社群媒體分析 Project2

黃仁勳

第 15 組

N104020016 楊世華

N104020008 游淑媛

N104020029 陳勇任

B096060020 黃湘安

目錄

圖目錄

表目錄

一、 選擇主題動機與目的

人工智慧（artificial intelligence, AI）的是出於對科技進步和人類生活改善的渴望所研發而成的系統，能夠正確解釋資料並靈活應用於特定任務與目標，眾人皆深信 AI 的潛力能夠解決許多現實世界的難題、帶來重大的社會和經濟影響。2023 年全球 AI 產業爆發性的百花齊放，自然語言模型 ChatGPT、AI 繪圖等被廣為使用，食衣住行育樂各個產業也紛紛投入 AI應用之中，物聯網產業的發展重心也從硬體轉移至軟硬整合的系統與應用服務，對於大量計算、高效能的晶片的需求倍增。

1993 年黃仁勳創立的輝達（Nvidia），核心業務為設計高效晶片，原先的主要客群為電玩市場，但隨著 AI 對於晶片計算能力的需求，以及 2022 年底 ChatGPT 的出世，各大企業與新創公司對於輝達生產的處理器趨之若鶩，進而帶動市場焦點，其股價從年初的每股 143 美元上漲超過 200%至 374 美元。2023 年 5 月 30 開始的台北國際電腦展（COMPUTEX），作為一年一度科技產業盛事，除了國內外大廠比拚最新技術，今年另一大矚目焦點，人工智慧教父、輝達(Nvidia)執行長黃仁勳也親自來台演講，掀起一波 AI 狂潮，並帶動近期台股走高，電子股又再度成為盤面重心。

故本組希望透過主題模型將相關文本提取出主題關鍵字，除了了解大家對於輝達或 AI 等關鍵字的看法，並透過社會網路圖來探討網民對於相關關鍵字的留言走向。

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，工作流程名稱為期末_黃仁勳，預計分別針對內文及留言進行資料處理、主題模型及留言之相關分析，分析流程如圖 1。

