

社群媒體分析 期末專案報告

指導老師：黃三益 教授

第二組

組員

N094220001	賀燕君
N094220017	林慶儒
N094220021	李宛珊

目錄

一、動機和分析目的	1
二、資料集描述	1
三、資料的分析過程	2
四、視覺化的分析結果與解釋	12
五、結論	19

一、動機和分析目的

分析動機：延續期中報告關於 iGroup 對於 K12 科技教育分析的需求，將範圍擴大到對最近三年(2019~2021)與教育相關的討論，以瞭解除資訊科技教育議題，民眾對教育相關的討論涵蓋哪些議題，又哪些議題討論熱度最高。

分析目的：從網友最喜歡發文和留言的 PTT 進行資料蒐集與分析，觀察社會輿論發展的趨勢，包括使用 LDA 主題模型來區分出幾個主題、繪製字詞網路圖了解關鍵詞彙相關連結、繪製社群網路圖了解教育議題的主要發言者及建立有關教育相關主題的情緒分類器。

二、資料集描述

(一)、使用工作流程平台 PTT 元件爬取 2019/01/01~2021/12/31 之自學版(learnef)與國高中版(juniorhigh)之關鍵字為:課綱、新課綱、資訊、科技、自學、遠距教學、教育、素養、老師、家長、學校、教育部等字詞，共計 742 篇文章。

(二)、將 PTT 爬取的 742 篇文章進行留言萃取，共計 1583 則留言，進行資料清理後，再使用 Lexicon_Based 情緒分析，計算每則留言的情緒值，然後與原始留言合併後匯出 csv 檔，在 Excel 程式中標註每則留言的正面、負面情緒(當作 target Y)，然後再匯入工作平台中建模。

(三)、額外分析之資料集：

由於八卦板(Gossiping)是 PTT 熱門議題的討論板，故除上述主要分析的資料集外，本組另外用上述 PTT 元件爬取前述期間八卦板

討論文章，設定的關鍵字為：課綱、新課綱、自學和遠距教學，八卦板有 2492 篇相關文章。與上述自學板與國高中板整合後一起萃取留言，共有 184701 筆留言。

三、資料的分析過程

資料分析過程於工作流程平台建立的 workflow 為 iGroup，以下為完整的工作流程：

- (一) 使用工作流程平台 PTT 元件爬取 2021/12/19~2021/12/20 之自學版與國高中版之關鍵字為:課綱、新課綱、資訊、科技、自學、遠距教學、教育、素養、老師、家長、學校、教育部等字詞，共計 742 篇文章。
- (二) 使用 PTT 元件爬取的 742 篇文章進行資料清理、轉成 DTM 及計算共現矩陣產生字詞網路圖。
- (三) 使用 LDA 主題模型元件產生 3 個主題匯出 csv 檔，與 742 篇文章進行合併然後進行分群彙總產製社會網路圖。
- (四) 使用 PTT 元件爬取的 742 篇文章，進行欄位篩選，與留言進行合併資料後，進行分群彙總，然後繪製社會網路圖。
- (五) 將 PTT 元件爬取的 742 篇文章進行留言萃取，共計 1583 則留言，進行欄位篩選，再與 Lexicon_Baed 情緒分析的情緒值進行合併資料，然後進行分群彙總，再繪製社會網路圖。
- (六) 對 1583 則留言萃取進行資料清理、轉成 DTM 及計算共現矩陣產生字詞網路圖。
- (七) 將 Lexicon_Baed 情緒分析的情緒值與留言資料合併後匯出 csv 檔，在 Excel 程式中標註每則留言的正面、負面情緒(即 target Y)，然後再匯入工作平台建模。

包括八卦板之額外資料集的分析流程：

- (一) 對清理後資料產生文字雲，並使用 LDA 主題模型元件產生 5 個主題進行分析，以及繪製社會網路圖。
- (二) 進一步衡量產生更多主題的效果，以確認是否能夠較全面涵蓋文章性質，故另外使用 LDA 主題模型元件產生 20 個主題進行分析，並分析這 20 個主題的留言數在這 2019~2021 這三年的趨勢變化(註)。

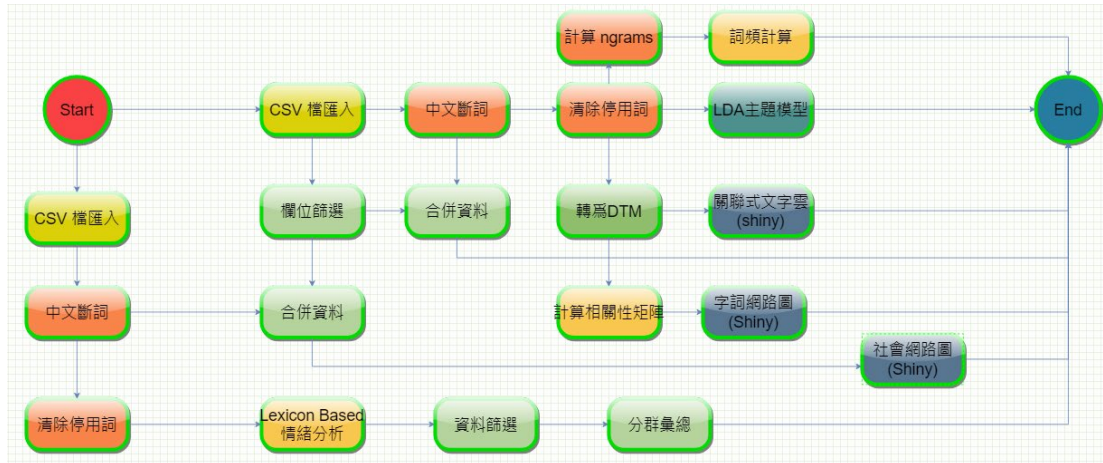
註：因為工作流程系統之數位儀錶板發生資料讀取錯誤的情形，故改用 Excel 樞紐分析圖進行分析。

