**第三次讀書會報告**

**組別**：11

**組員**：

邱昱榕 N114320004

朱怡樺 N114320005

林威呈 N114320011

董力慈 N114320019

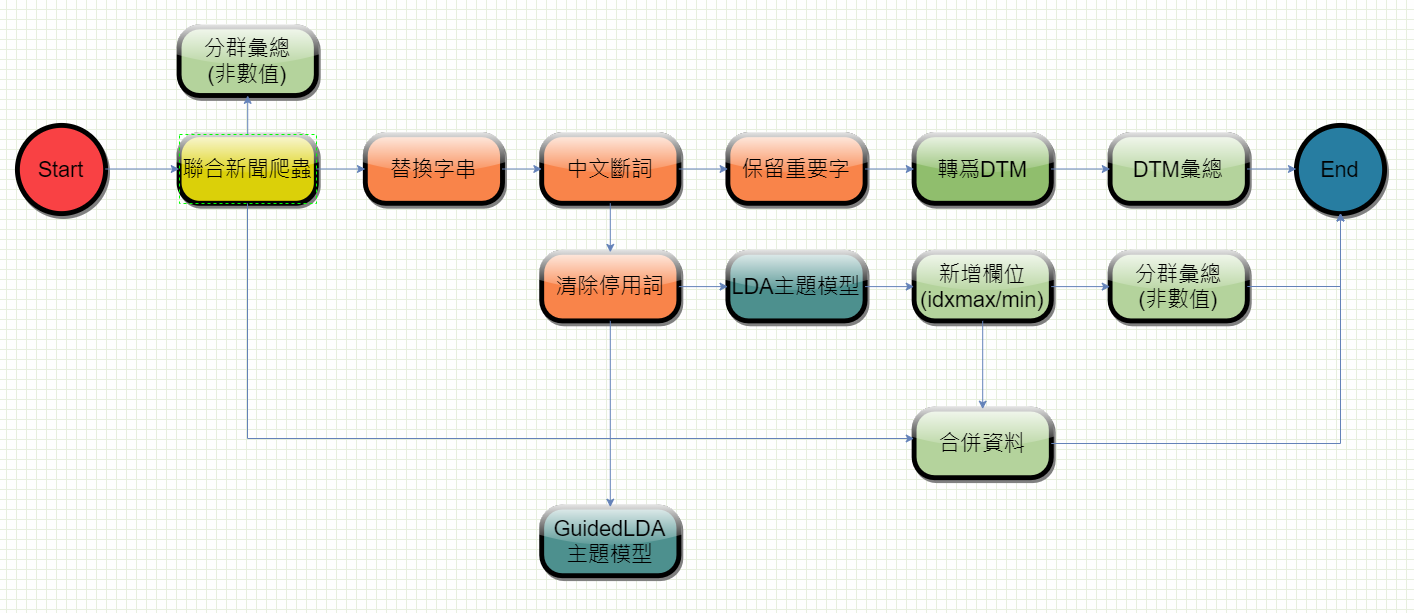
花秀旻 N114320021

趙丞德 N114320023

陳昱維 N114320026

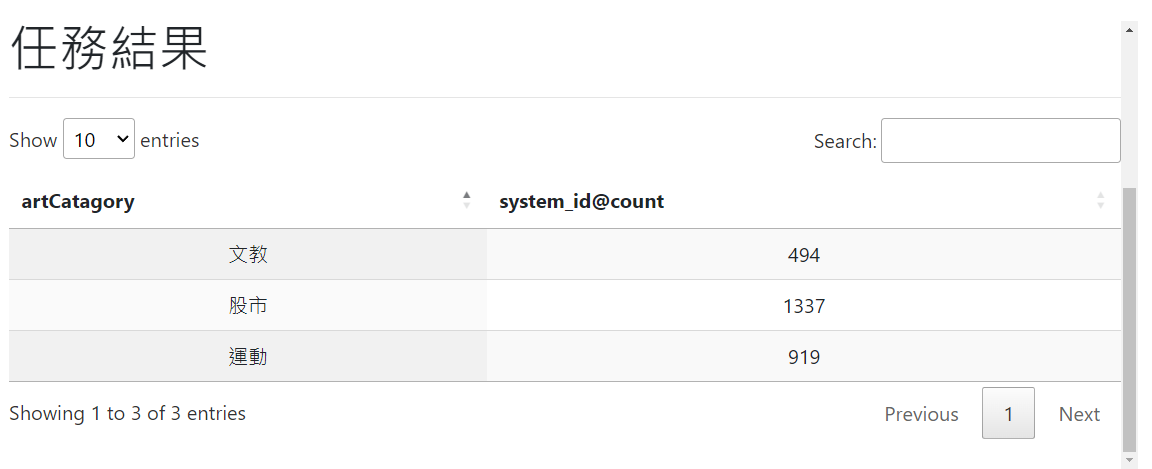
李晉安 N114320028

本次主要的是使用LDA and GuidedLDA 主題模型來判斷文本主題的分類屬性，其 Tarflow 的程序如下圖：

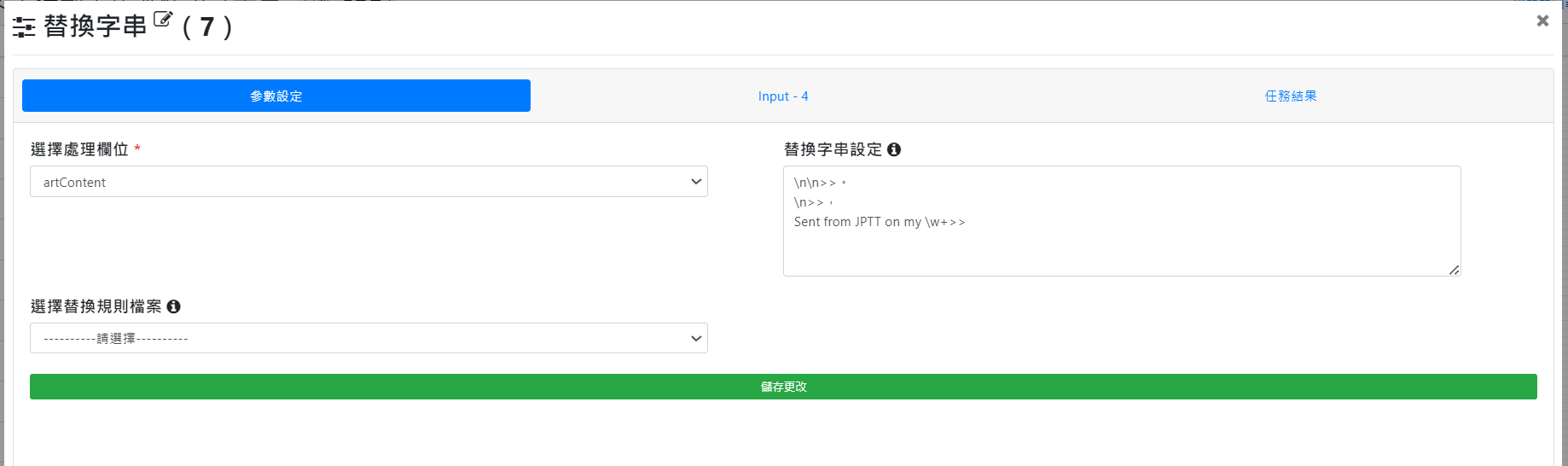


