<< 社群媒體分析 >>

指導教授：黃三益 教授

第二次讀書會報告_第十四組

一、分析主題 : 針對金秀賢事件的評價分析

二、組員名單 :

馮蕙芳_N124080003

高健芝_N124080006

廖胤翔_M134610022

Deepan_M134610032

孫郁琪_M136020020

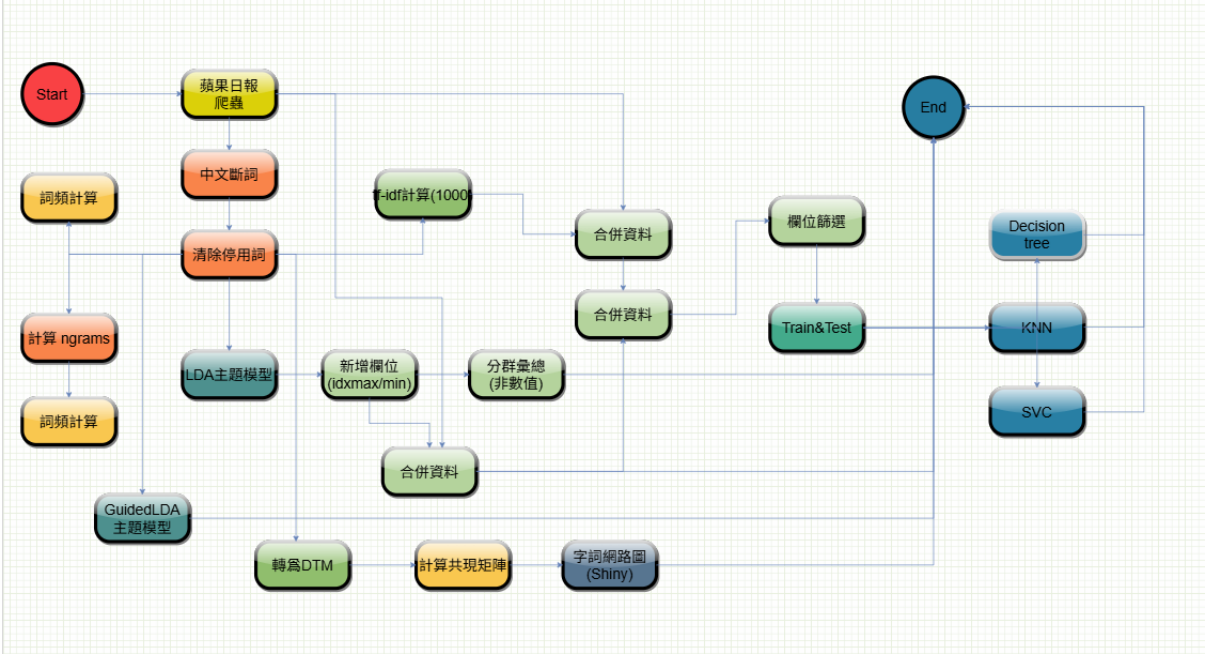
胡 賀_J136020007

三、分析工具 : • 中山大學工作流程平台 Tarflow • 工作流程名稱 : 第二次讀書會(1)