以下為主要資料集分析的工作流程圖設計：

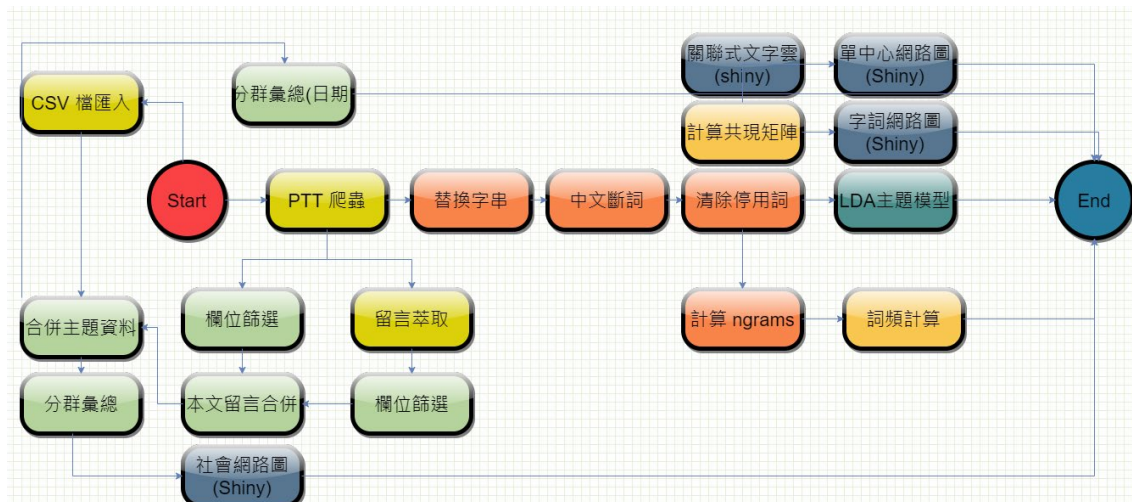


包括八卦板之額外資料集分析的工作流程圖設計：

(一)



(二)



➤ 主要資料集分析的主題模型建立與命名：

使用 LDA 主題模型元件將 PTT 的 742 篇文章分為三個主題

≡ LDA主題模型 (12)



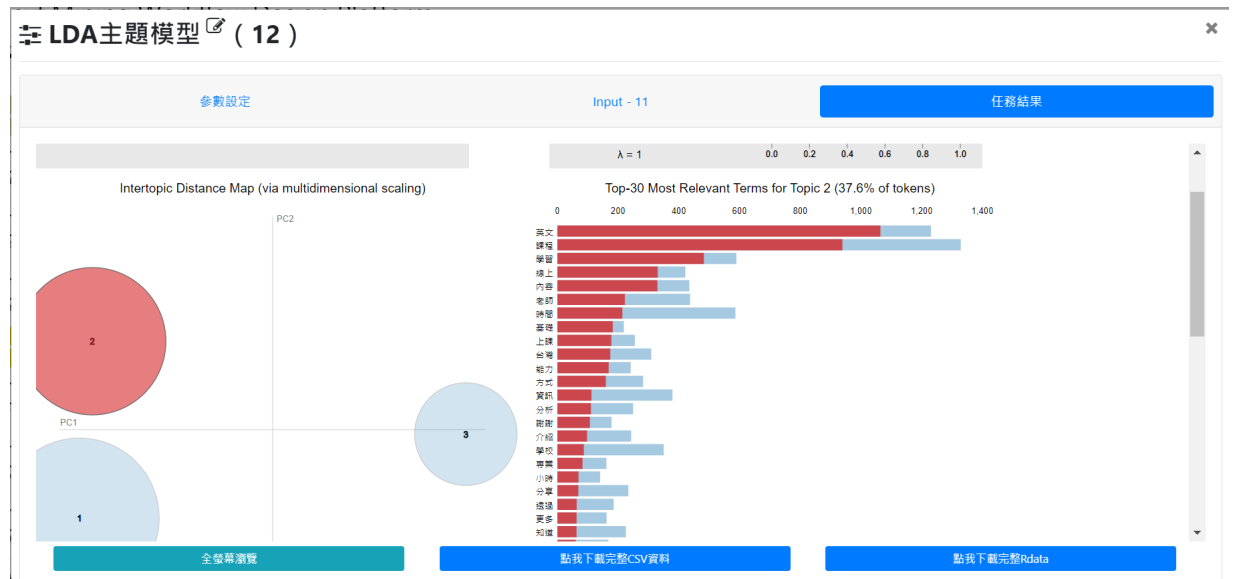
主題一：本組命名為「課程活動的報名」
(它是最多人討論的主題，可看到圈圈最大)

≡ LDA主題模型 (12)



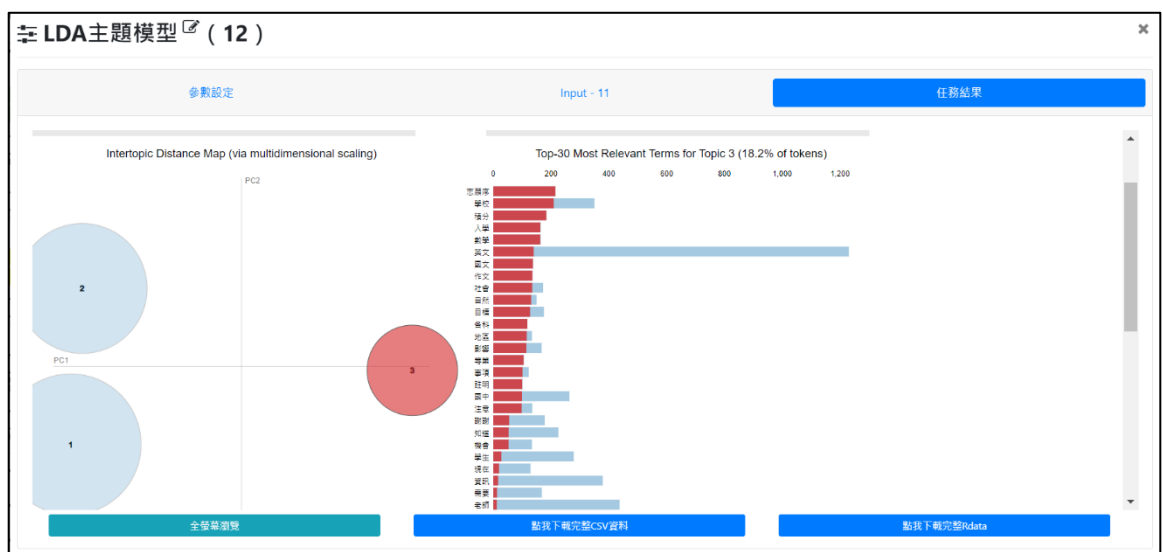
- 出現最高的前十名詞頻：報名、活動、課程、時間、講座、資訊、免費、地點、學生、老師。
- 從這些字詞可以看出民眾在乎課程是否和自己切身相關，包括時間方面，民眾主要關心是否有時間可以上課、是否免費、是否有相關活動及報名資訊等等，因此教育業者如果可以安排靈活的時間或有效率的學習平台，應該可以受到多數民眾的青睞。