本次取聯合新聞網看板「股市、運動、文教」

期間為2023/12/01-2023/12/15



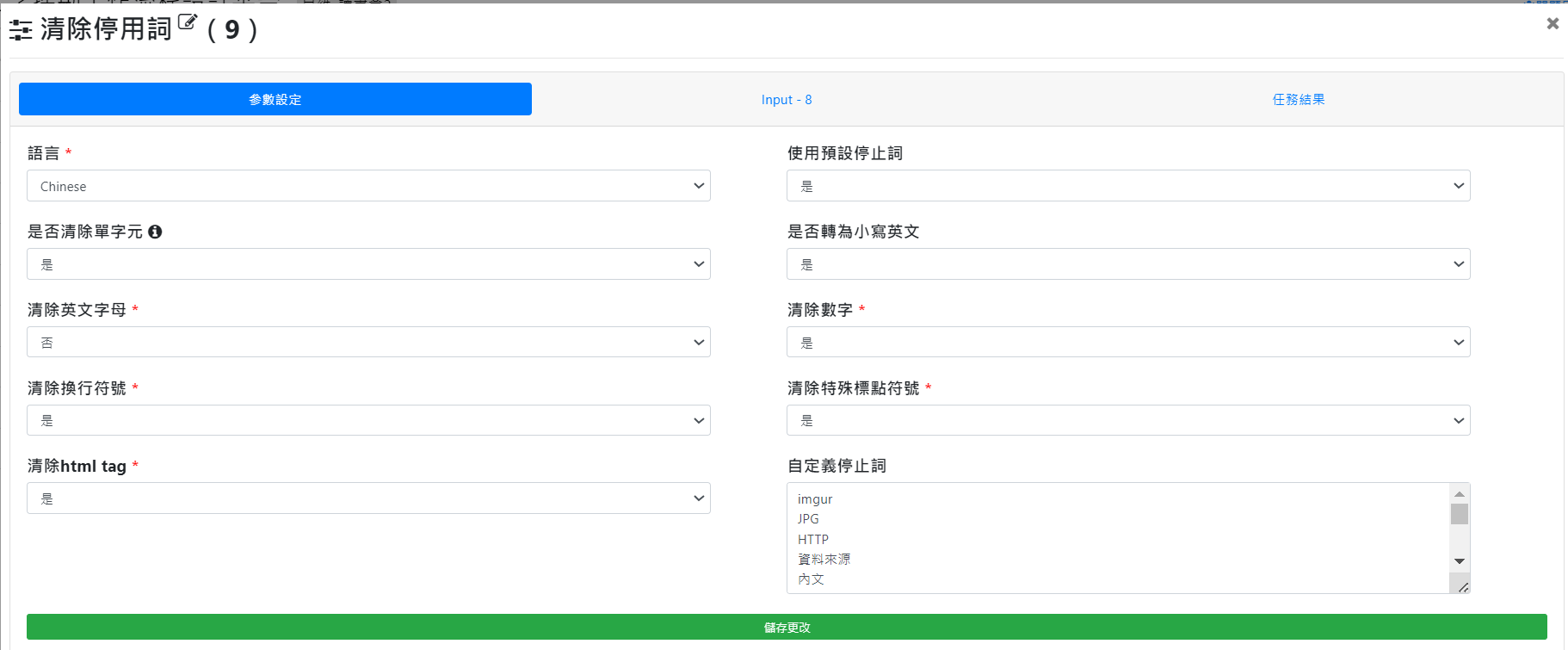
替換字串



中文斷詞

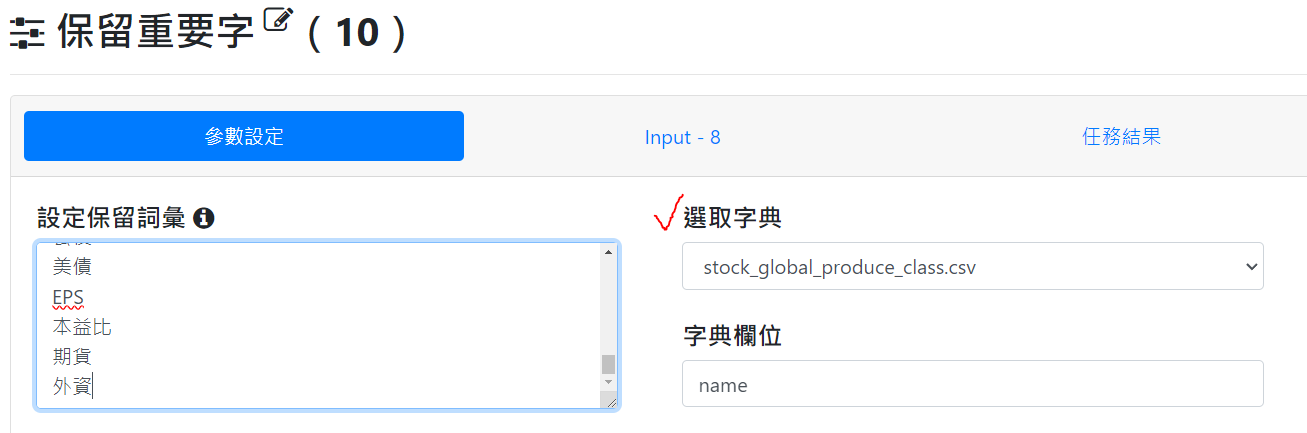


清除停用詞

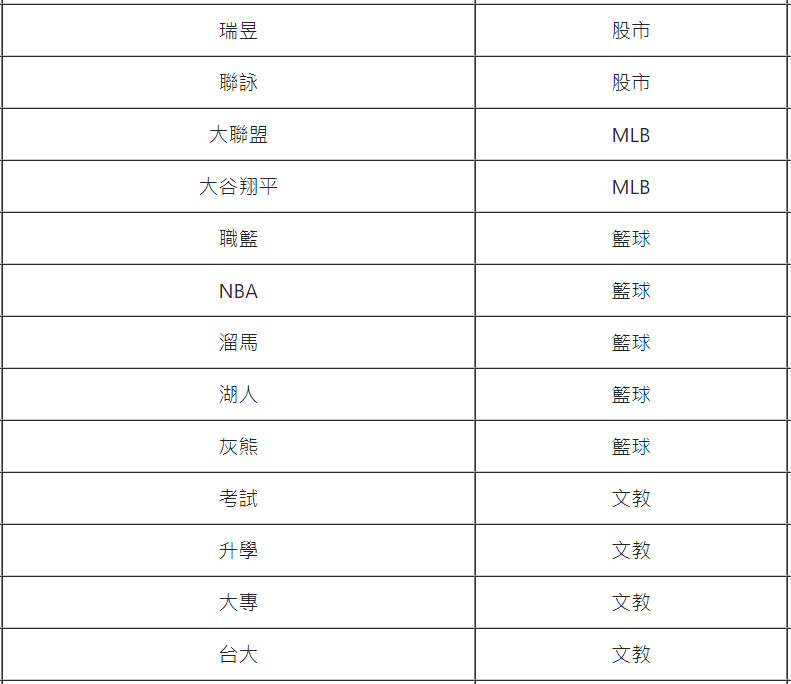


替換字串、中文斷詞、清除停用詞和以前的做法一樣，這次並沒有做什麼而外的特別處理

針對保留重要字部分，我們選擇了Stock\_global\_produce\_Class.csv的自創字典，讓一些關鍵字可以更精準的被分別出來，使用在我們所分析的文本當中。