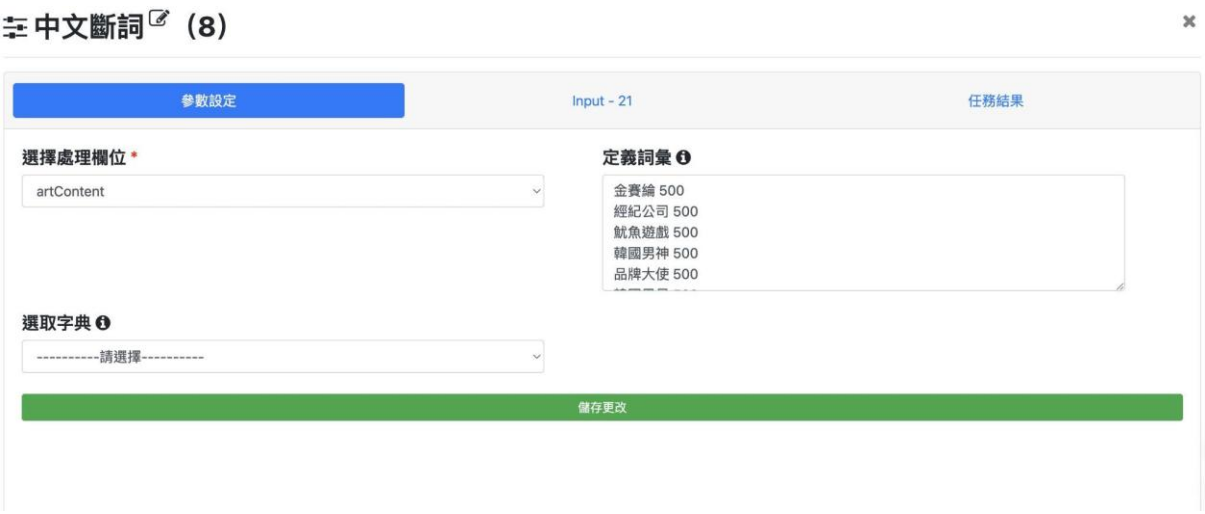
四、目的: 探究新聞媒體報導金秀賢事件本身所反映出的輿論導向與延伸的社會現象

五、資料來源：蘋果日報中時尚娛樂、國際共2個看板。

完整tarflow工作流程如下圖 ：

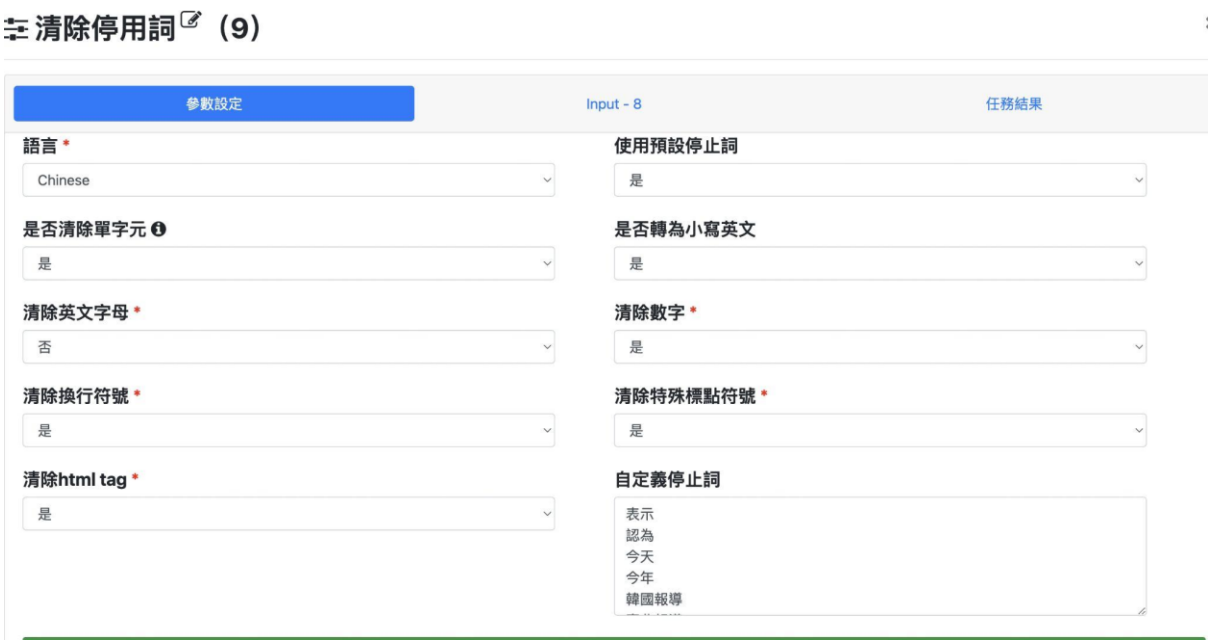


分析期間：2024/03/01 ~ 2025/04/13



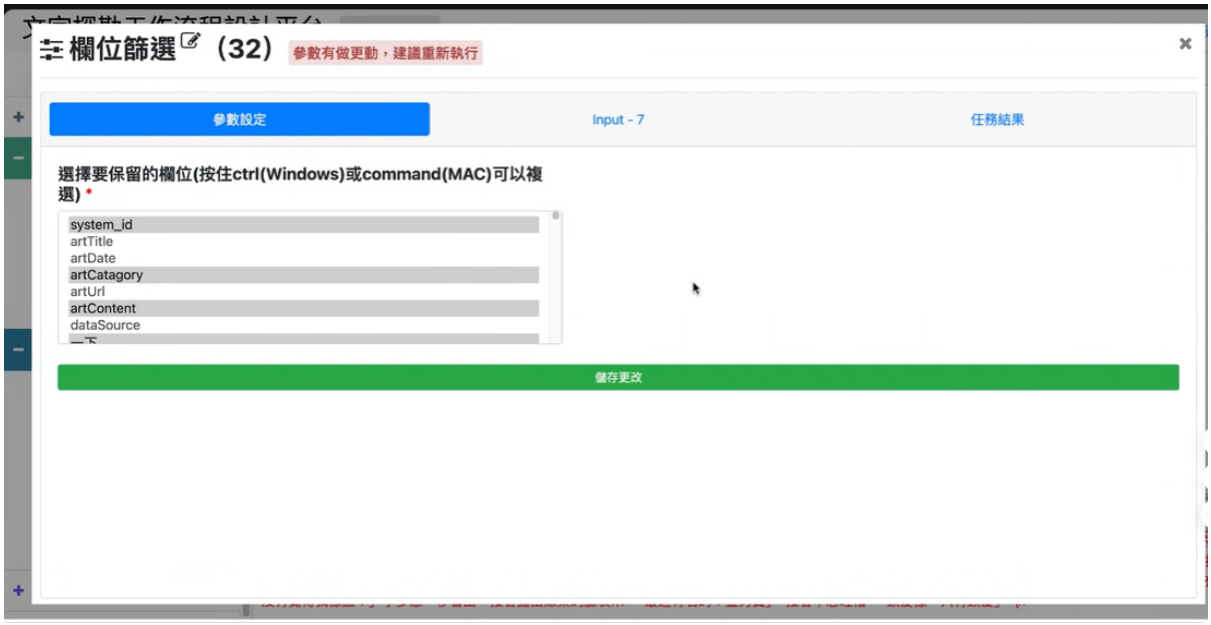
流程概述：

1. 指定特定關鍵字，爬取「韓國演藝圈、韓國娛樂圈、金秀賢、金賽綸、Gold Medalist、淚之女王、HYBE、山寨人生」，共 425 筆資料。
2. 進行資料清理，以「中文斷詞」將新聞內容分解成字詞單位。
3. 使用「清除停用詞」將不必要的符號、單字元去除；以及自定義停止詞如「表示、認為、今天、今年、韓國報導、臺北報導、綜合、報導、橫豎、看到、知道」
4. 設定停用字以過濾出現頻率高但無意義的字詞。
5. 使用「分群匯總(非數值）」工具，以得出爬蟲結果中6個文章類別的數量。
6. 將清理好的資料轉為DTM，進行文件分類的流程，並將完成的預測模型進行分類預測。
7. 計算相關性矩陣與共現矩陣，產生字詞網路圖、單中心網路圖跟關聯式文字雲分析。




tf-idf計算(1000) (26)
參數有做更動，建議重新執行
參數設定
Input - 9
任務結果
保留詞彙
以換行符號區隔，e.g.
國立中山大學
西子灣
壽山...
最多篩選詞彙數量
1000
儲存更改

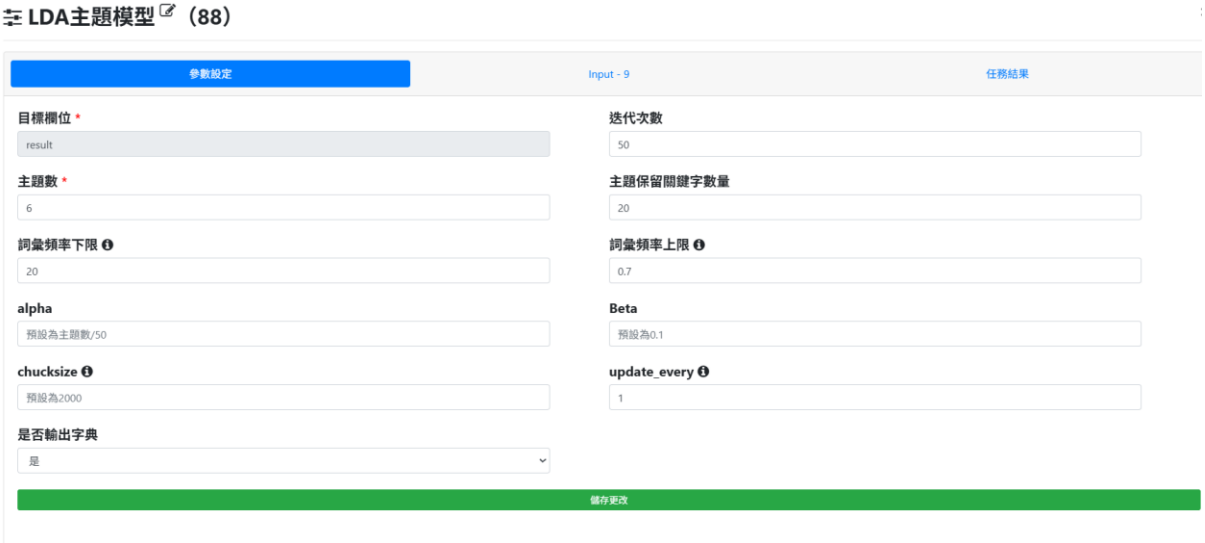

合併資料 (7)
參數有做更動，建議重新執行
參數設定
Input - 21
Input - 26
任務結果
JOIN規則
新增規則
刪除規則
任務一欄位
任務二欄位
system_id
system_id
----------請選擇----------
----------請選擇----------
儲存更改



主題模型: 我們使用了 LDA（Latent Dirichlet Allocation）主題模型，作為本次文本探勘的主要方法，來自動識別資料中隱含的討論主題。在 Tarflow 平台上進行設定時，我們將主題數設為 6，也就是希望模型能從文本中分出六類不同的潛在話題。在進行模型訓練之前，我們設置了詞彙頻率的上下限，過濾掉出現次數少於 20 次的低頻詞，以及出現比例高於 70% 的高頻詞，以提升主題辨識的準確度。

在模型參數方面，我們使用了預設的 alpha 與 beta 值，讓每篇文章能聚焦在少數幾個主題上，並讓每個主題具有明確的代表詞分布。迭代次數則設定為 50，讓模型有足夠的時間進行參數收斂與主題穩定；每次訓練處理的文本數量為 2000 篇，並設定每次處理完一個批次就更新一次模型權重（update_every = 1），以強化模型在訓練過程中的即時調整能力。

此外，我們保留了每個主題的前 20 個關鍵詞作為輸出，並選擇輸出字典資料，以利後續分析或交叉比對使用。整體而言，這些設定有助於我們在保有模型效能的同時，也確保主題結果具備可解釋性與一致性。



LDA 主題模型將文本分為六個主題，以下為前三個主題的視覺化結果與內容解讀：

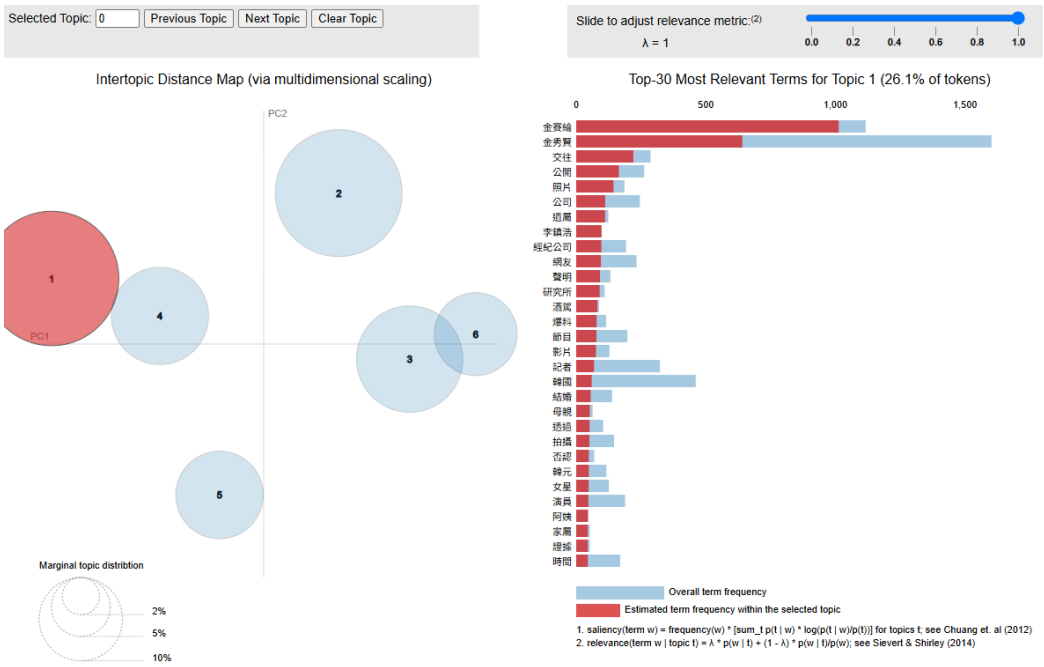
在主題一（Topic 1）中，模型擷取出最具代表性的關鍵詞包括「金賽綸」、「金秀賢」、「交往」、「公司」、「曝光」、「照片」等。從這些詞彙可以明顯看出該主題主要圍繞在事件的核心人物與爆料內容上，涵蓋媒體揭露、戀情傳聞與經紀公司相關的爭議。這也是我們整體文本中比例最高的主題，顯示事件本身的發展過程與新聞報導為討論的主幹。

