社群媒體分析第七組第三次讀書會作業

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](#_Toc134370554)

[二、 訓練流程 1](#_Toc134370555)

[(一) 分類模型 2](#_Toc134370556)

[(二) LDA主題模型 7](#_Toc134370557)

[三、 分析結果 12](#_Toc134370558)

[(一) 分類模型 12](#_Toc134370559)

[(二) LDA主題模型 15](#_Toc134370560)

1. 探討主題

此次我們選擇的探討主題為分類運動項目，由於課堂中助教所分類的為新聞的看板類型，例如：主題為運動、兩岸的新聞，藉由訓練模型，達到自動判斷文章的類別。也因此讓我們思考是否有辦法進一步以運動新聞為主題，區分運動的種類，經過討論後，我們最終決定以五大運動種類來進行訓練，分別是：棒球、足球、籃球、羽球以及其他。

1. 訓練流程

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，並預計分別針對類型、內文進行資料處理及相關分析，使用分類模型與LDA主題模型訓練資料，第三次讀書會分析流程如圖 1(分類模型)、圖 2(LDA主題模型)。

一張含有 圖表 的圖片

自動產生的描述

圖 1、分類模型分析流程

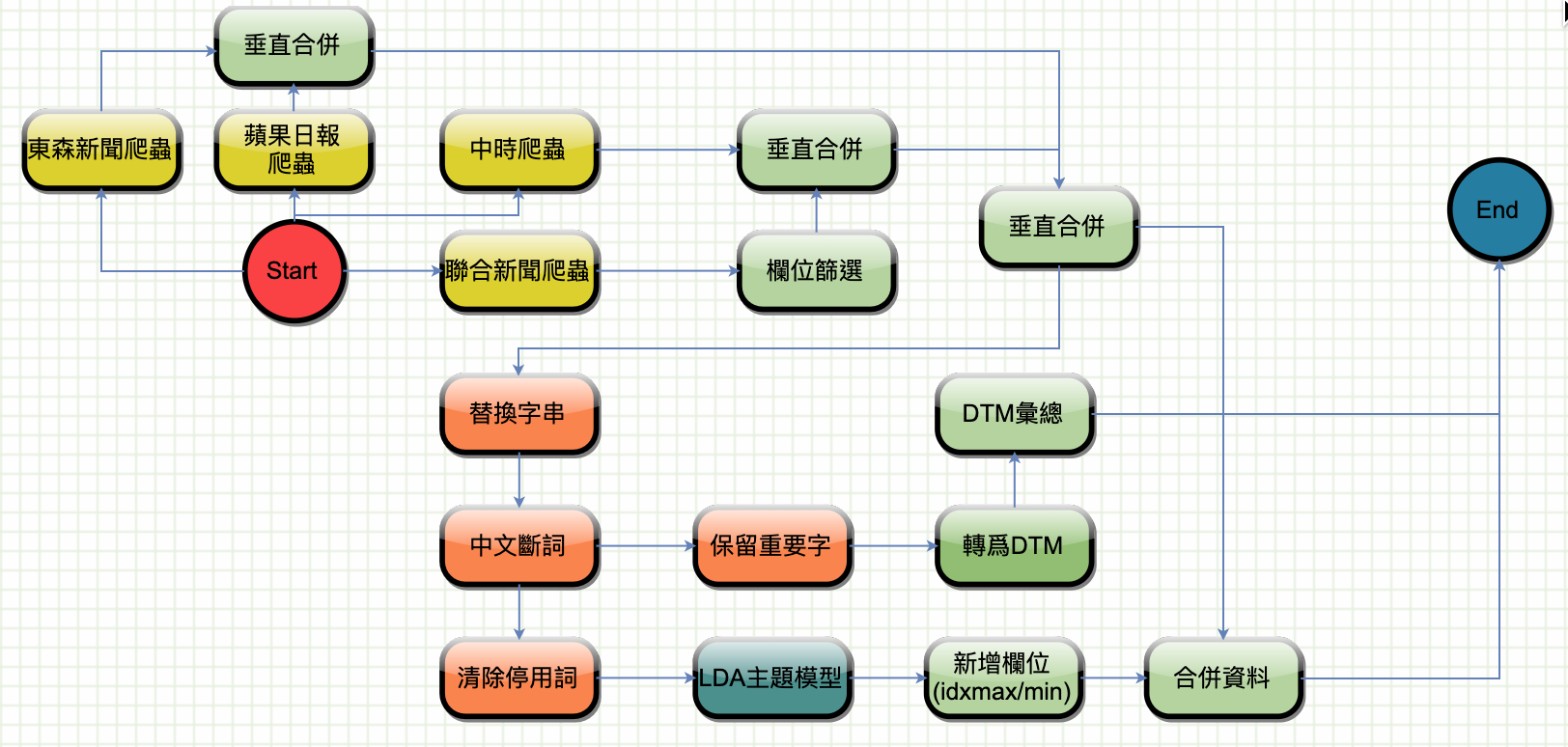


圖 2、LDA主題模型分析流程

* 分類模型與LDA主題模型分析步驟如下：

1. 分類模型
2. 資料爬取：我們選擇從蘋果日報、東森新聞、聯合新聞網分別爬取運動類的新聞。

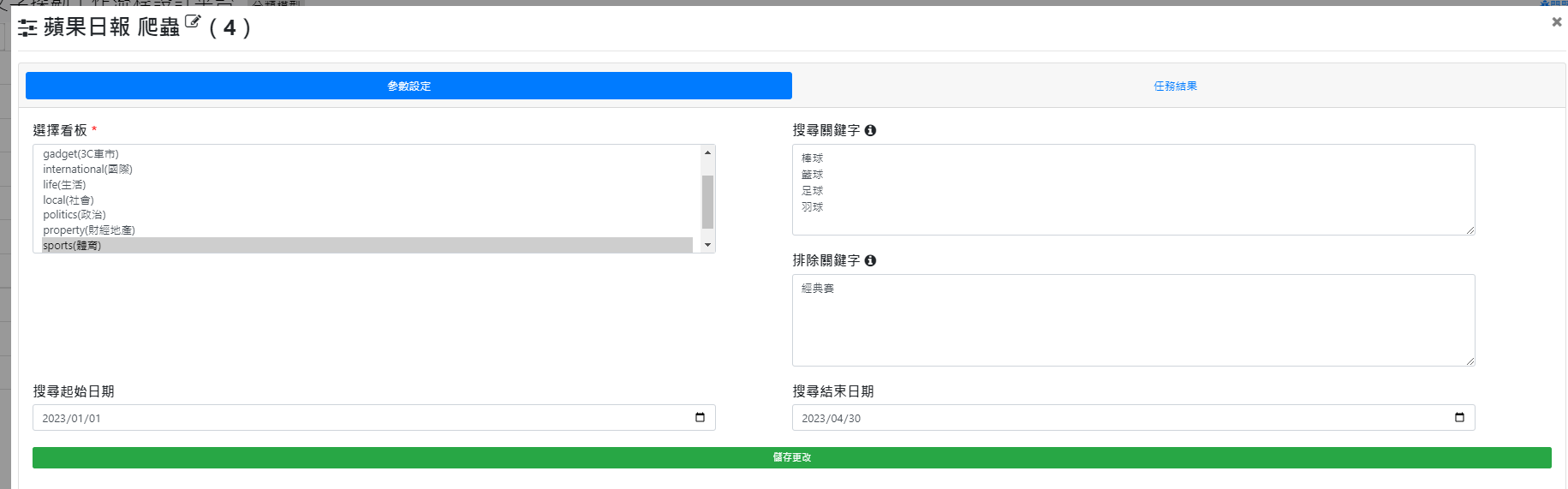


圖 3、蘋果日報爬蟲

看板選擇運動版，搜尋關鍵字選擇棒球、籃球、足球、羽球，由於前陣子經典賽新聞較多，擔心結果過於導向棒球，因此排除關鍵字放入經典賽。

1. 欄位篩選：保留了相關所需要的欄位。
2. 資料整理：透過中文斷詞、清除停用字、ngrams等方式，最終轉出了較具有意義的DTM矩陣。



圖 4、中文斷詞

一張含有 文字 的圖片

自動產生的描述

圖 5、計算ngrams

一張含有 文字 的圖片

自動產生的描述

圖 6、清除停用字

一張含有 文字, 白板 的圖片

自動產生的描述

圖 7、詞頻計算

通過計算ngrams、詞頻計算等輔助，來判斷中文斷詞與清除停用字的的內容關鍵字，去除較不重要的詞彙。

一張含有 日曆 的圖片

自動產生的描述

