

5月22日 週一 · 癸卯年四月初四

19:00

第八屆
第四次讀書會報告

PTT 日旅版之網絡分析

指導教授:黃三益教授

組 員:

N104020023林義行	N104020022凌瑩琪
N104020018謝明和	N104020024林品均
N104020020李宜臻	N104020025陳姿樺
N104020021馮慧嬌	

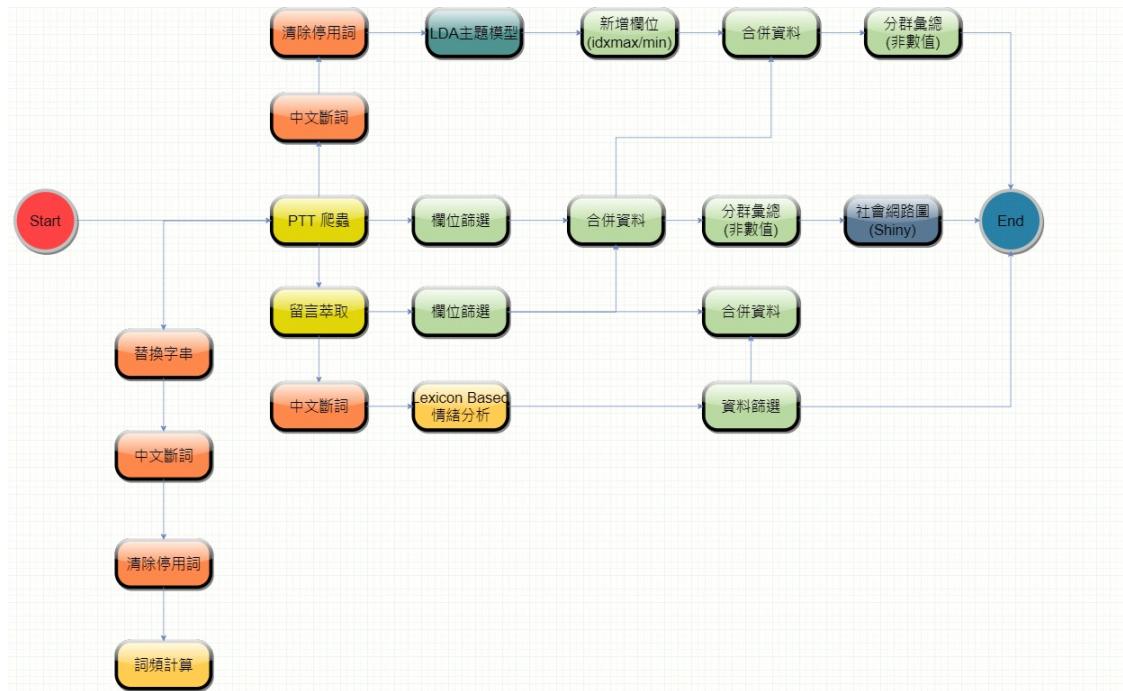
目錄

一、工作流程	5
二、資料前處理	6
(一) 替換字串	6
(二) 中文斷詞	7
(三) 清除停用詞.....	8
(四) 詞頻分析	9
三、文章主題分析	10
(一) 第 1 主題：	11
(二) 第 2 主題：	12
(三) 第 3 主題：	13
(四) 第 4 主題：	14

(五) 第 5 主題 :	15
(六) 第 6 主題 :	16
(七) 第 7 主題 :	17
(八) 第 8 主題 :	18
(九) 第 9 主題 :	19
(十) 第 10 主題 :	20
四、社會網路圖	21
五、Lexicon Based 情緒分析	24
六、視覺化工具 Gephi 應用	26
(一) 資料匯入 :	26
(二) 進行統計運算 :	27
(三) 繪製圖表 :	28

七、結論.....	31
-----------	----

一、工作流程



(流程檔案名稱 : 0522_8-4)

資料來源 :

選擇工作平台資料集「PTT 爬蟲」，期間設定為 2023/03/01 ~ 2023/03/31 .

近期隨著各國解除針對疫情的邊境管制措施，悶了許久的旅人們陸續出國旅遊，

其中日本是疫情之後最受旅客歡迎的國家之一，因此我們的資料以日本旅遊為設定參數。

二、資料前處理

(一) 替換字串

將空格與一些常出現較無意義的符號及網址做替換，讓搜尋出來的資料比較容易閱讀。

替換字串 (36)

參數設定

Input - 12

任務結果

選擇處理欄位 *

artContent

替換字串設定 ⓘ

```
\n\n>>
\n>>
\n>>
Sent from JPTT on my \w+>>
Sent from BePTT on my \w+>>
Sent from MoPTT on my \w+>>
my iphone \w+>>
on my \w+>>
iphone \w+>>
by ptt \w+>>
Sent from \w+>>
XD>>
xd>>
((http|ftp|https)://)[a-zA-Z-09.%~.]+[a-zA-Z][2.6)][0-9][1.3].[0-9][1.3].[0-9][1.3].[0-9][1.3)]([0-9][1.4)*
/[a-zA-Z-09.%~.]-*)>>
(\w+).(jpg|gif|png|html)>>
/^(?:https?:\/\/)?(?:www\.)?(?:youtu\.be|youtube\.com)(?:embed|v|watch)\?v=(\w{11})
(\w{5})?$/>>
from=udn-ch1_breaknews-1-0-news6.>>
```

選擇替換規則檔案 ⓘ

-----請選擇-----

儲存更改

(二) 中文斷詞

使用中文斷詞加大特定詞彙權重。



(三) 清除停用詞

使用清除停用字詞設定，將不具分析意義的詞彙列出清除。

■ 清除停用詞 (40)

參數設定		Input - 38
語言 *	Chinese	使用預設停止詞
是否清除單字元 ⓘ	是	是否轉為小寫英文
清除英文字母 *	否	清除數字 *
清除換行符號 *	是	清除特殊標點符號 *

(四) 詞頻分析

討論度較高的字詞，即為相對重要的參數。同時也對應並改善我們定義的字典詞彙，以下我們選取出現頻率最高的前 100 個詞組來產生文字雲。



三、文章主題分析

運用 LDA 主題模型我們將主題數的參數設定為 10，透過發文者的發問來了解所討論的前 10 大主題是什麼？

LDA 主題模型 (48)