主題二（Topic 2）則偏向劇情相關的討論，關鍵詞如「金秀賢」、「金智媛」、「飾演」、「角色」、「電視」、「收視」、「浪漫」等，顯示許多文本將焦點放在兩位演員在《淚之女王》中的形象延伸至現實的情感想像。觀眾對戲劇角色的情感投入進一步加深了對演員私生活的關注，也形塑了討論中戲裡戲外模糊邊界的特徵。

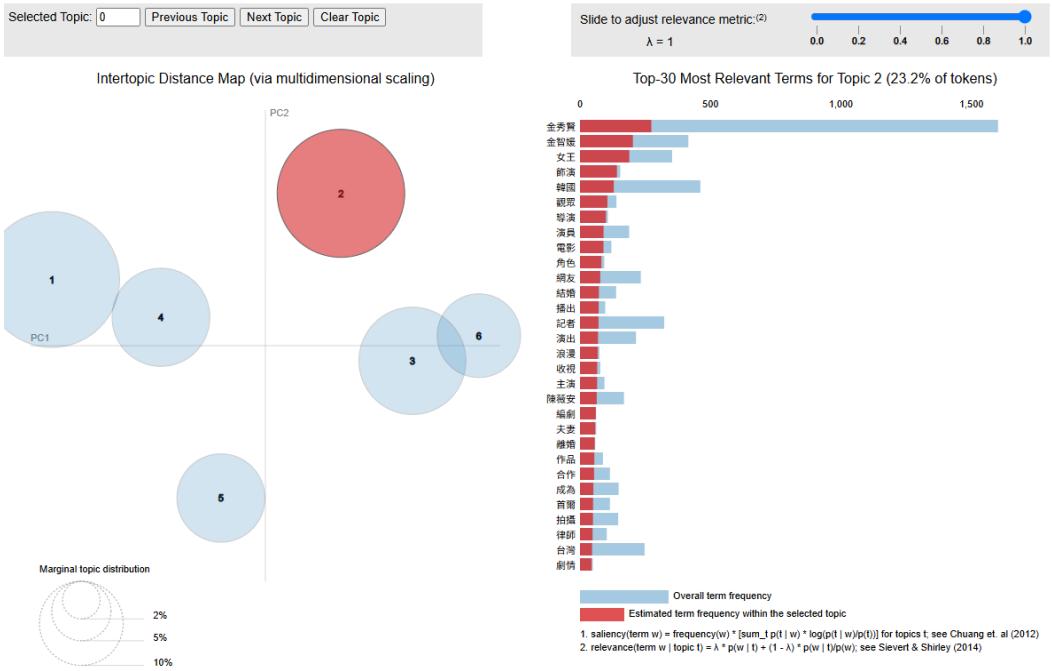
至於主題三（Topic 3），較多詞彙與粉絲活動和見面會相關，包括「粉絲」、「見面會」、「台北」、「韓國」、「活動」、「現場」、「亞州」、「主持」、「出場」等。此主題所呈現的是演員與粉絲之間的互動情境，尤其針對實體活動的報導或反應，是較偏向支持性與參與感的內容類型。

這三個主題分別代表了事件的核心事實、情感想像與偶像文化互動，彼此在語意圖中有相對明確的分布，顯示模型有良好的區辨能力，也反映出社群對事件的多面向討論。

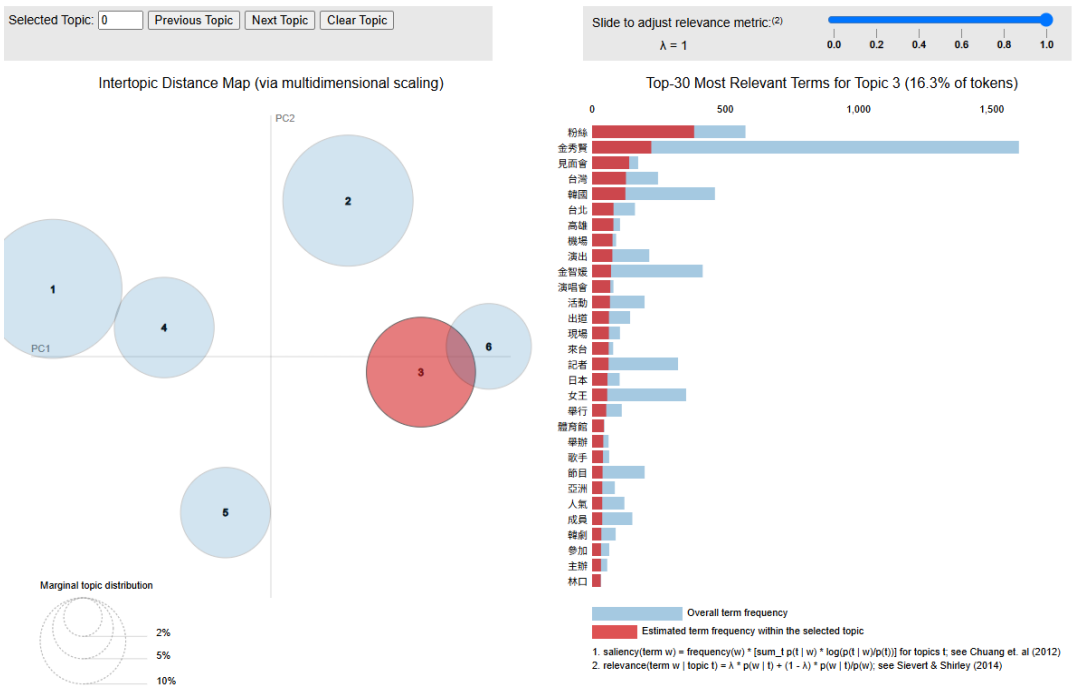
LDA Vis


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
Marginal topic distribtion
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (26.1% of tokens)
0
500
1,000
1,500
金賽綸
金秀賢
交往
公開
照片
公司
追蕭
李鎮浩
經紀公司
網友
聲明
研究所
酒駕
爆料
節目
影片
記者
韓國
結婚
母親
透過
拍攝
否認
韓元
女星
演員
阿姨
家屬
提議
時間
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

LDA Vis

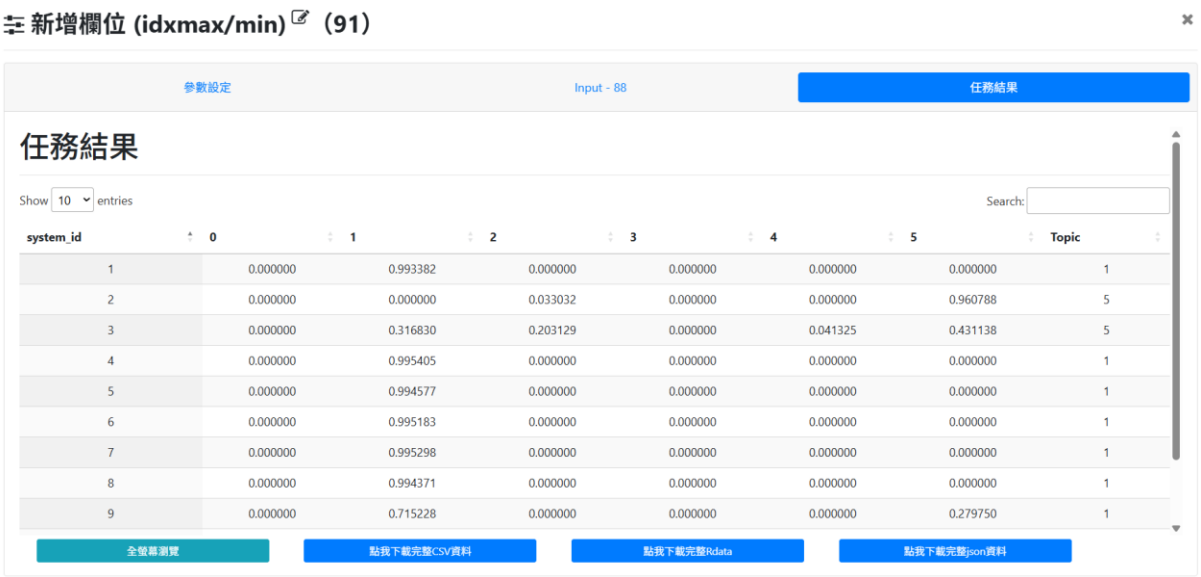


LDA Vis



為了進一步分析每篇文章在主題模型中的主要歸屬，我們使用了 Tarflow 中的「新增欄位（idxmax）」功能。透過這項設定，我們讓系統從六個主題機率中，自動找出每篇文章中占比最高的主題，並將該主題的編號記錄下來，新增為一個名為「Topic」的欄位。這樣的設定有助於我們後續將每篇文章明確分類，並針對各主題進行數量統計。

需要特別說明的是，系統預設的主題編號是從 0 開始，因此我們將 0 視為主題一，1 為主題二，依此類推，最終總共對應六個主題。透過這個步驟，我們建立了文章與主題之間一對一的最大關聯，有效簡化後續的圖表製作與分群觀察。



我們首先使用 Tarflow 中的「分群彙總（非數值）」模組，將前一步透過 LDA 建立的主題欄位（Topic）作為分群依據，並以 system_id 為計算欄位。彙總函數部分，我們選擇使用「count」，目的是計算每個主題對應的文章數量，也就是每個主題在整體語料中出現的次數。