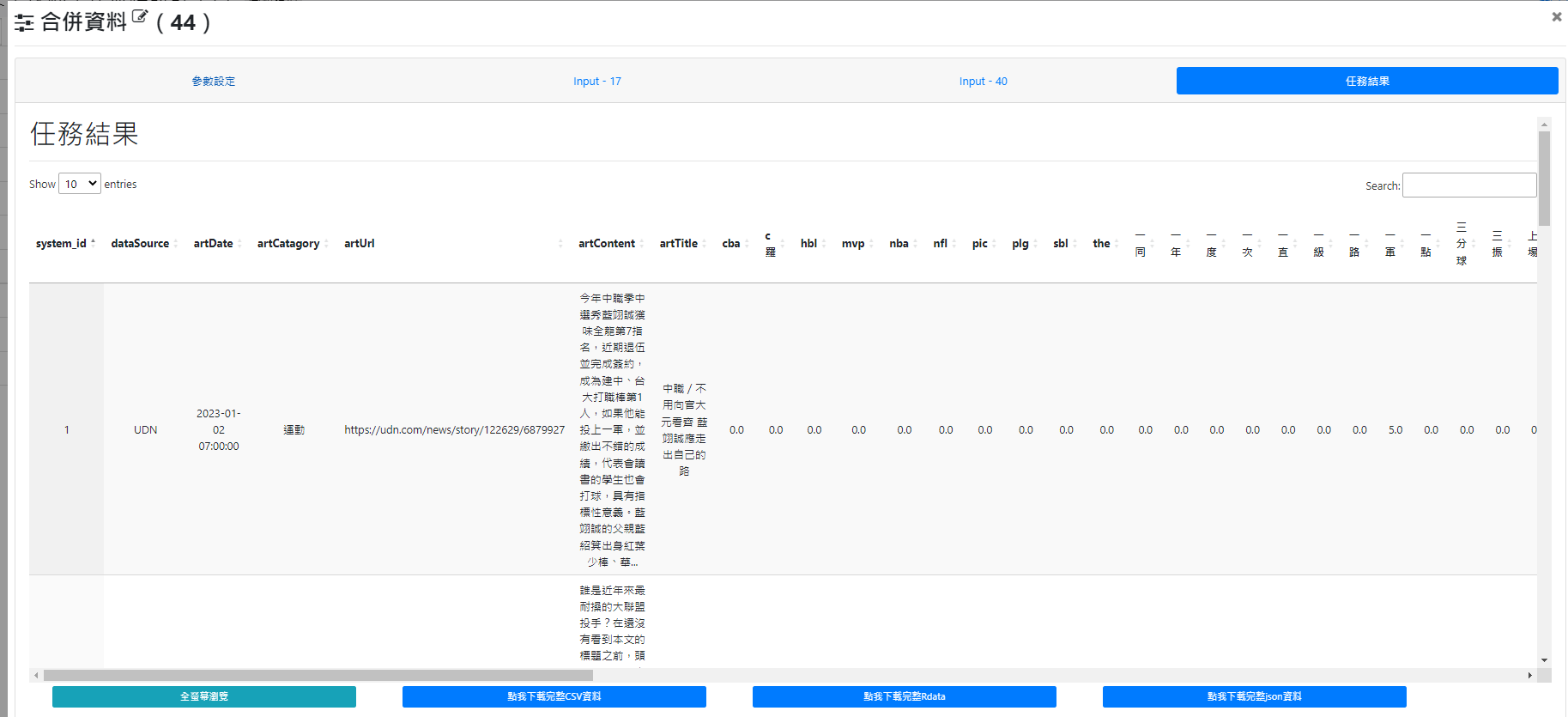
圖 8、轉為DTM

圖 9、合併資料

1. 人工分類：合併資料後，匯出CSV檔，並以人工的方式將新聞以籃球、棒球、羽球、足球、其他五大分類，進行運動項目細分。

一張含有 圖表 的圖片

自動產生的描述

圖 10、人工分類的CSV檔

1. 模型訓練：將人工細分後的CSV檔匯入後，進行Train&Test，再執行羅吉斯迴歸，獲得最終結果。

一張含有 日曆 的圖片

自動產生的描述

圖 11、篩選CSV欄位

利用匯入分類好的CSV檔，進行欄位篩選，只保留類別與DTM中的欄位。

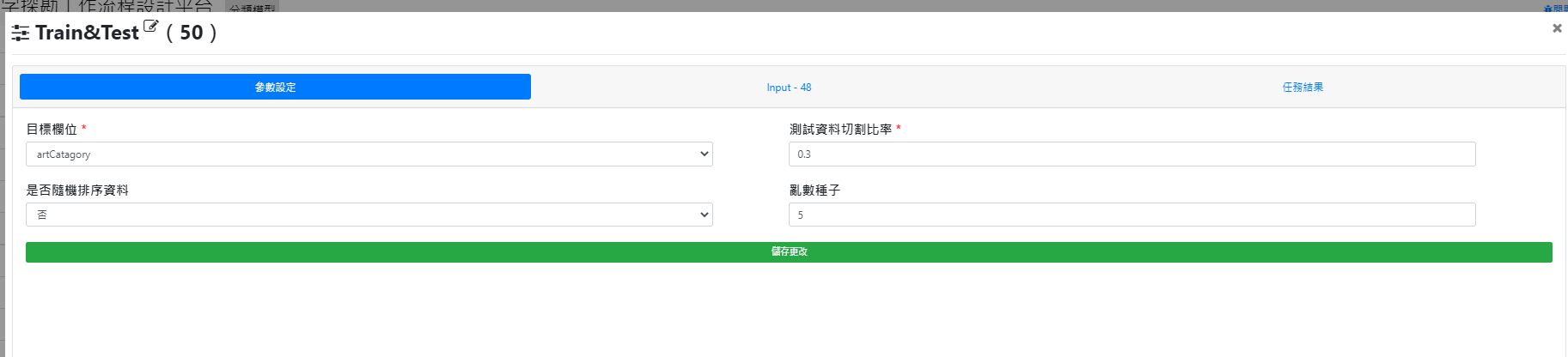


圖 12、Train & Test

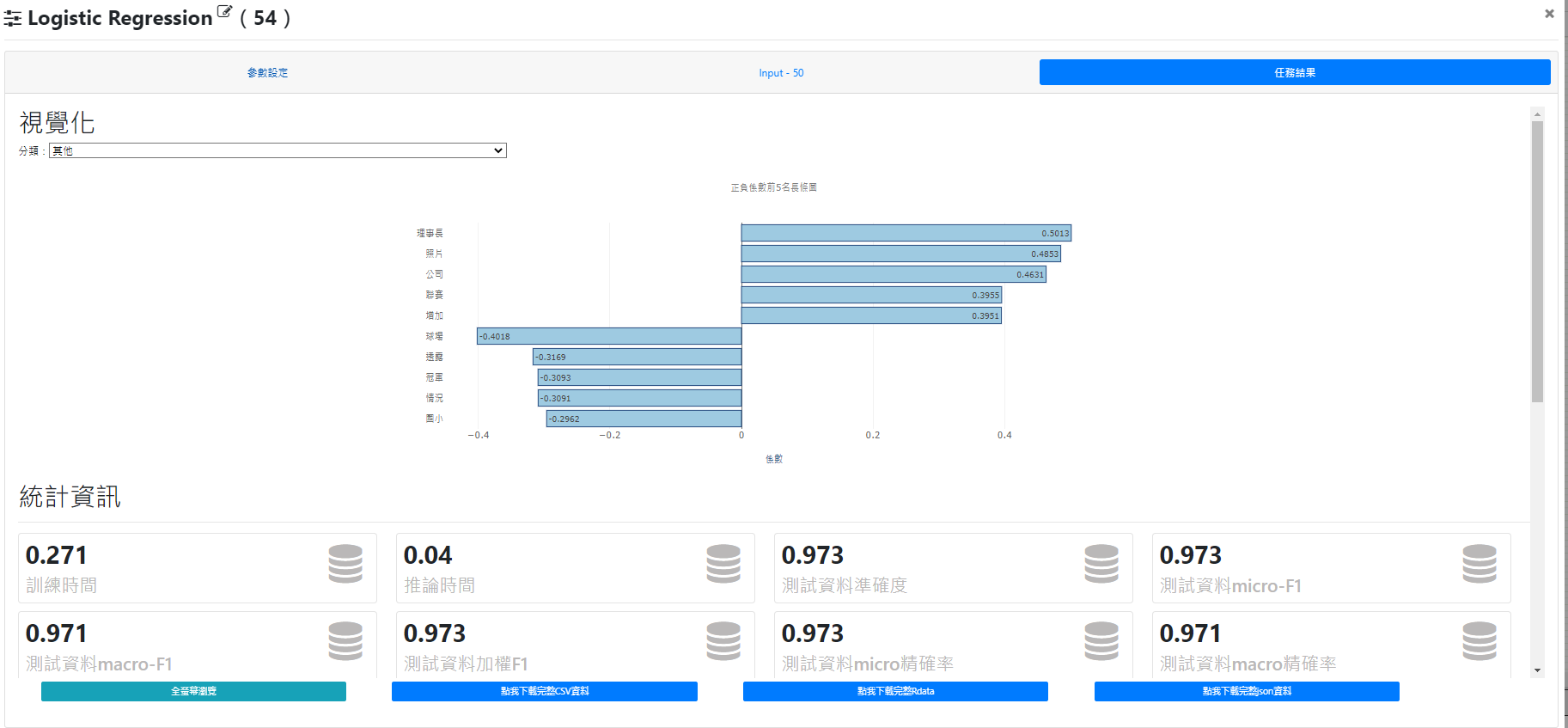


圖 13、羅吉斯迴歸

將資料經過Train & Test，再跑羅吉斯迴歸，獲得結果。

1. LDA主題模型
2. 資料搜集：以聯合新聞、中時新聞、蘋果日報、東森新聞四種的運動(體育)版，時間區間為2023-01-01至2023-05-06，透過垂直合併，總筆數為12,904筆。
3. 自訂字典：以籃球、棒球、桌球、羽球、網球、跆拳道、柔道、啦啦隊...等幾類，總共定義了383筆。

