**社群媒體分析期中專案報告**

主題：線上教育的趨勢分析

組員：

N074220001鍾承玫

N094220005謝宜蓁

N094220010吳貞儀

N094220016林映汝

1. **研究問題**
2. 線上教育是否隨著科技趨勢發展，逐漸成為新穎的學習趨勢?
3. 線上教育平台使用客群是否隨著時間的推演，而有不同的變化趨勢?
4. 台灣常見的自主學習平台有什麼共同特色?
5. 使用線上自主學習平台的消費者樣貌?
6. **取用的資料集**
   1. **資料區間：2017/1/1-2018/12/31**
      * **資料來源**：使用工作流程平台內建資料PTT
      * **資料看板**：studyteacher、medstudent、graduate
      * **搜尋關鍵字**：自主學習、教育相關、語言學習 (英文)、程式語言、營隊、線上課程、親子、自學、共學
      * **任務結果**：10個欄位數、188筆資料
      * **資料清理做法**：文章斷句、替換字串、中文斷詞、清除停用詞、計算ngrams、詞頻計算
      * **情緒分析**：Lexicon Based、Corpus Based、LIWC情緒分析
      * **文辭和文件分析**：轉為DTM、計算相關性矩陣、計算共現矩陣、字詞網路圖、關聯式文字雲
   2. **資料區間：2019/1/1 – 2019/12/31**
      * **資料來源**：使用工作流程平台內建的「聯合新聞」
      * **資料看板**：digit, education, global, life
      * **搜尋關鍵字**：自主學習、教育相關、語言學習、程式語言、線上課程、自學、共學
      * **任務結果**：7個欄位數、382筆資料
      * **資料清理做法**：中文斷詞、清除停用詞、計算ngrams、詞頻計算
      * **情緒分析**：Lexicon Based、Corpus Based、LIWC情緒分析
      * **文辭和文件分析**：轉為DTM、計算相關性矩陣、計算共現矩陣、字詞網路圖、關聯式文字雲
7. **資料清理過程**
   1. **2017/1/1-2018/12/31資料篩選：**

使用工作流程平台提供的資料集，包含PTT、Dcard、蘋果日報、中時、東森新聞、聯合新聞。

【蒐集資料】

* 步驟一：
  + - * 資料區間設定：2017/1/1-2018/12/31。
      * 關鍵字：「自主學習、教育相關、語言學習 (英文)、程式語言、營隊、線上課程、親子、自學、共學、HTML、CSS、JavaScript、Scratch、Minecraft、Python」。
      * 初步篩選結果：資料偏向工程師否此之間討論工作內容、針對程式語言提出疑問，與自主學習沒有關聯性。
* 步驟二：
  + - * 資料區間設定：2017/1/1-2018/12/31。
      * 關鍵字：「自主學習、教育相關、語言學習 (英文)、程式語言、營隊、線上課程、親子、自學、共學」。
      * 初步篩選結果：資料偏向工程師否此之間討論工作內容、針對程式語言提出疑問，與自主學習沒有關聯性。
      * 初步篩選結果：只有PTT有適合的資料筆數共188筆，其餘皆顯示沒有相關資料筆數。
* 步驟三：
  + - * 資料區間設定：2017/1/1-2017/12/31。
      * 關鍵字：「自主學習、教育相關、語言學習 (英文)、程式語言、營隊、線上課程、親子、自學、共學」。
      * 初步篩選結果：初步判斷是否資料量過於龐大，所以先以2017年一整年的資料為篩選對象，但最終呈現0個搜尋結果。
* 步驟四：
  + - * 資料區間設定：2017/1/1-2017/12/31。
      * 關鍵字：「自主學習、教育相關、語言學習 (英文)、程式語言、營隊、線上課程、親子、自學、共學」。
      * 初步篩選結果：考量2017年距離現在有一段時間，有可能因為年代久遠，所以工作流程平台沒有相關資料量。故調整資料區間為2018/1/1-2018/12/31，但篩選結果還是顯示沒有相關資料筆數。
* 總結論：使用PTT資料筆數共188筆。