完成這項設定後，我們進一步視覺化主題分佈的狀況，運用視覺化儀表板分別製作出長條圖與圓餅圖。

在長條圖中，我們將 X 軸設為主題，Y 軸為文章數量，圖中清楚呈現各主題在資料集中被分配的頻率。由圖可見，主題四的文章數最多，其次依序為主題一、主題六、主題二、主題三，而主題五則最少。這顯示使用者對某些特定主題有明顯的集中關注，尤其是與**復出演藝活動**相關的主題四。

另一方面，圓餅圖則透過將彙總函數改為「mean」，呈現各主題在單篇文章中所佔的平均比例。從結果來看，主題四除了文章數最多，其在文章中出現的佔比也最高（32.2%），顯示該主題在輿論中具有強烈主導性。

值得一提的是，Tarflow 將主題由 0 開始編號，因此我們實際上的主題順序應該為：「Topic 0」對應主題一、「Topic 1」對應主題二，以此類推至主題六（Topic 5）。根據 LDA 模型中各主題的代表關鍵詞，我們歸納六個主題的內容如下：

主題一：與戀情事件本身有關，包含「金賽綸」、「金秀賢」、「交往」、「照片」、「公司」等，是整起事件的爆發起點。

主題二：屬於戲劇角色與情感想像延伸，關鍵詞如「角色」、「夫妻」、「收視」、「浪漫」、「tvN」，觀眾將戲劇中的形象套用至現實情境。

主題三：聚焦在粉絲與偶像互動情境，例如「見面會」、「舞台」、「出場」、「活動」、「現場」等。

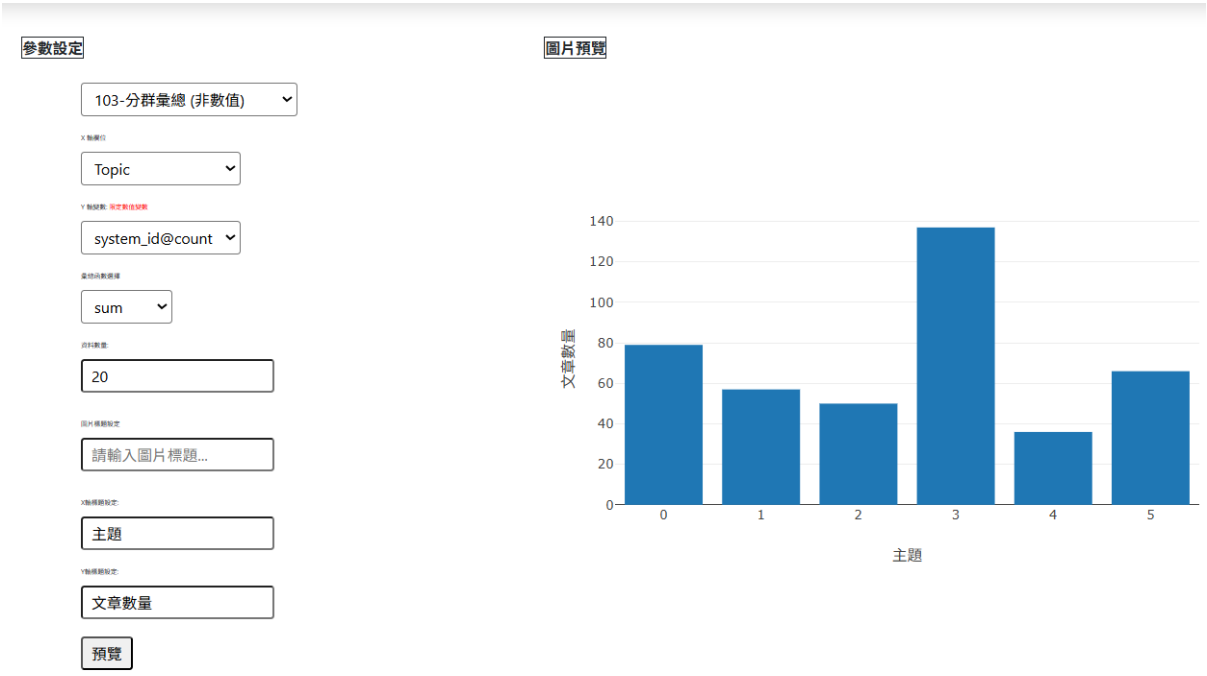
主題四：最多文章數與最高關注度的主題，關鍵詞包括「記者會」、「時間」、「作品」、「回歸」、「節目」、「復出」等，顯示社群廣泛關心事件是否會影響藝人的

職涯與未來發展。

主題五：涵蓋事件可能波及的外部單位，如「NewJeans」、「HYBE」、「合約」、「成員」、「法律」、「聲明」等，帶出連鎖效應與擴散式擔憂。

主題六：圍繞藝人形象與品牌價值，詞彙如「品牌」、「穿搭」、「風格」、「設計」、「代言」、「系列」等，反映公眾對於藝人商業形象與時尚風格的高度關注。

從長條圖與圓餅圖的主題分布來看，我們可以觀察到輿論焦點高度集中在「主題四」，這類內容主要與藝人是否復出、未來節目安排、形象修復等後續發展相關，顯示大眾對事件後續影響的關注已超越事件本身。其次為主題一與主題六，分別對應到事件初期的戀情爆料與藝人商業形象的討論，代表除了新聞本身，粉絲與消費者也在意藝人品牌價值的變動。相較之下，關於戲劇角色（主題二）與粉絲活動（主題三）的討論熱度則相對中等，僅作為輿論中的延伸補充。而涉及外部公司的主題五則出現



次數最少，反映出社群並未將責任擴散至其他無直接關聯的單位。整體來看，此次事件的社群反應呈現出從爆料關注轉向形象與未來發展的移動趨勢，且多以主角本人為核心，討論相對集中且具現實導向。

參數設定

103-分群彙總 (非數值)

Label 欄位:

Topic

Values 欄位: 設定數值變數

system_id@count

彙總函數選擇

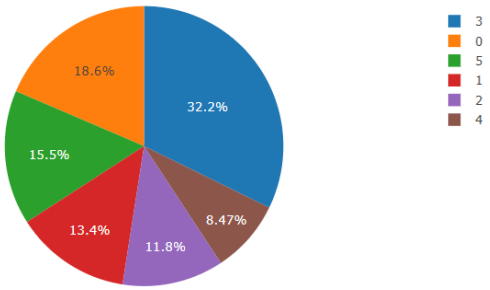
mean

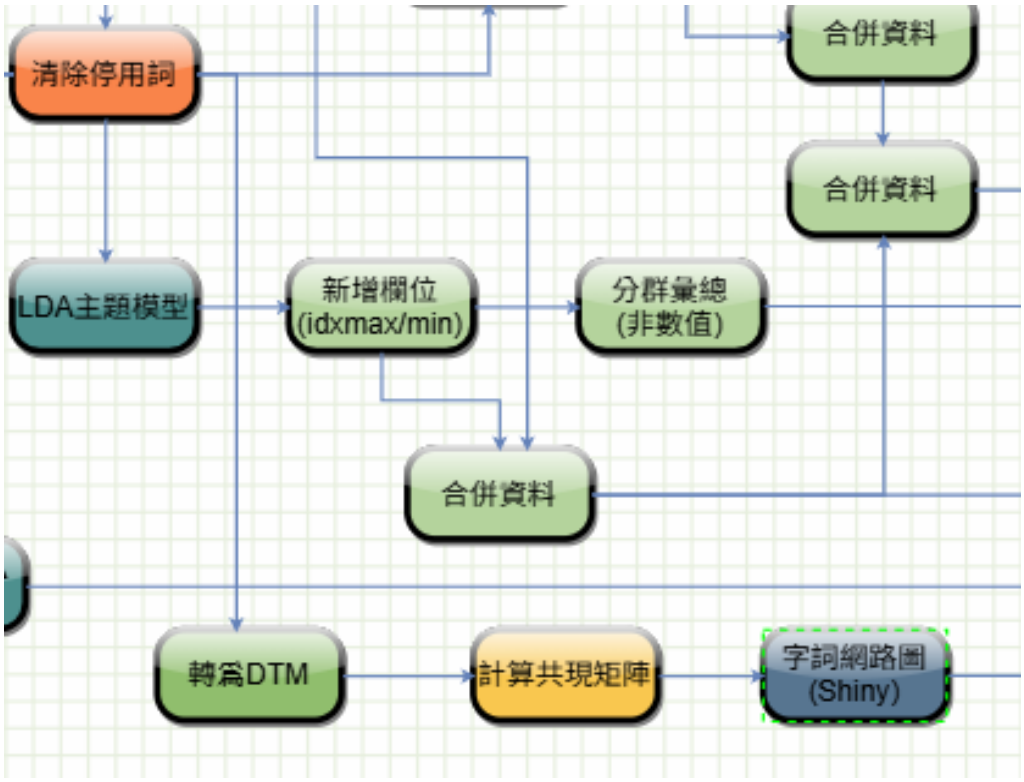
圖片標題

請輸入圖片標題...

