社群媒體分析 Project2

疫情解封後的報復 - 日本旅遊主題分析

指導老師：黃三益 教授

組別：第十六組

組員：

N074220025 陳昱廷

N104020009 蔡雨臻

N104020010 盧貴聰

N104220017 胡冠甄

目錄

1. 動機和分析目的---------- 3
2. 資料集描述 ---------- 4
3. 資料的分析過程---------- 5
4. 視覺化的分析結果與解釋---------- 14
5. 結論 ---------- 23

1. 動機和分析目的

a. 主題：「疫情解封後台人出國首選日本，最愛景點大公開」

b. 議題發想動機：

去年 10 月開放出境後，累積兩年一系之間大爆發，台人出國首選國家就屬日本，是哪個地區景點讓台灣人最為想念，一解封就迫不及待想要去呢？許多航空推出短程快閃日本行程、降價促銷航班，吸引消費者蜂擁而至，更有不少旅行社趁機推出誘人賞櫻、賞富士山行程等，限定班次及機位讓消費者搶破頭也想搶到。是什麼樣的名勝景點，讓人甚是想念壓抑不住想出國的慾望呢？

因此，我們想要知道在 PTT 日本旅遊版上，蒐集討論熱度最高的景點文章，取得鄉民們最有興趣的日本聖地，以了解日本讓人風靡也癡迷的箇中原因，故此本研究為 -「疫情解封後台人出國首選日本，最愛景點大公開」。

c. 分析目的：

規劃行程需要蒐集很多旅遊部落客或鄉民的經驗分享，常有旅遊前輩會提供懶人包以供新手或有意願前往該地的鄉民可以快速了解當地旅遊注意事項及購物攻略。藉由鄉民們的分享等相關文章言論的情緒分析，從發言中得到關於日本旅遊寶典，使得日本行能夠不只玩得用力、買得用力、內心超快樂。

d. 使用平台：文字探勘工作流程平台

e. 報告影音連結：https://youtu.be/Bp6lKdKZJ5I

2. 資料集描述

a. 工作流程：project2_N010

b. 資料來源：PTT-Japan_Travel

c. 資料時間：2023/01/01 - 2023/06/05

d. 排除關鍵字：政治、徵人、讓售

e. 資料筆數：4,894 筆

f. 流程概述：

i. 爬取 PTT_Japan Travel 今年度發布之文章內容，進行資料清理。

ii. 以「替換字串」將詞彙斷詞分行、去除多餘空格，同一個字詞。

iii. 再以「中文斷詞」將內容分解成字詞單位，並使用「清除停用詞」將不必要的符號、單字元去除，最後設定停用字以過濾出現頻率高但無意義的字詞。

iv. 將清理好的文章字詞使用 Lexicon Based 情緒分析，期望能看出大家今年對日本旅遊的體驗與期待如何。

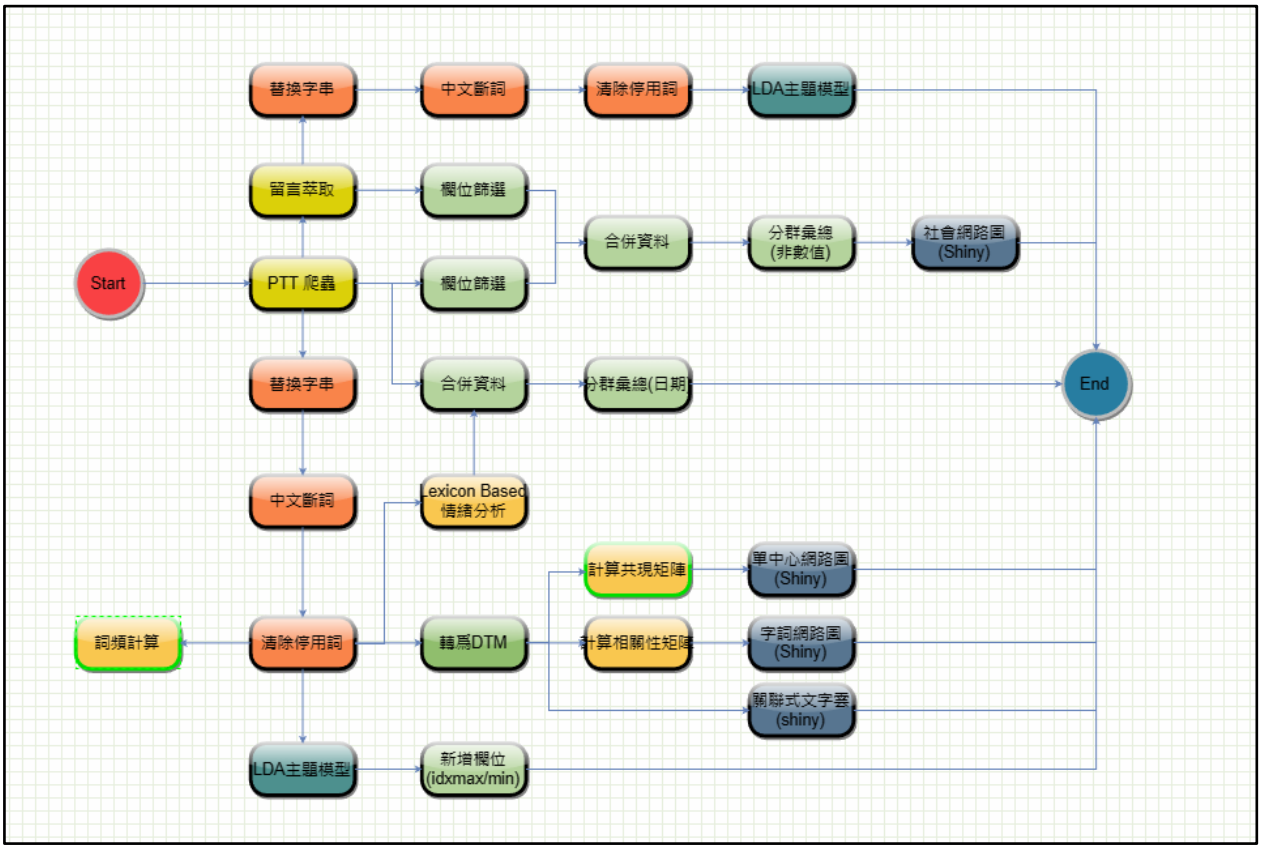
v. 將原爬蟲資料與留言萃取資料篩選適當欄位且進行合併， 為了進行之後的社會網路圖，需先進行分群彙總。