主題二：本組命名為「英文學習相關討論」



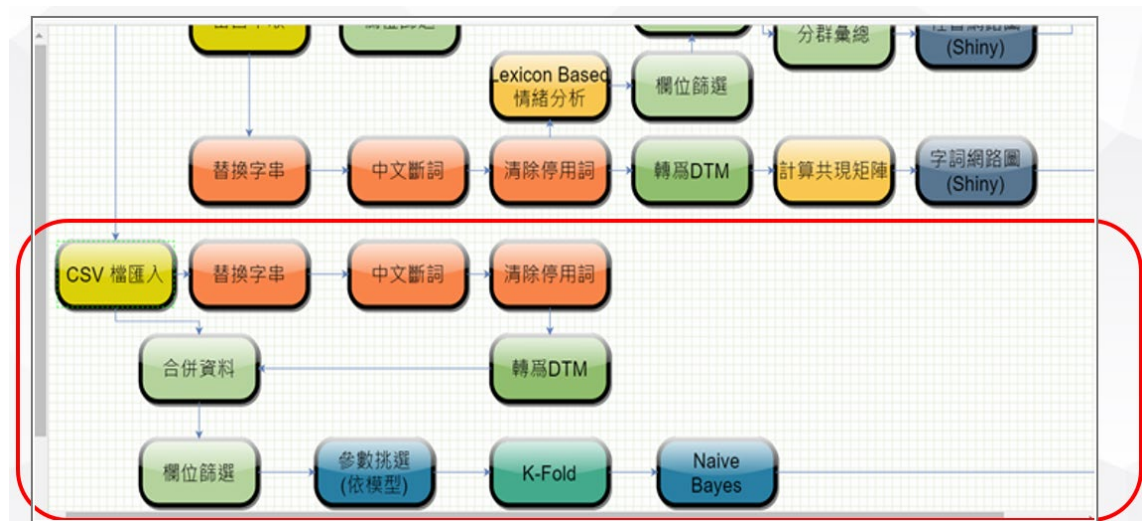
- 出現最高的前十名詞頻：英文、課程、學習、線上、內容、老師、時間、基礎、上課、台灣。
- 台灣的資訊教育政策非常強調英文能力，所以在 742 篇的文章中，英文課程的學習討論熱度一直不減，近來因為 COVID-19 的關係，線上學習的方式與內容，大家也逐漸關心及討論。

主題三：本組命名為「國高中升學考試」

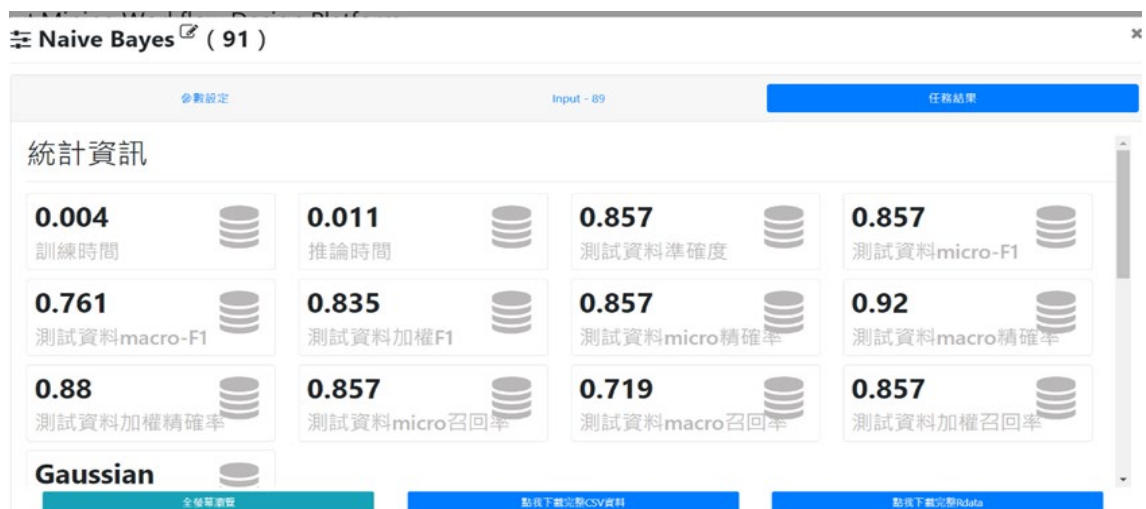


- 出現最高的前十名詞頻：出現最高的前十名詞頻：志願序、學校、積分、入學、數學、英文、作文、社會、自然。
- 根據國高中生主題的關鍵字，能不能考上理想學校、志願學校排序的填寫仍是許多學生關心的焦點。此外，推薦入學需要看學校特定學習的積分，因此教育業者若能推出促進學習積分的活動，將會受到許多學生和家長的歡迎。