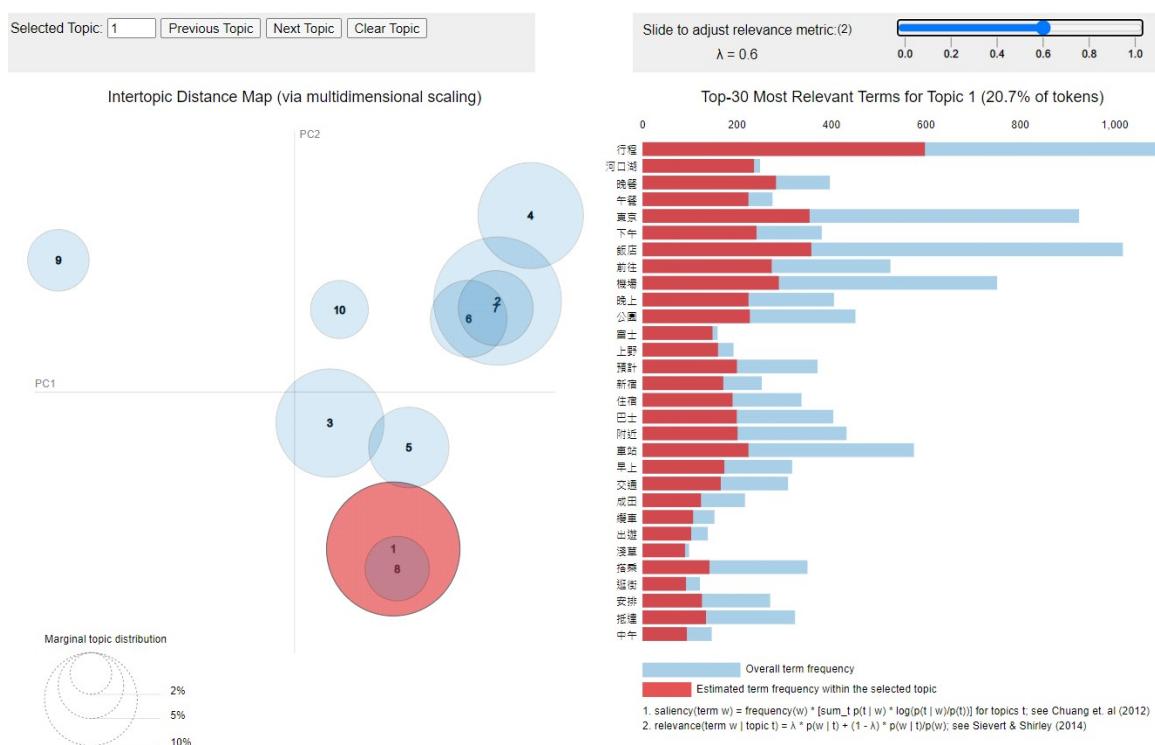
參數設定		Input - 46
目標欄位 *	result	迭代次數
主題數 *	10	主題保留關鍵字數量
詞彙頻率下限 ⓘ	40	詞彙頻率上限 ⓘ
alpha	預設為主題數/50	Beta 預設為 0.1

統計資訊

200 字數		10 主題數		-1.611 主題連貫性(UMass)		-0.418 主題連貫性(PMI)	
0.469 主題連貫性(Cv)		323.36 混淆度					

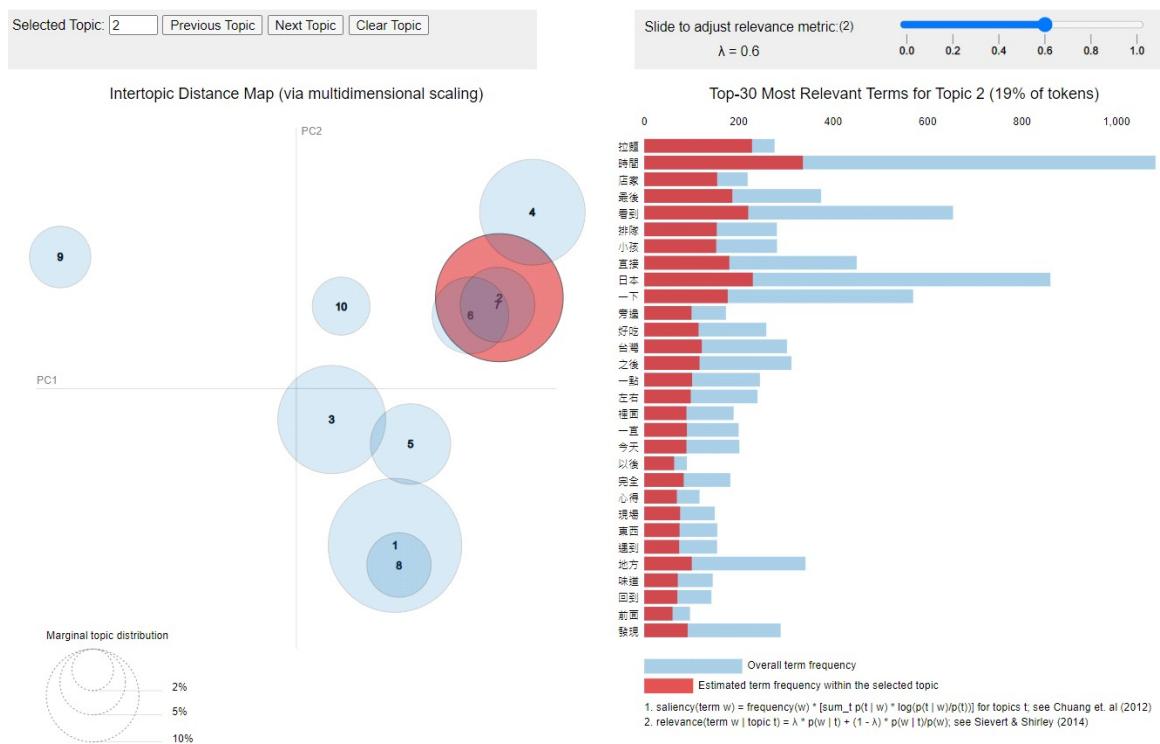
(一) 第 1 主題：

我們發現這類發問的內容中，出現較多的關鍵字如「行程」、「東京」、「飯店」、「晚餐」、「機場」等，這些字彙跟我此次所搜尋日本旅遊都相關，且其他字彙所出現的次數也不少，因此我們可以從這些出現的字彙中分類，分類第 1 群主題是跟「東京旅遊」相關主題的內容。