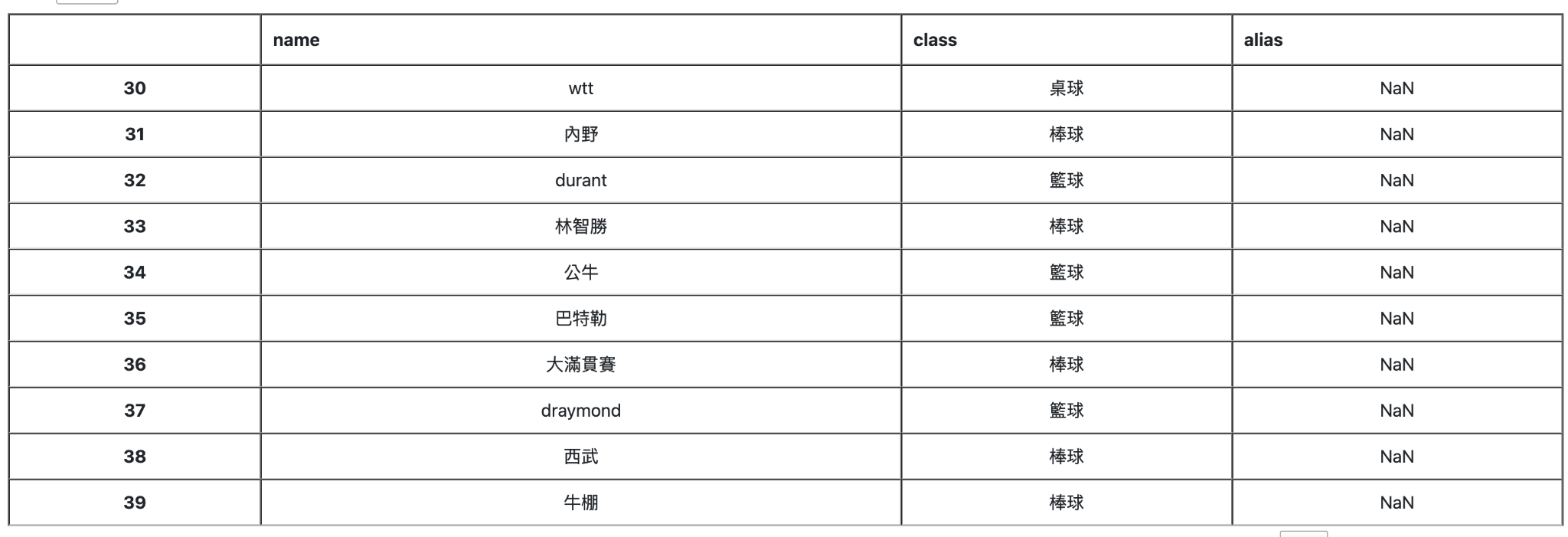


圖 14、自訂字典

替換字串時選擇字典，其他設定一樣。



圖 15、替換字串設定

中文斷詞先前設定方法相同，共定義126組詞彙。



圖 16、中文斷詞定義

清除停用詞：今天、沒想到、臉書、朋友、學校......等633個詞彙。



圖 17、清除停用詞設定

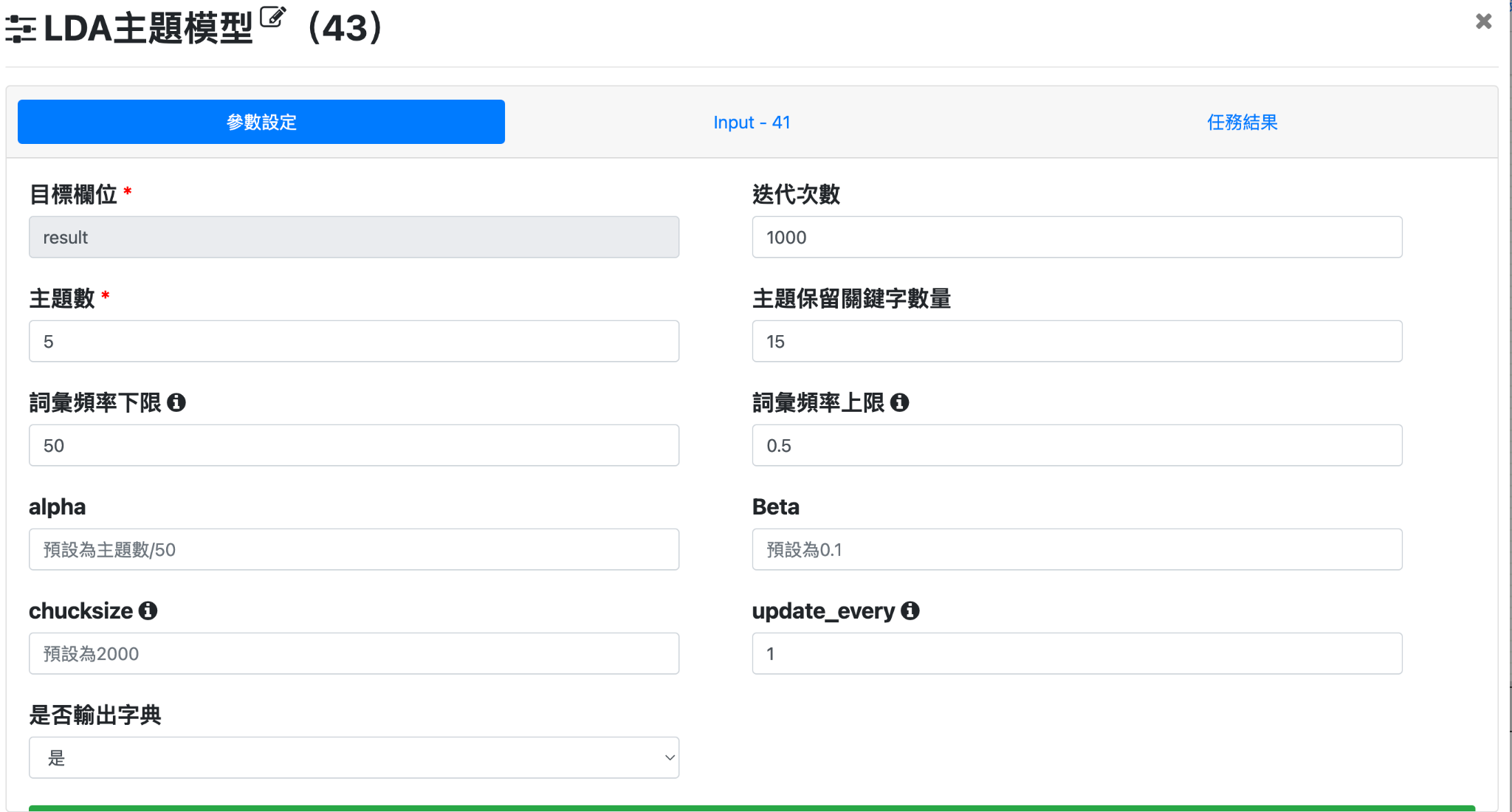
1. 主題模型LDA設定如下，反覆測試，最後將主題數分成五群，詞頻下限設定為50，保留主題關鍵字為15個，迭代次數拉高至1000次，詞彙頻率上線設定為0.5。