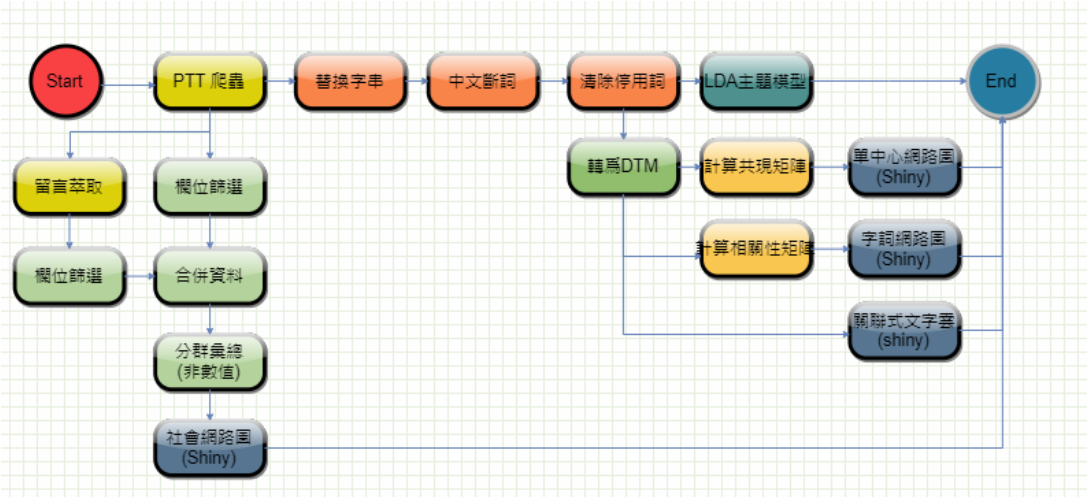


圖 1:tarflow 工作流程圖

二、 資料蒐集

(一) 爬蟲

(1) 目的：獲取分析標的。

(2) 資料抓取標的：PTT 金融業、外匯、保險、貸款、期權、軟體工作、股票、科技工作和八卦板，共 9 個。

(3) 搜尋關鍵字：黃仁勳、NVIDIA 和輝達。

(4) 排除關鍵字：無。

(5) 抓取資料區間：2023 年 1 月 1 日~2023 年 6 月 5 日，共 639 筆。

圖 2:爬蟲結果

(6) 留言萃取：抓取每筆文章底下之留言，共取得 28,332 筆留言，平均每篇 53.26 文章有筆留言

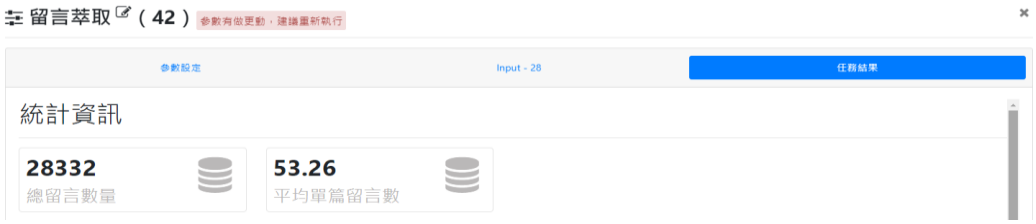


圖 3:留言萃取結果

(二) 替代文字

(1) 目的：將特殊符號或特定字串替換，以利後續分析。

(2) 替換字串設定：去除標點符號、特殊字元及超連結，共 6 種符號變化需替換為空白。

圖 4:替代字串設定畫面

(三) 中文斷詞

(1) 目的：將句子進一步切割成適當詞語，以利後續分析。

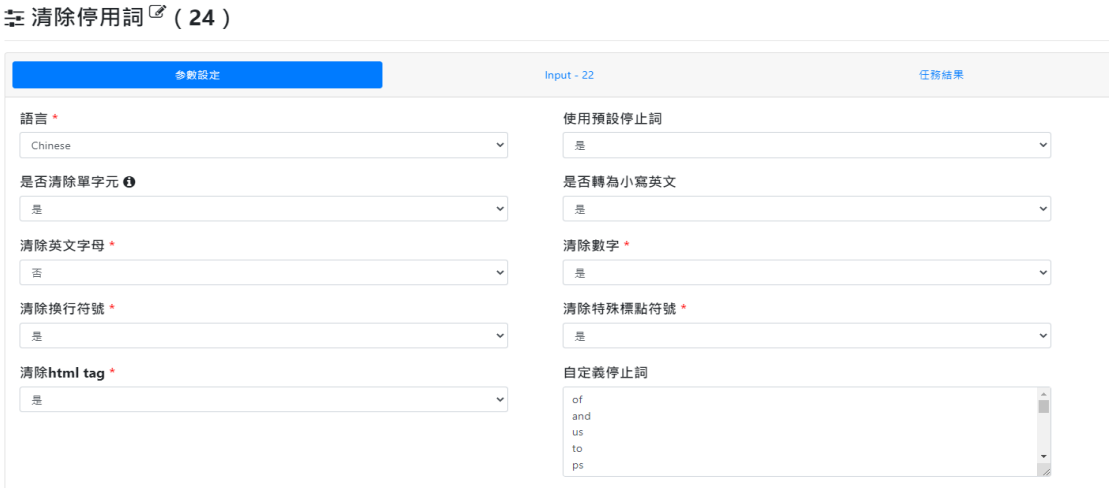
(2) 定義詞彙：等具相關性詞語進行權重設定...等 10 個詞彙。

圖 5:中文斷詞設定畫面

(四) 清除停用詞

(1) 目的：過濾掉無實際含意或非分析主體等之詞彙，以利後續分析。

(2) 定義詞彙：使用預設停止詞且轉換小寫英文，並清除換行符號、html tag、數字、特殊標點符號，除了系統設定外，亦加入自定義停止詞，內容包括較無意義的語助詞、轉貼新聞報導的詞彙，停用 149 個詞彙，清除 126,197 筆。