vi. 將清理好的資料轉為 DTM，針對前 200 名詞彙結果，以熱圖查看兩種維度之間的關聯性。

vii. 將清理好的資料進行 LDA 主題模型，針對主題出現字詞分析。

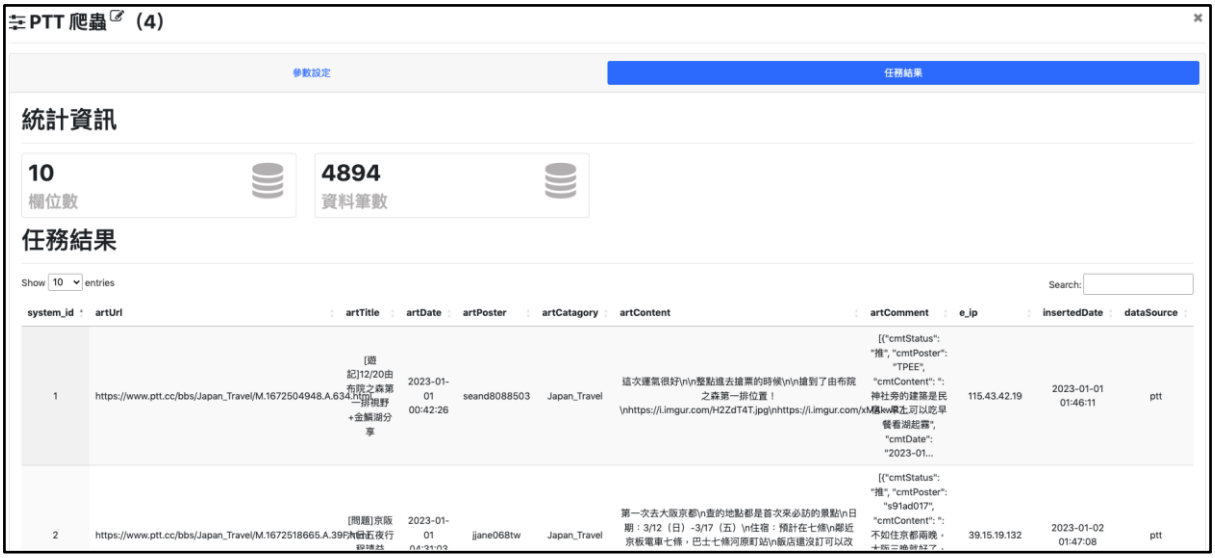
3. 資料的分析過程

a. 工作流程設計：



b. 爬蟲與資料清理

資料來源：選擇 PTT 爬取 2023/01/01 - 2023/06/05 期間，Japan Travel 版文章資料，共 4,894 筆資料。



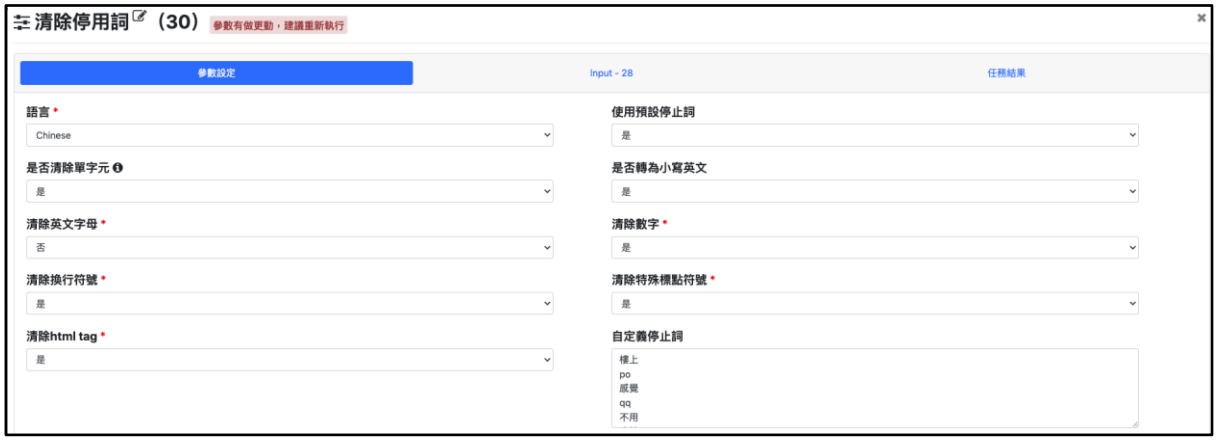
c. 資料前處理(替換字串、中文斷詞、清除停用字)：

i. **替換字串**：去除換行符號、網址、html 標籤、底線與多餘的空格

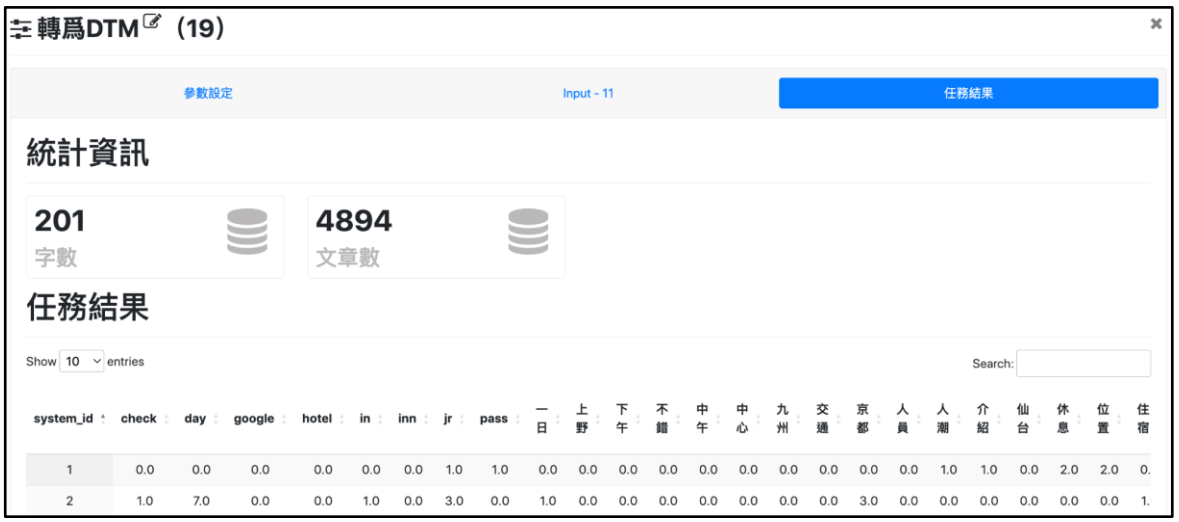
ii. 中文斷詞：


中文斷詞 (9)
參數有做更動，建議重新執行
參數設定
Input - 7
任務結果
統計資訊
5978
最大字數
1934143
總字數
0
最小字數
395
平均字數
任務結果
Show 10 entries
Search:
system_id
result
1
[這次, 運氣, 很, 好, 。, 整點, 進去, 搶票, 的, 時候, 。, 搶到, 了, 由, 布院, 之森, 第一排, 位置, ！, 、, 、, 、, 這台, 由, 布院, 之森, 很, 明顯, 外國, 觀光客, 超級, 多, 。, 尤其, 是, 韓國人, 跟, 台灣人, XDDDD, 。, 接下來, 進入, 正題, 。, 上車, 後, 我, 原本, 想, 先到, 販賣部, 領, 預定, 的, 便當, 。, 和, 購買, 限定, 甜點, 和, 啤酒, 。, 結果, 發現, 販賣部, 大排長龍, 居然, 排, 了, 半節, 車廂, 。, 果斷, 先, 放棄, 回, 自己, 的, 位置, 、, 。, 也, 建議, 大家, 選位, 的, 時候, 。, 不要, 選太, 靠近, 販賣部, 的, 座位, 。, 不然, 會, 被, 排隊, ...]
2
[第一次, 去, 大阪, 京都, 、, 查, 的, 地點, 都, 是, 首次, 來, 必訪, 的, 景點, 、, 日期, ：, 3, /, 12, （, 日, ）, -, 3, /, 17, （, 五, ）, 、, 住宿, ：, 預計, 在, 七條, 、, 鄰近, 京板, 電車, 七條, 、, 巴士, 七條, 河原, 町, 站, 、, 飯店, 選沒釘, 可以, 改, 。, 、, DAY1, 、, 17, :, 45, 關西, 機場, 、, 搭, haruka, 至, 京都, 車站, 、, 拖, 行李, 至, 飯店, check, in, (, 走路, 15, 分鐘,) 、, 放完, 行李, 、, 搭, 巴士, -, 四條, 河原, 町, -, 鴨川, 納涼, 床看, 夜景, 、, 吃, 燒肉, -, 巴士, ...]
3
[造訪, 日期, ：, 2022, /, 12, 。, 圖文, 網誌, 版, ：, 、, 、, 因為, 疫情, 隔, 了, 快, 三年, 沒到, 日本, 、, 這次, 打算, 來, 完成, 一些, 願望, 清單, 、, 其中, 一項, 就是, 想到, 搖曳, 露營, 中, 出現, 過的, 露營場, 露營, 看看, ！, 、, 、, 這次, 計畫, 前往, 的, 是, 在, 搖曳, 露營, 第一, 集中, 出現, 的, 浩, 庵, 露營場, ：, 。, 浩庵, 牛, ヤ, ン, ブ, 場, 、, 山梨, 県, 南, 巨, 摩, 郡, 身, 延, 町, 中, ノ, 倉, 2926, 、, TEL, :, 0556, -, 38, -, 0117, 。, 無奈, 日本, 冬天, 的, 晚上, 實在, 太冷, 了, 、, ...]
4
[想, 詢問, 日本, 的, 磯, 丸, 水產, 大概, 晚上, 21, :, 00, 之後, 、, 能不能, 帶, 2, 歲, 小朋友, 過去, 用餐, ？, 、, 因為, 去, 完, 樂園, 可能, 都, 超過, 21, :, 00, 了, 、, 或者, 鳥, 貴族, 以及, 其, 他, 居酒屋, 類型, 能不能, 帶, 孩童, 進去, 呢, 、, ？, 。, 友人, 曾經, 被, 拒絕, 帶, 孩童, 進入, 磯, 丸, 水產, 居酒屋, 之類, 的, 、, 感謝, 大家, 解惑, 一下, 、]