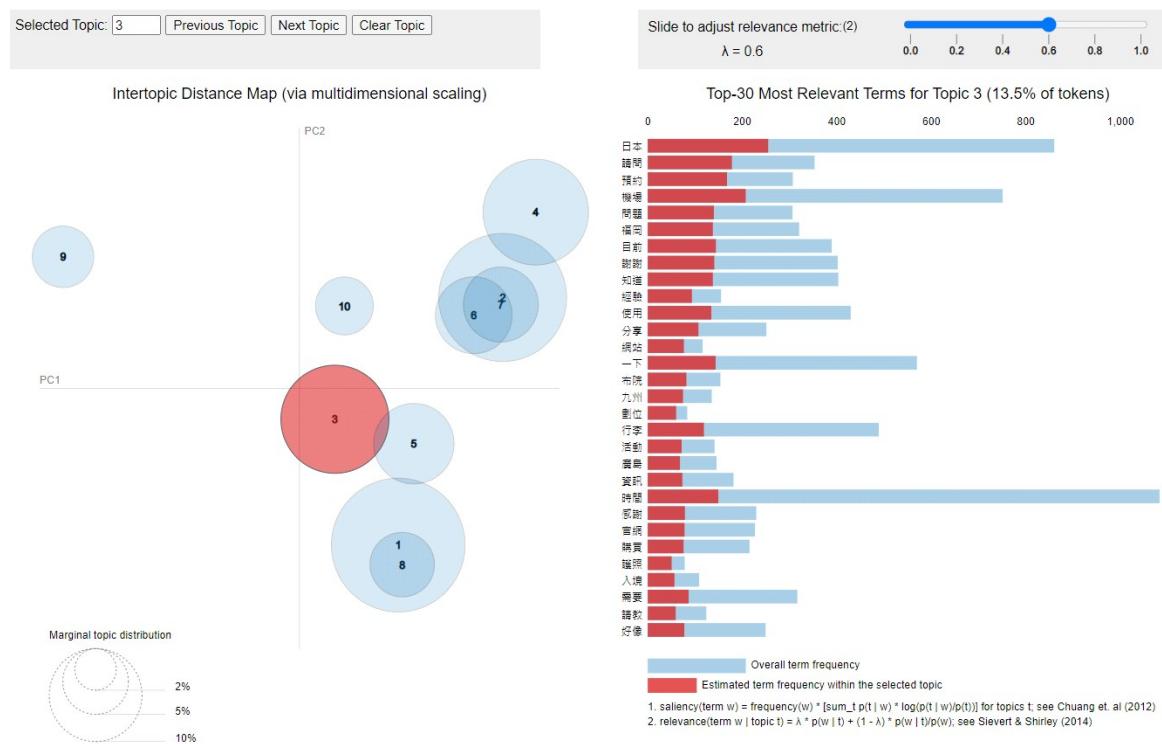
(二) 第 2 主題：

出現較多的關鍵字如「時間」、「拉麵」、「日本」、「店家」等，這些字彙主要是討論吃拉麵或是等待排隊的時間有相關，而其他字彙並無太多任何關係，因此我們可以從這些出現的字彙中分類，分類第 2 群主題是跟「日本拉麵」相關主題的內容，有可能是要詢問知名拉麵店用餐所需要的排隊時間。



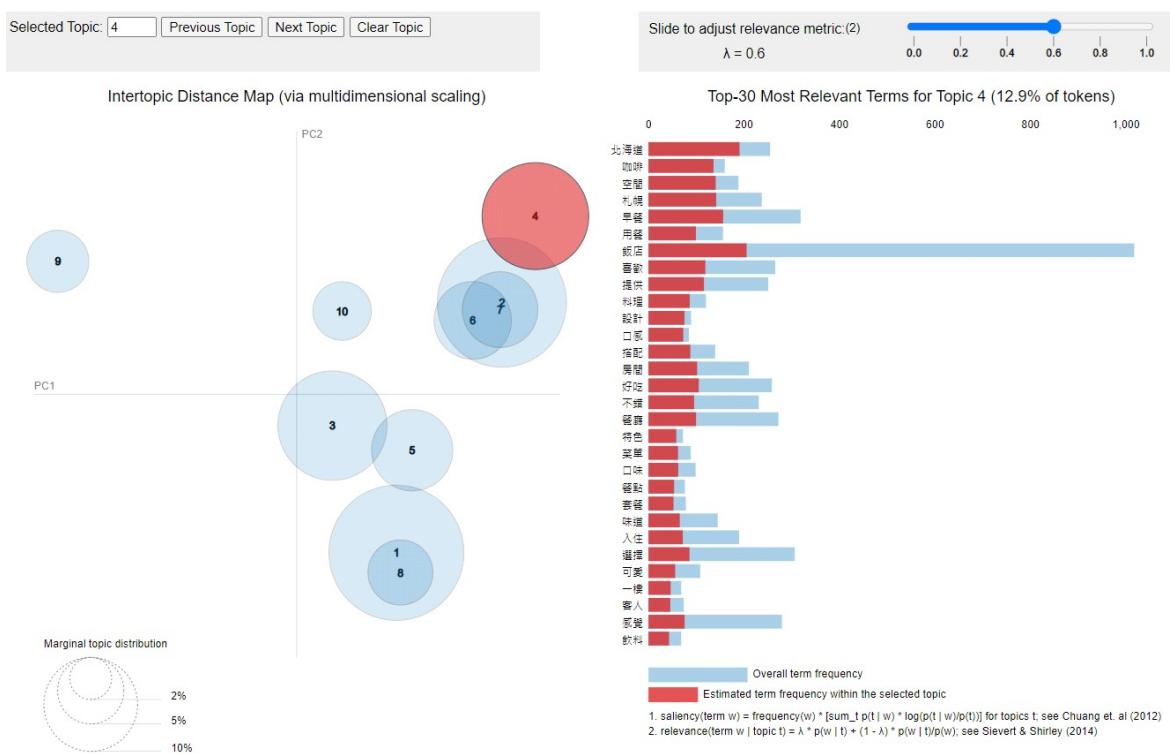
(三) 第3主題：

出現較多的關鍵字如「日本」、「機場」、「時間」、「福岡」等，這些字彙主要是討論福岡旅遊等有相關，因此我們可以分類第3群主題是跟「福岡旅遊」相關主題的內容，有可能是要詢問到日本福岡旅遊等相關的主題資訊。