【資料清理】

* 步驟一：以網路上所提供共200多個停用詞做為清理字詞。
* 步驟二：檢視字詞網路圖，出現「xd、qq、一次、不用、一點、一場」，故將上述字詞納入清理字詞中，進行第二次停用詞整理。
* 步驟三：檢視字詞網路圖，出現「以下、之後、一直、兩個、一遍、不錯、主要、一下、今年」，故將上述字詞納入清理字詞中，進行第三次停用詞整理。
  1. **2019/1/1 – 2019/12/31資料篩選：**

使用工作流程平台提供的資料集進行分析，包含PTT、Dcard、蘋果日報、中時、東森新聞、聯合新聞。

【蒐集資料】

* 步驟一：
  + - * 資料區間設定：2019/1/1-2019/12/31。
      * 關鍵字：「自主學習、教育相關、語言學習 (英文)、程式語言、營隊、線上課程、親子、自學、共學、HTML、CSS、JavaScript、Scratch、Minecraft、Python」。
      * 初步篩選結果：

關鍵字「親子」跟「營隊」會找出跟自主學習無關聯性的文章，故將「親子」及「營隊」的關鍵字移除，並調整「語言學習(英文)」為「語言學習」。

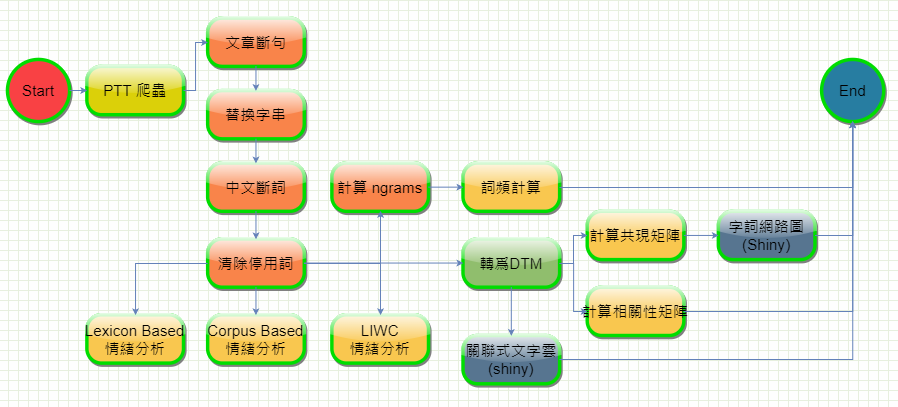
關鍵字「HTML、CSS、JavaScript、Scratch、Minecraft、Python」會找到偏軟體開發相關的文章，跟自主學習無關聯性，故將這些關鍵字移除。

* 步驟二：
  + - * 用步驟一關鍵字及日期區間進行資料蒐集、中文斷詞、清除停用字、詞頻計算，發現「聯合新聞」的資料品質較符合本次探討的主題，故後續分析皆採用「聯合新聞」的資料。
      * PTT：BabyMother, teacher, studyteacher; 共162筆資料
      * Dcard：沒有符合的資料
      * 蘋果日報：沒有符合的資料
      * 中時：life, opinion, technology, word; 共175筆資料
      * 東森新聞：living, story, world; 共6筆資料
      * 聯合新聞：digit, education, global, life; 共382筆資料
* 總結論：使用「聯合新聞」資料筆數共382筆。

【資料清理】

* 透過「詞頻計算」「bigram後的詞頻計算」調整「中文斷詞」及增加「停用詞」。
* 中文斷詞清單(權重皆為100)：自主學習，程式語言、線上課程、應用程式、實驗小學、人工智慧、學習模型、十二年國教、108課綱、實驗教育、自我成長、學習歷程、終身學習
* 停用詞清單：表示、也是、提供、贊助、該篇、文章、新聞、登入、帶回、心中、查詢、當您、看到、中獎、機率、金幣、當您、支持、仟元、現金、會員、獲得、越多、越高、即可、目前、使用、優質、活動、點按、按鈕、期間、經濟日報、新聞網、中心、狀況、網點、攝影、分享、住院、醫師