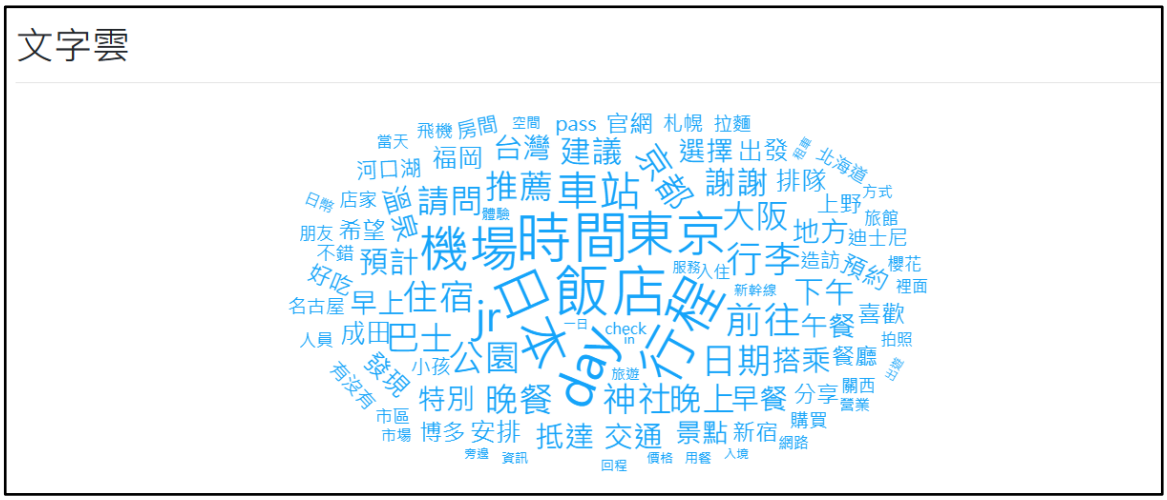
iii. **清除停用詞**：除了預設的常用停用字外，在進行了幾項分析後增加清除一些可能干擾分析結果的停用字(例如：樓上、po、感覺、qq、不用、分鐘、小時、這天、光光、啊啊等) 。



iv. **轉為 DTM**：將清理後的資料轉為 DTM，設定保留最多使用前 200 個詞彙。

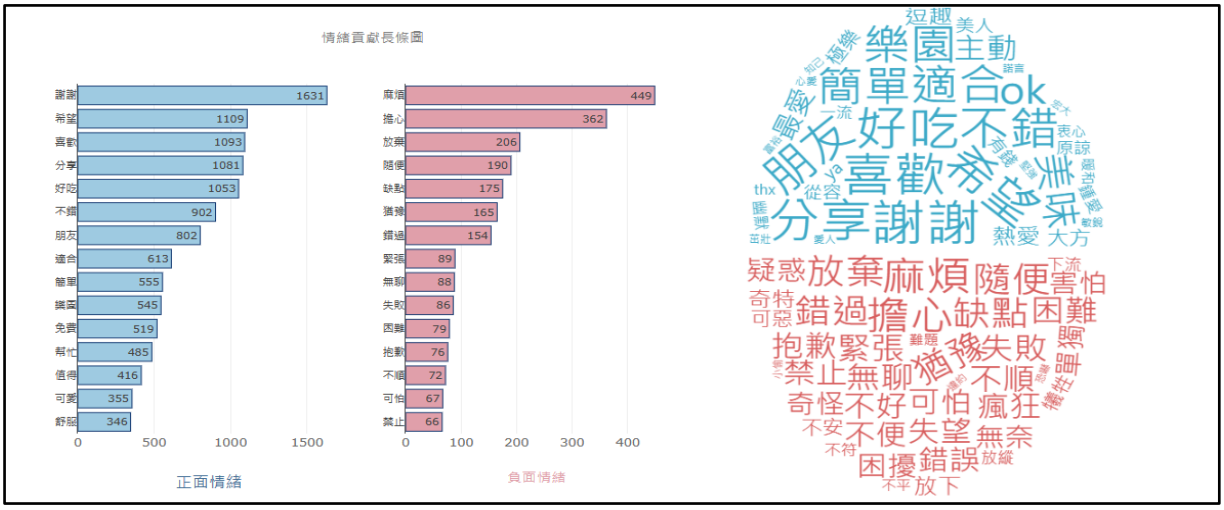


v. **<u>詞頻計算(文字雲)</u>**：透過文字雲我們可以清楚得知於文章中較常出現的討論字詞如：行程、飯店、神社、迪士尼、機場、溫泉、餐廳、景點等等。

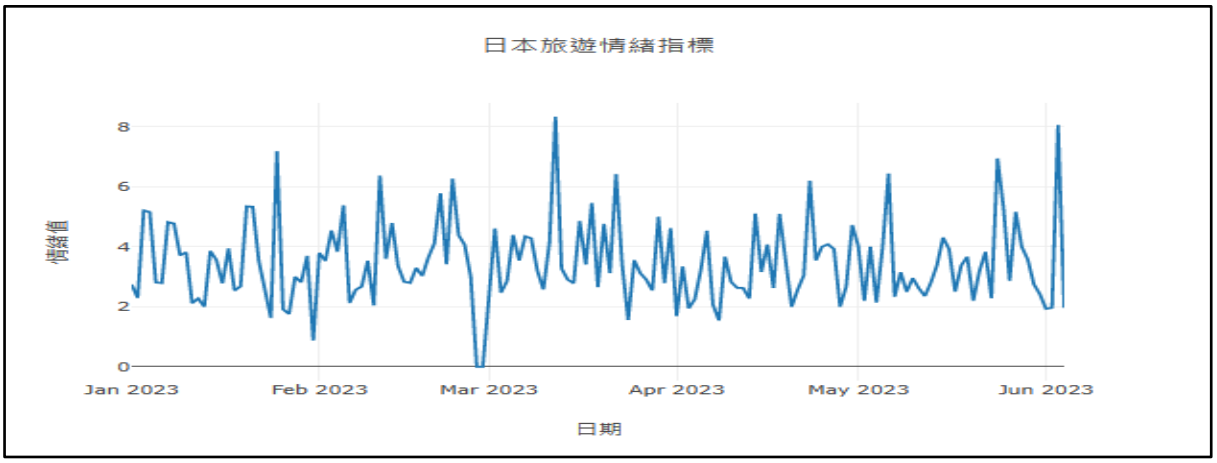


d. 情緒分析

i. **Lexicon Based 情緒分析**：為了找出網友對於日本旅遊的期待與情緒反應，故採用「Lexicon Based」情緒字典。由下圖可得知，對於日本旅遊，網友們的正向詞彙遠遠大於負向詞彙，從中得知包含：樂園、美味、好吃等等可以猜測網友們對於遊樂園(環球、迪士尼等等)與美食都有相當正面的評價呢。