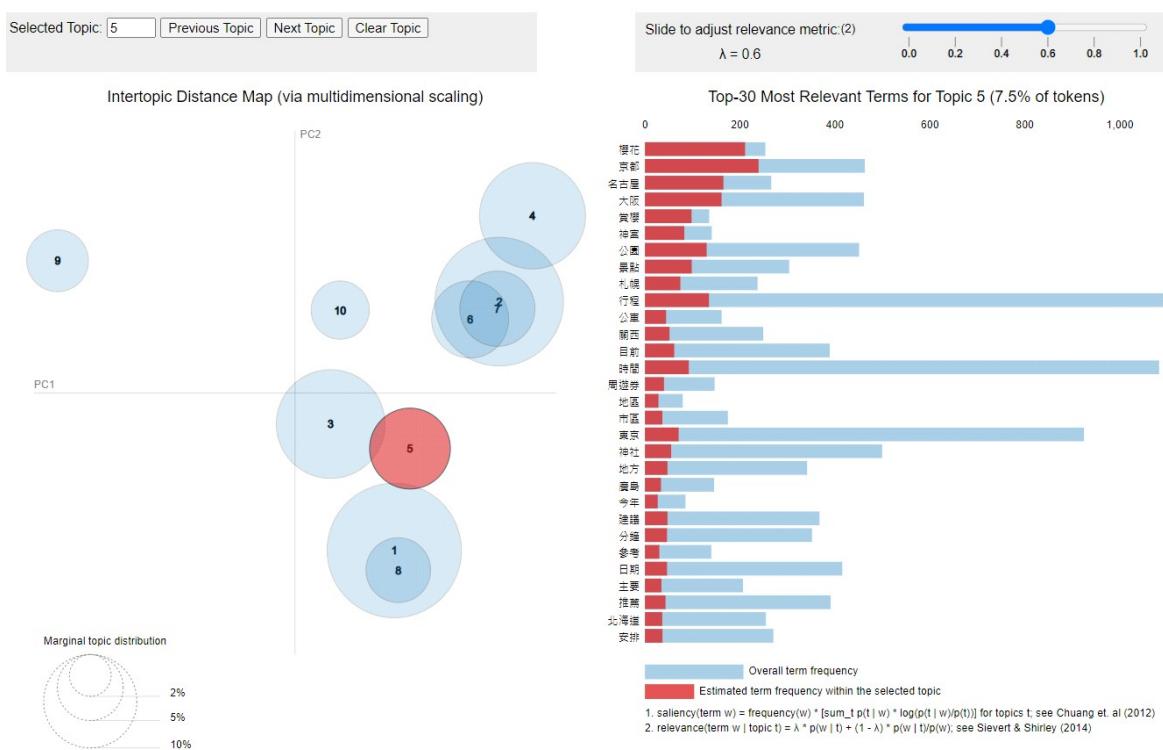
(四) 第 4 主題：

出現較多的關鍵字如「北海道」、「飯店」、「札幌」、「早餐」等，這些字彙主要是討論日本北海道住宿遊等有相關，因此我們可以分類第 4 群主題是跟「北海道住宿」相關主題的內容，有可能是要詢問到日本北海道飯店住宿等相關的主題資訊。



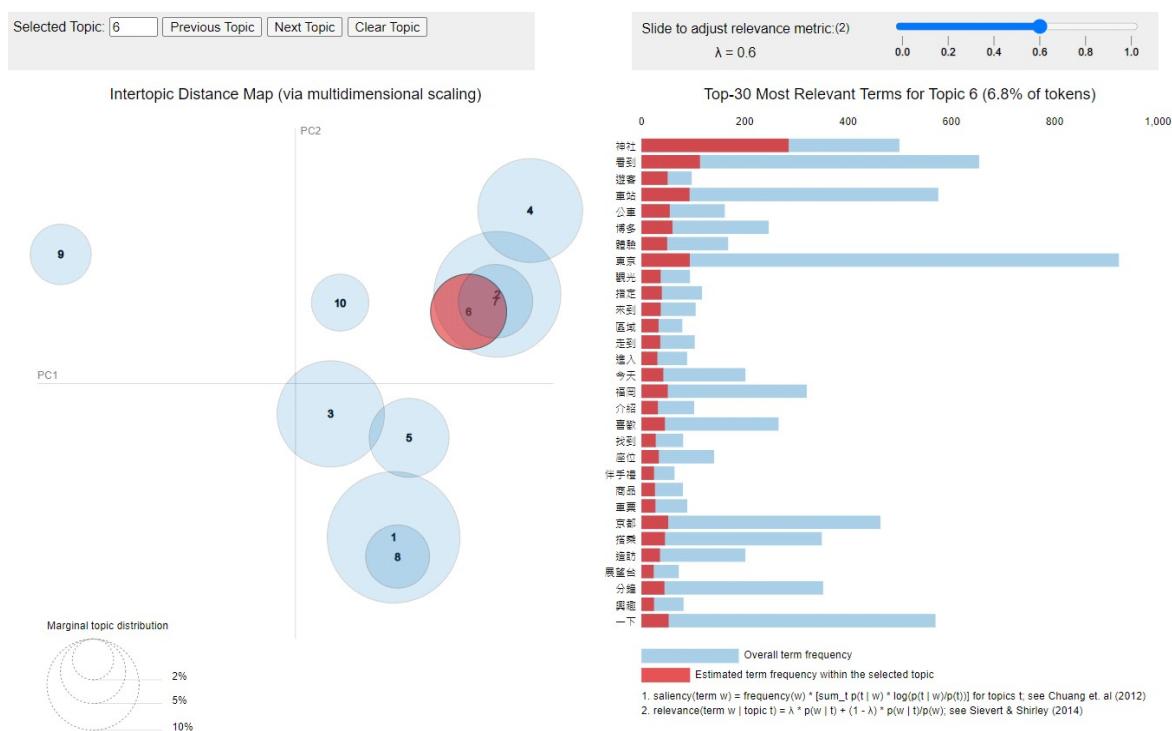
(五) 第 5 主題：

出現較多的關鍵字如「東京」、「名古屋」、「京都」、「大阪」、「賞櫻」等，這些字彙主要是討論日本城市與賞櫻花的行程有相關，因此我們可以分類第 5 群主題是跟「日本城市賞櫻」相關主題的內容，有可能是要詢問到日本主要城市櫻花盛開的時間，以期能夠搭配旅遊行程。