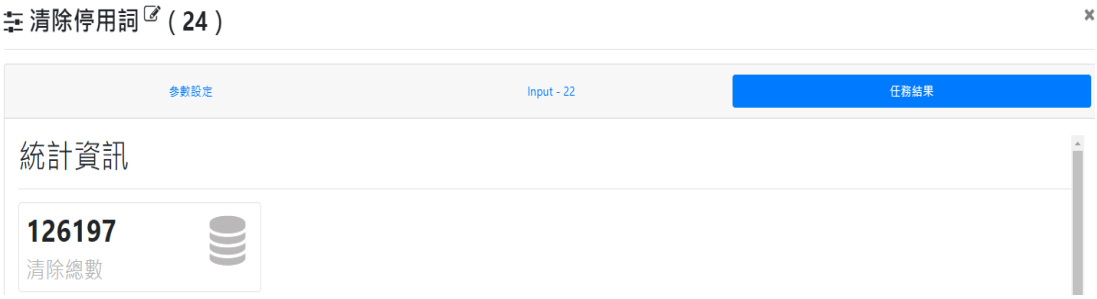


圖 6:清除停用詞設定及結果

三、 分析流程

(一) 主題模型

1-1 LDA 主題模型

(1) 目的：從大量的文本數據中自動發現和提取主題。

(2) 處理方式：反覆測試，最後將主題數分成 3 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

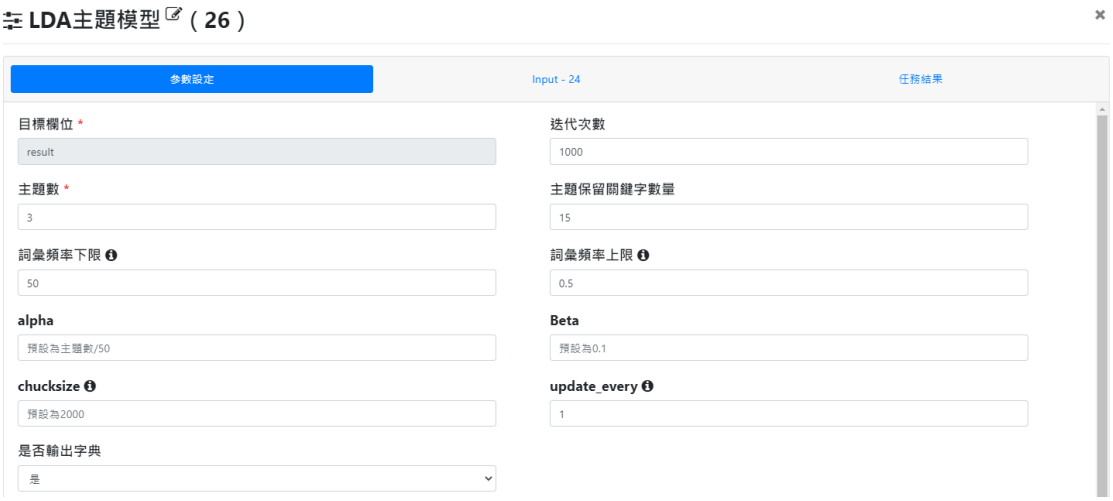


圖 7: LDA 主題模型設定

(二) 視覺化分析

2-1 轉為 DTM

(1) 目的: 接續清除停用詞處理，將斷詞後結果轉製成 200 維度的矩陣。

2-2 單中心網路圖

(1) 目的: 接續 2-1 處理，進行詞彙間關聯性。

(2) 處理方式:選擇字詞自定義類型的字典檔(dis.csv)，其他設定為預設。

圖 8: 單中心網路圖設定

2-3 字詞網路圖

(1) 目的: 接續 2-1 處理，進行文本中的字詞之間的相關性和關聯性。。

2-4 關聯式文字雲

(1) 目的: 接續 2-1 處理，進行詞彙分群計算共線關聯性。

(2) 處理方式:以關聯系最高的前 300 個字詞，分成 20 群顯示，迭代次數設定為 1000 次。

圖 9: 關聯式文字雲設定

2-5 欄位篩選

(1) 目的：針對爬蟲結果與留言萃取結果進行篩選所需欄位，如「system_id、artPoster、comment_idx、cmtPoster、cmtContent」。