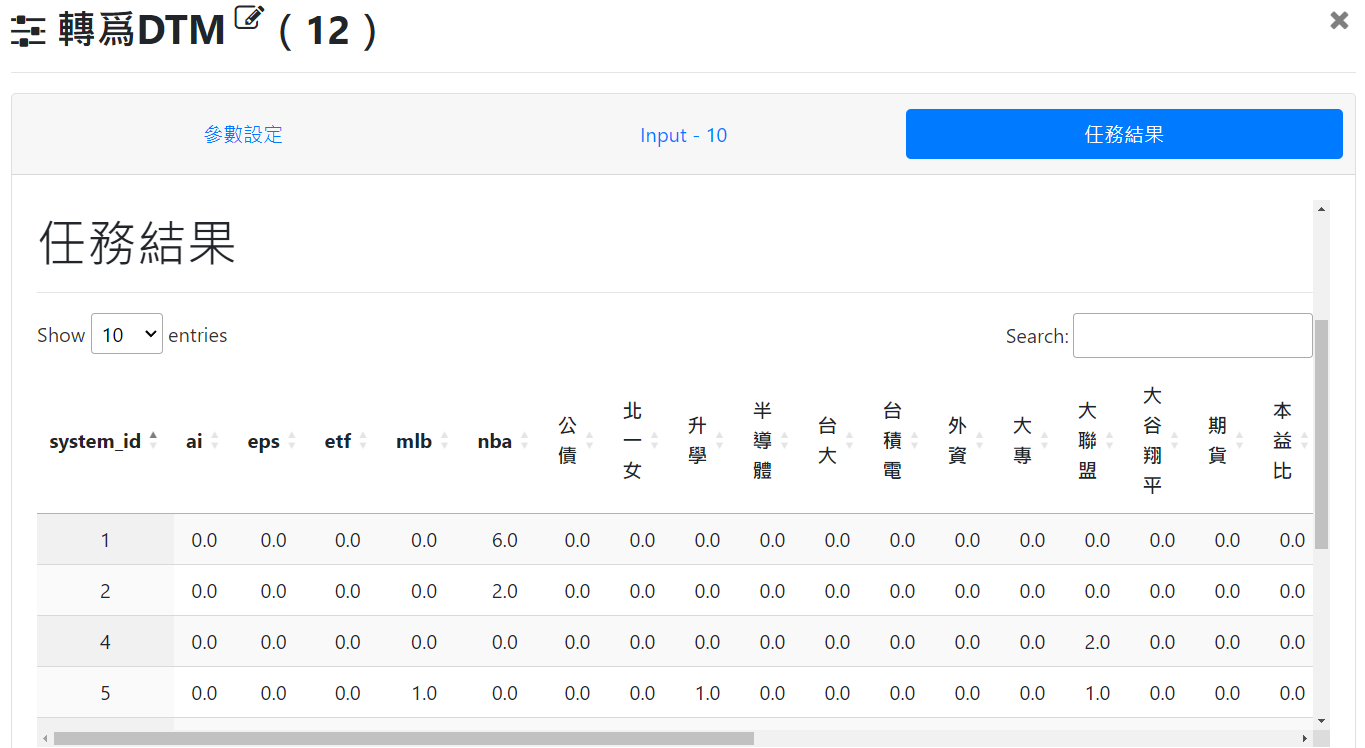


使用字典，字典內部的文字定義分類資訊如下：

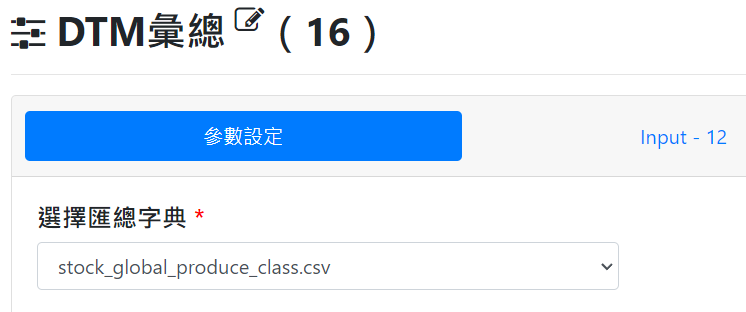


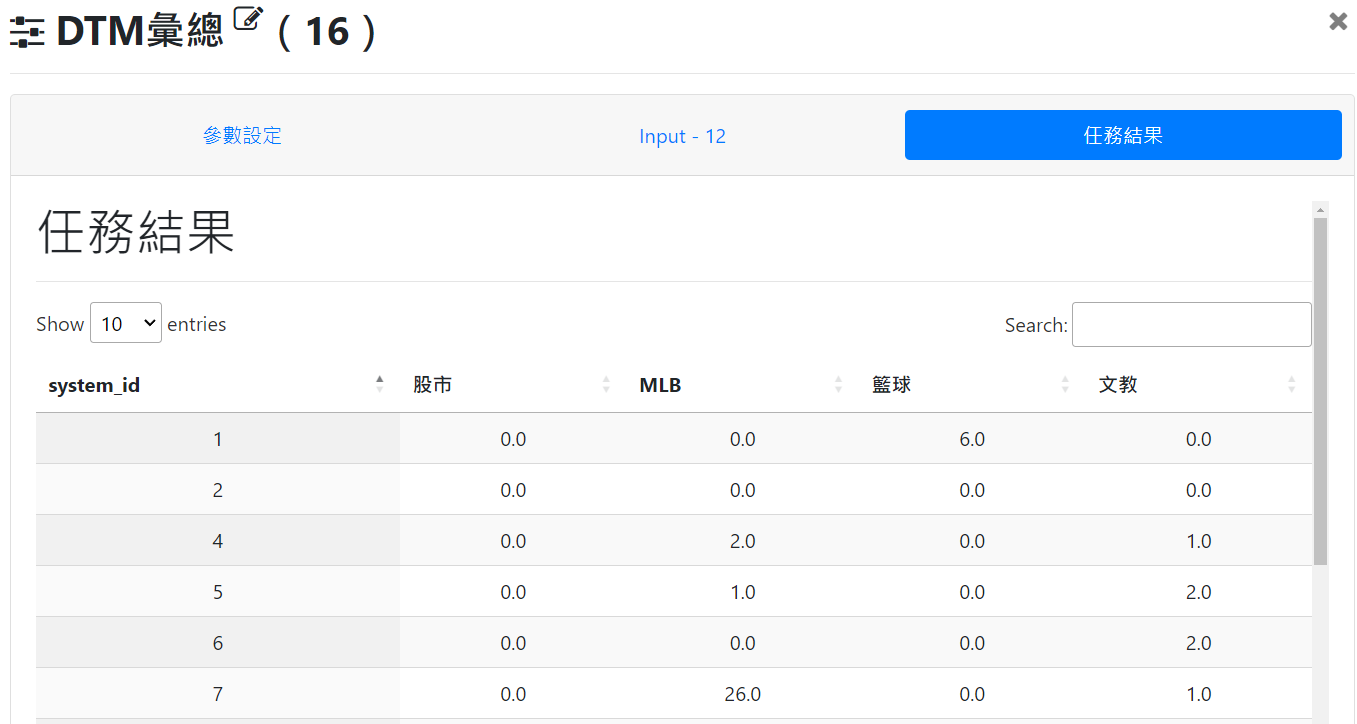


轉為DTM