圖 18、主題模型LDA設定

1. 第一群為「棒球混其他資訊」

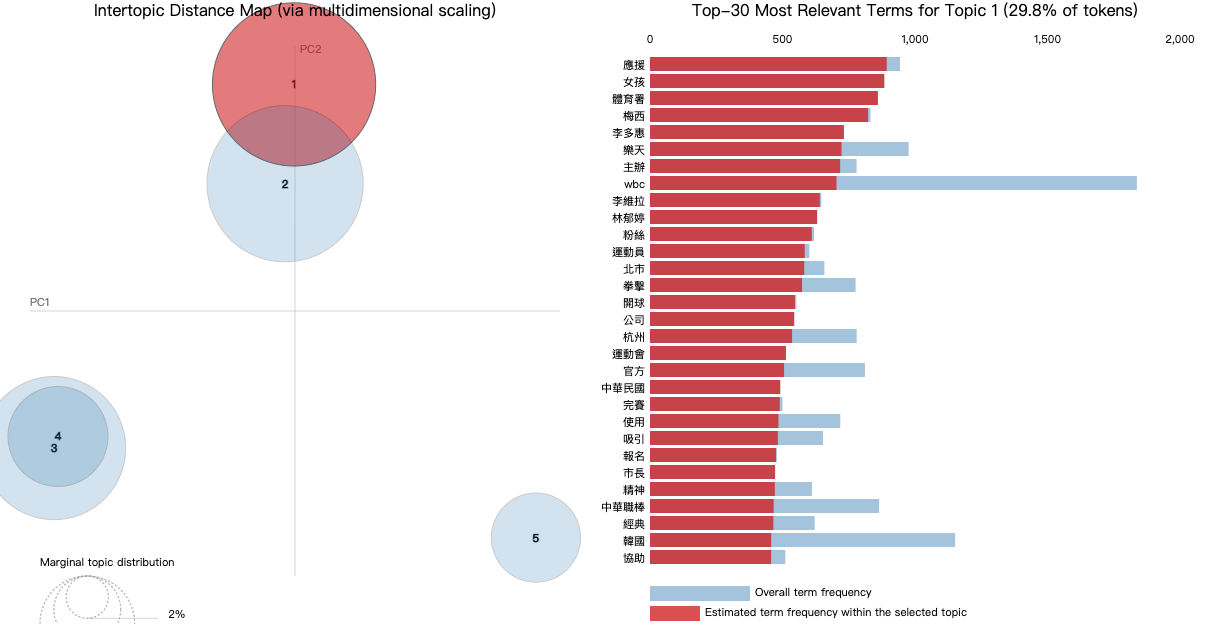


圖 19、第一類主題

1. 第二群為「棒球」

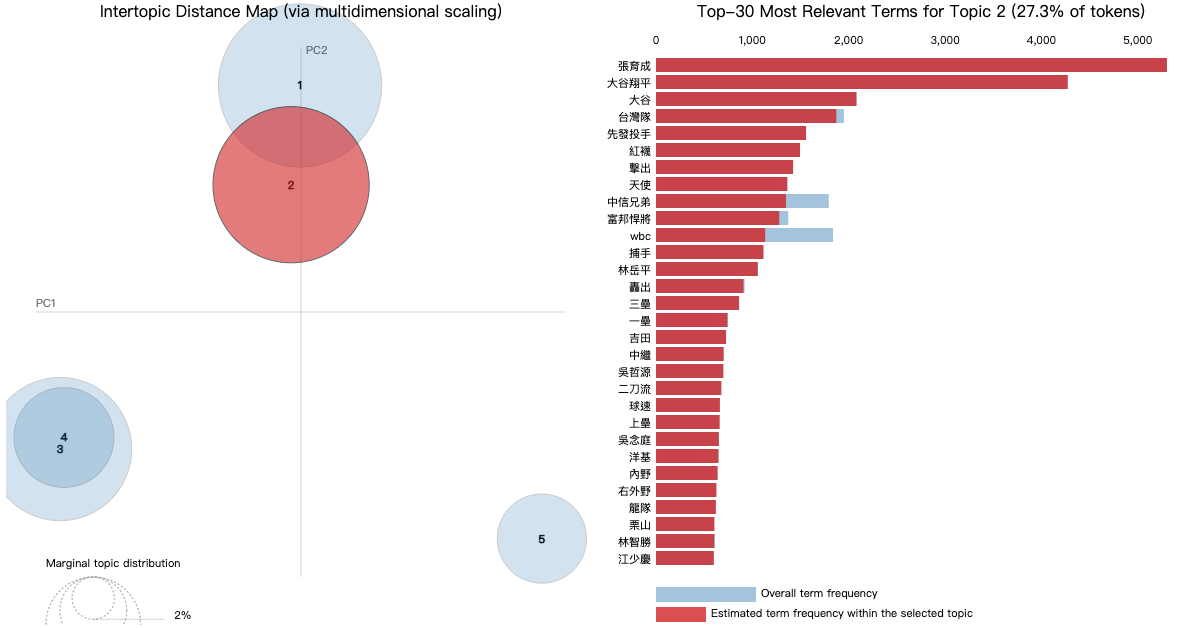