➤ 接著匯入 312 則有標註正面和負面文章留言的 CSV 檔到工作流程平台建立情緒分類模型



➤ 檢視模型的統計資訊可看到這個分類模型的測試資料精確度、召回率及 F1 皆大於 80%



➤ 以下為建立好的分類預測元件

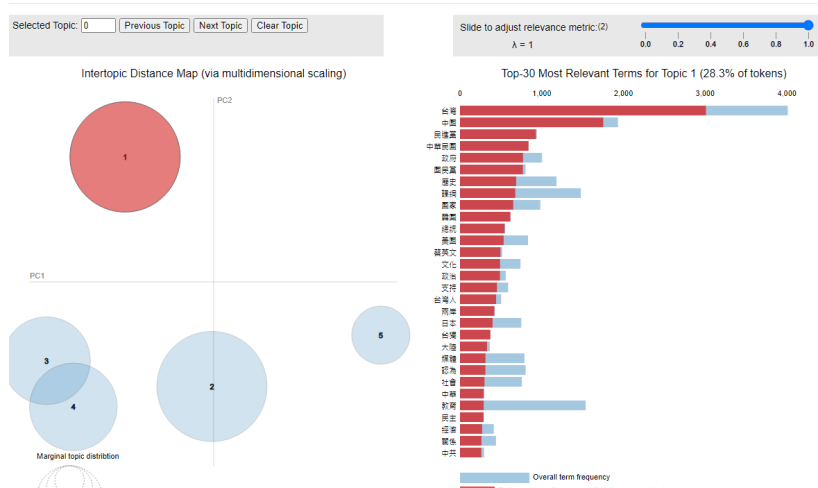


※對於含八卦版文章之額外資料集建立主題的說明：

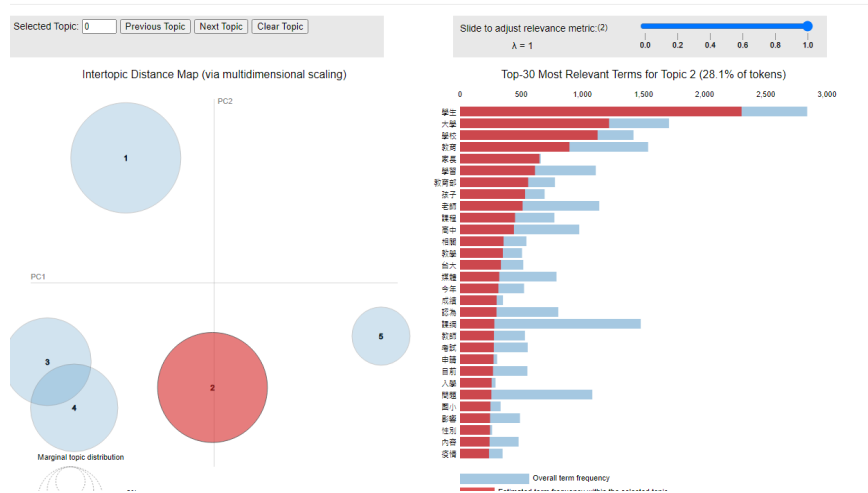
(一)首先以 LDA 主題模型元件產生五個主題，其分別為：

(1)主題一為「出現談論課綱的國家政黨新聞」

關鍵字有台灣、中華民國、民進黨、國民黨、課綱

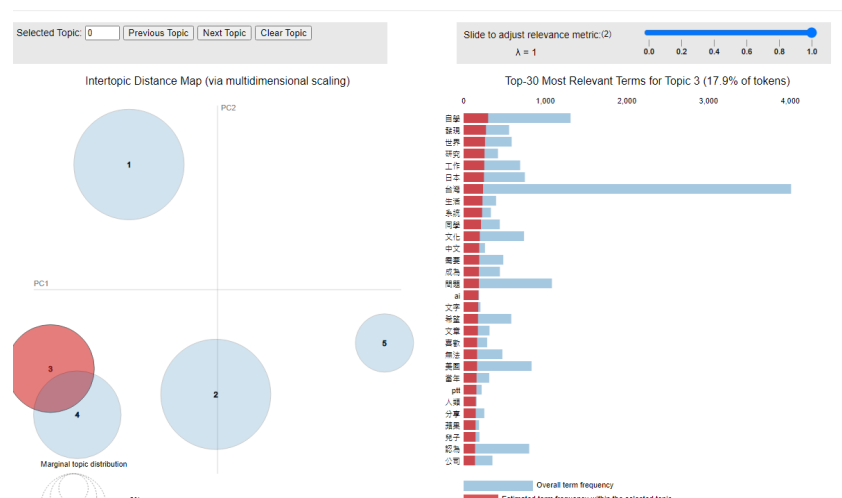


(2)主題二為「學校大學教育為主的討論群」



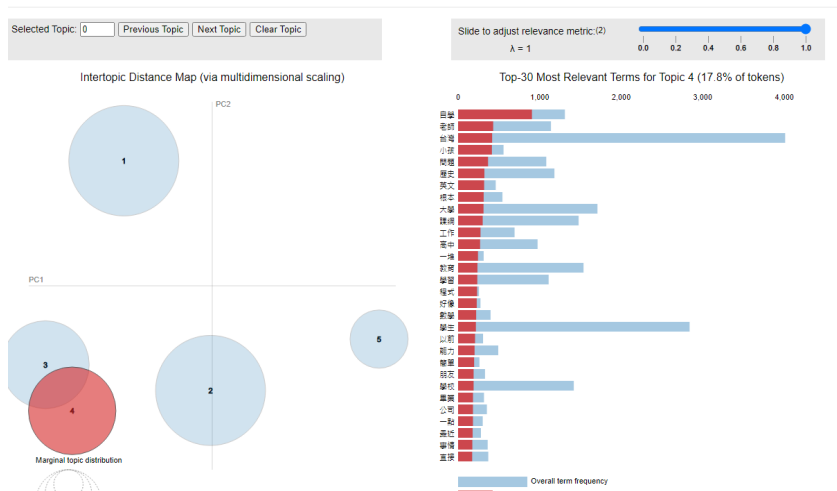
(3)主題三為「以國外自學為主的討論群」

但主要針對國外狀況進行討論，如：日本美國。



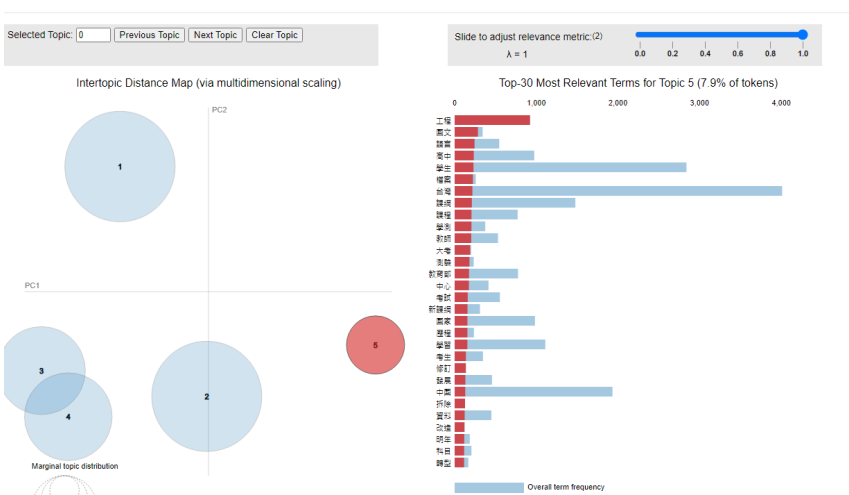
(4)主題四為「以國內 自學為主的討論群」

該群以針對台灣的課程為主，例如歷史、程式、老師等。



(5)主題五為「聚焦在課綱及課程的討論群」

例如工程、國文、語言、課綱、課程及學測。



(6)檢查統計資訊可看到 LDA 之主題連貫性與混淆度，有不錯的表現



(二)接著再以 LDA 主題模型元件產生二十個主題：

這次的主題區分可以較細的看出整題文章分布。

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Topic	Topic_Summary
工作	台灣	時間	自學	問題	朋友	需要	經濟	每天	關係	學歷	事情	投資	兩個	生活	公司	東西	機會	正常	以後	T01	自學的問題與迷惘	
老師	學生	上課	學校	同學	遠距教學	教授	教學	影片	時間	照片	家長	作業	討論	直播	問題	教室	學期	讀書	認真	T02	遠距教學上課問題	
台灣	中國	中華民國	美國	國家	大陸	兩岸	蔡英文	日本	政府	台灣	台灣	主權	民進黨	國民黨	獨立	臺灣	關係	國際	中共	T03	兩岸關係議題	
疫情	防疫	遠距教學	停課	肺炎	感染	學校	確診	台灣	病毒	指揮	口罩	措施	宣布	群眾	線上	中心	社區	民眾	三級	T04	防疫停課措施	
韓國	警方	高雄	香港	語言	線上	警察	男子	網址	教學	時間	現場	網友	前往	下午	台南	成員	大學	總統	民眾	英語	T05	提到自學的社會新聞
英文	學習	課程	語文	線上	網站	教學	網路	時間	使用	平台	數位	適合	資訊	賣場	外語	方法	方式	免費	英語	T06	英語學習	
疫苗	醫院	確診	死亡	案例	問題	醫療	病毒	醫師	狀況	調查	團體	治療	出現	台灣	國家	人員	增加	病種	造成	T07	疫苗與疫情	