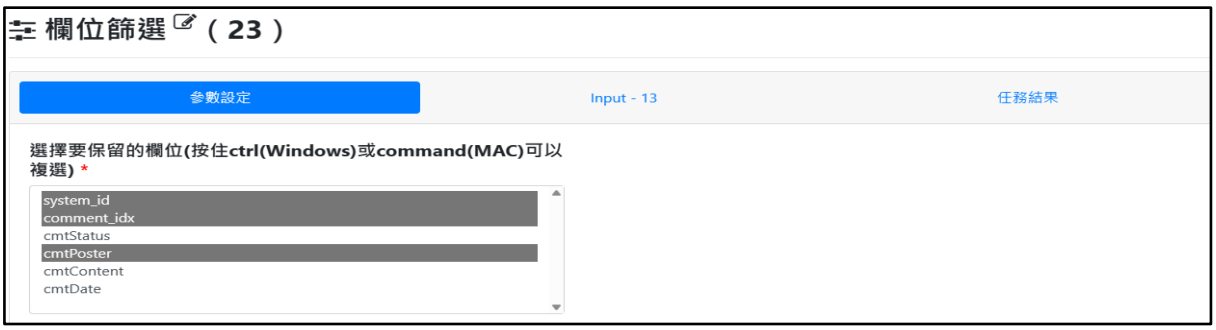


ii. **日本旅遊情緒日波動**：由下圖可以得知，網友們今年度於日本旅遊版討論關於日本旅遊的文章，幾乎每日的平均情緒分數皆為正向，可見國人對於日本旅遊的期待與喜愛。

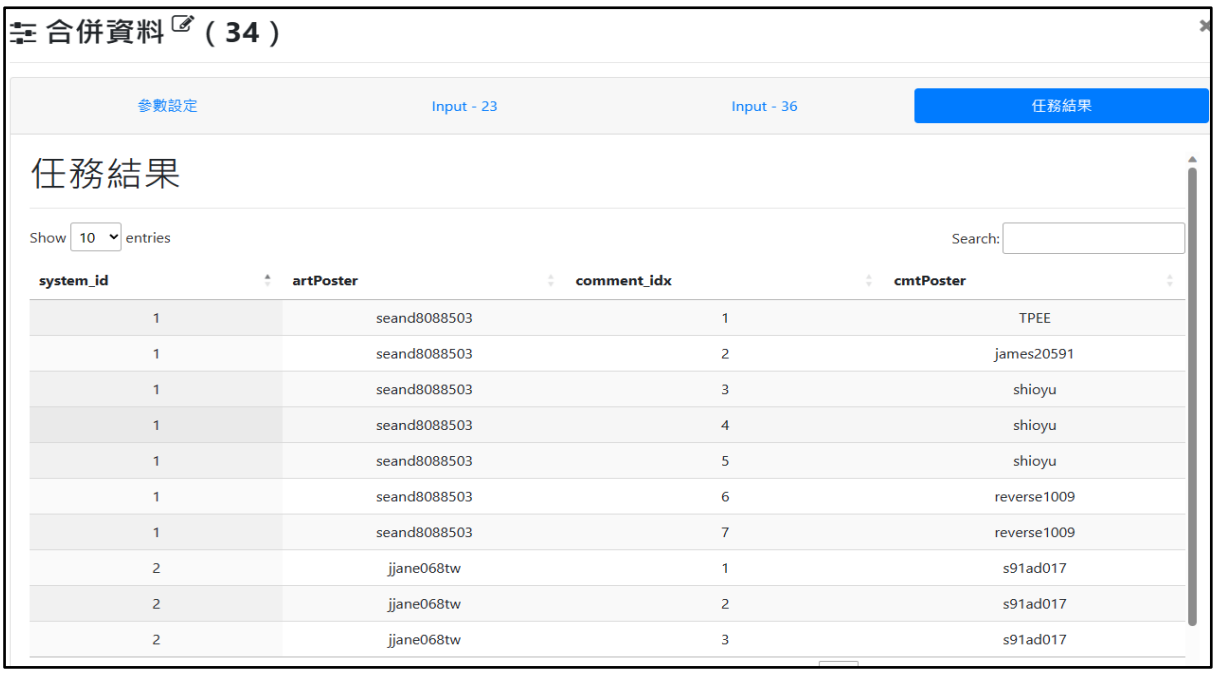


e. 發文者與留言者關係-社會網路圖

i. **資料合併**：將 PTT 文章資料與留言萃取後結果，經過欄位篩選留下需要欄位後，再將之依 system_id 進行合併。

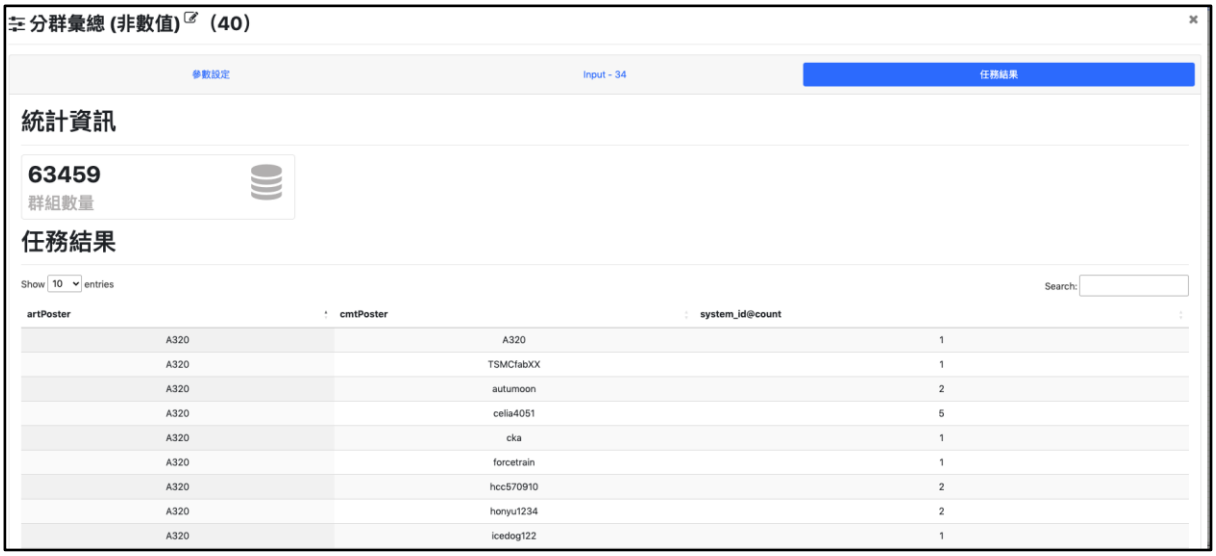


ii. **合併後結果**：將「PTT 爬蟲」與「留言萃取」兩者進行欄位合併，可綜合出文章 ID 及發文者與留言者資料。



system_id	artPoster	comment_idx	cmtPoster
1	seand8088503	1	TPEE
1	seand8088503	2	james20591
1	seand8088503	3	shioyu
1	seand8088503	4	shioyu
1	seand8088503	5	shioyu
1	seand8088503	6	reverse1009
1	seand8088503	7	reverse1009
2	jjane068tw	1	s91ad017
2	jjane068tw	2	s91ad017
2	jjane068tw	3	s91ad017