DTM彙總



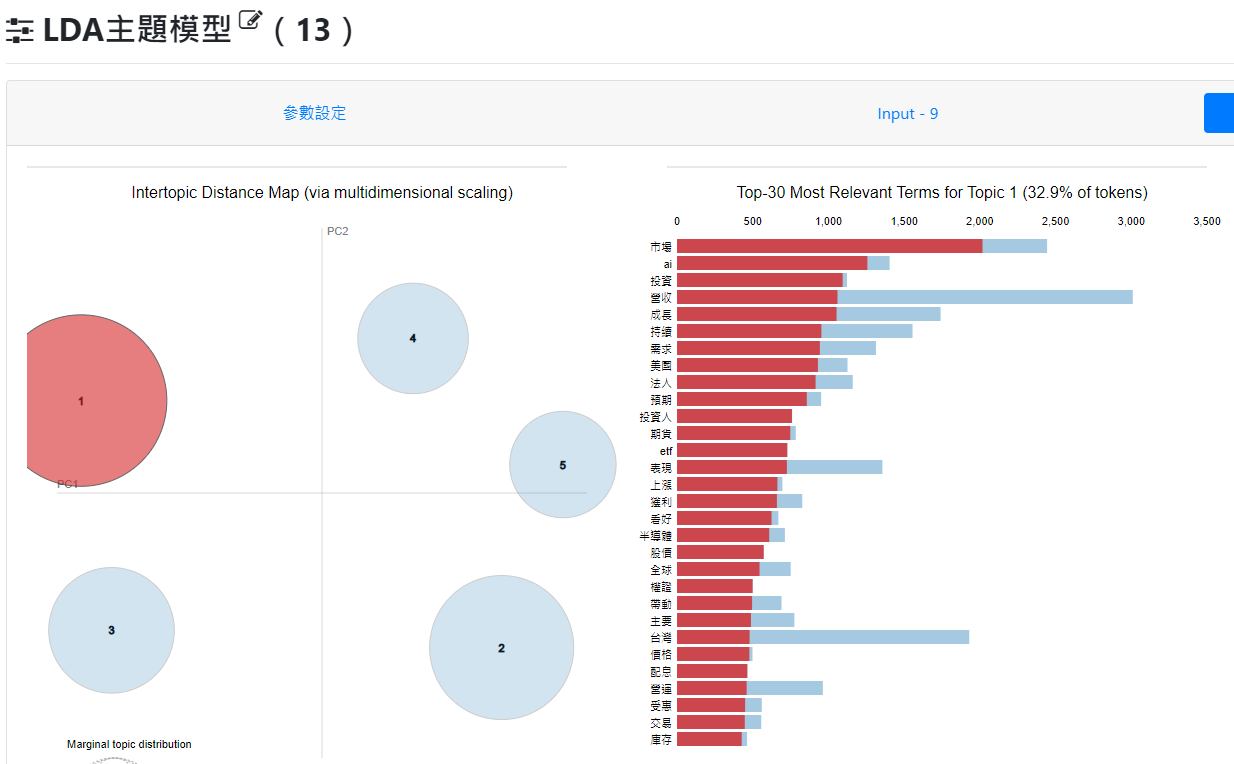


LDA 主題模型，

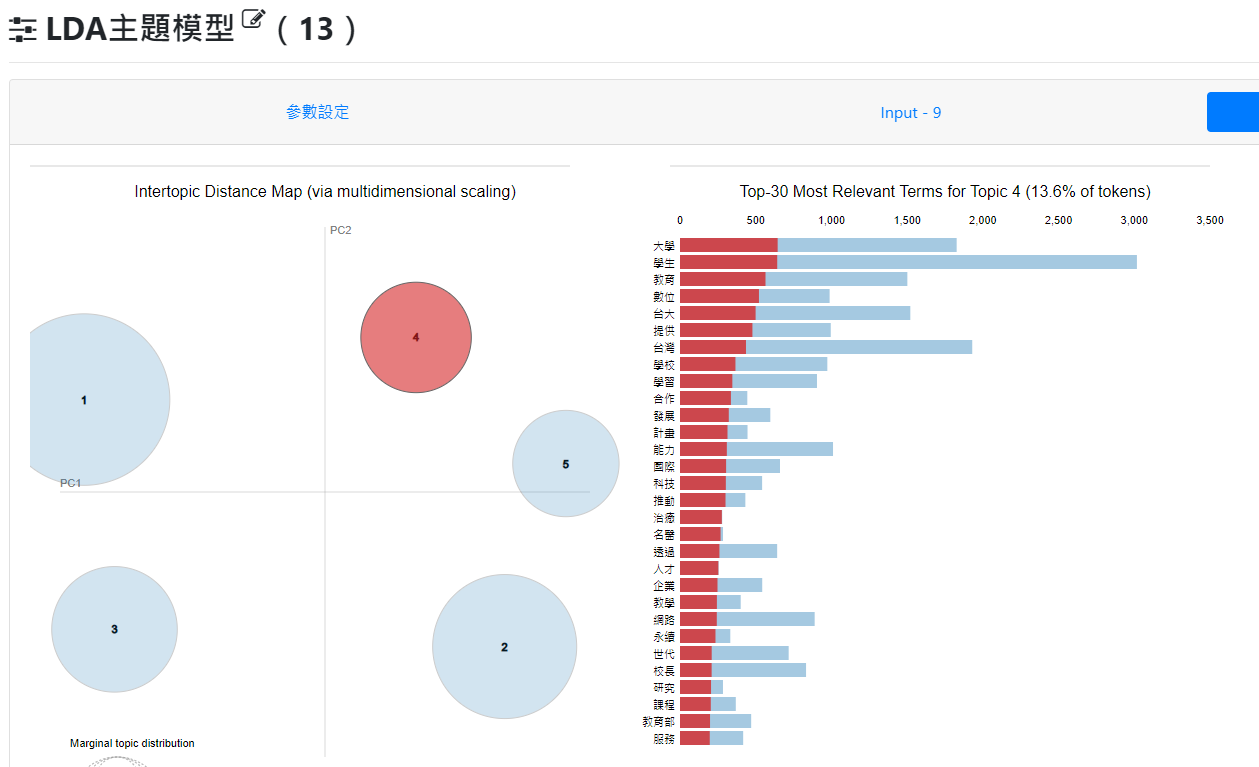
主題數=5；詞彙頻率下限=20；主題保留關鍵字數量=20；詞彙頻率上限=0.7；迭代次數設定為 500