1. **初步分析方法和結果**
   1. **資料區間：2017/1/1-2018/12/31**



* Ngramg數量：69,622

關鍵字詞：準備考試、備審資料、程式語言

一張含有 文字 的圖片

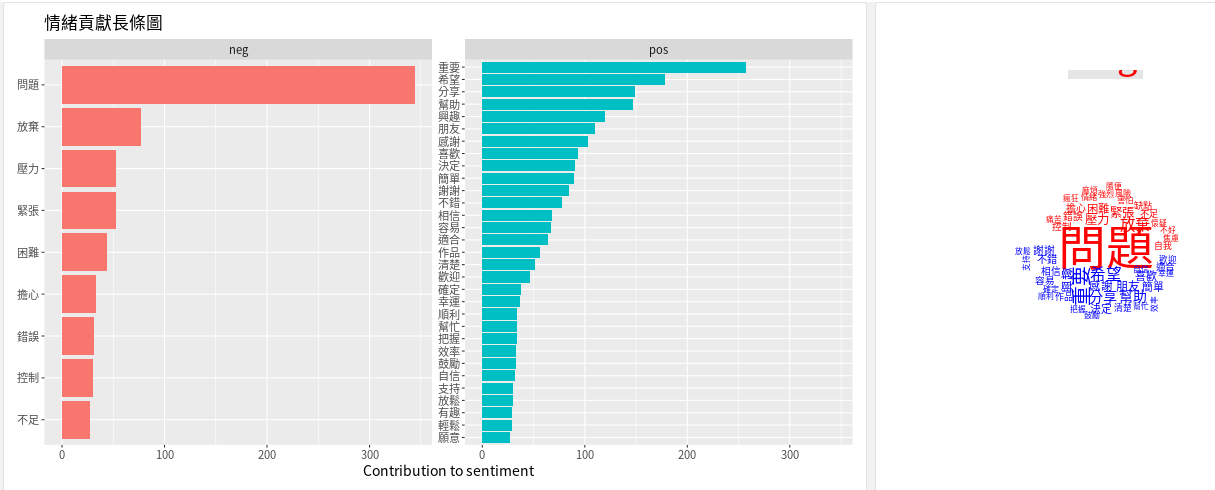
自動產生的描述

* Lexicon Based情緒分析

清除總數：107,521；最大正向情緒6、最大負向情緒4

負向關鍵字詞：問題、放棄

正向關鍵字詞：希望、分享

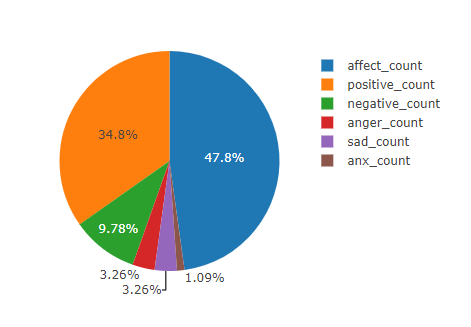


* Corpus Based情緒分析