高中	學生	數學	課程	國文	考試	能力	學測	成績	科目	新課綱	問題	教師	國中	老師	測驗	指導	高中生	題目	科系	T08	新課綱與學測考試	
學生	學校	教育部	教育	家長	相關	協助	校園	委員	性別	調查	團體	代表	參與	政策	會議	教育局	提供	學習	國防	T09	性別議題	
日文	文法	意見	課程	日本	教材	解讀	日語	閱讀	翻譯	現代	台灣	能力	記憶	豐富	網路	理解	價格	邏輯	精準	T10	日文學習	
民進黨	國民黨	支持	韓國	總統	台灣	立委	政治	課綱	投票	選舉	市長	政黨	兩岸	執政	年輕人	蔡英文	未來	候選人	做	T11	與課綱有關的選舉新聞	
自學	大學	教育	東西	中文	畢業	知識	簡單	台灣	台大	美國	問題	大學生	程度	不僅	肥宅	學校	社會	課綱	工程師	T12	自學與畢業出路	
歷史	文化	課綱	課本	台灣	教育	語言	教科書	台語	本土	國家	母語	中華	國語	國中	中國	中國化	新課綱	三國	中國史	T13	母語學習及歷史新課綱	
大學	學生	課程	高中	教育	學校	教師	台大	人學	申請	學習	教育部	學習歷程	學院	領域	實驗	教學	檔案	畢業	未來	T14	學習歷程準備	
政府	社會	運動	學生	自由	民主	網路	政治	女性	成為	開放	抗議	參與	公民	影響	未來	議題	系統	價值	無法	T15	社會運動與公民議題	
孩子	小孩	家長	國小	教育	學校	國中	父母	媽媽	家庭	小學	政府	公立	兒子	小朋友	環境	外省人	長大	台灣	補助	T16	國中國小教育問題	
自學	世界	人類	遊戲	影片	生活	成為	音樂	喜歡	分享	推薦	作品	興趣	工作	參加	人生	美國	獲得	比賽	活動	T17	喜歡自學與興趣分享	
公司	科技	軟體	女兒	技術	企業	電腦	自裁	工作	網路	app	使用	自學	員工	分析	研究	資訊	系統	蘋果	開發	T18	上班族自學軟體開發	
工程	台灣	中國	轉型	修訂	孫中山	拆除	發展	正義	改造	覺青	二二八	國家	菁英	國民黨	國歌	真相	加入	銅像	國父	T19	轉型正義議題	
學生	大學	中心	遠距教學	考生	台大	考試	教師	大考	學測	學校	師生	校方	公告	教育部	自主	實施	台北	宿舍	健康	T20	防疫相關的學校考試公告	

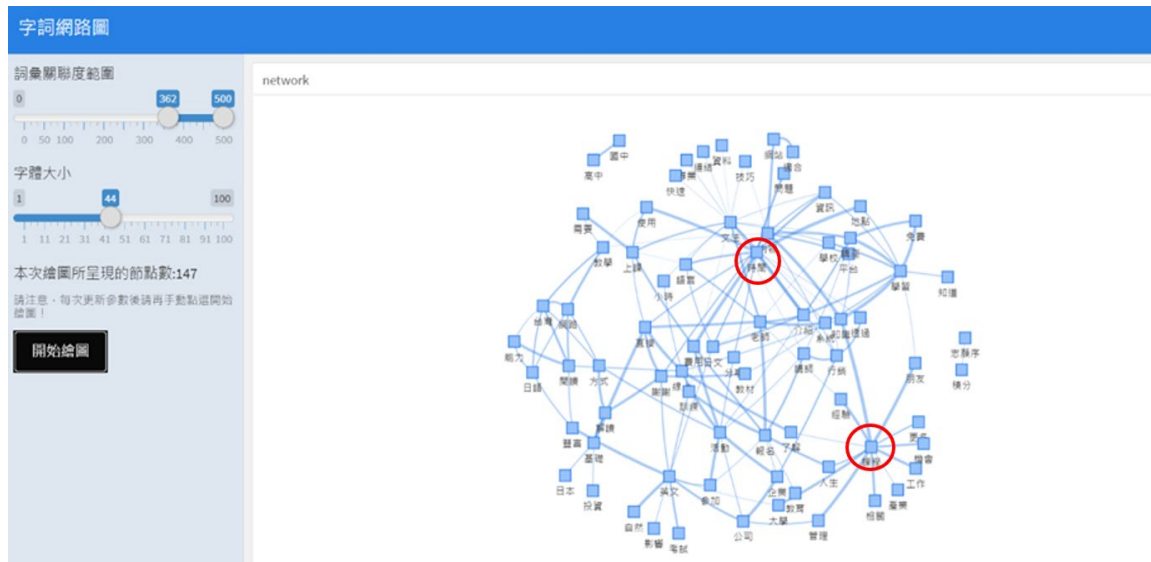
經過以上每個主題前 20 個關鍵字的判斷，所命名的主題分別是：

Topic	Topic_Summary
T01	自學的問題與迷惘
T02	遠距教學上課問題
T03	兩岸關係議題
T04	防疫停課措施
T05	提到自學的社會新聞
T06	英語學習
T07	疫苗與疫情
T08	新課綱與學測考試
T09	性別議題
T10	日文學習
T11	與課綱有關的選舉新聞
T12	自學與畢業出路
T13	母語學習及歷史新課綱
T14	學習歷程準備
T15	社會運動與公民議題
T16	國中國小教育問題
T17	喜歡自學與興趣分享
T18	上班族自學軟體開發
T19	轉型正義議題
T20	防疫相關的學校考試公告