2-6 合併資料

(1) 目的：接續 3-1 處理，合併原資料與逐筆文章之留言。

(2) 處理方式：join 規則為「system_id」。

2-7 分群匯總__留言者之負向留言數

(1) 目的：接續 3-2 處理，以文章標題與留言者為計算標的，計算該筆文章之留言數。

(2) 處理方式：使用「artTitle、cmtPoster」作為分群依據，並計算「system_id」欄位之數量。

2-8 社會網路圖

(1) 目的：接續 3-2 處理，以文章標題與留言者為計算標的，計算該筆文章之留言數。

(2) 處理方式：使用「artTitle、cmtPoster」作為分群依據，並計算「system_id@count」欄位之數量。

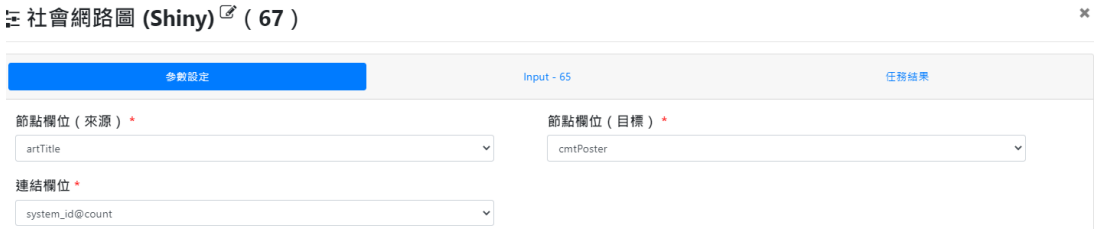


圖 10: 社會網路圖設定

四、 分析結果

(一) LDA 主題模型結果

圖 11: 主題模型分析結果

(1) 第一群為「nvidia 產品/技術應用」

主題一前 30 名相關詞彙，佔了該主題總詞彙的 42.3%；其中相較於其他主題，主題一特有的詞彙為：「應用」、「GPU」、「ChatGPT」、「生成」、「AI」、「遊戲」、「模型」、「軟體」、「平台」等等；而來源文章之關鍵字為 nvidia 相關，因此推測主題一為「nvidia 產品/技術應用」。

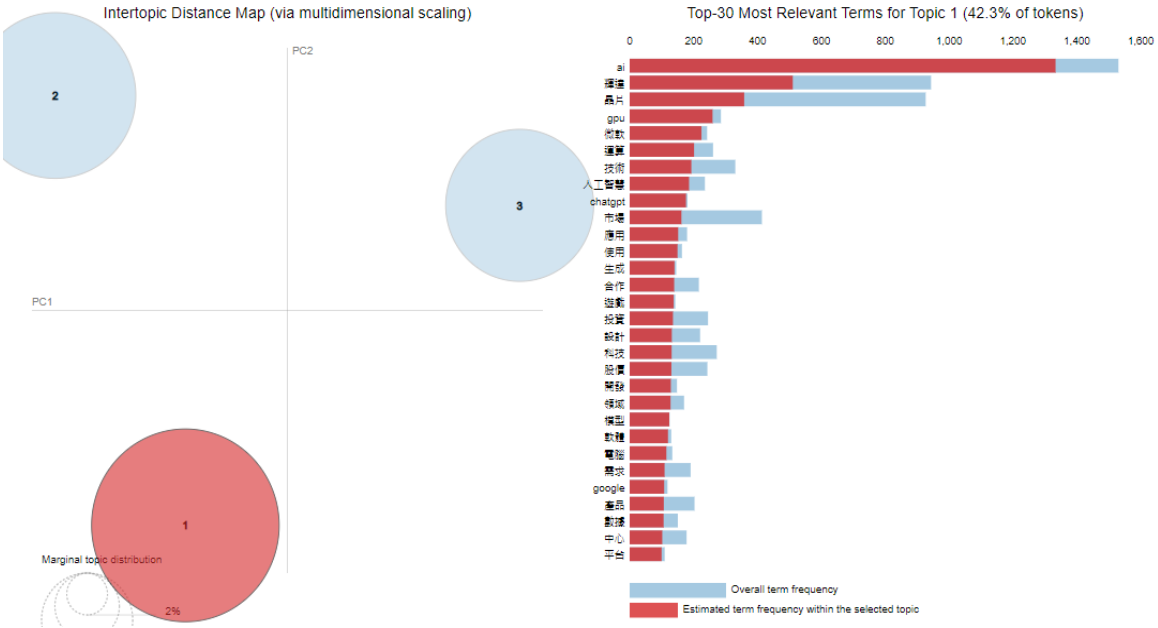


圖 12: 第一類主題

(2) 第二群為「nvidia 產品/技術開發」

主題二前 30 名相關詞彙，佔了該主題總詞彙的 31.3%；其中相較於其他主題，主題二特有的詞彙為：「製程」、「台積電」、「半導體」、「英特爾」、「生產」、「代工」、「晶圓」、「先進」、「超微」、等等；而來源文章之關鍵字為 nvidia 相關，因此推測主題二為「nvidia 產品/技術開發」。