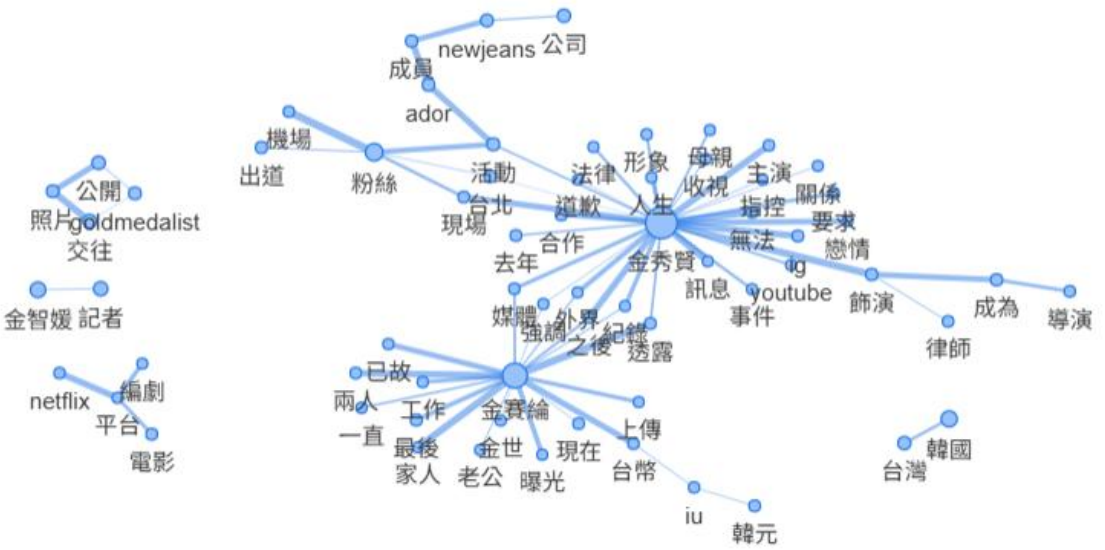
預覽

圖片預覽





我們清除停用詞後也使用了轉為DTM、計算共現矩陣(參數為原先設定)，並製作Shiny字詞網路圖。我們發現整張共現圖呈現出多個彼此有關聯但語意不同的群集，而位於圖中央的主核心，正是這起事件所帶動的輿論焦點。



在中央群集中，輿論最密集的討論區域圍繞著「事件」、「法律」、「消息」、「形象」、「媒體」、「情緒」、「關係」等詞彙，顯示出社群最關心的是事件的發展脈絡、法律責任歸屬與藝人形象受損的情況。像是 IG、YouTube、評論、粉絲等詞語也呈現出社群平台在此事件中扮演的放大與傳播角色，加速了輿論情緒的擴散與對立。

隨著事件延燒，我們觀察到話題逐漸向外擴散，延伸出其他具代表性的討論脈絡。在圖的上方區塊，出現一群與 K-pop 產業高度相關的詞彙，包括「NewJeans」、「ADOR」、「出道」、「合作」、「粉絲」等，反映出此事件引發的外溢效應，牽動未直接涉案的藝人或公司也被波及，顯示出產業生態高度連動的現象。

圖的右側則聚焦於演員形象與國際觀感，像是「演員」、「飾演」、「韓國」、「台灣」、「韓元」等關鍵詞，暗示著事件的影響已從個人層次延伸至品牌代言、地區形象與其他藝人的連帶風險，IU 等知名人物也被輿論納入評議之中。

在圖的下方，我們辨識出以「過世」、「老公」、「戀愛」、「曝光」、「家人」等詞彙為主的情感型群集，呈現出網友對金賽綸私生活與感情狀態的高度關注，甚至帶有哀悼與八卦色彩，顯示這場討論已不僅止於公眾事件，更進一步滲入個人情感與隱私層面。

最後，在左側群集，我們發現以「Goldmedalist」為核心的話題，延伸至「公關」、「記者」、「Netflix」、「電影」等詞，顯示出事件觸及了娛樂產業背後的操作層面與商業平台，輿論開始聚焦於公司應對策略、公關危機與媒體處理，牽動整體娛樂生態的信任與利益結構。

總結上述可以看到，媒體在報導這起事件時，已不再只是聚焦於娛樂八卦，而是透過不同看板呈現出更廣泛的社會關注。從時尚娛樂到國際新聞，相關報導不僅涉及藝人個人行為與形象爭議，也延伸到法律責任、商業影響與國際觀感等議題。可以看出，這場事件所引發的討論，已不再只是單一藝人新聞，而是反映出媒體如何透過語言與版位安排，引導輿論關注，形塑出一場涵蓋社會價值、形象認知與集體情緒的多層次公共對話。

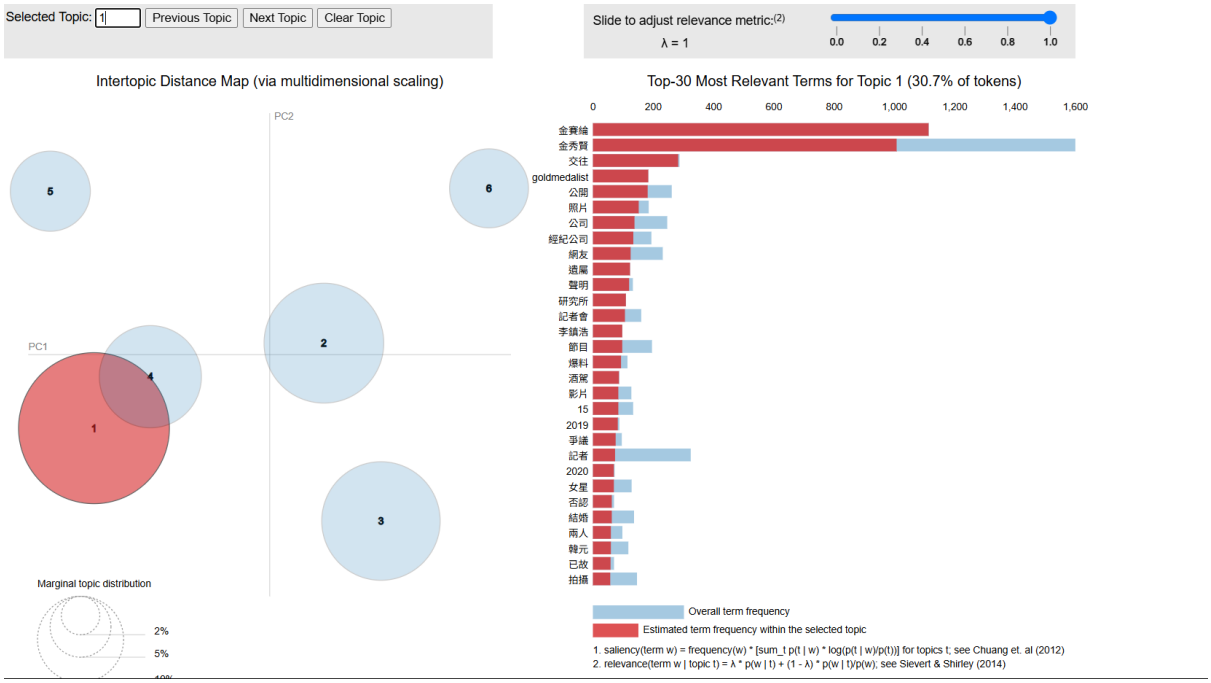
另外我們使用 GuidedLDA 模型，設定 6 個主題進行建模，並依據事前設計的種子字（法律，官司，訴訟，證據，誹謗，法院，律師，禁令，傳喚，刑事，民事，提告，拘留，辯護，形象，公眾，道德，責任，風波，丟臉，代言，損害，信任，公關，演員，藝人，聲譽，名氣，戀愛，交往，緋聞，七年，兩人，老公，戀情，約會，私生活，情侶，關係，曝光，照片，影片，粉絲，留言，討論，推文，評論，攻擊，支持，聲援，酸民，網友，dcard，ig，youtube，熱搜，過世，離世，哀悼，懷念，已故，痛心，悲劇，自殺，家屬，追思，安息，消息，新聞，震驚，記者，報導，轉載，平台，網路，爆料，視頻，輿論，網紅，節目，頻道，傳播，點閱，goldmedalist，公司，合約，賠償，品牌，代言，合作，解除，收入，商業，損失，活動，停播）作為語意引導，確保模型能聚焦在事件相關的討論面向上。這樣的設計有助於提升主題的可解釋性，避免出現模糊或無意義的分類。

從主題視覺化圖中，我們可以觀察主題之間的語意距離。圖中每一個圓形代表一個主題，大小表示該主題佔據語料的比例，圓與圓之間的距離則反映它們的語意相似度。整體來說，六個主題的分布具有良好的區隔性，其中主題 1（紅色）所佔比例最大（約 30.7%），顯示它是整體語料中最主要的討論焦點。