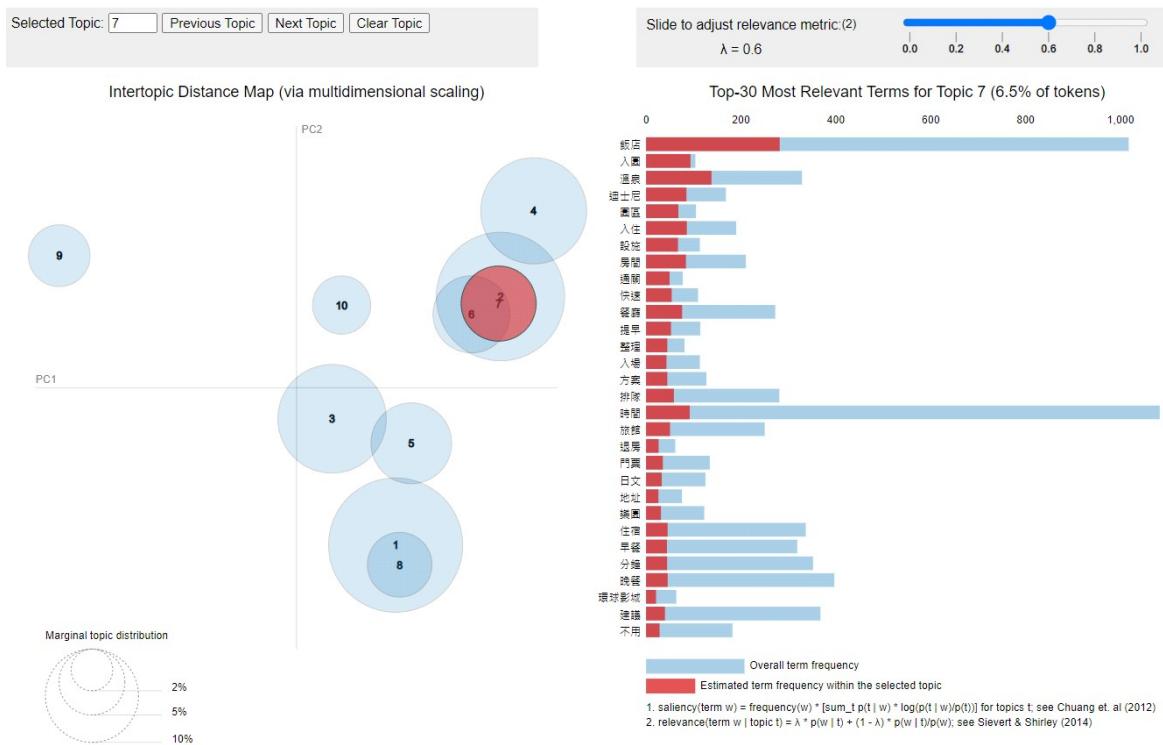
(六) 第 6 主題：

出現較多的關鍵字如「東京」、「神社」、「車站」、「博多」等，這些字彙主要是討論日本主要是安排交通去東京都內的景點觀光的規劃，因此我們可以分類第 6 群主題是跟「東京自由行」相關主題的內容，詢問東京都內著名景點以及主要的交通方式，以期能夠順利規劃旅遊行程。



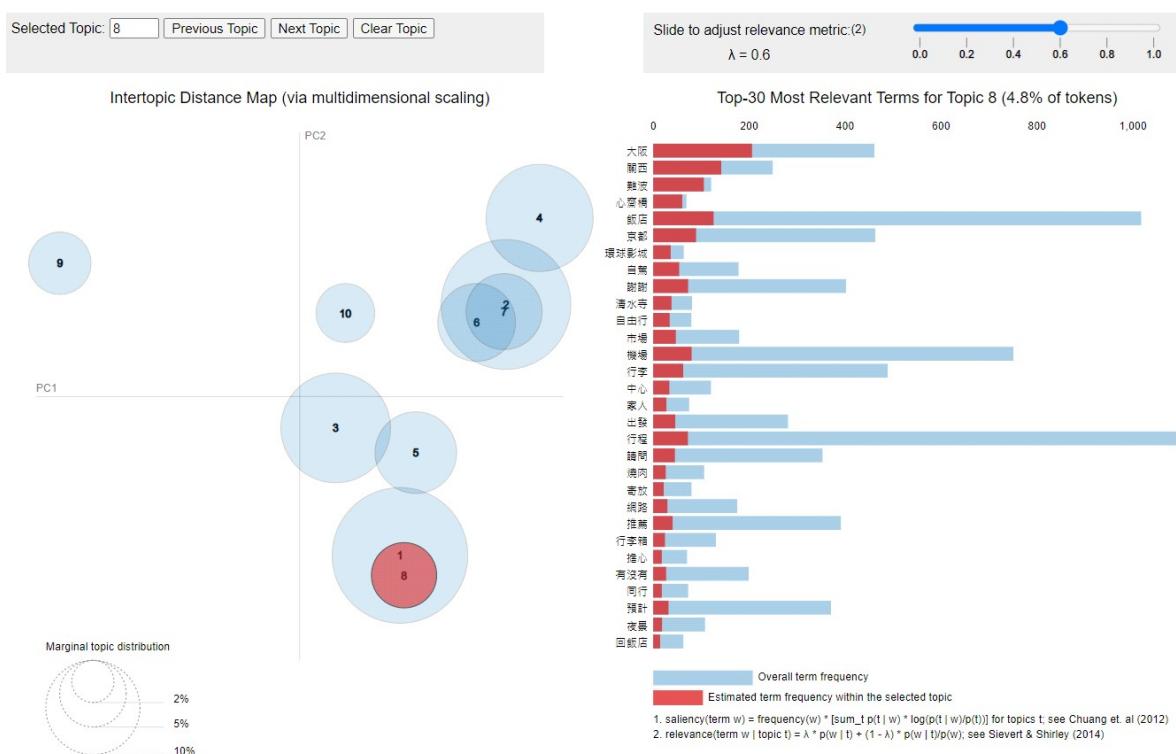
(七) 第 7 主題：

出現較多的關鍵字如「飯店」、「溫泉」、「迪士尼」、「入園」、「時間」等，這些字彙主要是討論相關飯店安排或是東京迪士尼主題樂園的行程有相關，因此我們可以從這些出現的字彙中分類，分類第 7 群主題是跟「安排住宿飯店」相關主題的內容，不同的客群會有泡溫泉、家庭客層會帶小孩去迪士尼遊園等問題。