四、視覺化的分析結果與解釋

以下為主要資料集的視覺化分析結果：

- 從 PTT 擷取之 742 篇文章的字詞網路圖



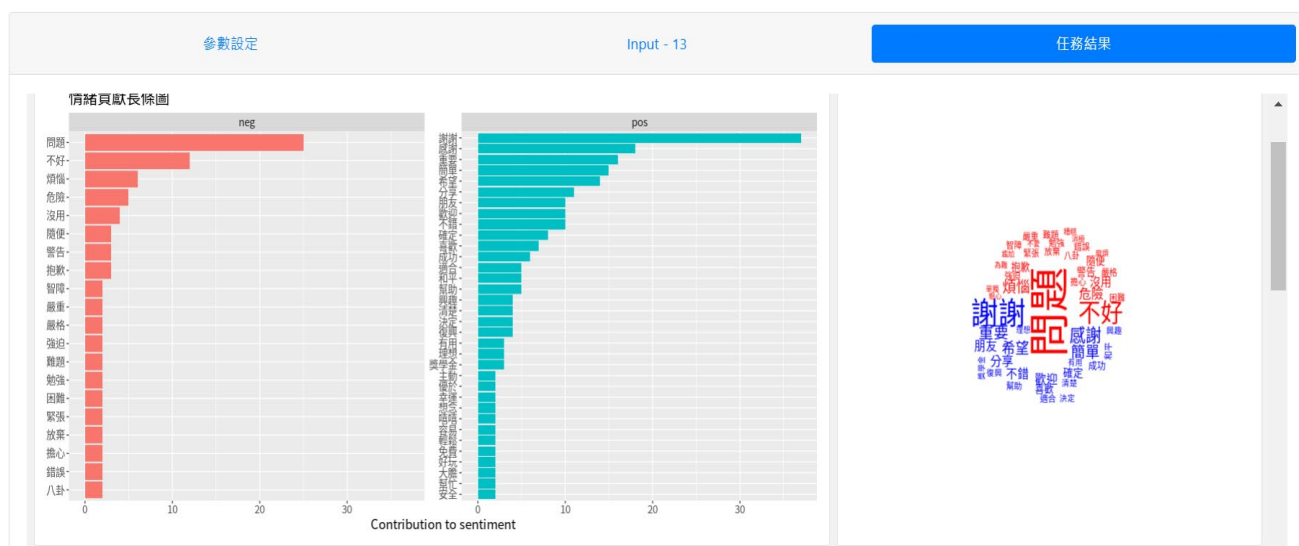
從字詞網路可以看到"課程"和"時間"串聯最多的關聯詞，表示民眾在乎課程是否和自己切身相關，包括連結到工作、產業、單位，以及自己是否有基礎或經驗來參加這個課程。

- 接著繪製社會網路分析，其中網路節點的定義與邊的權重說明如下：
 - 網路節點分別是發文者(artPoster)和留言者(cmdPoster)。留言者為來源節點，發文者為目標節點，
 - 邊的權重則分別依照三個主題(LDA 元件產生的 3 個主題)文章數量、發文數量及情緒值為正(大於 0)的留言數量

➤ 關於留言的情緒分析

從 Lexicon_Based 情緒分析元件繪製的視覺化圖形，看得出來網友留言的正面情緒是大於負面情緒。

≡ Lexicon Based 情緒分析 (19)



※對於含八卦版文章之額外資料集的視覺分析說明：

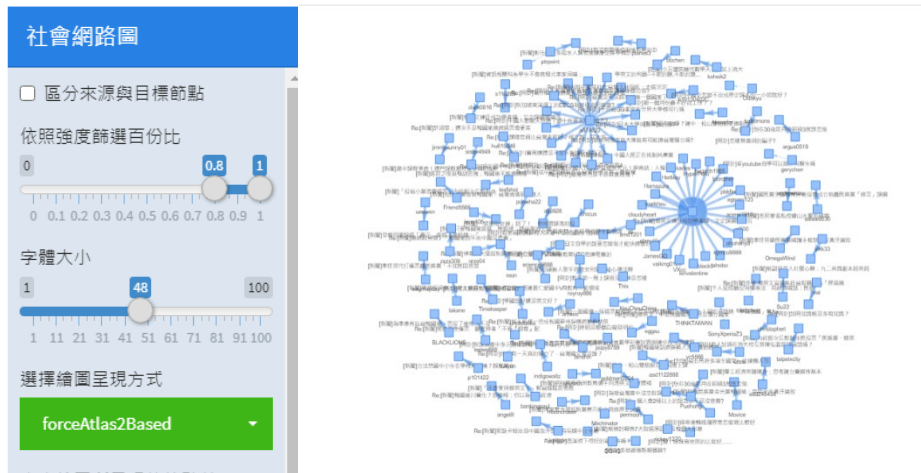
1. 文字雲詞頻計算結果：

可看到「學習 歷程」、「大考 中心」都屬於熱門字詞，顯示國家大考中心發布內容，以及學習歷程如何準備是民眾會去發文討論的焦點，

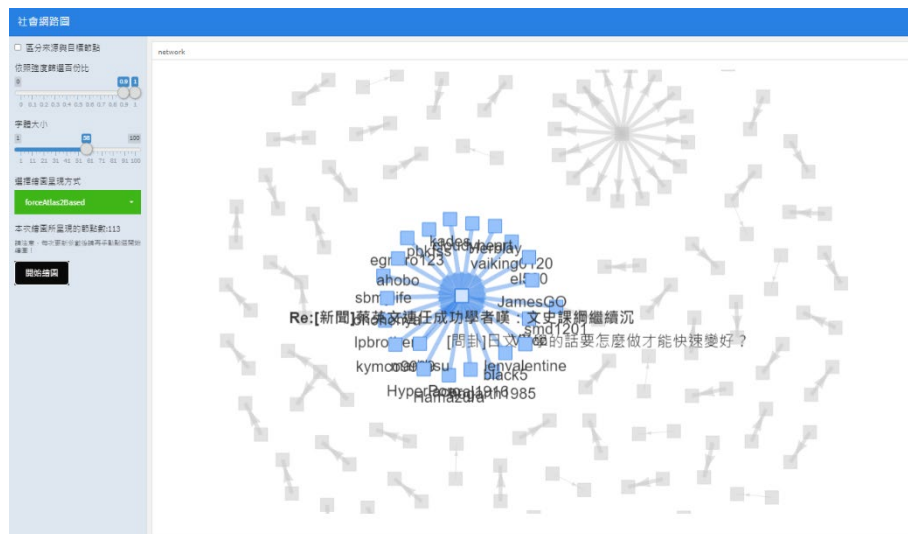
≡ 詞頻計算 (138)