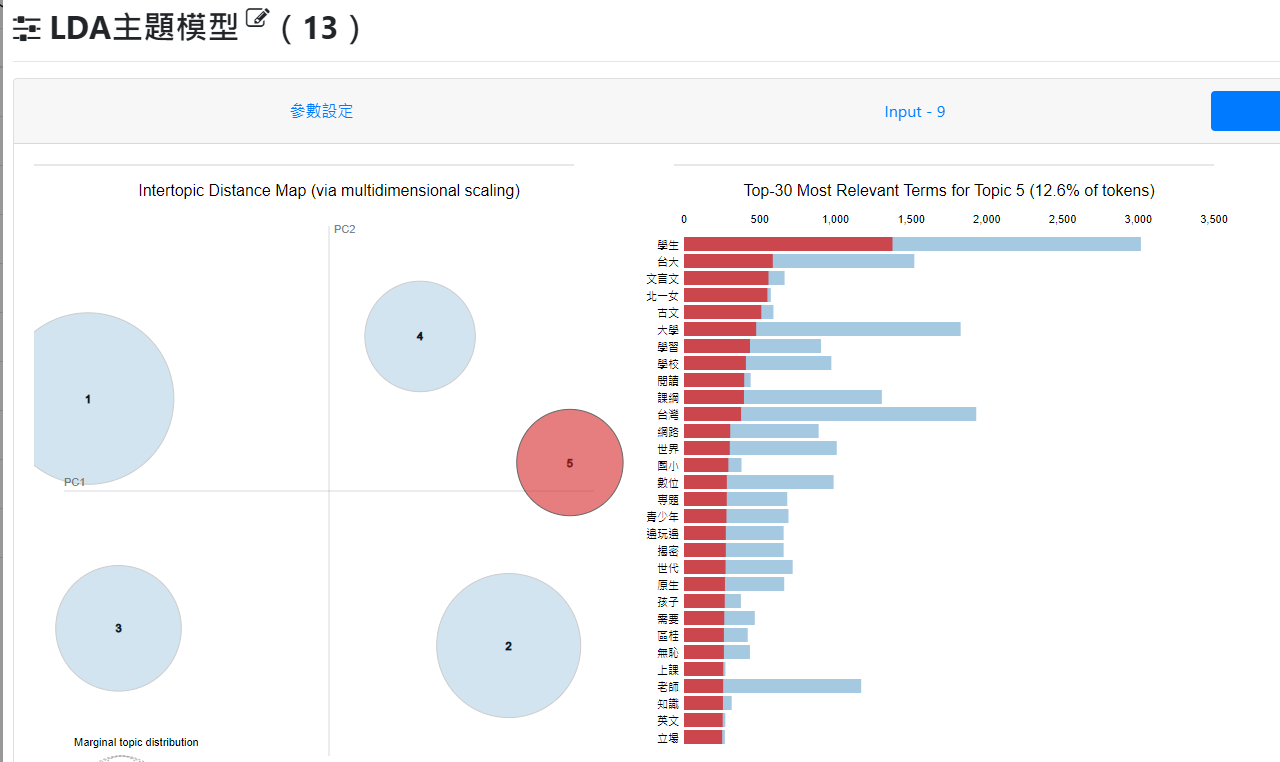
#1 主要為 股市 文本主題



#4 這是屬於 文教 文本主題

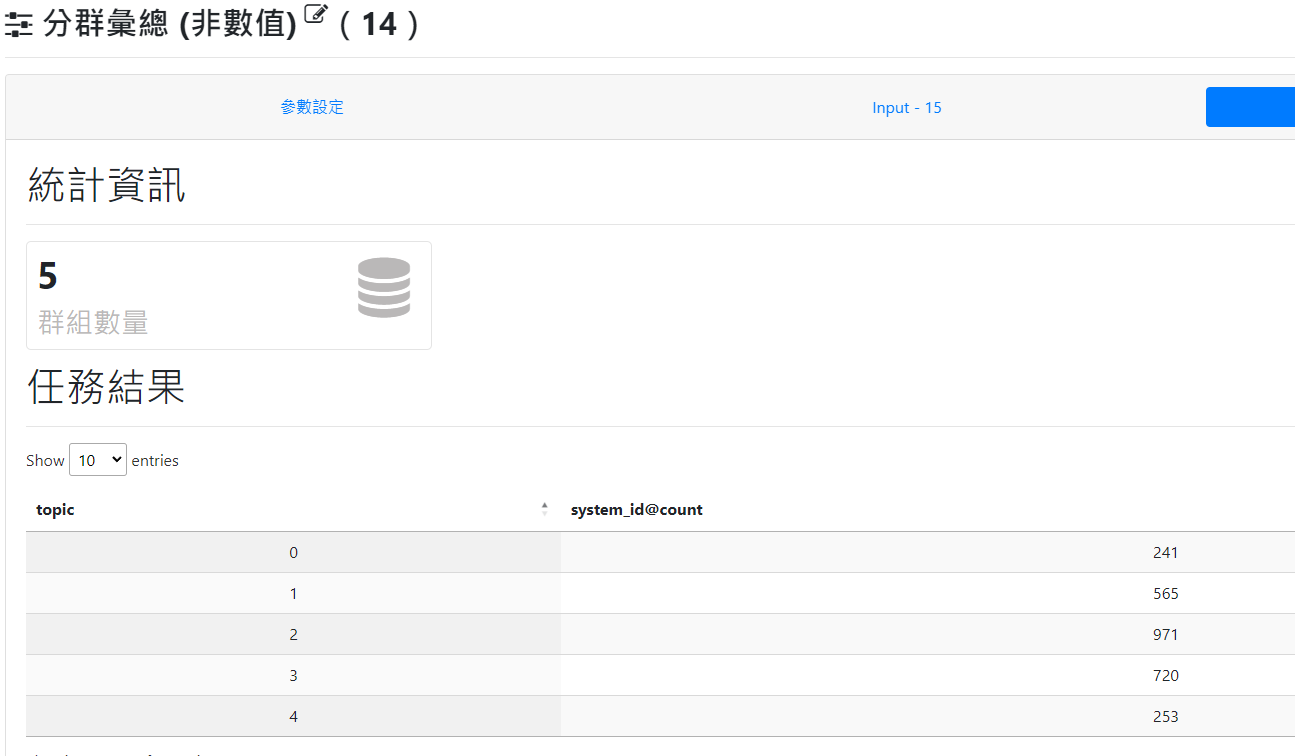


#5 這是屬於 文教 文本主題



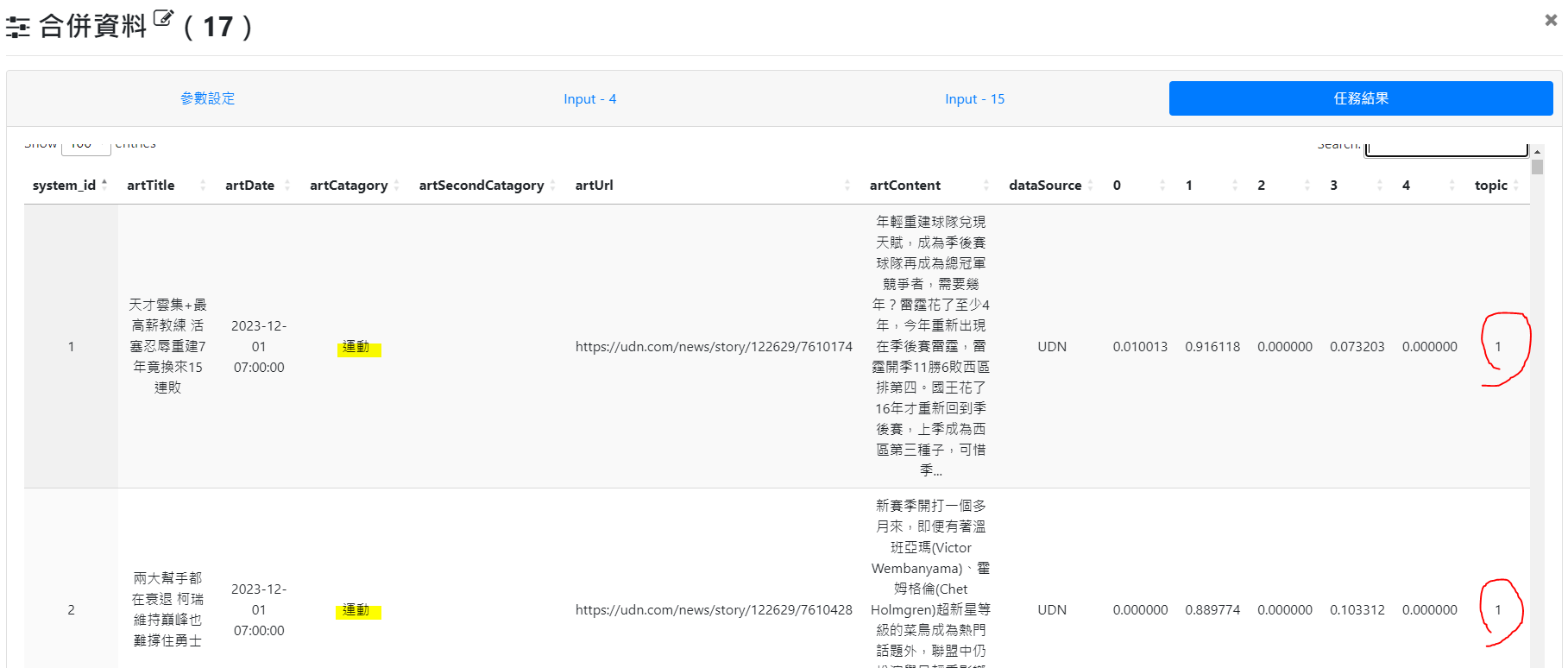
分群彙總，可以觀看各文本主題群組各有多少文章篇數