iii. **分群彙總**：選取「發文者」跟「留言者」欄位進行分群，統計結果共有 63,459 筆留言數。

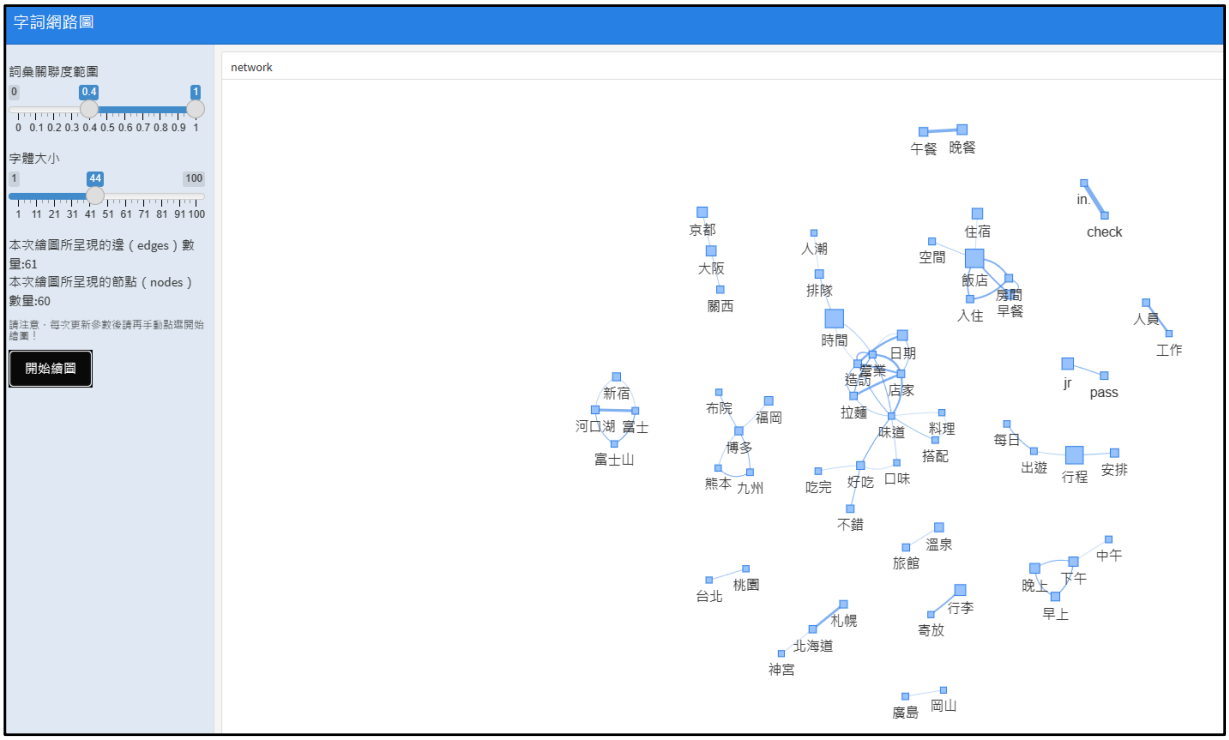


iv. **社會網路圖**：於視覺化分析結果章節詳述。

4. 視覺化的分析結果與解釋

a. 字詞關聯度：

將文章中較常出現的前 200 的字詞依照各自關聯度產出字詞網路圖，關聯度範圍設定為 0.4~1。從中可以大致推測出較常討論的有美食、地名景點、行程安排、住宿等相關主題。