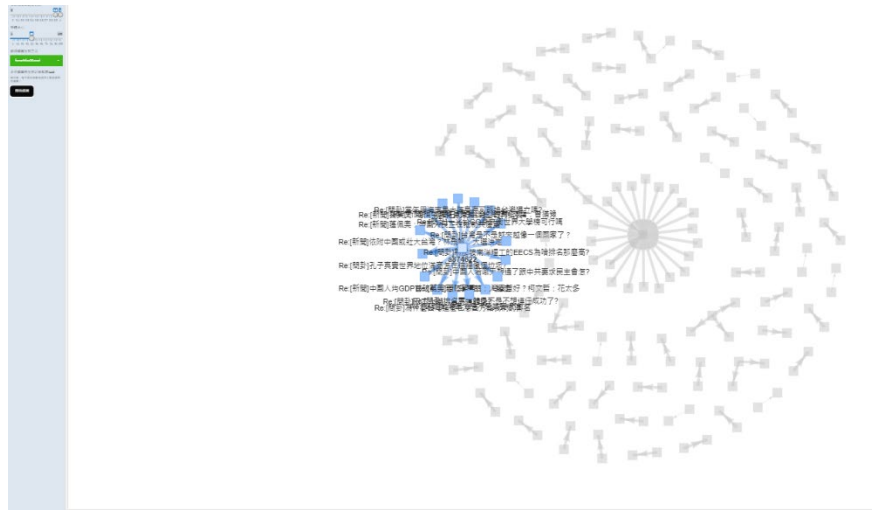
2. 社會網路圖：
設定強度為 0.8-1 後，可得到兩個主要的社會網路圖



- (1) 第一個是以討論蔡英文連任成功, 文史課綱的走向為主



(2)第二群是以 A74522 發文者為主,延伸而出的討論,針對該發文者的留言回應,有部分跳出政黨議題及做比較深入的討論。

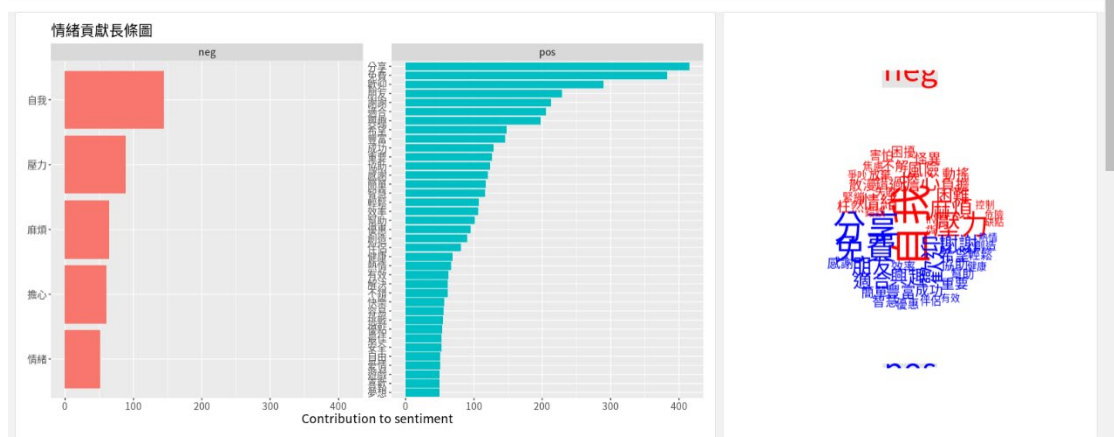


3. 情緒分析的結果

從討論文章的情緒分析可看到自學或學習並不總會帶來快樂，「快樂」在正向情緒排名相當後面，但自學有成果後，許多人會樂於分享，可看到「分享」在正向情緒排名第一。

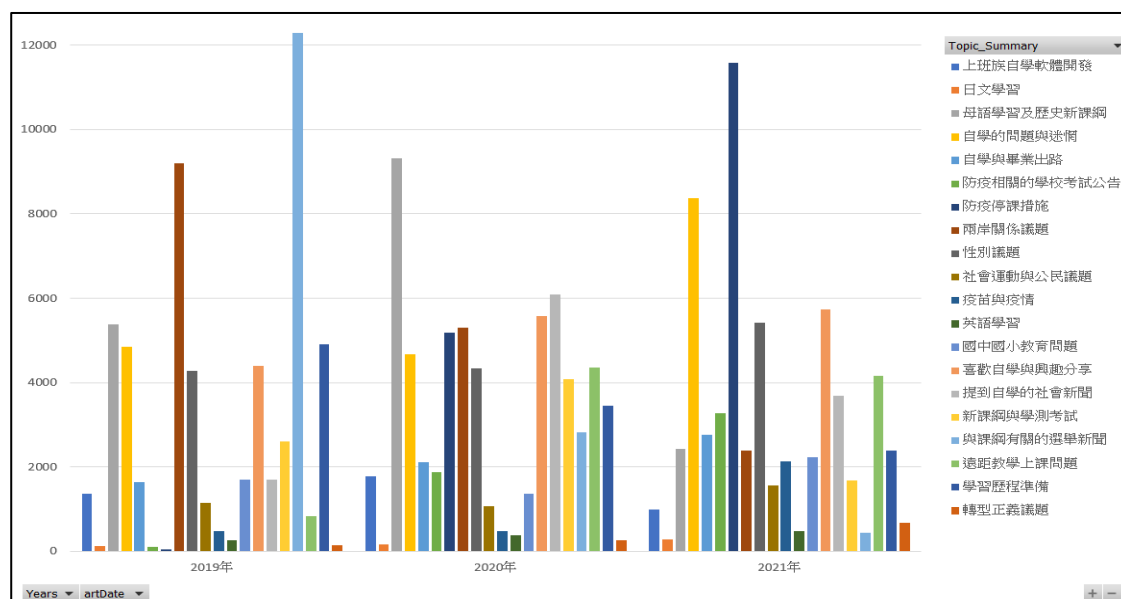
而負面情緒先排開和懷疑或苦惱的自我情緒詞彙，許多人在學習上普遍的情緒詞彙是「壓力」，可能自於考試或工作單位。