被分到 0 和 4 的類別較少，比較集中在 1 ~ 3 類



合併資料

聯合新聞網的運動版被分類為第1類



GuidedLDA主題模型

主題數=5；詞彙頻率下限=10；主題保留關鍵字數量=15；詞彙頻率上限=0.5

主題種子字=

(1)大聯盟, MLB, 大谷翔平

(2)職籃, NBA, 溜馬, 湖人, 灰熊,

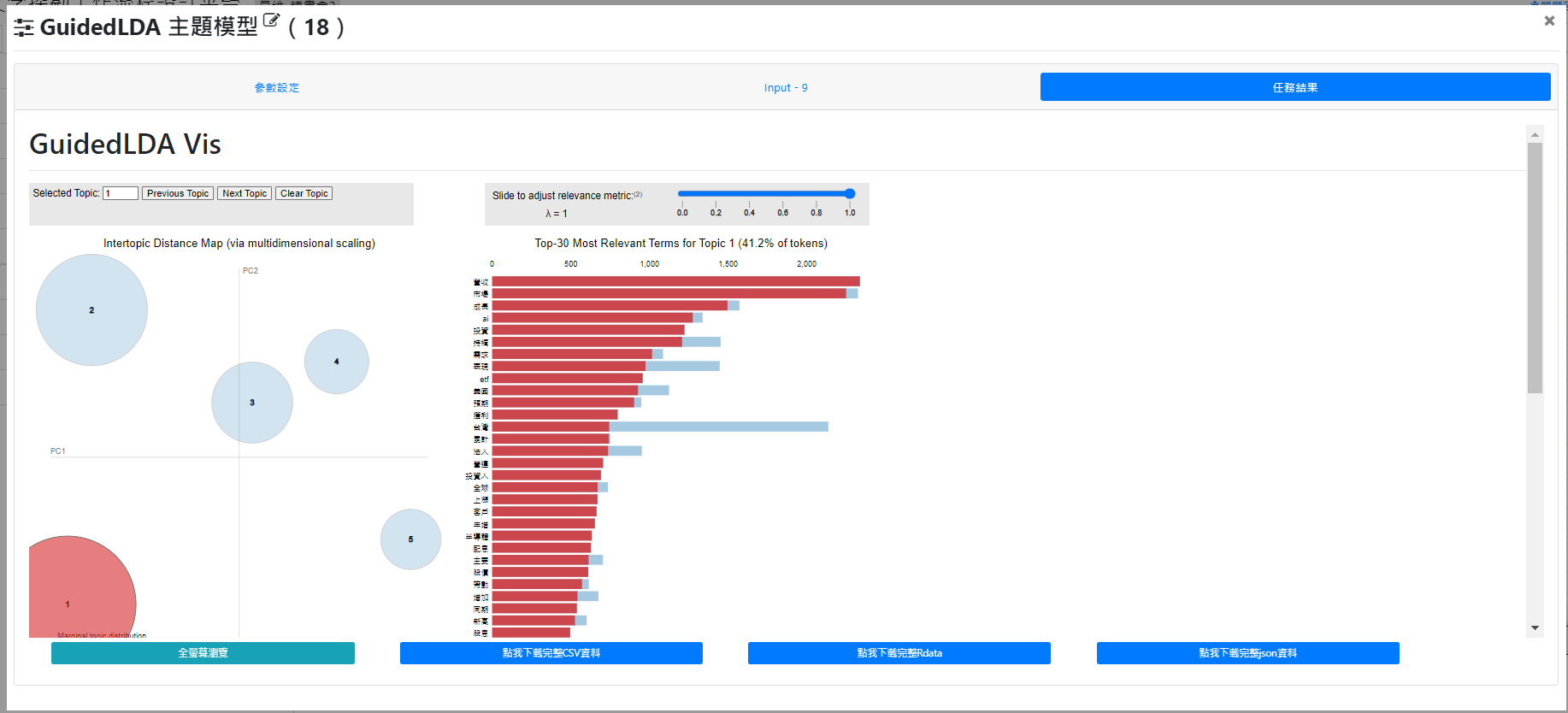