圖 20、第二類主題

1. 第三群為「籃球-NBA」

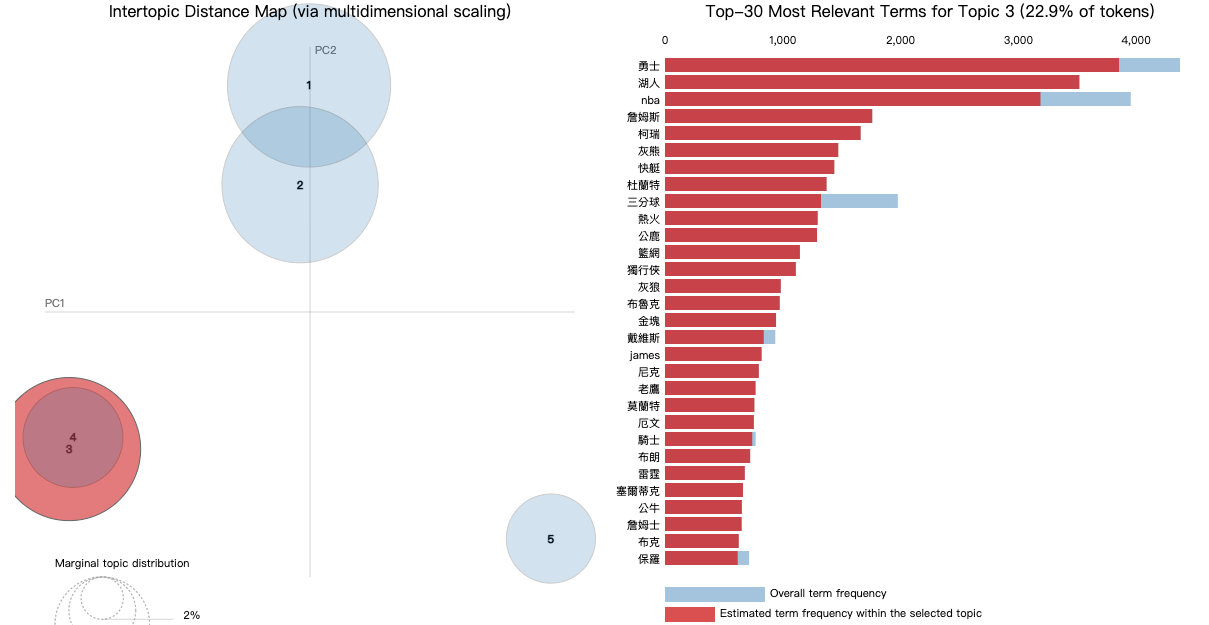
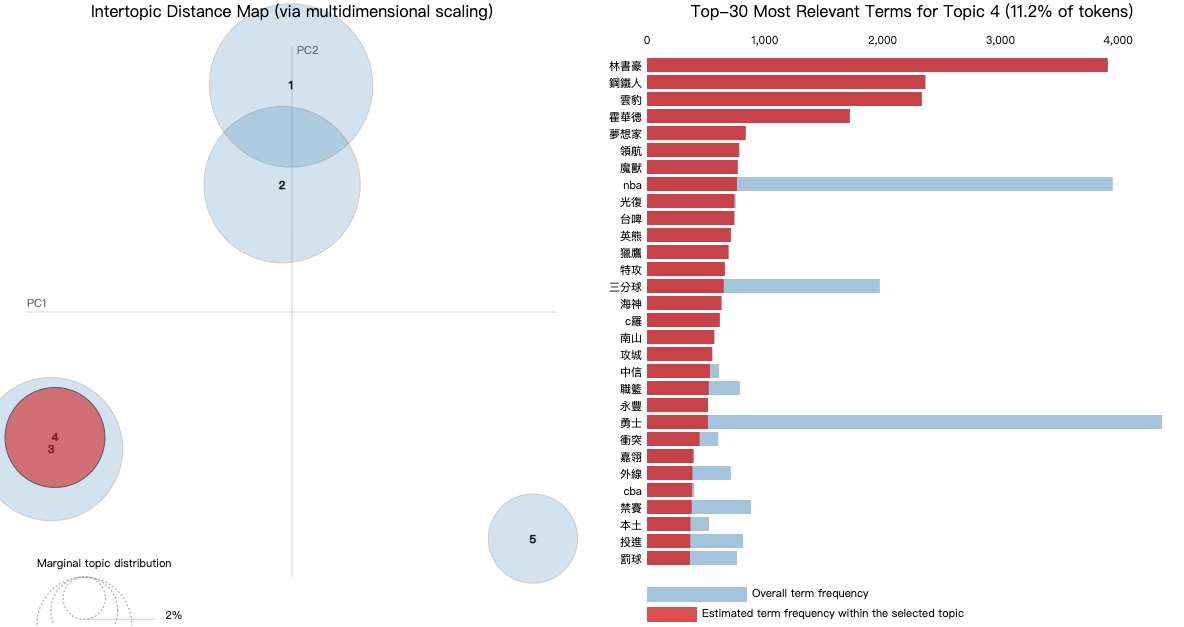