視覺化

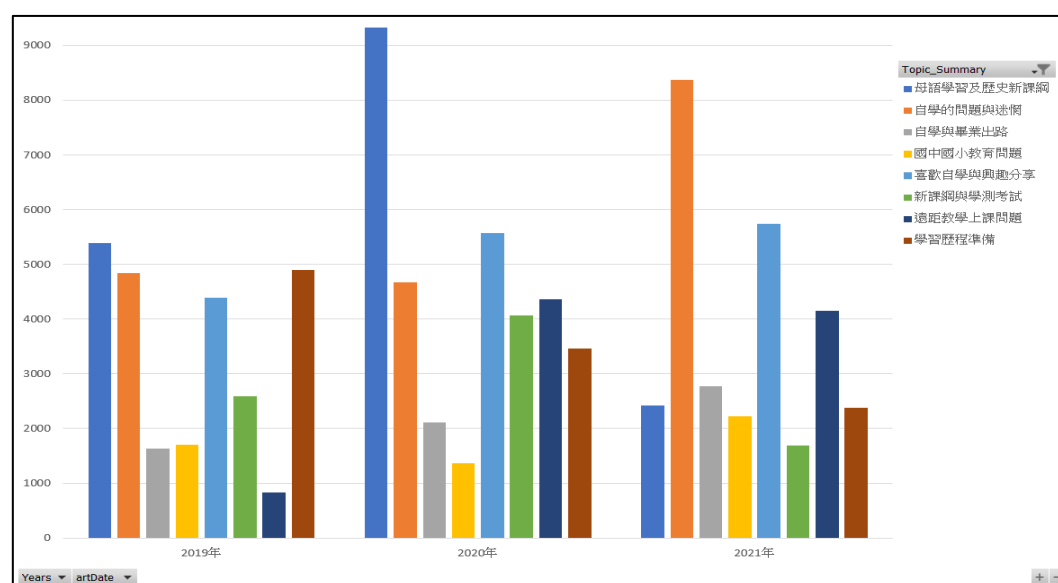


4. 使用樞紐分析圖分析 20 個主題三個年度留言數彙總

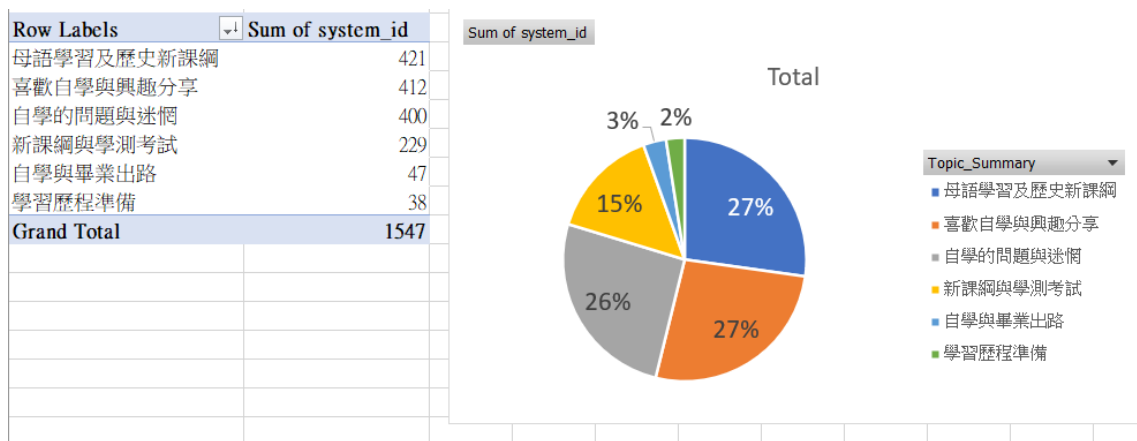
由於 2019 年剛好遇到選舉，也是新課綱提出的年度，所以與課綱相關的政黨選舉新聞留言數最高，次高為兩岸關係，第三高則是母語學習與歷史新課綱。



排除和教育不攸關的議題，可看到母語學習與歷史新課綱在 2020 討論度增加(有網路名人發言特定歷史為什麼沒有教)，而可能是疫情依舊持續，自學的問題與迷惘這個主題在 2021 年增加討論。



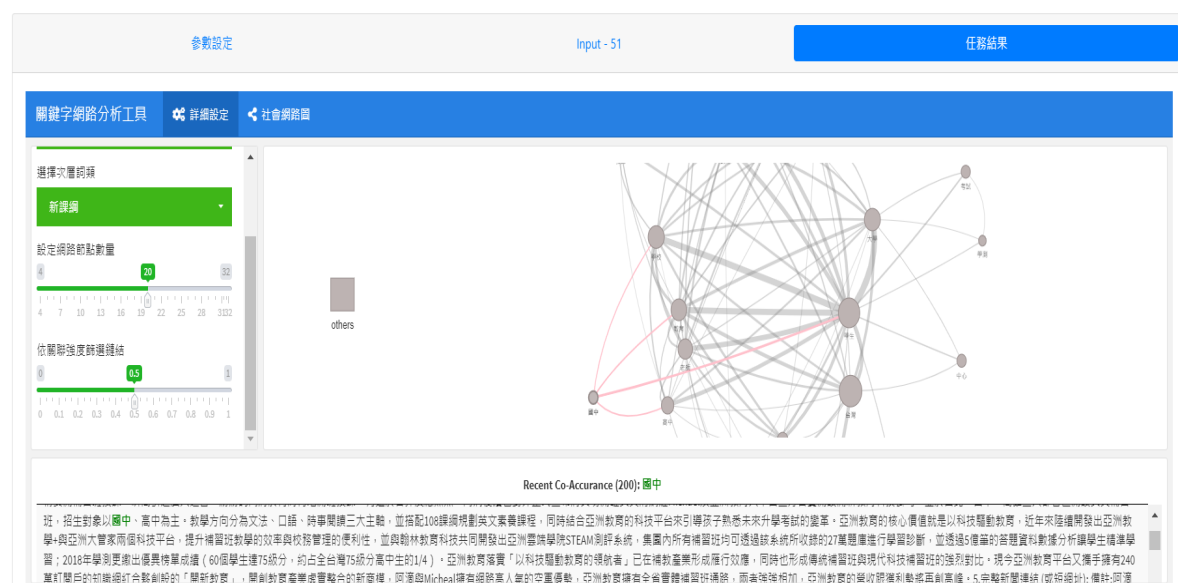
另外，針對 2019~2021 年留言數超過 30 的文章，並排除選舉、社會新聞等較不攸關的議題，發現母語學習與歷史新課綱、喜歡自學及興趣分享、以及自學的問題與迷惘為前三大討論的主題。



5. 單中心網路圖之分析結果

在新課綱與國中相關且討論度高的新聞裡面，可看到網紅與教科書業者翰林成立教學平台的新聞是民眾會想要討論與關心的消息。

單中心網路圖 (Shiny) (57)



五、結論

- 教育相關議題不外乎報名的課程、活動、時間和老師等等資訊，其中還包括是否有提供免費課程和相關費用，就網友關心的議題，經營資訊教育業者應提供方便的報名方式，並常常舉辦活動，將課程相關資訊在網路上發布。
- 資訊科技與英文學習是息息相關，從主題分析的結果來看，2019 年到 2021 年英文仍是許多人討論的焦點，。
- 國高中學生對於考試相關科目，仍有課外學習的需要，因此教育業者若能推出促進學習廣度的相關課程，將會受到許多學生和家長的歡迎。
- 另外，從字詞網路我們看到“時間”串聯許多的關聯詞，表示民眾對於課程時間的安排有彈性的需求，因此教育業者可以朝線上學習或其他有效率的學習方式進行規劃。
- 最後建議相關業者以鄉民討論熱度最高的議題調整內部營業方向，以創新的教學方法，例如設計漫畫或卡通吸引學生接受遠距教學，增加營收。

※補充關於額外分析八卦板文章的結論：

- 對於新舊課綱的討論，鄉民討論方向著重在政治因素，其普遍認為黨政因素將影響文史課綱修訂方向，未看到對真正課綱內容做深入討論。
- 對於課程的討論熱度依序為：文史、國語、語言、英文，之後才有程式、AI 等出現，顯見鄉民關心的仍以傳統相關科目為主。
- 母語學習與歷史新課綱是 2019~2020 許多人討論的焦點。許多民眾開始關心台語和新住民語的學習，並且也開始正視與討論哪些歷史值不值得納入教材或從教材中刪除。更多人的討論與參與有助於未來相關教育業者對教材內容設計考慮得更細，以產生更好的教材，值得令人期待未來教育的發展。

- 疫情持續延續使得民眾對目前傳統教育不肯定，而讓“自學的問題與迷惘”這個主題的相關文章在 2021 年討論度開始提高。許多人會開始評估需不需要花錢投入自學等涉及個人經濟的問題。若教育業者能提供安定人心的學習解方，將有助於相關課程的推廣。