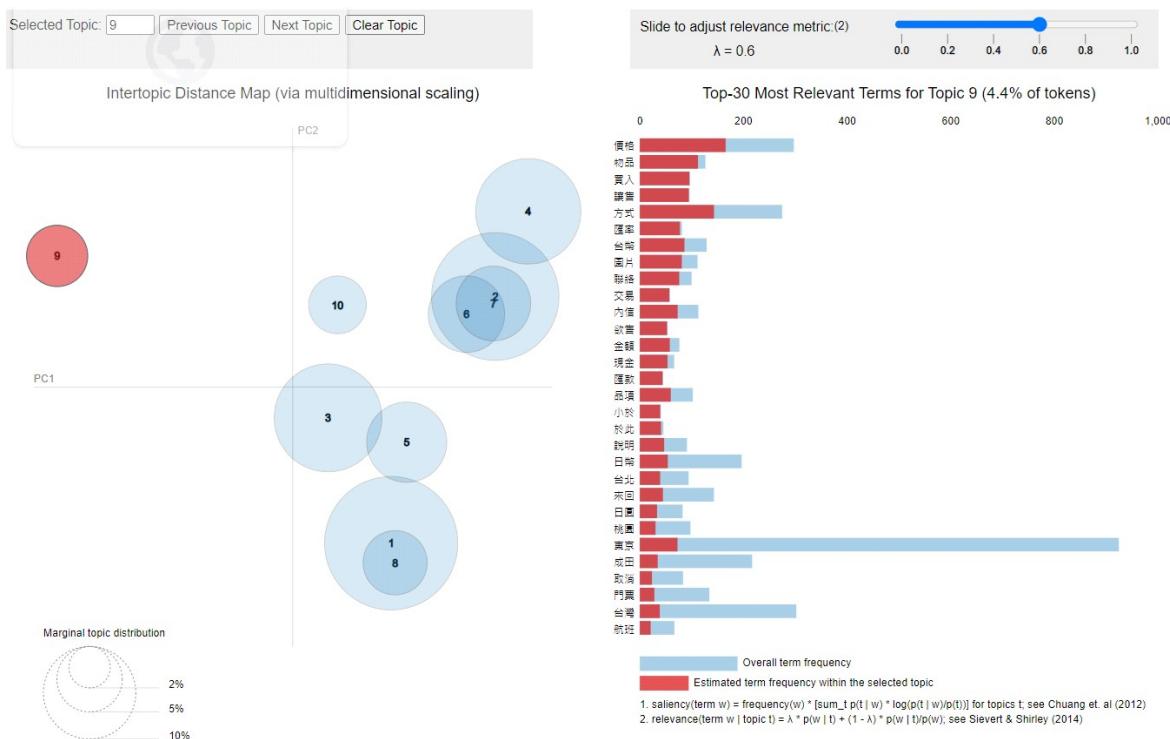
(八) 第 8 主題：

出現較多的關鍵字如「大阪」、「關西」、「京都」、「飯店」、「機場」等，這些字彙主要是討論進出日本關西主要城市的行程，因此我們可以從這些出現的字彙中分類，分類第 8 群主題是跟「關西旅遊」相關主題的內容，有可能是要詢問日本關西主要城市與住宿的主題。



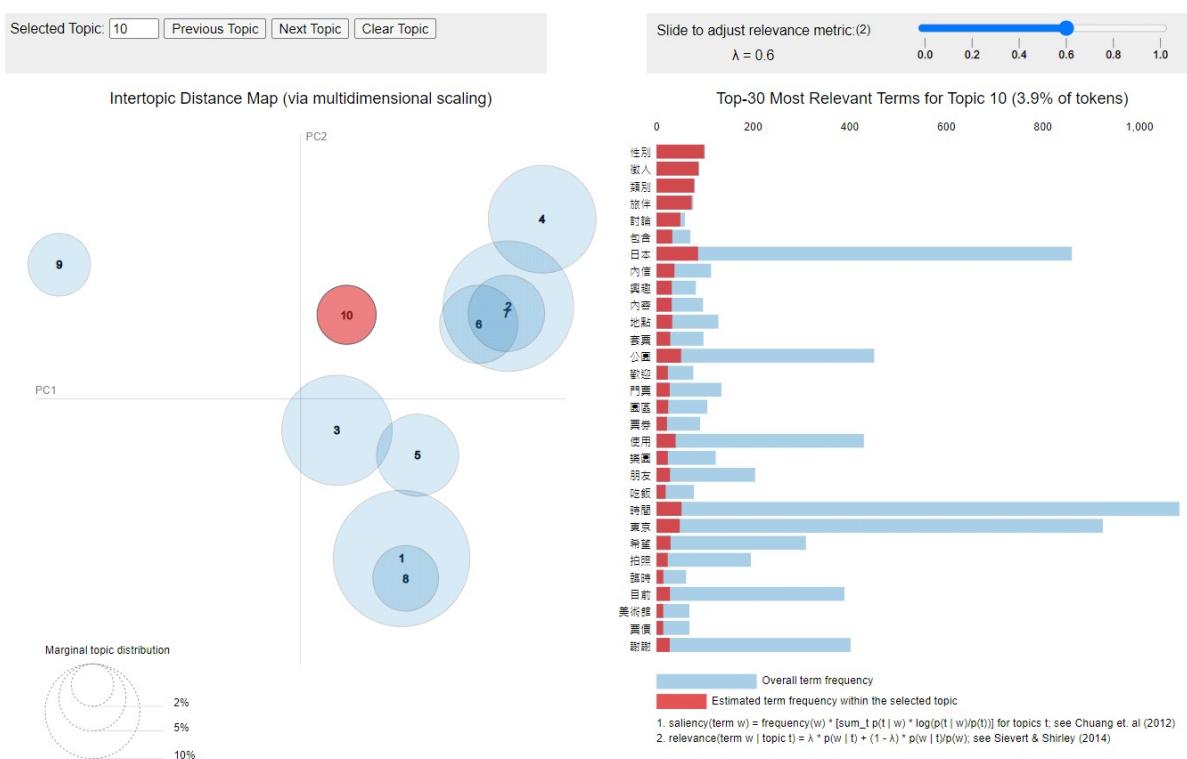
(九) 第 9 主題：

出現較多的關鍵字如「價格」、「方式」、「物品」、「匯率」等，且此主題與其他 9 個主題相去甚遠，可以歸類為在台灣事先詢問到日本購物與價格等問題，與其他主題並無太多任何關係，因此我們可以分類第 9 群主題是跟「購物行程」相關主題的內容，了解到日本如何購物、購買甚麼商品，甚至是換算匯率後是否有比在台灣便宜等計畫。



(十) 第 10 主題：

出現較多的關鍵字如「日本」、「性別」、「類別」、「時間」等，這些關鍵字彼此之間似乎沒有太多的關聯性，因此我們分類第 10 群主題是「其他」，並無法明確地歸納其主題。