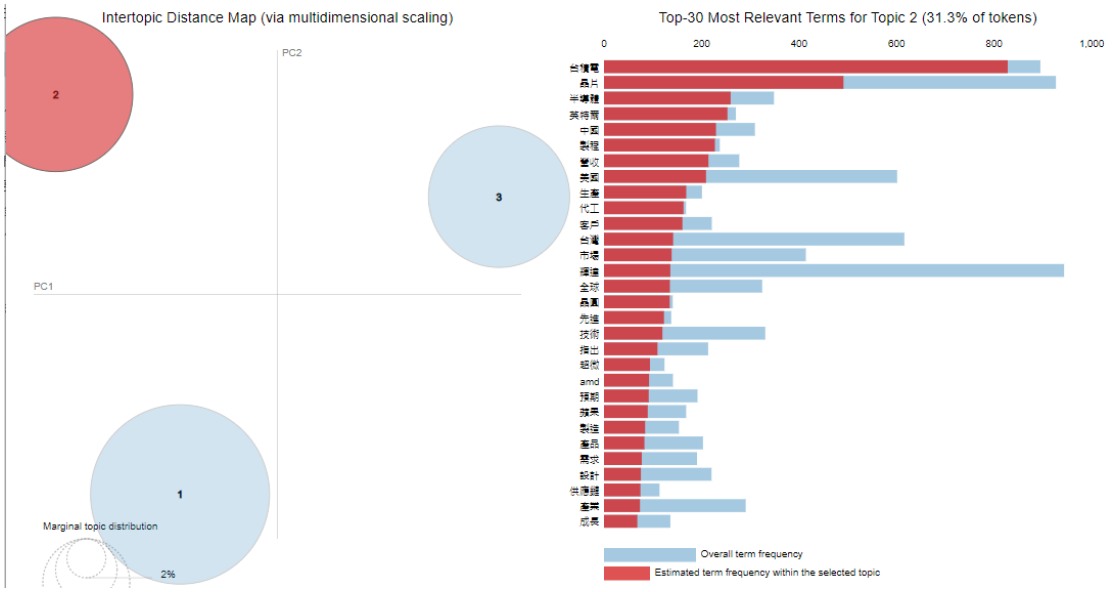


圖 13: 第二類主題

(3) 第三群為「黃仁勳台大演講」

主題三前 30 名相關詞彙，佔了該主題總詞彙的 26.3%；其中相較於其他主題，主題三特有的詞彙為：「台大」、「黃仁勳」、「演講」、「創辦人」、「指數」、「大漲」等等；而來源文章之關鍵字為黃仁勳相關，因此推測主題三為「黃仁勳台大演講」。

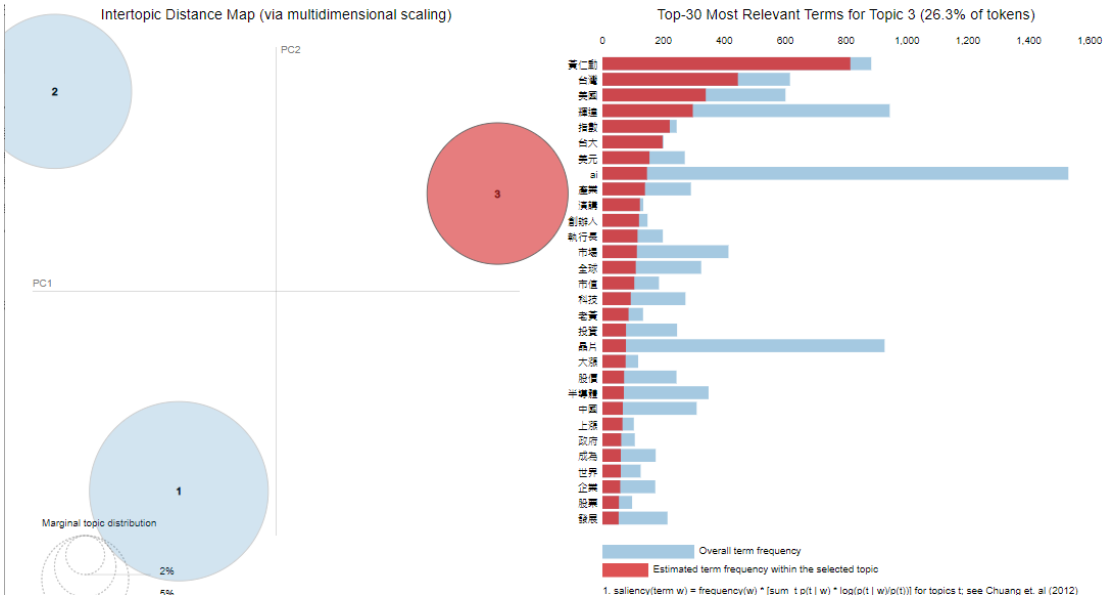


圖 14: 第三類主題

(4)各主題關鍵字彙總(λ= 1，TOP15)

如下表 1，統計 LDA 三類主題中出現次數前 15 名之關鍵字，共 45 個關鍵字中，僅 9 個關鍵字重複，其中重複的字詞為：美國(2)、全球(2)、AI(2)、市場(3)、台灣(2)、輝達(3)、晶片(2)等泛用詞彙；因此，LDA 將數篇文章濃縮為數個主題，且各主題的文章內容及議題具有差異性。