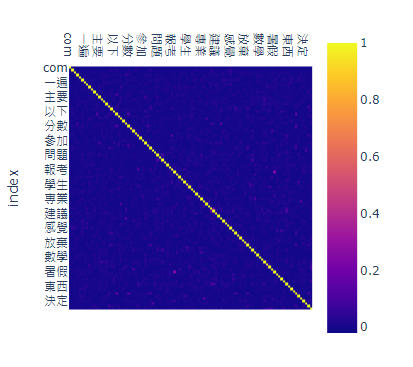
清除總數：107,521；負向情緒數14,627

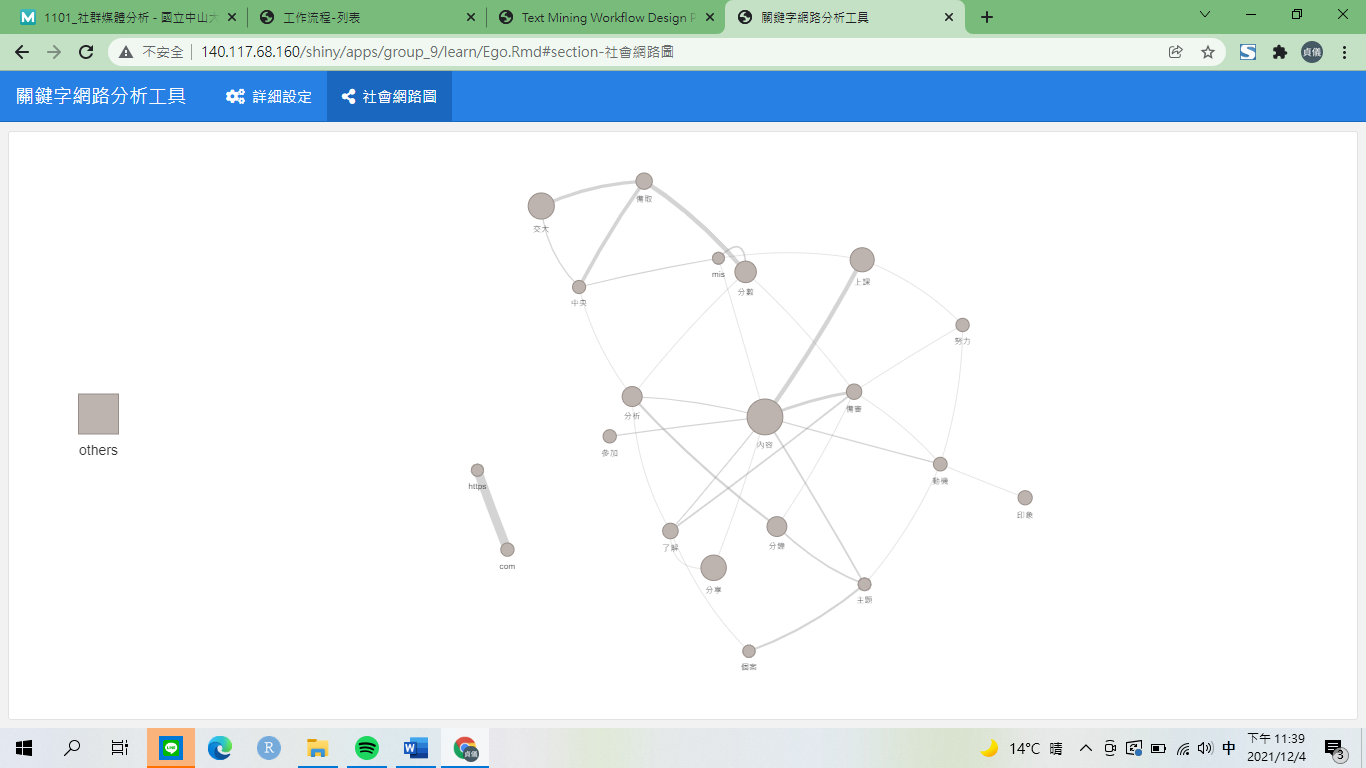
* LIWC情緒分析

清除總數：107,521；負向情緒數14,627



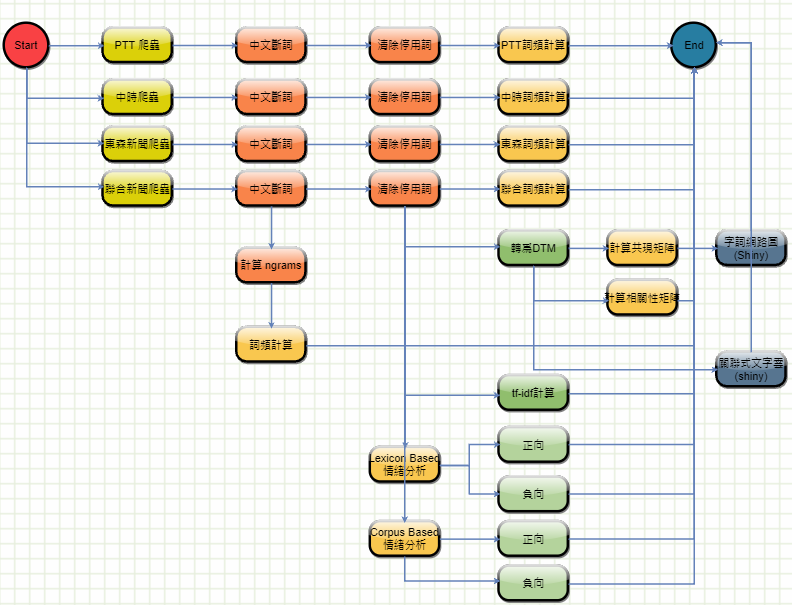
* 計算相關性矩陣



* 字詞網路圖
* 關聯式文字雲

一張含有 文字, 螢幕擷取畫面, 膝上型電腦, 電腦 的圖片

自動產生的描述

* 1. **資料區間：2019/1/1 – 2019/12/31**