主題一：事件核心與金賽綸爆料風波

關鍵詞包含：金賽綸、金秀賢、交往、goldmedalist、公關、爆料、照片、影片、節目、媒體、記者、事件、曝光

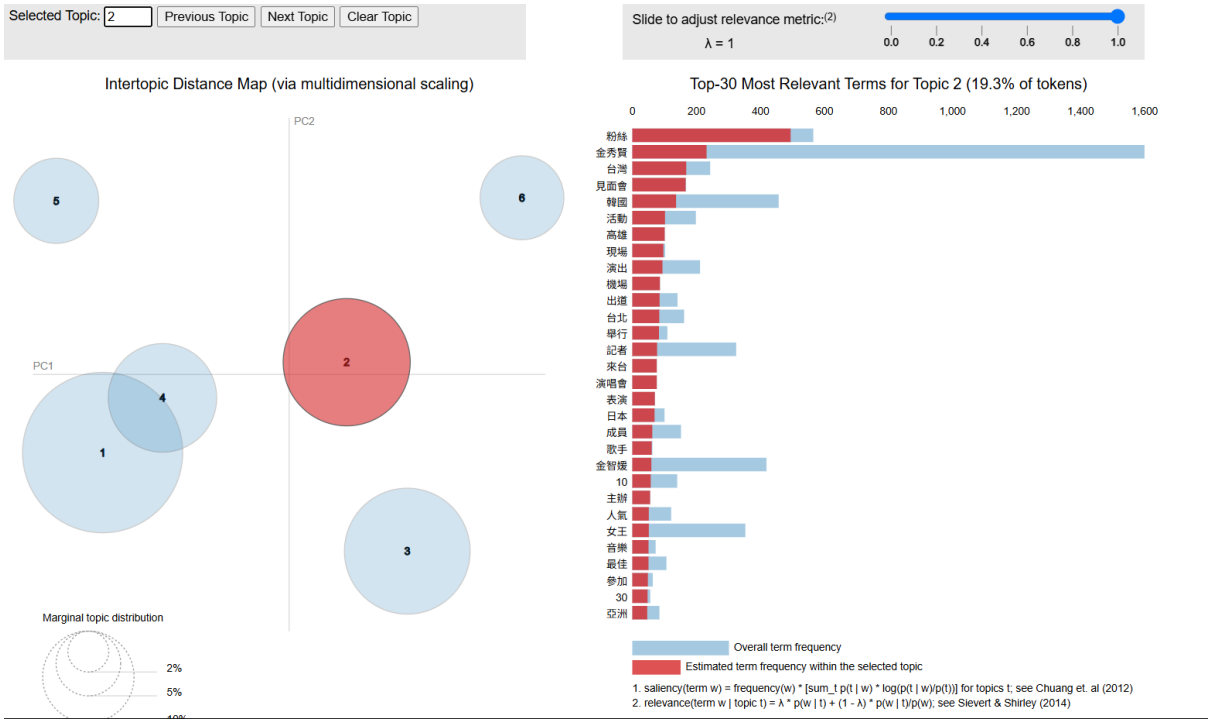
這是整場輿論風暴的核心，聚焦在金賽綸與金秀賢的交往爆料，以及經紀公司 Goldmedalist 相關爭議。新聞中所流出的照片與影片也成為網友討論與再製的熱點，引發大量情緒反應與媒體關注。



主題二：粉絲互動與在東亞國家的形象影響

關鍵詞包含：粉絲、見面會、台灣、韓國、活動、現場、演唱會、橫幅、女王、亞洲、參加

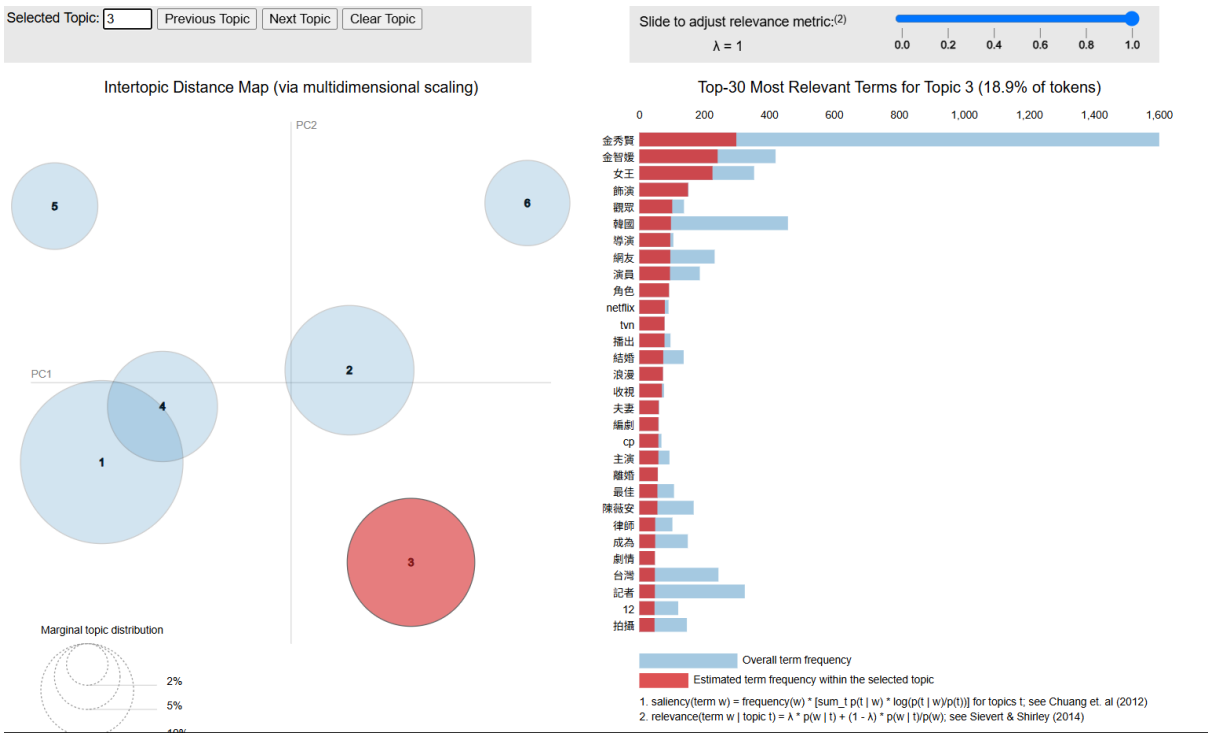
此主題聚焦在粉絲與實體活動的互動，包括金秀賢、金智媛等人在台灣、日本等地的見面會與演出。討論也延伸至粉絲現場應援、橫幅、媒體報導等反應，顯示藝人形象如何在亞洲文化中被放大。



主題三：戲劇角色與戀情投射幻想

關鍵詞包含：金秀賢、金智媛、女王、飾演、角色、tvN、結婚、夫妻、浪漫、收視、主演、律師、劇情

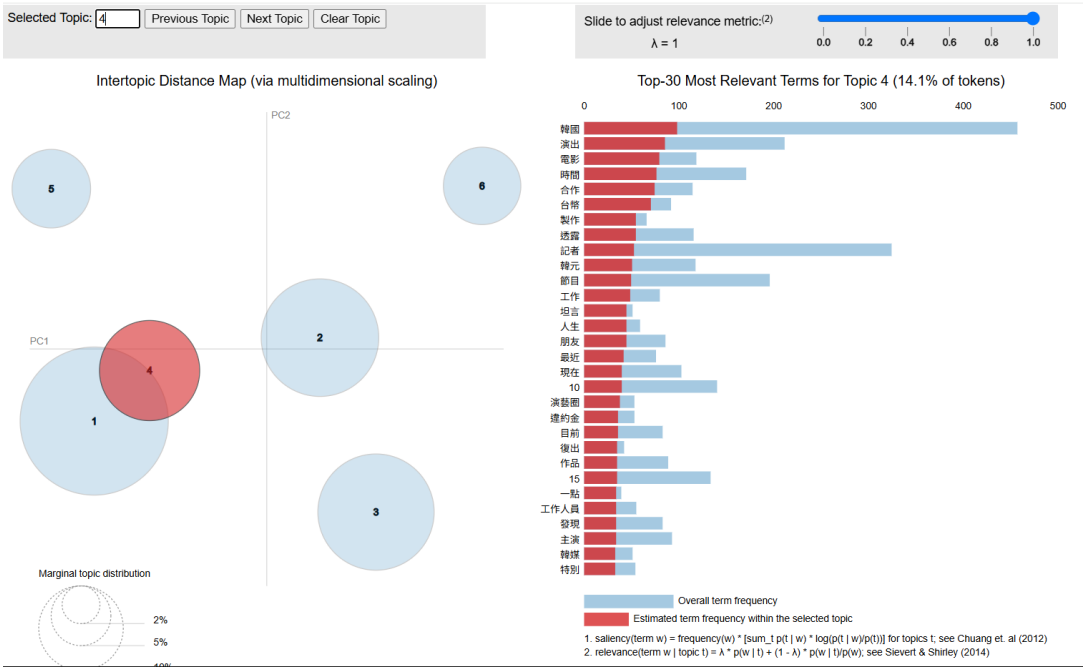
這是一個高度情緒化的主題，反映觀眾將《淚之女王》中角色關係投射到現實生活的傾向。網友不僅討論劇情，更將兩位主角的螢幕形象與真實戀情做連結，產生戲裡戲外模糊地帶的討論熱潮。