表 1:主題關鍵字結果

Order	TOPIC 1	TOPIC 2	TOPIC 3
1	AI	台積電	黃仁勳
2	輝達	晶片	台灣
3	晶片	半導體	美國
4	GPU	英特爾	輝達
5	微軟	中國	指數
6	運算	製程	台大
7	技術	營收	美元
8	人工智慧	美國	AI
9	ChatGPT	生產	產業
10	市場	代工	演講
11	應用	客戶	創辦人
12	使用	台灣	執行長
13	生成	市場	市場
14	合作	輝達	全球
15	遊戲	全球	市值

(二) 視覺化分析結果

(1) 單中心網路圖

下圖為與「AI」字詞相關的詞彙，關係強度最強的是「晶片」、「人工智慧」、「GPU」等。以晶片來看，關係強度最高的有「台積電」、「輝達」等半導體公司。

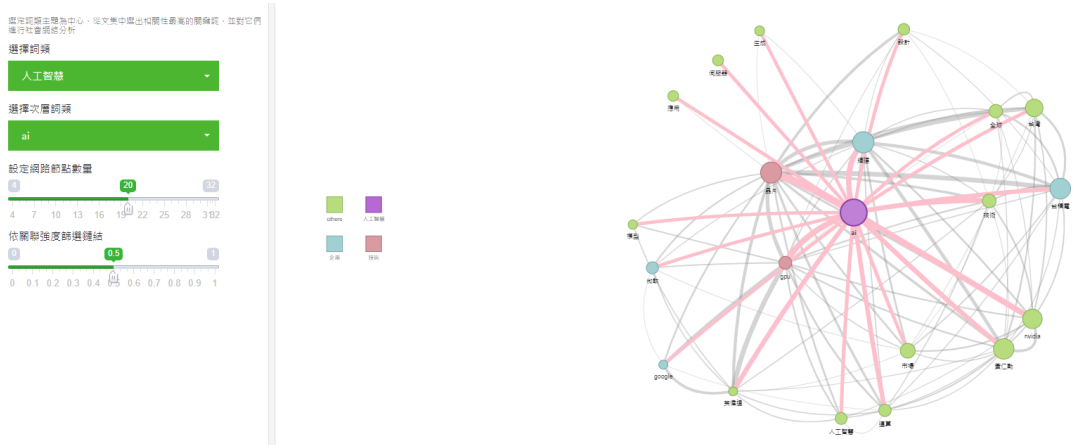


圖 15: 以 AI 為中心的網路圖

(2) 字詞網路圖

以「GPU」這個詞彙為中心，相關連的字詞有計算、架構、Google、英偉達(同nvidia，中國大陸譯為英偉達)、網路、訓練等等；GPU 為 nvidia 之核心產品，而上述詞彙皆與 nvidia 在 AI 人工智慧、雲端運算、機器學習等領域之應用相關，上述字詞反映了近期網友討論關於「GPU」相關應用領域。

以「台大」這個詞彙為中心，發散出相關連的字詞有畢業典禮，本組認為是因黃仁勳於台灣大學的畢業典禮上擔任致詞嘉賓，黃仁勳的致詞內容引起網友廣泛討論，連帶使得此次台大畢業典禮受到大家關注。