(3)考試, 升學, 大專, 台大, 北一女, 高中, 補習班, 課綱

(4)台積電, ETF, AI, 營收, 殖利率, 公債, 美債, EPS, 本益比, 期貨, 外資

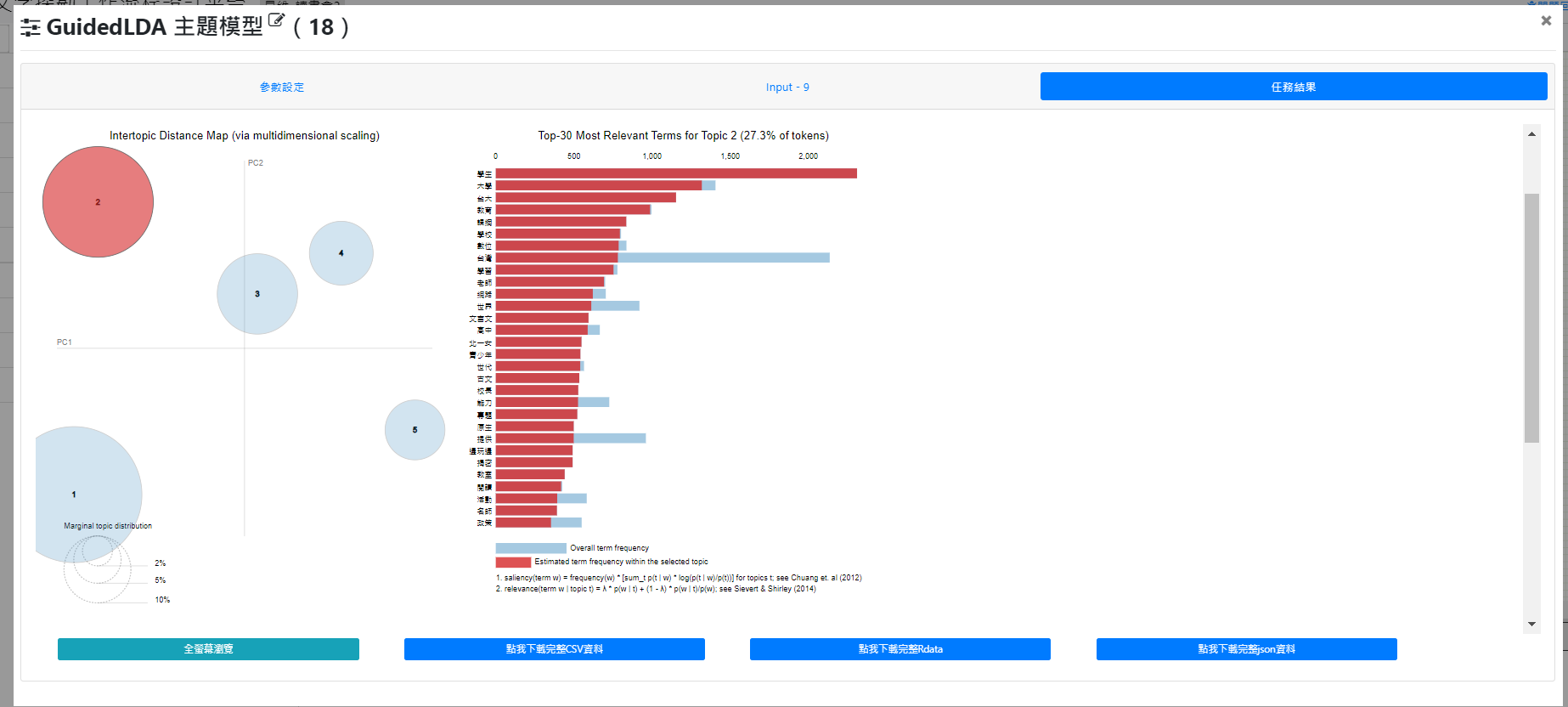


GuidedLDA 主題模型

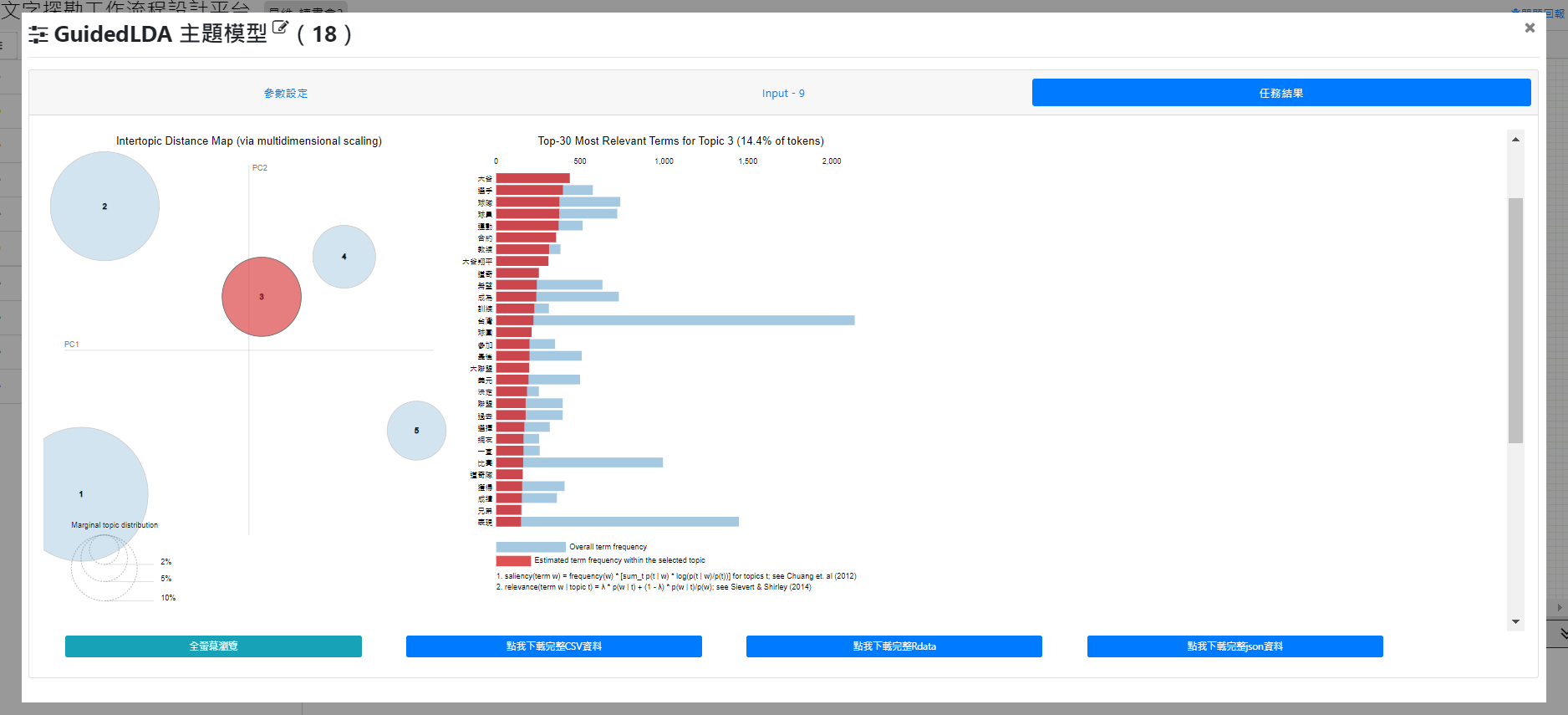
#1 這是屬於 股票 文本主題