圖 21、第三類主題

1. 第四群為「籃球-台灣職籃」

圖 22、第四類主題

1. 第五群為「其他(羽球、網球、桌球、足球等)」

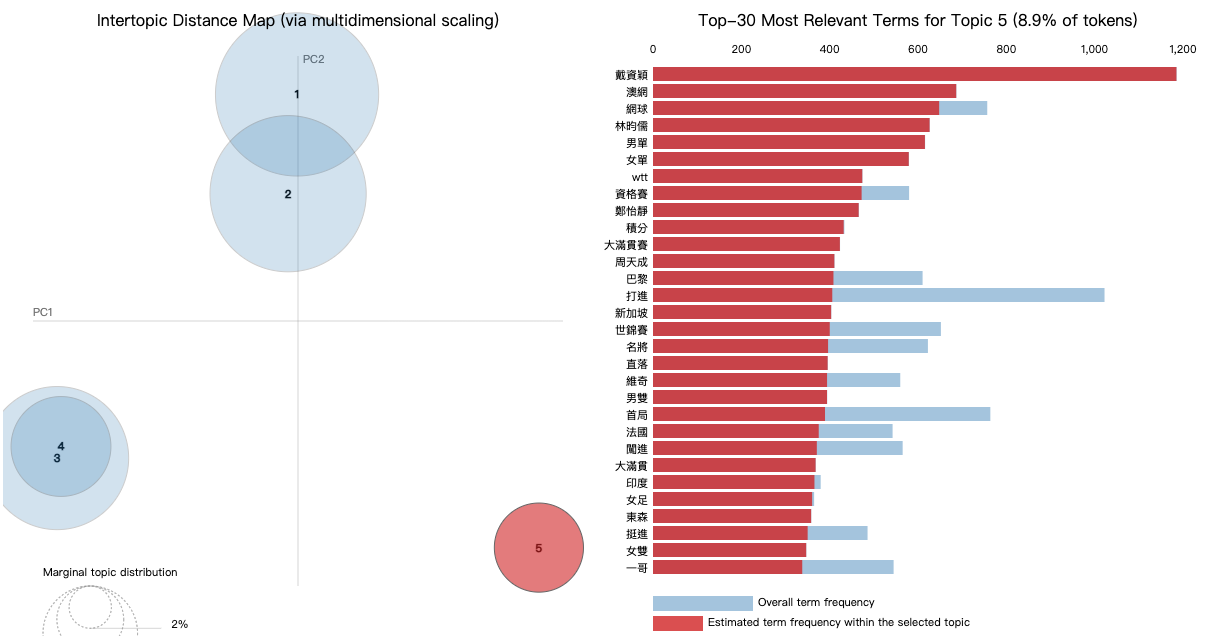


圖 23、第五類主題

1. 分析結果
2. 分類模型

使用分類模型整理完資料，並獲得一個分類模型後，接著我們將用中時新聞網的運動新聞，來進行我們結果的驗證，最終得到分析結果。

一張含有 圖表 的圖片

自動產生的描述

圖 24、結果驗證流程圖

* 驗證步驟如下：

1. 資料前處理：一樣先進行爬蟲、中文斷詞、清除停用字等步驟，將資料轉為DTM後，並匯出CSV檔，一樣經由人工分類，獲取最正確的分類答案。
2. 分類預測：將分類過的CSV檔匯入後，進行答案比對。