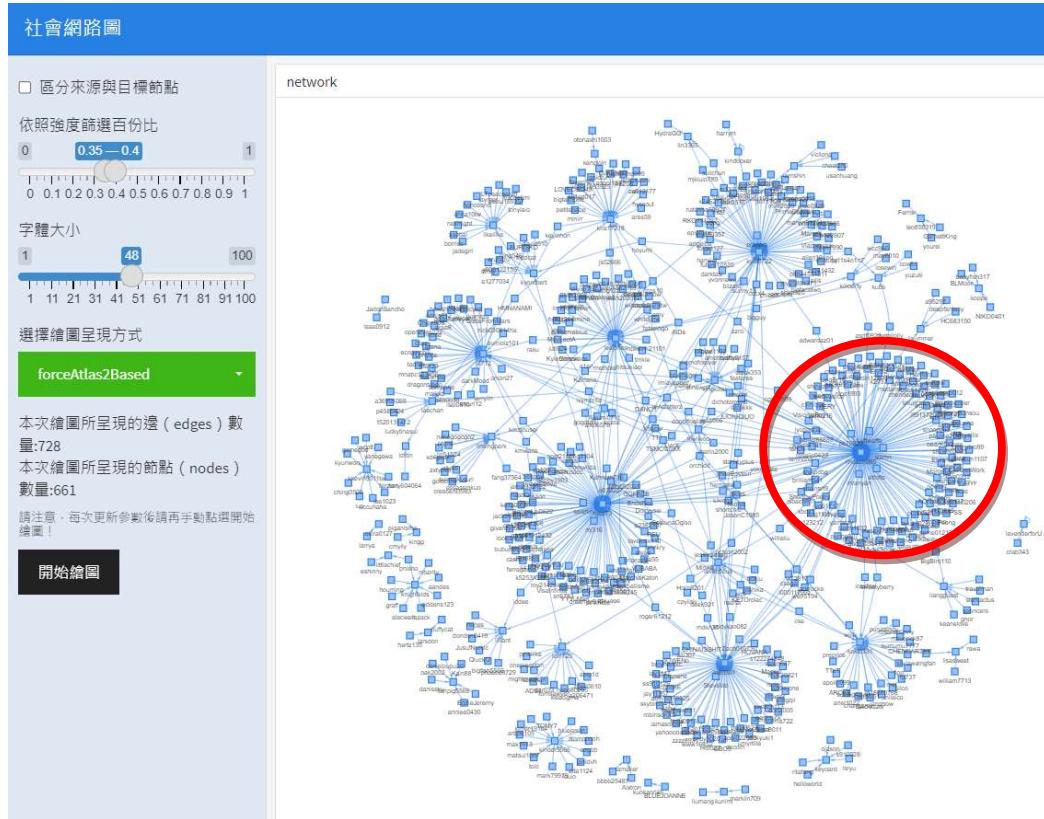
四、社會網路圖

利用 ptt 爬蟲功能找出發文者、發文內容，再用留言萃取找出留言者、留言內容後合併資料並分群匯總，保留發文者、留言者，以及留言者共回覆幾則留言，再利用 shiny 社會網路圖，看出發文者、文章留言者及留言數量關係。

■ 留言萃取 (13)



根據社會網路圖，可以找出日本旅遊版有幾個帳號發的文被回文最多次，其中獲得最多的回文的是網友「lindach095」，只有一篇文章也是被回文最多的文章，主題是第 2 主題，共獲得 171 個回覆，比單篇平均留言數高出許多。



lindach095 發的唯一一篇文，關於分享神社婚禮，留言大部分以恭喜為主。

批踢踢實業坊 > 看板 Japan_Travel

作者 lindach095 (竹)
標題 [遊記] 名古屋舉辦神前式婚禮
時間 Thu Mar 16 00:32:48 2023

聯絡資訊 關於我們 看板 Japan_Travel

造訪日期：2023年2月

文長慎入
為了讓有興趣的人能參考，前面寫較多細節，不喜可往下拉跳到〔神前式當日〕。

——前情提要——
和男友交往了11年，前陣子搶到了名古屋888特價機票，排行程時看到日本婚禮66,000円的廣告....

我：好便宜耶要不要去神社婚禮
男友：看你，都可以啊
於是我們就結婚了....

因為也沒宴客的打算，告知父母後就雙方家長一起吃頓飯，在交往紀念日去戶政事務所登記，名古屋旅行時再排一個上午去神社舉辦神前式啦～

五、Lexicon Based 情緒分析

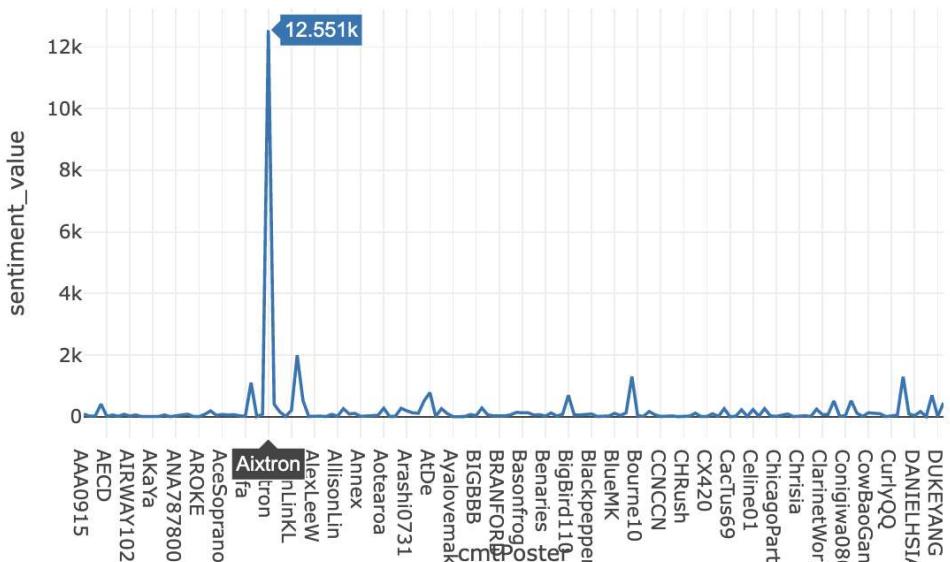
從情緒分析中可以看出正面情緒詞彙中有包含「分享」、「好吃」、「免費」，

負面詞彙像是「哭鬧」、「麻煩」在前幾名。

Lexicon Based 情緒分析 (16)