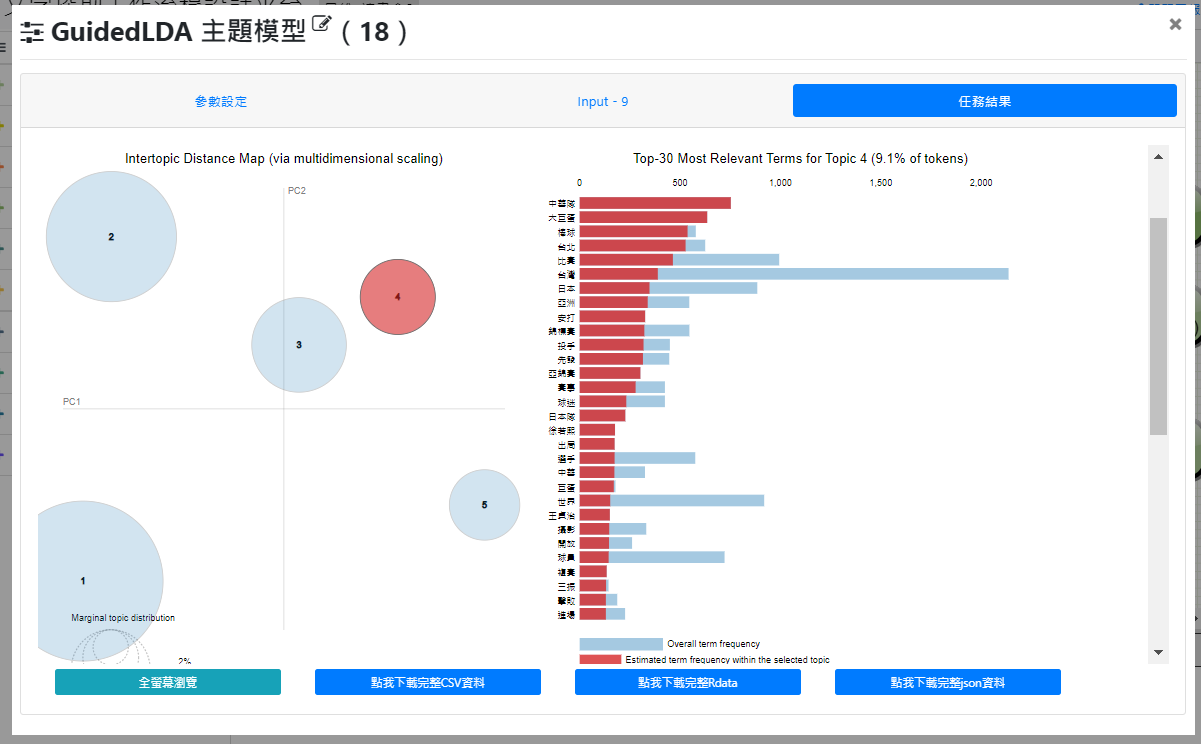
#2 這是屬於 文教 文本主題



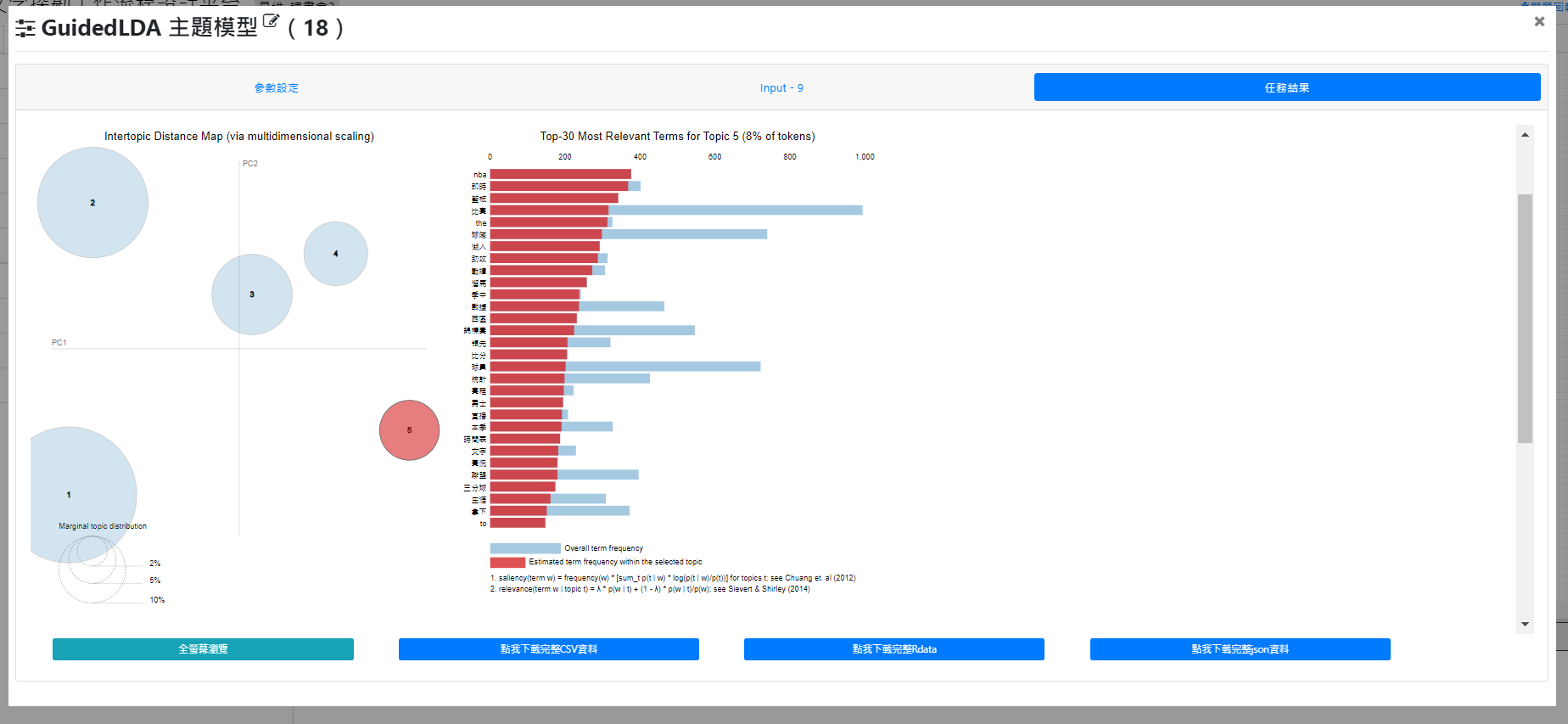
#3 這是屬於 大聯盟運動 文本主題



#4 這是屬於 中職運動 文本主題



#5 這是屬於 NBA 運動 文本主題



使用視覺化儀表板，圓餅圖畫出各文本主題所佔此次分析的比例，Topic and 值 system\_id@count