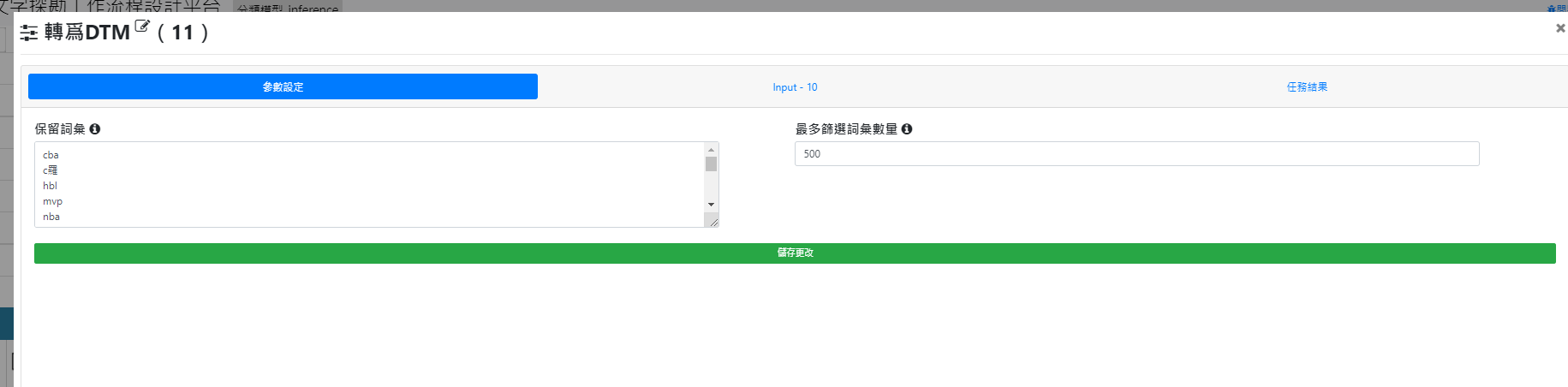


圖 25、轉為DTM

此處的保留詞彙，我們與訓練模型時，轉為DTM的關鍵字一樣。

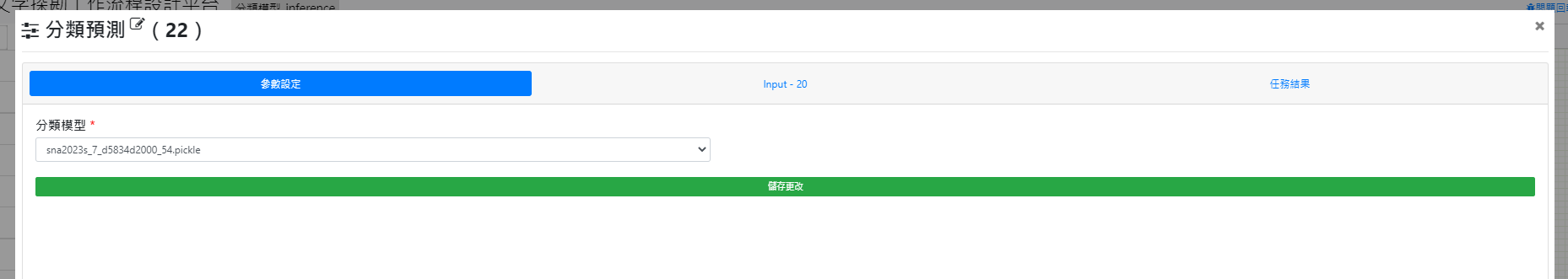


圖 26、分類預測

一張含有 文字 的圖片

自動產生的描述

圖 27、分類預測結果

先選擇我們訓練的模型作為分類模型，然後獲得每篇新聞，模型所判斷的分類結果。

一張含有 資料表 的圖片

自動產生的描述

圖 28、與正確答案比對

此處，藉由經過人工分類的CSV檔合併資料後，可以一目了然的看到由模型訓練的分類結果與正確答案的比對。

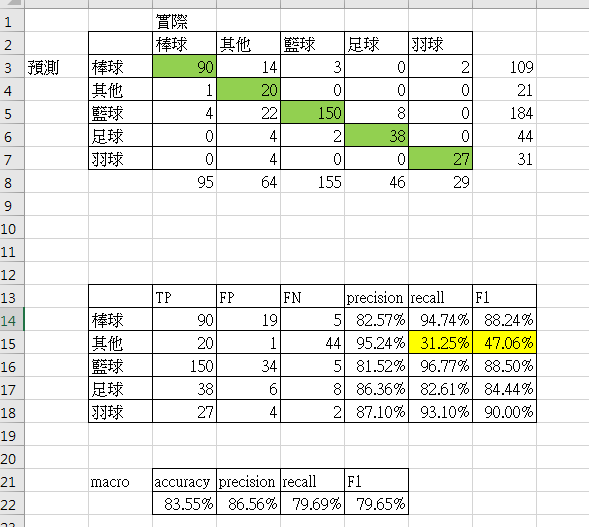


圖 29、最終結果統計

從結果來看，除了「其他」這個類別的準確率較低外，其餘準確率基本都有達到80%以上，這個結果算是不錯的。

而關於「其他」準確率較低的原因，推測是因為許多運動新聞在運動項目的區分上，較為模糊。例如：關於奧運、世大運等重要賽事，新聞的內容常常並非只報導單一運動，而是進行綜合報導與介紹，因此較難進行分類；而許多新聞也許主要並非在報導運動，例如：運彩新聞，但也許因為新聞中出現了某項運動的相關關鍵字，而可能導致模型有錯誤的分類。

1. LDA主題模型

第一群與第二群皆為棒球，但第一群參雜了較多不相關的資訊，而第二群滿精準地歸類為棒球相關的主題。

第三群與第四群分析皆為籃球，但又有更細分為 NBA 與台灣職籃，主題算是分得滿明確的。

而第五群為其他運動項目，像是羽球、桌球或網球等，大致上也算是滿準確的，少有一些不相關的字詞出現。

整體而言，四份新聞網所爬出來的資料，這個期間，籃球與棒球佔了大多數的新聞，而跑完 LDA 主題模型後，雖然有些不相關的詞彙，但以整體結果來說，確實有分出正確的運動主題。