* Ngramg數量：112,051

關鍵字：自主學習、教育相關、語言學習、程式語言、線上課程、自學、共學



* 詞頻計算

一張含有 文字, 白板 的圖片

自動產生的描述PTT

中時

一張含有 文字 的圖片

自動產生的描述

東森新聞：跟此次探討議題無關

一張含有 文字 的圖片

自動產生的描述

聯合新聞

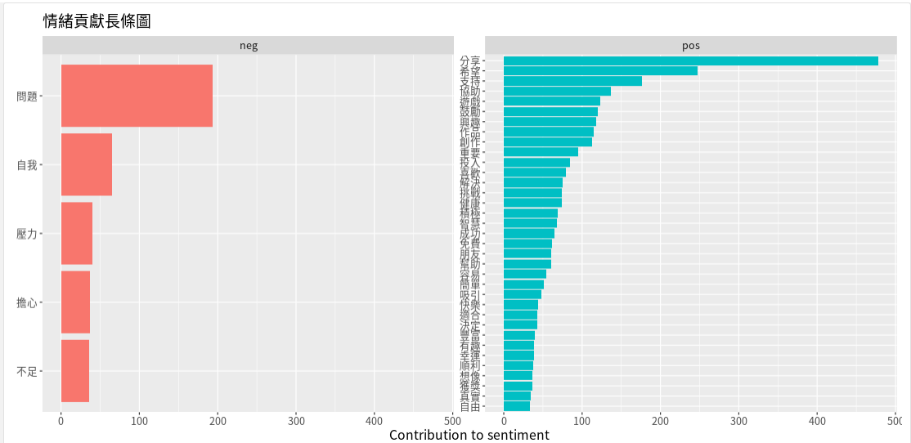


* 情緒分析 🡪以聯合新聞的文章進行

Lexicon based

* + 1. 一張含有 文字 的圖片

       自動產生的描述正向文章數：341篇 🡪sentiment\_value > 0