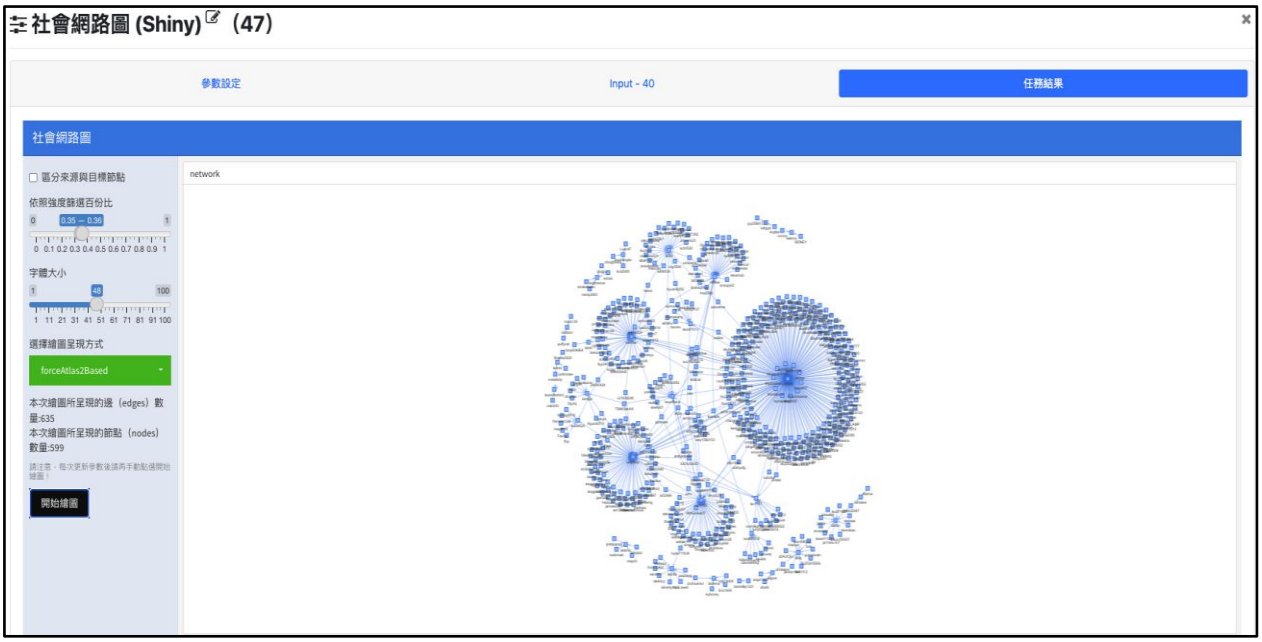
b. 社會網路圖（Shiny）：發文者與貼文者之間關係。

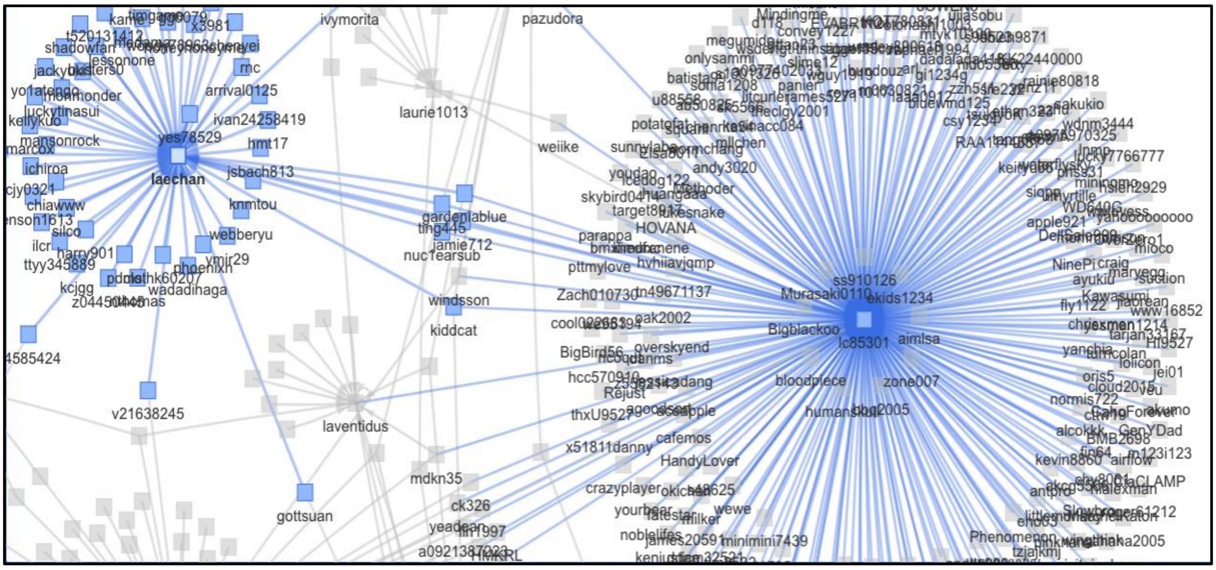
i. **節點欄位（來源）：artPoster**

ii. **節點欄位（目標）：cmtPoster**

iii. **連結欄位：system_id@count**

iv. **節點聚落：**觀察發文者與留言者之間的互動連結關係，以下圖可知道以發文者「lc85301」、「laechan」為核心發展出一個關係聚落。



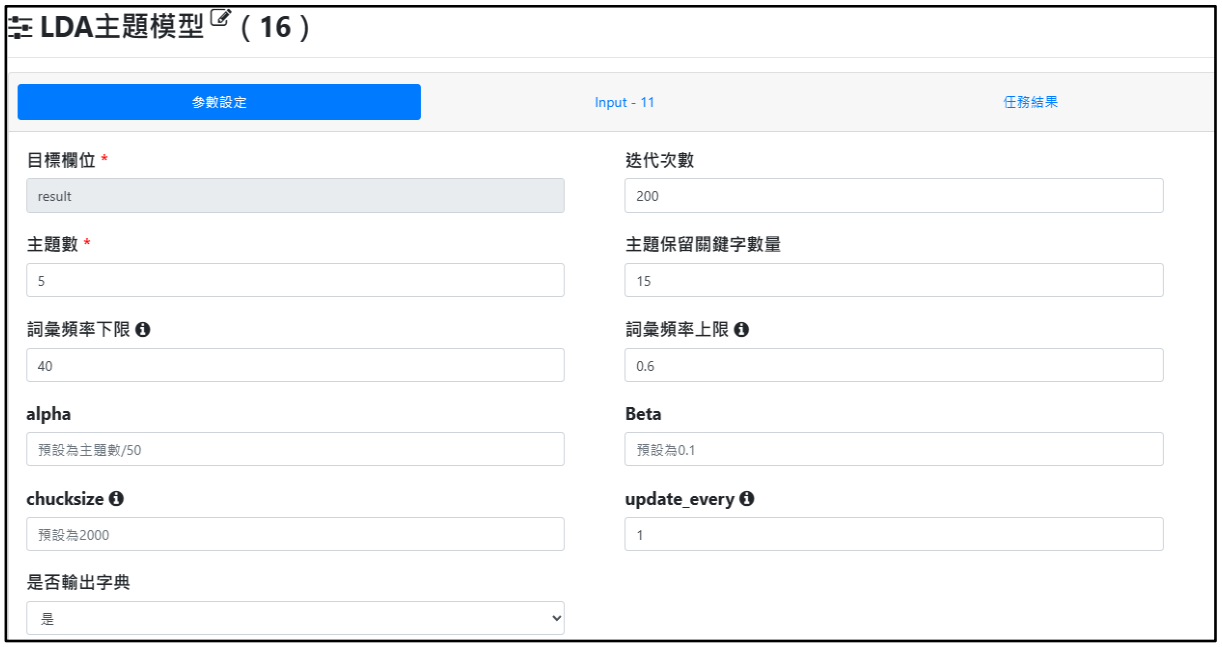


laechan
yes78529
shadowfan
lessonone
momonder
arrival0125
ivan24258419
hmt17
jsbach813
knmtou
webberyu
ymir29
wadadihaga
mansonrock
ichiroa
chiawww
silco
harry901
ttyy345889
kcjgg
v21638245
4585424
gottsuan
ivymorita
pazudora
laurie1013
weiike
laventidus
gardenlablue
ting445
jamie712
nuc1earsub
windsson
kiddcat
mdkn35
ck326
ss910126
Murasaki0110
okids1234
Bigblackoo
lc85301
aimlsa
bloodpiece
zone007
andy3020
HOVANA
pttmylove
Zach01073
dn49671137
oak2002
BigBird56
overskyend
Rejust
x51811danny
cafemos
HandyLover
crazyplayer
minimini7439
wewe
megumidc
onlysammi
slime12
panier
batistage
dadalada
rainie80818
sakukio
wdnm3444
lucky7766777
apple921
NinePi
ayukiu
Kawasumi
fly1122
www16852
yanchia
oris5
cloud2015
normis722
BMB2698
alcokkk
GenYDad
kevin8860
airflow
Phenomenon
tzjajkmj

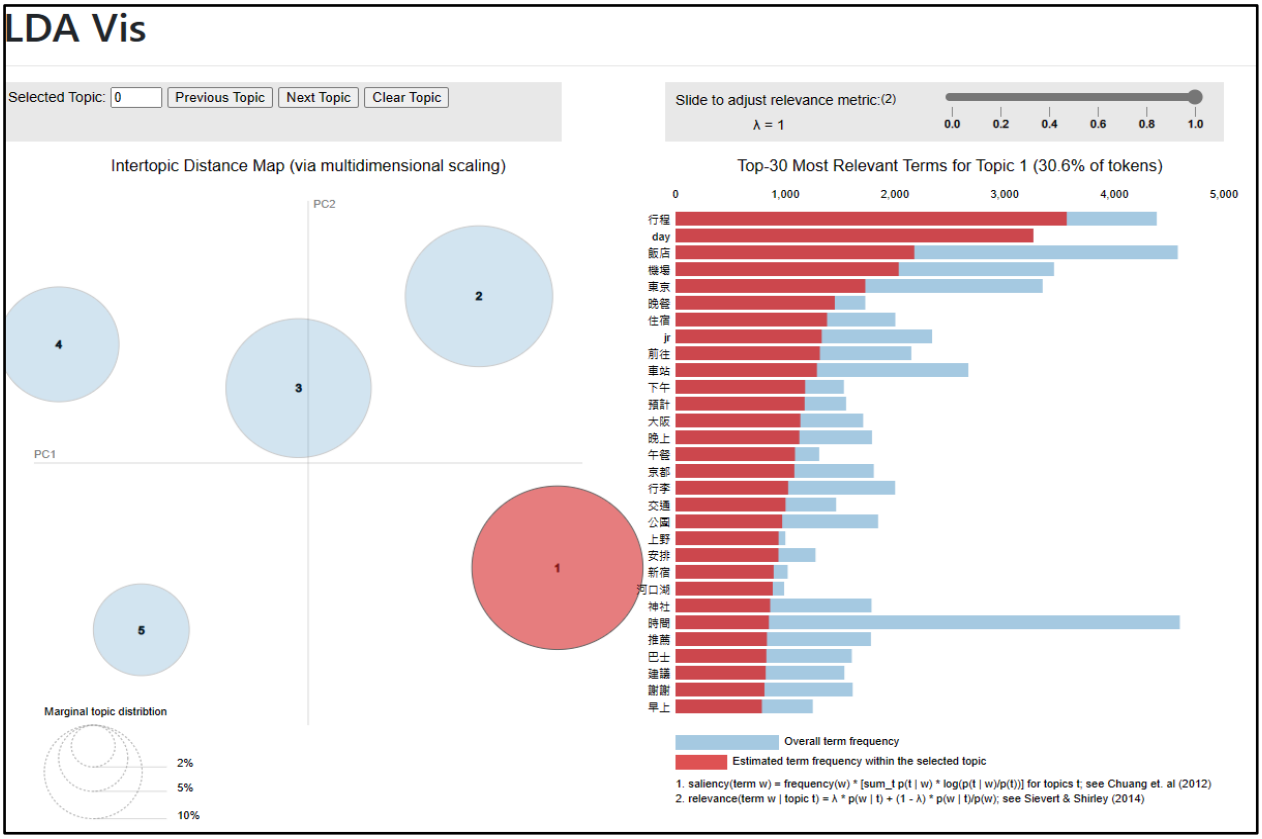
c. **LDA 主題模型**：針對 Result 欄位，設定主題數 5 個，保留關鍵字數量為 15 個。