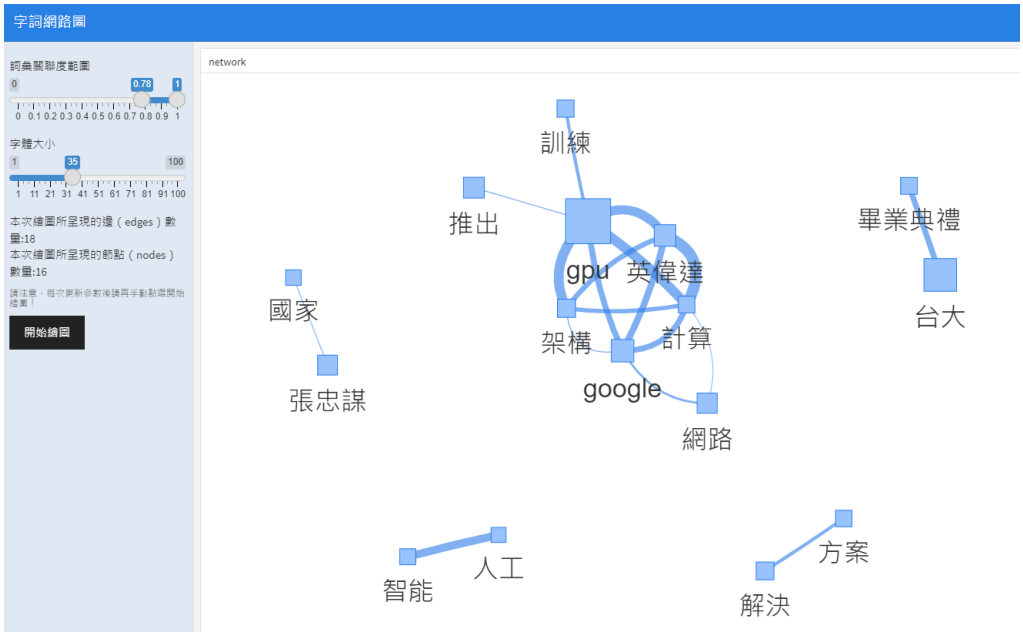


圖 16: 字詞網路圖結果

(3) 關聯式文字雲

關聯式文字雲可以反映個別字詞之出現次數，而字詞間之距離及顏色反映共同出現之頻率，以「台積電」為例，出現頻率高，其最為相近的字詞為「製程」、「奈米」及「訂單」等，反映了 NVIDIA 是台積電的大客戶，因此相關文章中經常連帶討論台積電之製程，及訂單量之變化；而「輝達」出現頻率也高，其最為相近的字詞為「股價」及「市值」，反映了近期 AI 的相關發展及應用，及 NVIDIA 超乎預期的財測，使得 NVIDIA 股價的市值變化引起網友的廣泛討論。

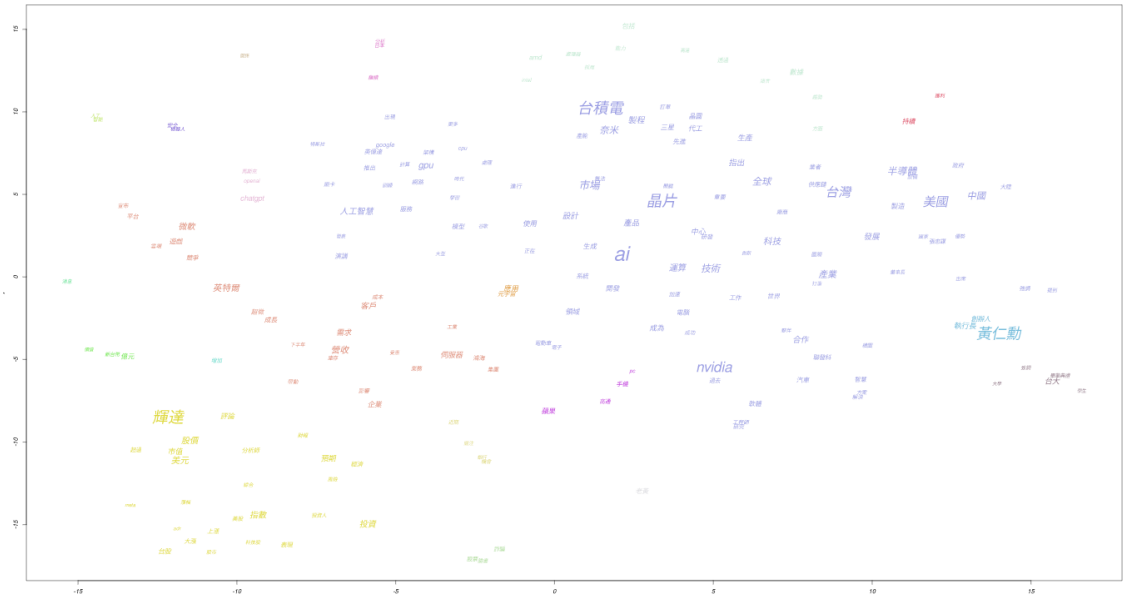


圖 17: 關聯式文字雲結果

(4) 社會網路圖

將被討論之文章標題和留言者分群做成社會網路圖，可以取得文章標題和留言者之間的網路關係， 從下圖可見，有六筆文章被高度討論，其中最高的是"[新聞] 黃仁勳太忙沒逛夜市讚蔡英文「她太棒了！」"，從留言區可看到許多網友是偏向負面的角度在回應，另外還有幾篇討論度高的文章是黃仁勳受邀參加台大畢業典禮擔任嘉賓時的致詞，因為口譯員翻譯的問題，引發好幾篇文章的討論。也可以看到其中有很多留言者不單單針對某幾篇文章做回應，也可以從下圖中看到他們之間的關係。