    2. 一張含有 文字 的圖片

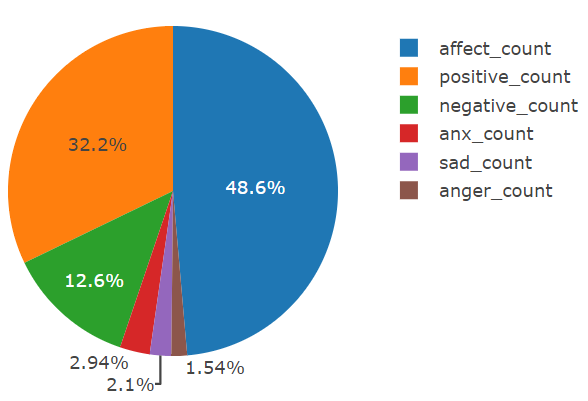
       自動產生的描述一張含有 文字 的圖片

       自動產生的描述負向文章數：30篇 🡪sentiment\_value < 0
  + Corpus based
    1. 一張含有 文字 的圖片

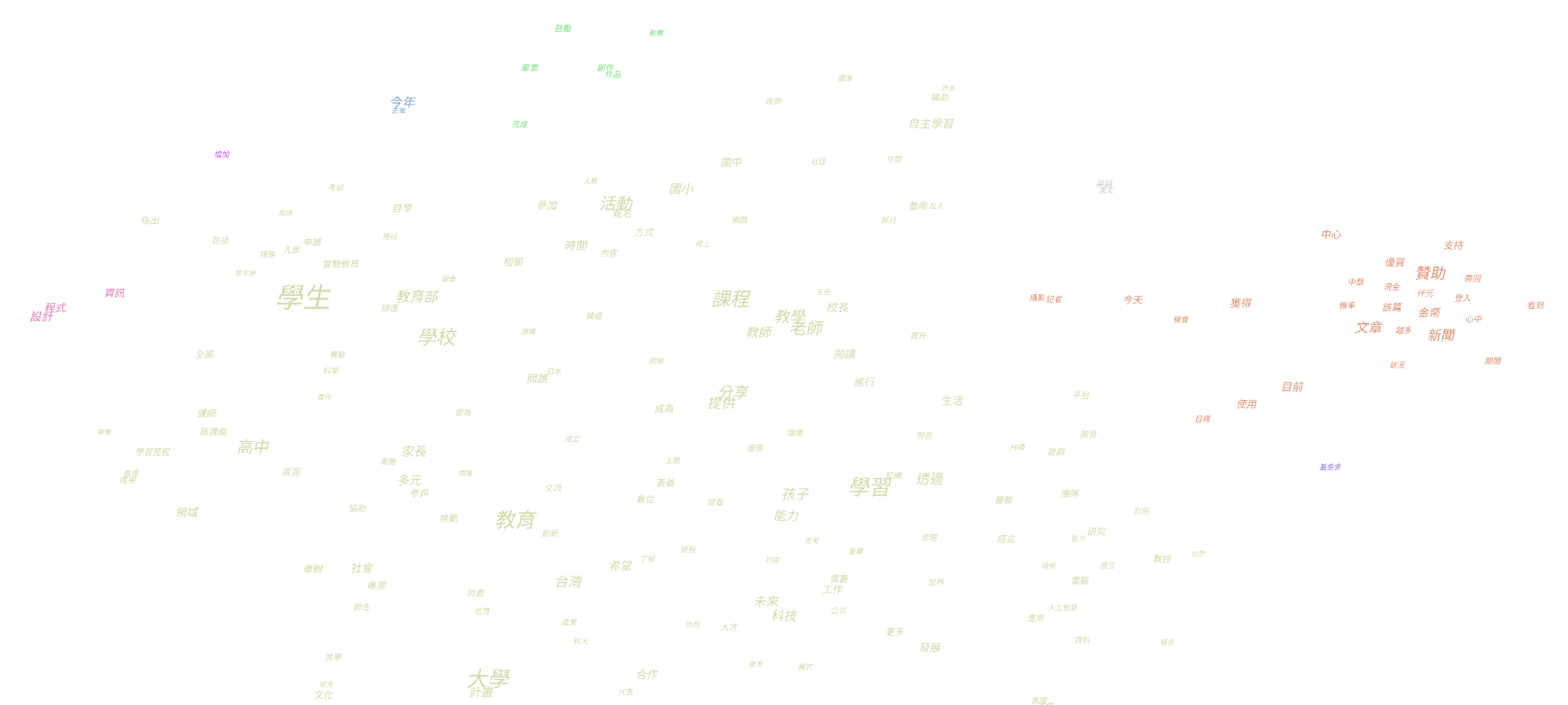
       自動產生的描述正向文章數：296篇 🡪sentiment\_value >= 0.5
    2. 一張含有 文字 的圖片