主題四：韓國影視產業與藝人回歸動態

關鍵詞包含：韓國、演出、電影、節目、製作、合作、朋友、近期、復出、主流、記者

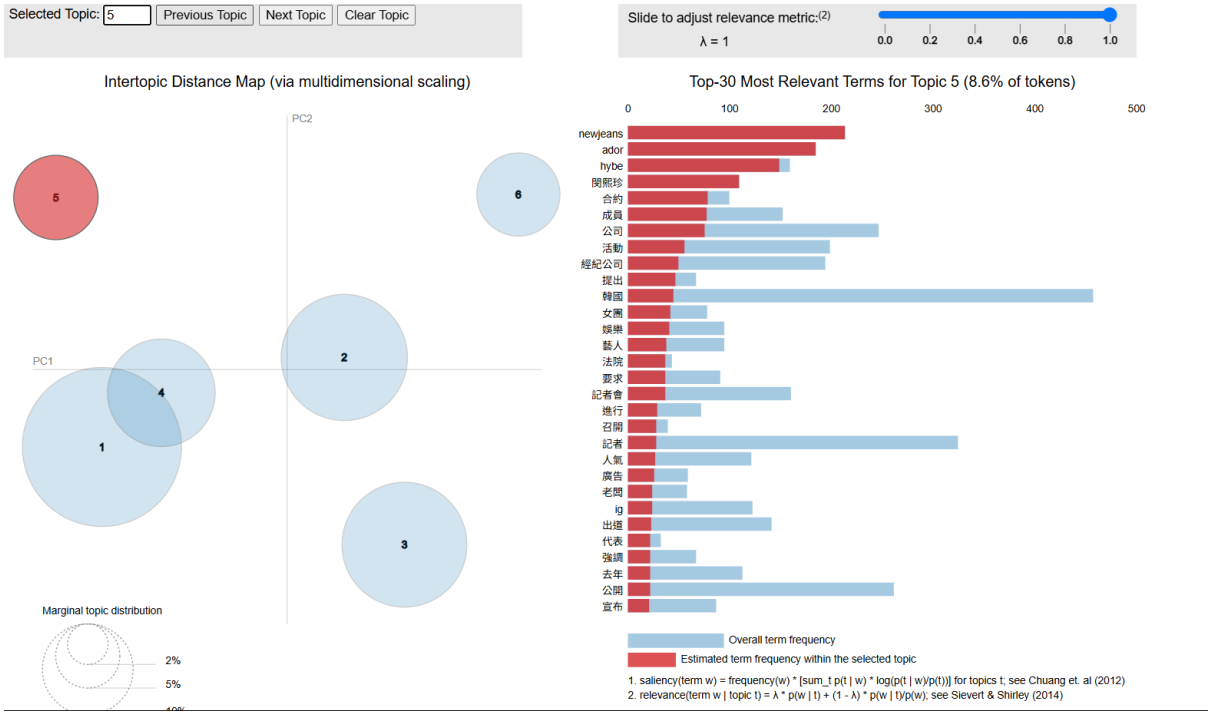
這個主題關注的是影視產業對事件的回應與未來規劃，包括藝人是否能如期復出、節目製作是否受到影響，以及與其他團隊或國家合作的變化。新聞媒體與製作人角色頻繁出現，顯示輿論延燒至幕後圈層。



主題五：娛樂圈外溢與 NewJeans 牽連

關鍵詞包含：newjeans、ador、hybe、經紀公司、藝人、活動、記者會、合約、要求、公告、代表

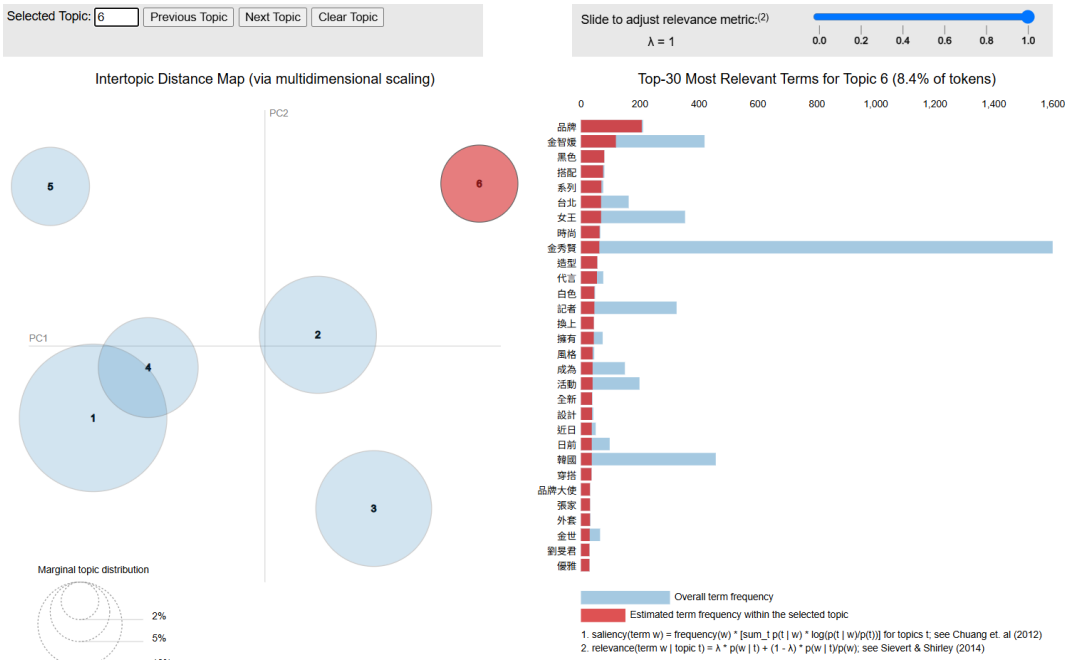
該主題呈現輿論外溢效應，NewJeans 與 HYBE 公司雖未直接涉案，卻因關聯詞彙或粉絲間的對比而被捲入話題。這反映了娛樂圈高度連動的風險結構，也暴露出社群如何透過簡單連結製造「連帶責任」或「錯誤關聯」。



主題六：形象風格與商業代言影響

關鍵詞包含：品牌、金秀賢、黑色、風格、代言、穿搭、優雅、系列、記者、拍攝、設計

此主題則屬於商業與時尚層面的探討，聚焦在事件對金秀賢形象與代言品牌可能帶來的影響。是否穿著象徵哀悼的服裝？品牌是否與其形象切割？都成為輿論討論的一部分，也牽動著背後龐大的廣告與時尚利益鏈。



從主題比例來看，主題一的占比最高，顯示民眾對「事件本身」的真相與後續發展最為關心。主題三與主題一距離較近，反映出「戀情幻想」與「事件爆料」之間語意相連，網友常將戲劇角色與現實混為一談。而主題六則比例最低（8.4%），顯示儘管商業影響在討論中存在，但情緒與八卦色彩的主題仍為主流。

由於本次資料來自《蘋果日報》的「時尚娛樂」與「國際」兩大新聞看板，並非直接蒐集自社群互動，因此雖無法推論實際的留言或分享行為，但從詞彙如「留言」、「IG」、「記者」、「粉絲」、「平台」、「youtube」等高頻出現來看，仍可推估此事件於社群間具有高度討論度與媒體擴散性。

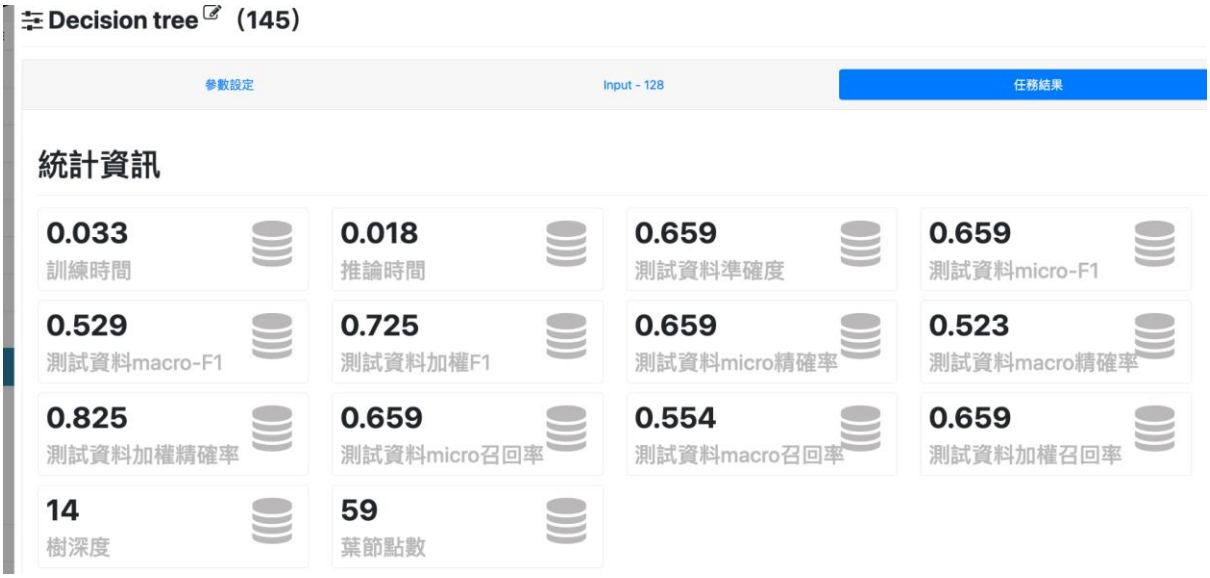
整體而言，我們也觀察到同樣的現象: 這場輿論風暴已超出單純的名人八卦層次，演變成了一場跨媒介的社會性集體討論。一開始南韓女演員金賽綸的離世消息震驚各界，隨後媒體陸續揭露男星金秀賢在金賽綸未成年時曾與她交往的內幕，而相關影片與細節也在社群平台上瘋狂轉傳，引發社會大眾強烈反應。