i. 為減少對於無為字詞的判斷，並將詞彙頻率上下限分別設定 0.6 與 40。

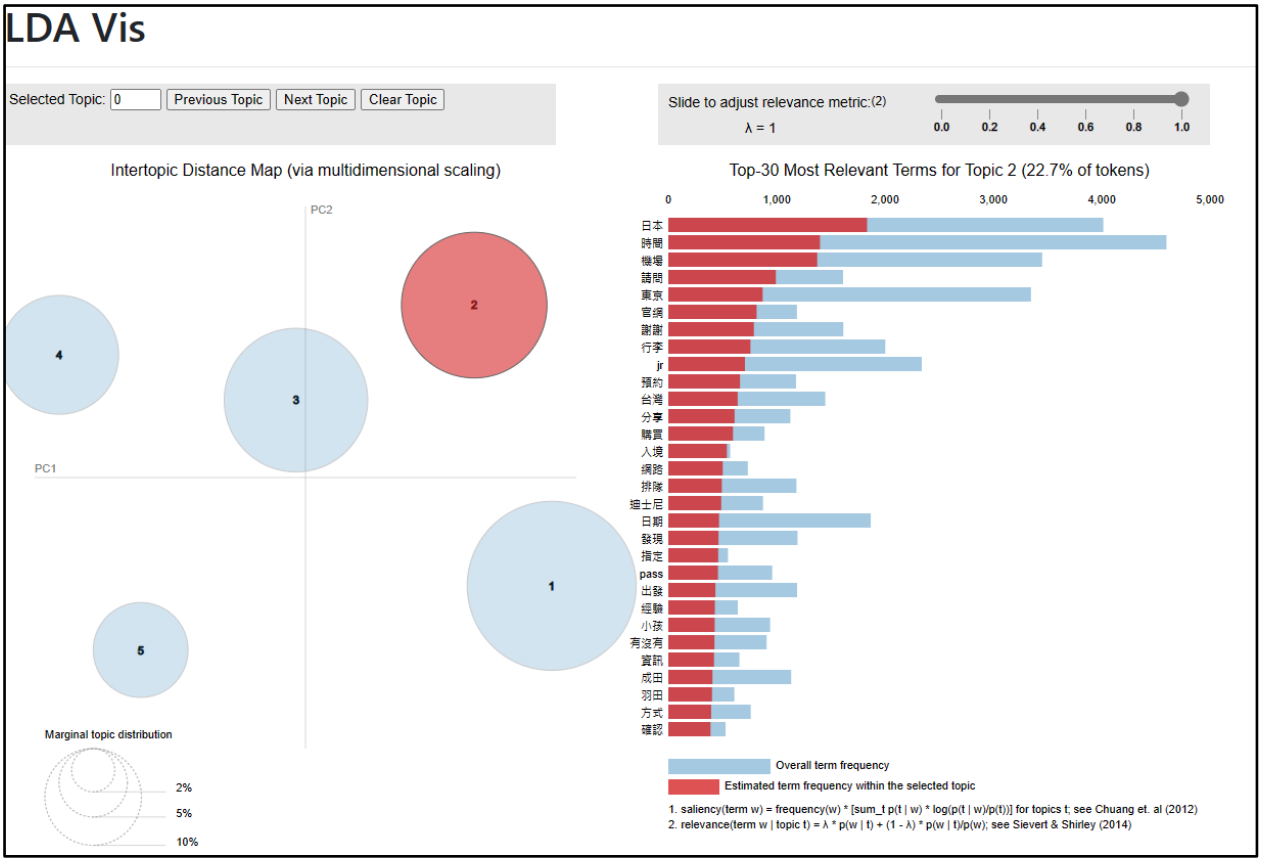
ii. 增加模型運算迭代次數至 200 次。



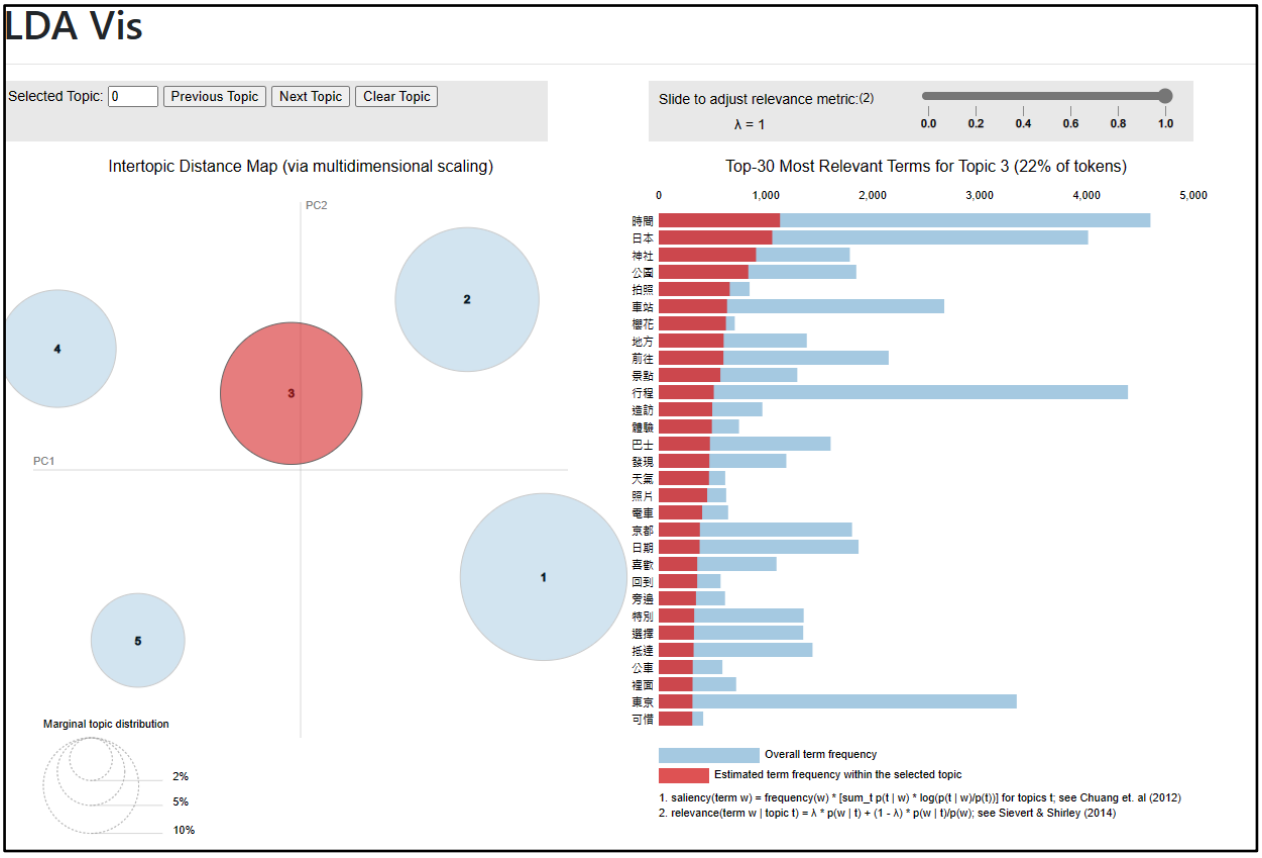
iii. 主題 1：高頻次排名前 10 的詞彙為「行程、day、飯店、機場、東京、晚餐、住宿、jr、車站」 => 其中還有包含[推薦、安排]等字眼，推測應該主要為[行程安排]的主題。



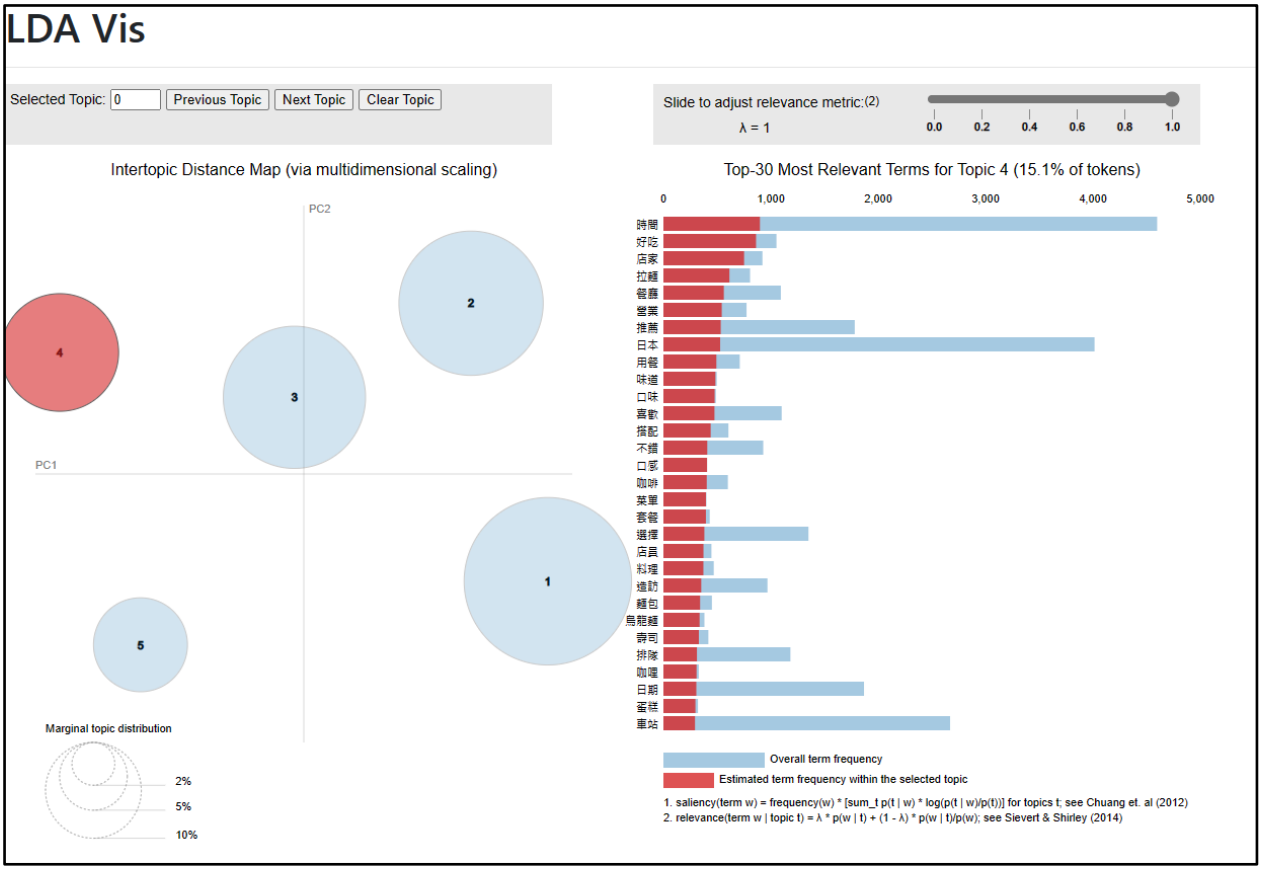
iv. <u>主題 2</u>：高頻次排名前 10 的詞彙為「日本、時間、機場、請問、東京、官網、謝謝、行李、jr、預約」=> 其中還有包含[預約、入境、網路]等字眼，推測應該主要為[入境前後相關事項]的主題。