       自動產生的描述負向文章數：86篇 🡪sentiment\_value < 0.5
  + LIWC情緒分析

清除總數：108,948；最大正向情緒：63 最大負向情緒：25

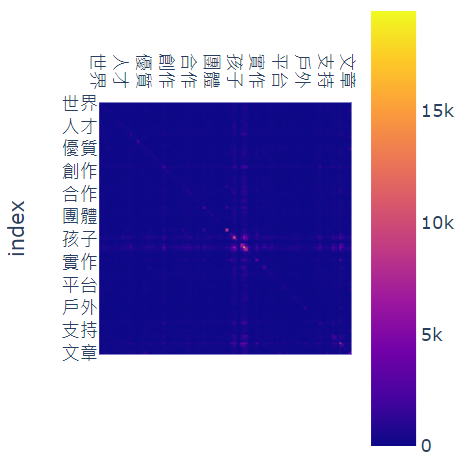


* + 關聯式文字雲

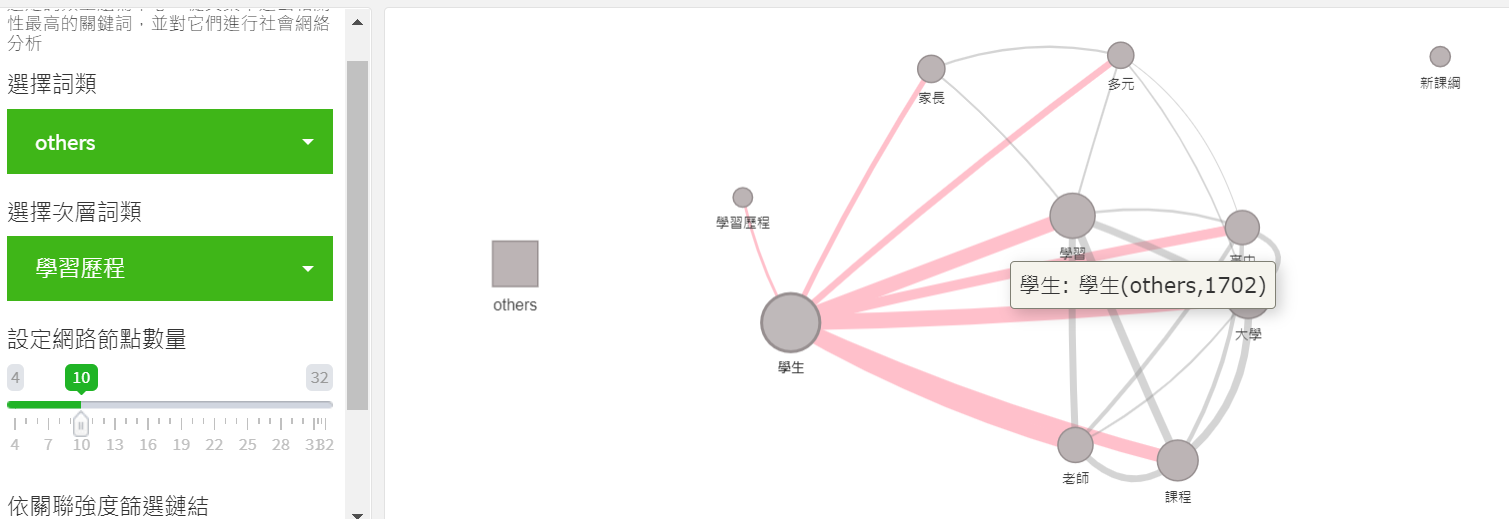


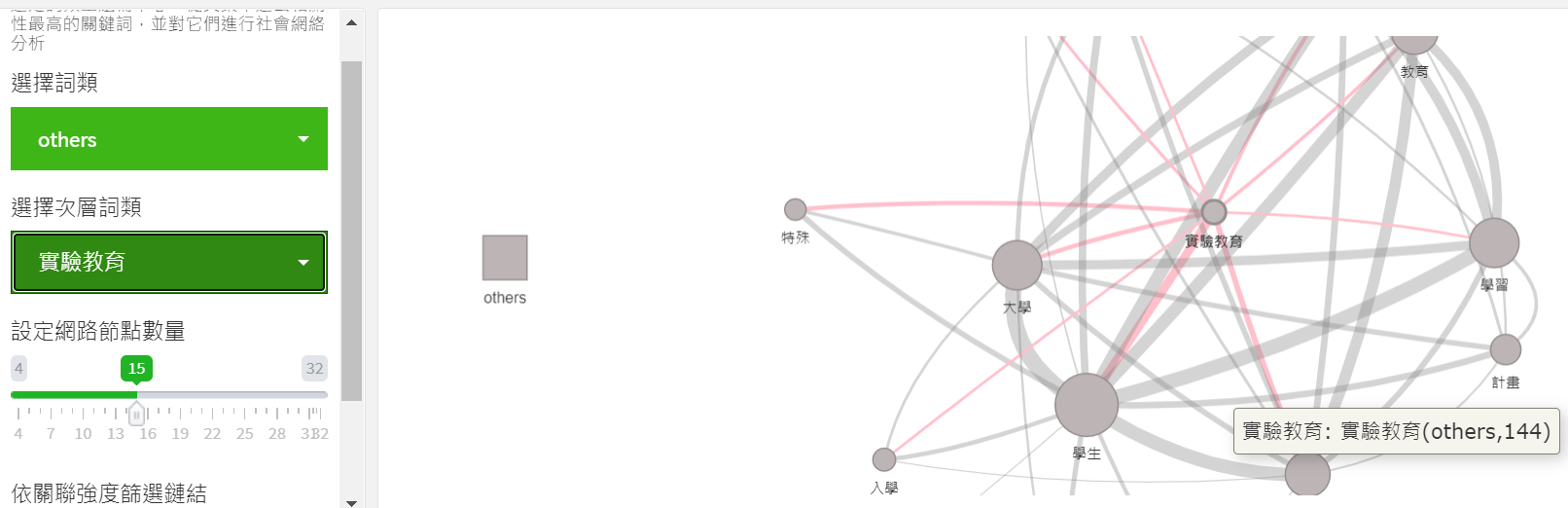
* + 共現關係(共現矩陣)

Heat map



* + 字詞網路圖





1. **資料分析結果**

**自主學習族群**：有進修計畫之學生、提升工作技能之上班族

**線上學習課程偏好**：程式語言、資料結構分析、作業系統…等資訊科技能力。

* 經過資料蒐集與分析，發現抱持著進修計畫的學生族群多數都擁有自主學習念頭。為了彌補自身欠缺的技能以順利進入想要就讀的研究所，而選擇線上學習程式語言、數據分析…等科技技術。
* 另外資料結果也顯示，有些上班族為了彌補自身於工作上所欠缺的技術能力，也會於閒暇時進行自我進修，以滿足職場的需求。

1. **資料分析心得**

* 依據廠商需求，我們分析的資料區間範圍為2017/1/1-2020/10/31，但2019-2021所篩選的資料內容無法針對疫情前後的自主學習做出明顯的差異分析。
* 針對分析結果，後續我們將與廠商進行更細部的溝通，以更符合廠商需求的資料集進行相關分析。資料集來源預計會從「政府資料開放平臺」尋找。

1. **後續計畫**

* 觀察自主學習平台的聲量及相關字詞。
* 觀察疫情前後對於自主學習的趨勢。
* 從其他討論版(e.g.Soft\_job)抓取資料進行比較。
* 根據相關字詞整理找出一般大眾對於自主學習平台期待的教育模式和主要學習目標。