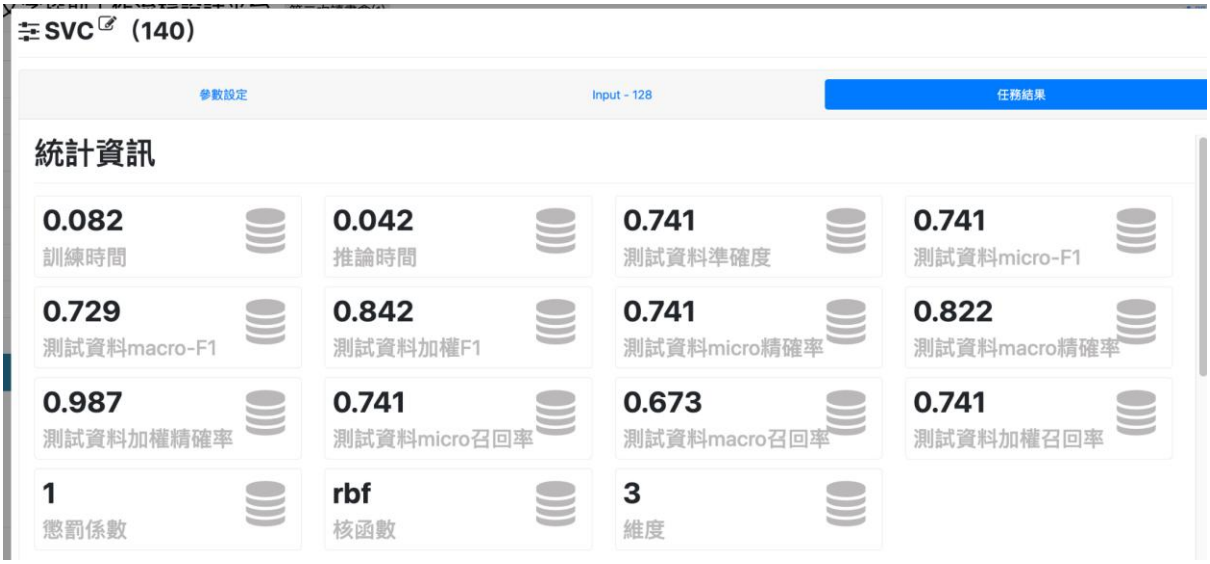
討論主軸也隨之從戀情爭議擴展到粉絲文化、戲劇角色幻想、藝人形象維護與品牌代言風險，甚至進一步波及跨國娛樂產業與經紀公司制度。這些主題彼此交錯，反映出大眾對於「藝人應承擔的社會責任」、「私生活與公眾形象的界線」以及「媒體與社群平台在輿論形成中所扮演的角色」等議題的持續關注與討論。從新聞報導到社交媒體的擴散與詮釋，我們也能看見，大眾如何透過這些平台參與討論，進一步建構出對這起名人爭議的共同理解與價值判斷。

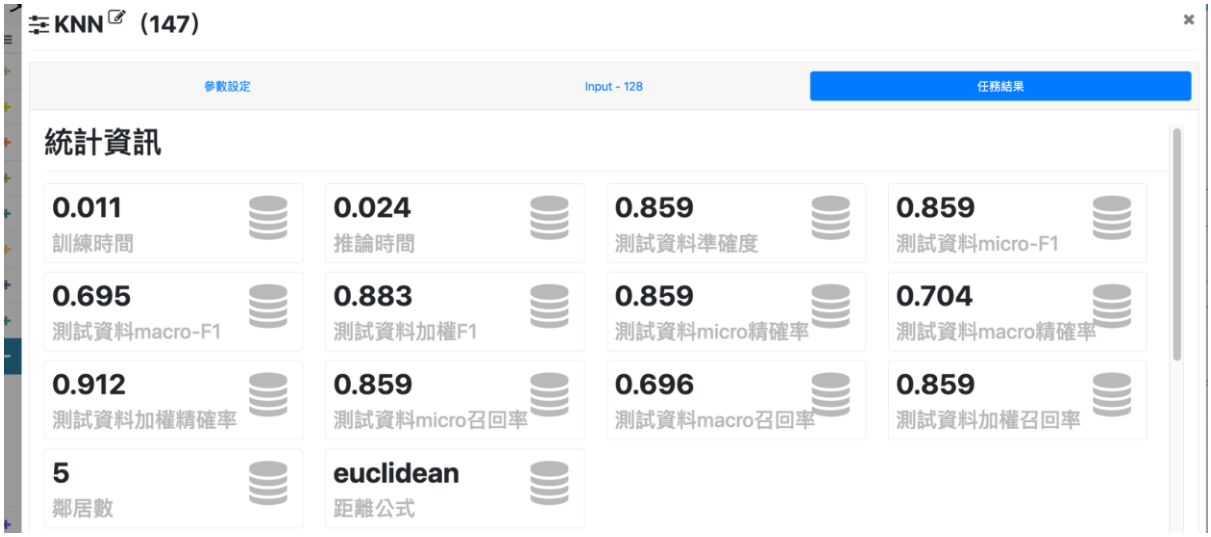
分類器結果：

我們將主題模型分類結果與原始資料進行合併後，進行分類器訓練。分別為decision tree 、SVC、KNN


Train&Test (128)
參數設定
Input - 124
任務結果
目標欄位 *
Topic
測試資料切割比率 *
0.2
是否隨機排序資料
否
亂數種子
777
儲存更改


Decision tree (145)
參數設定
Input - 128
任務結果
統計資訊
0.033
訓練時間
0.018
推論時間
0.659
測試資料準確度
0.659
測試資料micro-F1
0.529
測試資料macro-F1
0.725
測試資料加權F1
0.659
測試資料micro精確率
0.523
測試資料macro精確率
0.825
測試資料加權精確率
0.659
測試資料micro召回率
0.554
測試資料macro召回率
0.659
測試資料加權召回率
14
樹深度
59
葉節點數


SVC (140)
參數設定
Input - 128
任務結果
統計資訊
0.082
訓練時間
0.042
推論時間
0.741
測試資料準確度
0.741
測試資料micro-F1
0.729
測試資料macro-F1
0.842
測試資料加權F1
0.741
測試資料micro精確率
0.822
測試資料macro精確率
0.987
測試資料加權精確率
0.741
測試資料micro召回率
0.673
測試資料macro召回率
0.741
測試資料加權召回率
1
懲罰係數
rbf
核函數
3
維度



結果如下表:

衡量指標	Decision tree	SVC	KNN
測試資料準確度	0.659	0.741	0.859
測試資料加權F1	0.725	0.842	0.883
測試資料加權精確率	0.825	0.987	0.912
測試資料加權召回率	0.659	0.741	0.859

- **KNN 的整體表現最佳**，在測試資料準確度（0.859）、加權 F1 值（0.883）與加權召回率（0.859）等多項指標中皆取得最高分，顯示其在辨識各類別上具有穩定且均衡的表現。
- **SVC 緊追在後**，其加權精確率達到 0.987，為三者中最高，顯示它在預測結果的正確性上表現優異，適合用於對錯誤預測較敏感的任務。
- 相較之下，**Decision Tree 表現較弱**，尤其在加權召回率（0.659）和加權 F1（0.725）上與其他兩者有明顯落差，可能存在過度簡化資料結構的情況。

反思與限制

1. **語料來源偏新聞導向，缺乏用戶互動語言**
   資料僅來自《蘋果日報》的時尚娛樂與國際看板，缺乏社群資料（如 Dcard、Twitter 留言），導致模型學習的語言風格偏新聞化，與真實輿論情緒略有落差。

2. **主題模型仍有語意模糊空間**
   雖透過 GuidedLDA 提高解釋性，但個別主題仍存在交疊。例如「戀情幻想」與「粉絲活動」中的部分詞彙重複，可能導致分類邊界模糊。

3. **輿論變化的動態性未充分掌握**
   雖涵蓋時間範圍逾一年，但未進行主題趨勢隨時間變化的追蹤分析，未能呈現事件討論熱度如何隨社會議題演變。

4. **未進行負面與正面情緒的分層**
   雖然主題涵蓋情緒性內容，但未對文本情緒極性做區分，無法明確得知每主題中正面支持與負面批評的比例與趨勢。