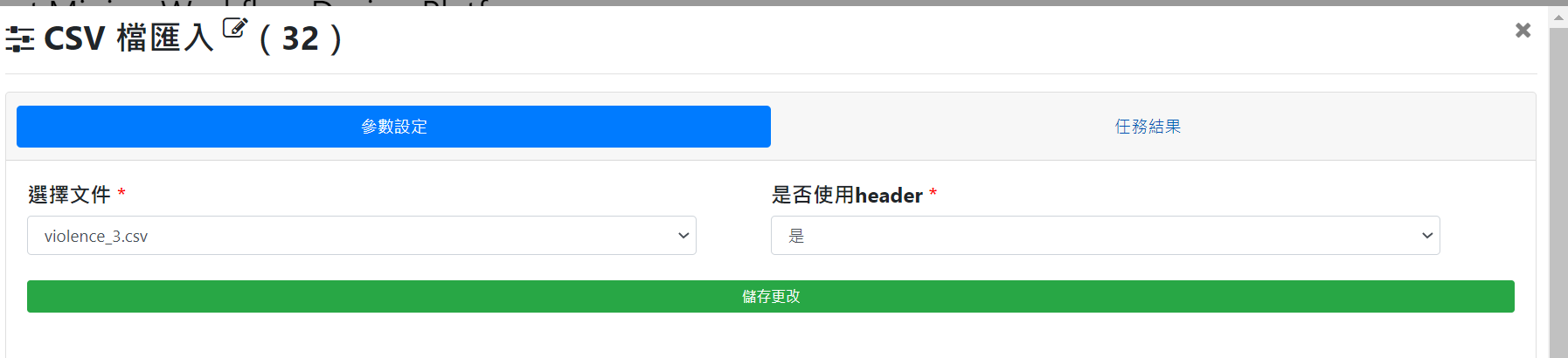


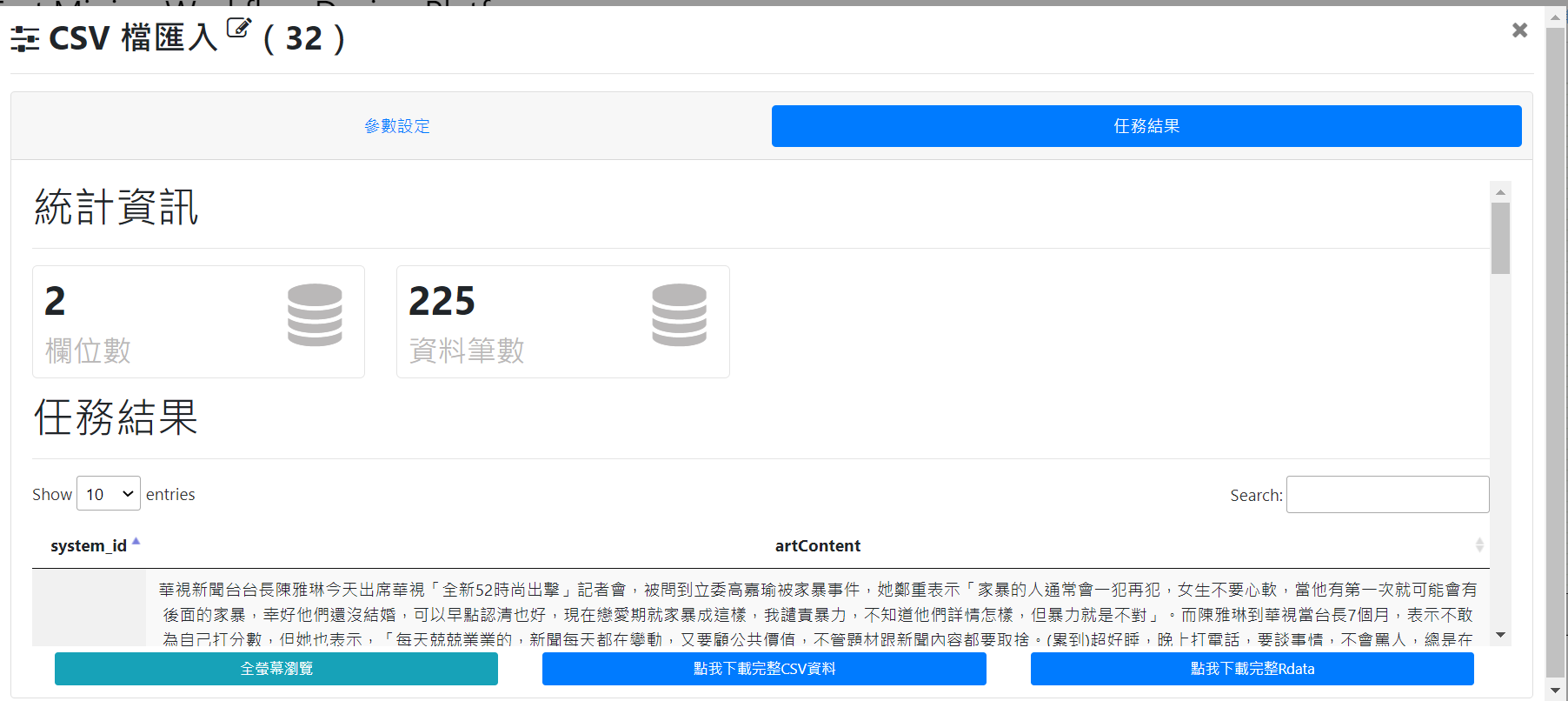


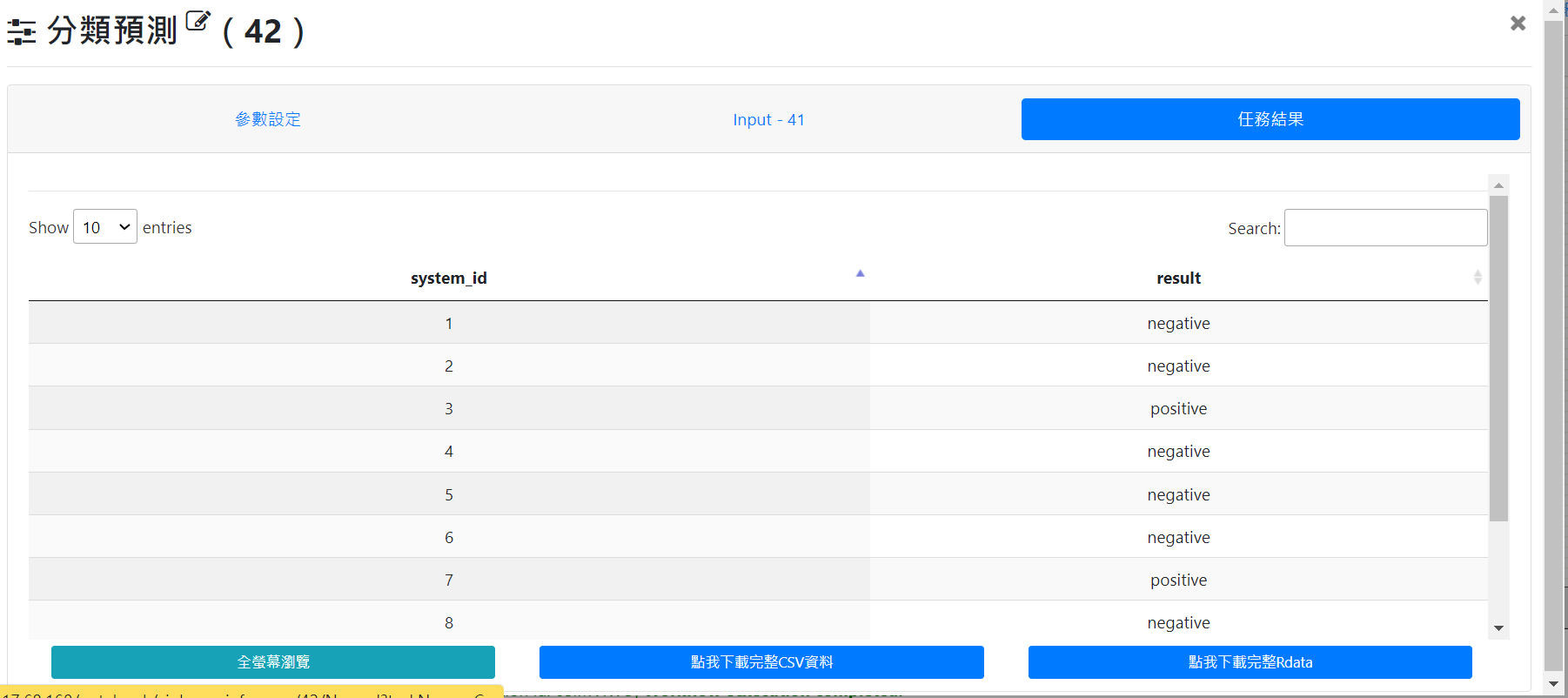
1. Corpus Based情緒分析:使用從蘋果日報爬蟲元件爬取的277篇文章，進行替換字串、中文斷詞、清除停用字後，輸入Corpus Based情緒分析元件進行分析，得到225篇負向情緒數、0篇正向情緒數、52篇與情緒無關的文章，結果如下，將這225筆輸出為violence\_3.csv，並丟入我們建好的情緒分類器進行分類預測，結果225篇文章，預測124筆為負面文章、101筆為正面文章，與Corpus Based情緒分析元件判定有225篇負面文章有所差異。

一張含有 文字 的圖片

自動產生的描述







1. 結論
2. 本次作業本小組擬訓練一個判斷正負面情緒的分類器，使用PTT爬取的1128文章，由於每篇文章並未標記是正面、負面或中性文章，所以我們使用Lexicon-based情緒元件字典法計算情緒值，情緒值有小於0、等於0和大於0，一開始將大於0設為positive、等於0為neutral、小於0為negative，將刪除neutral的文章當作測試資料丟進工作流程平台去訓練模型，結果得到模型的精確度、精確率、召回率及F1都不高，後來將匯出的csv檔案的情緒值欄位，採用統計四分位數進行標記positive、neutral及negative，將刪除neutral的文章再當作測試資料丟進工作流程平台去訓練模型，結果得到模型的精確度、精確率、召回率及F1大大提高，高於80%。
3. 在建模的過程中，發現使用Tree的模型，模型的精確度、精確率、召回率及F1較SVC模型表現更好。
4. 從字詞網路圖可以清楚的看出家暴主角的人際關係、職業及發生的事件(家暴)。
5. 從關聯式文字雲來看，除了事件男女主角的相關字詞外，還可以看出家暴常與親密關係同時出現，顯示親密暴力常發生在共同生活的家庭中，因為種種原因，外力防治介入有其困難，值得深入研究探討。