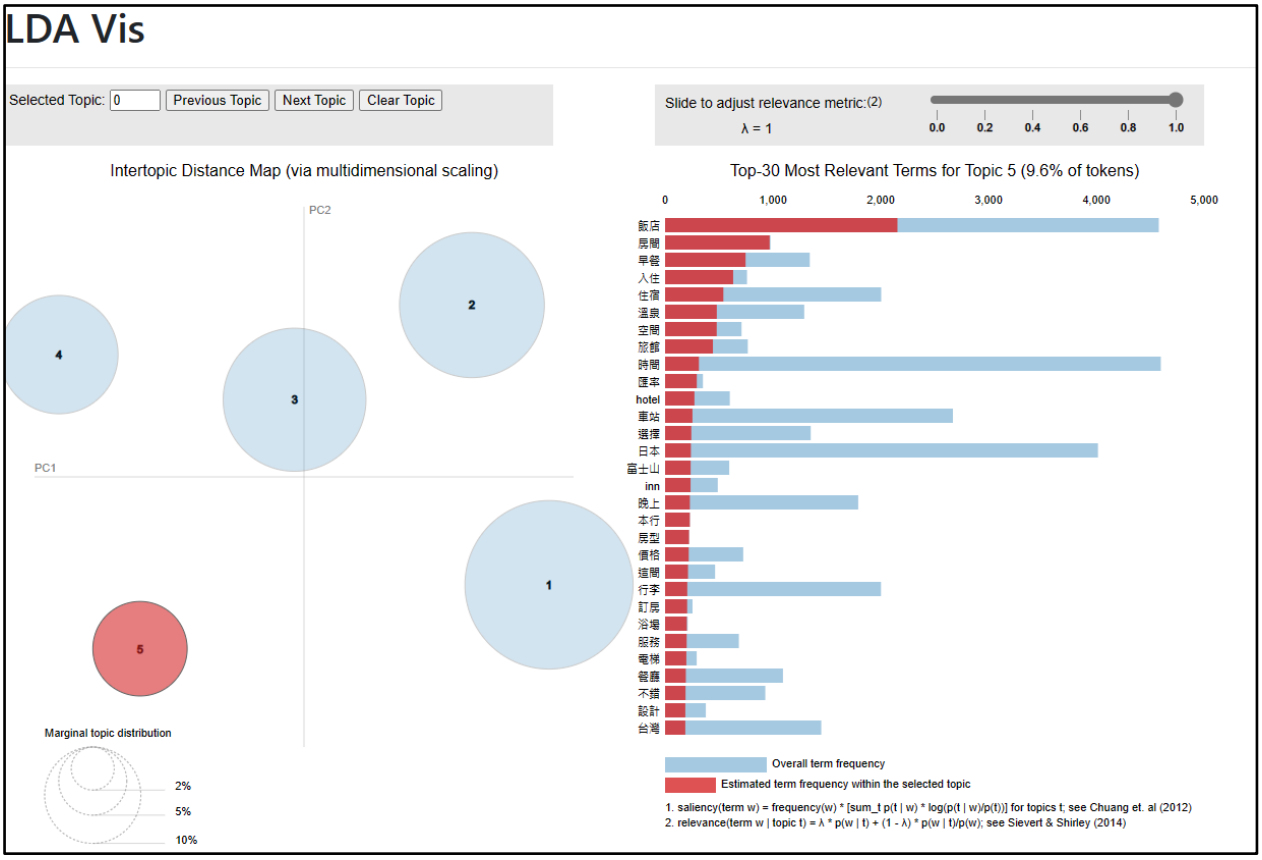
v. <u>主題 3</u>：高頻次排名前 10 的詞彙為「時間、日本、神社、公園、拍照、車站、櫻花、地方、前往、景點、行程」=> 其中還有包含[造訪、抵達]等字眼，推測應該主要為[景點遊玩]的主題。



vi. <u>主題 4</u>：高頻次排名前 10 的詞彙為「時間、好吃、店家、拉麵、餐廳、營業、推薦、日本、用餐、味道、口味」=> 此項光就排名前 10 詞彙即可推測出主要為[美食]的主題。



vii. 主題 5：高頻次排名前 10 的詞彙為「飯店、房間、早餐、入住、住宿、溫泉、空間、旅館、時間、匯率、hotel」=> 其中還有包含[房型、訂房]等字眼，推測應該主要為[住宿]的主題。



5. 結論

我們好奇台灣人瘋往日本旅行的原因，究竟是哪個景點吸引人們朝聖呢？依據 LDA 主題模型分析結果，擷取主題 1 關鍵字為「行程、day、飯店」；主題 2 關鍵字為「日本、時間、機場」；主題 3 關鍵字為「日本、神社、公園」；主題 4 關鍵字為「好吃、店家、拉麵」；主題 5 關鍵字為「住宿、溫泉、空間」，由上述五個主題模型可推論台灣人到日本旅遊需求的關鍵字為交通、景點、住宿、美食等，於 PTT_JapanTravel 版可搜尋到相關需求主題分類之文章。

資料區間設定為 2023/1-2023/6，透過主題可以知曉日本之所以吸引台灣人前往的誘因，以時間點來看有季節關係，正值春天櫻花盛開之際，誘使台人前往賞櫻，需知道如何前往故查詢當地乘車班次及交通方式，蒐集旅遊文章以作為日本行參考。有了目的地則進一步想知道當地有何美食推薦值得一吃的，尤其是日本拉麵於文章留言出現頻次相當高，故可推測台灣人對於日本拉麵有極高的品嚐意願。

賞櫻、吃拉麵以外，春天氣候怡人，最適合泡溫泉，溫泉也是日本相當有特色的一項必體驗項目。最後，有關於台灣人最愛旅遊勝地非屬迪士尼，受歡迎且具高討論度，於疫情解封後，迪士尼陸地的城堡也修建完成，終於露出全貌，因此吸引大量觀光客於早上開門前就前往排隊迫不及待等候入園大玩特玩。

總結，由 PTT_JapanTravel 版之社會網路圖結果，想要查找日本旅遊經驗的熱門文章，可以搜尋帳號 ID 為「Ic85301」與「laechan」」等版內核心人物，該文與留言者互動關係高，故搜尋其文章可獲取較多旅遊攻略資訊。