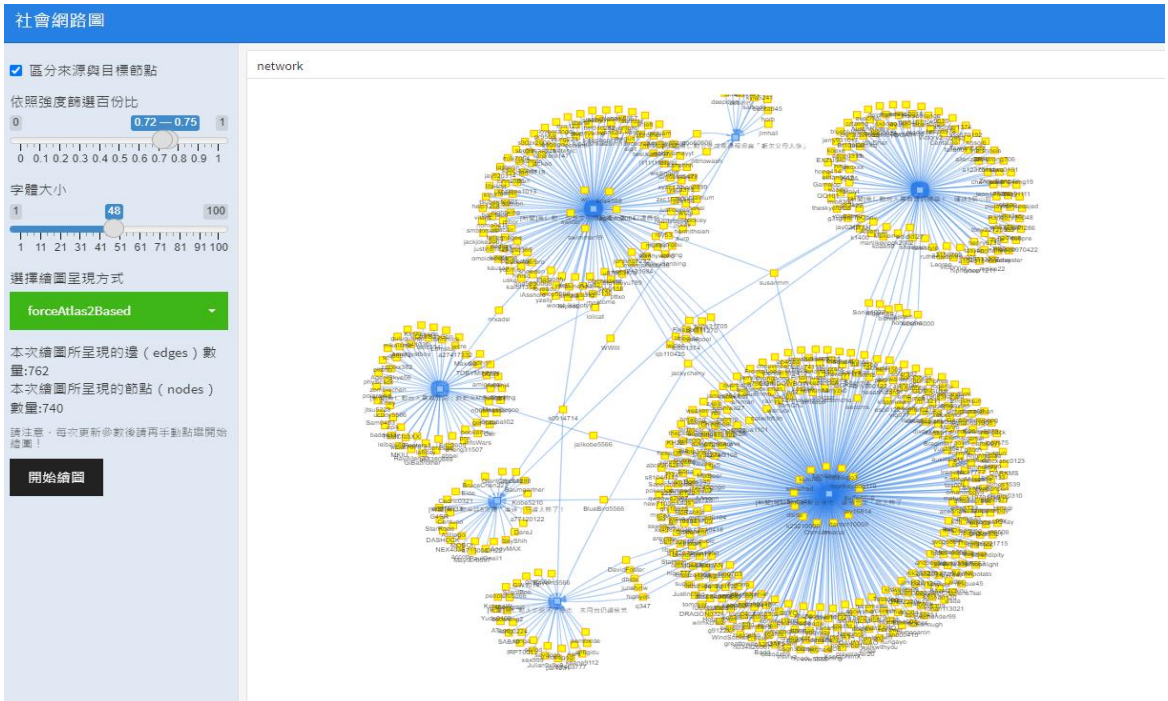


圖 18:文章標題與發文者之社會網路圖結果

五、 結論

2023 年，AI 應用如雨後春筍般在各項產業中蓬勃發展，關鍵晶片的製造廠－輝達，以及其臺裔美籍的創辦者－黃仁勳，在 PTT 論壇中獲得廣大討論。根據本組以「黃仁勳」、「NVIDIA」、「輝達」為關鍵字，在 PTT 股市、科技與八卦版的爬蟲結果，投入 LDA 主題模型後發現網友主要討論的面向有：nvidia 產品／技術應用、技術開發與黃仁勳至台大演講三大面。細究各主題關鍵字發現雖每個主題有「晶片」、「AI」、「輝達」、「市場」等字詞重疊，但仍可輕易判讀出模型生成的差異之處。另外，LDA 模型結果的主題連貫性（PMI）為 -0.304，混淆度為 74.59，也顯示此模型結果能顯著區分留言內容與其指向的議題。

第二部分以視覺化工具將網友留言內容進行分析，以單中心網路圖發現字詞的使用相當緊密集中，以「AI」為中心觀察，發散的結果與人工智慧、晶片、GPU、黃仁勳與輝達等關聯程度都相當高；以字詞網路圖可看出留言使用的詞語之間的關聯，包含「解決」、「方案」、「台大」「畢業典禮」，以及「GPU」與「英偉達」、「計算」與「架構」等，可透過成群的字詞看出其於留言內容情境中的使用情況；以關聯式文字雲則可看出詞語的頻率與相關性，AI、NVIDA、晶片、黃仁勳、台積電等本次分析著重之關鍵字皆有明確的的呈現，顯示抓取內容與實際討論情況相符；以社會網路圖則可看出與討論度高的特定文章互動頻繁的留言者，值得注意的是，在本組選取的強度百分比（0.72~0.75）之間，最大的六個留言互動文章皆屬於與科技內容無關的新聞，「黃仁勳逛夜市」與「黃仁勳於台大畢典致詞」相關的分別有二與三篇，顯示網友對於非科技相關的議題討論較為熱絡。

六、 參考影片連結