回文者中，分析數量 150 筆中，網友「Aixtron」的情緒分數最高，如下圖所示。

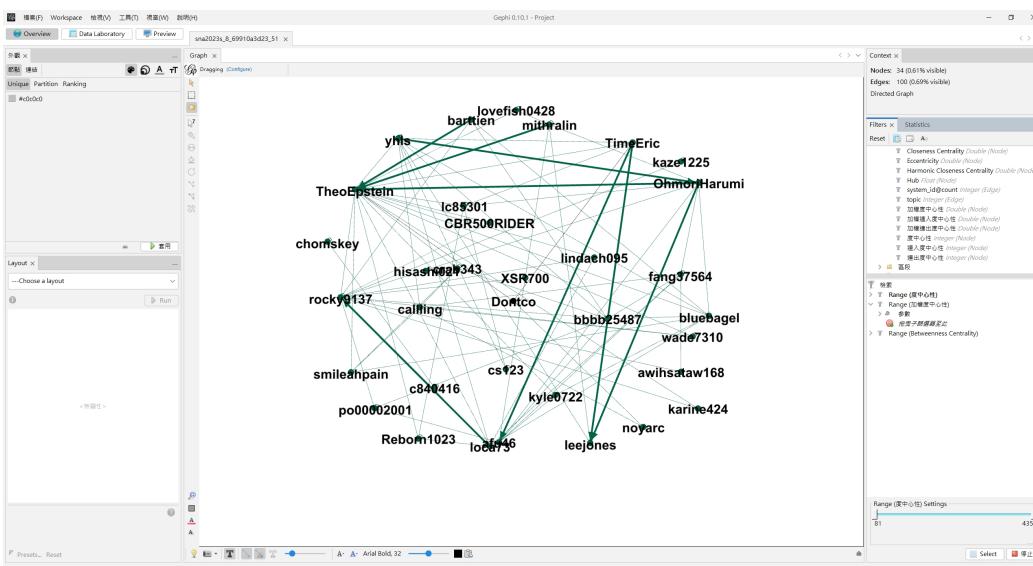


(三) 繪製圖表：

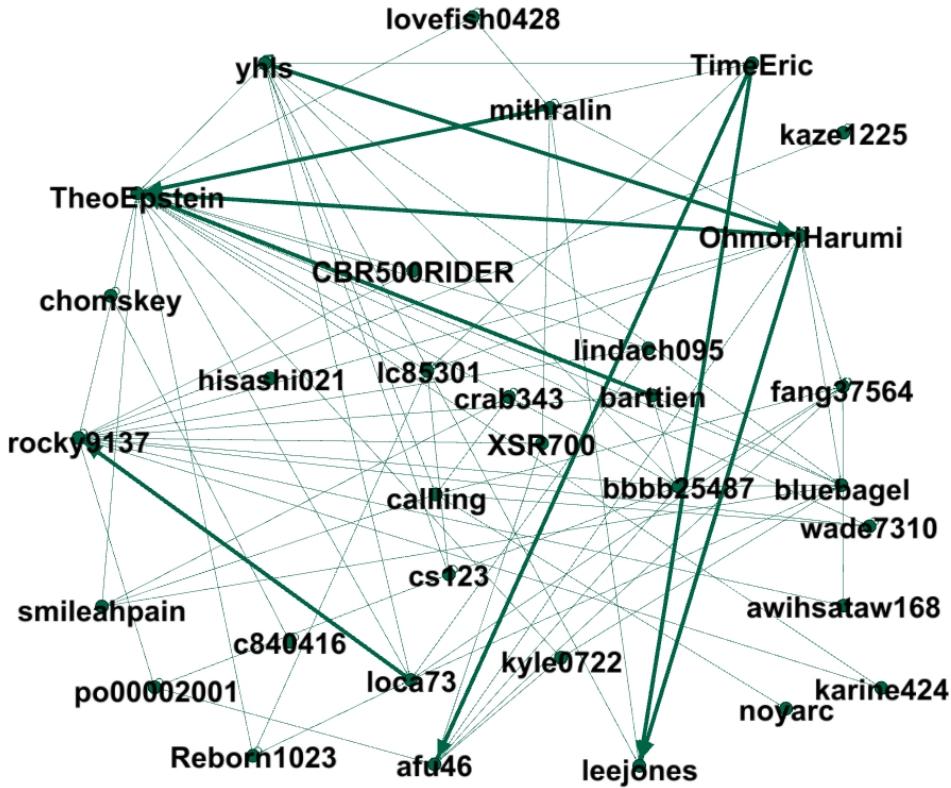
將資料檔進行繪製，運用 EDGE WEIGHT，並且將節點、線條與 ID 做顏色的調整，完成後可從線條方向及粗細看出 PO 文者與回文者的關聯，運用三種不同的方式繪製圖形如下：

(1) 度中心性，參數設定 81 ~ 435 進行篩選，Gephi 從線條的深淺且有方向性，

可看得出 po 文者跟回文者之間的互動強度：

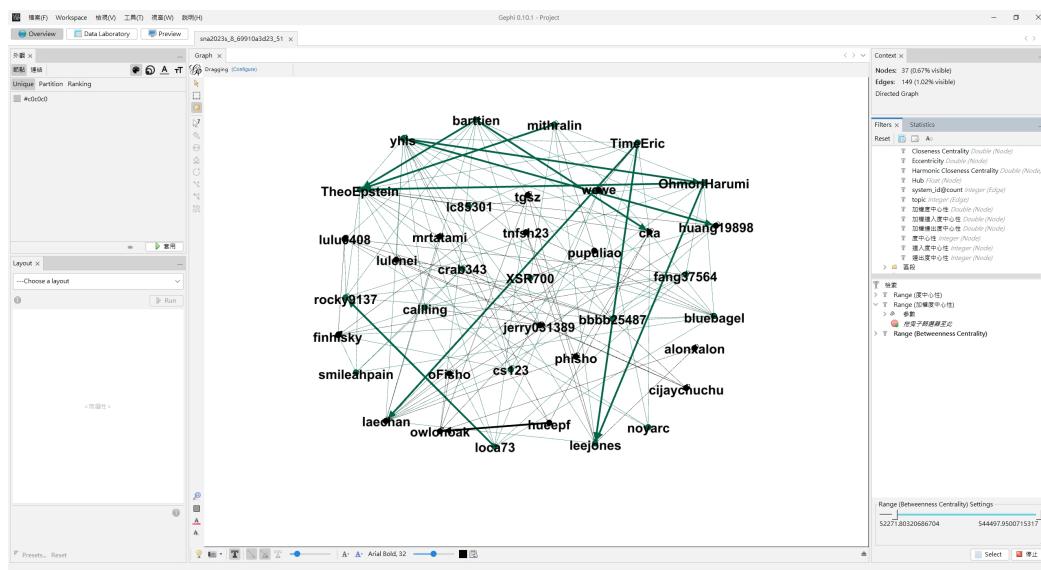


(2) 加權度中心性，參數設定 82 ~ 438 進行篩選，Gephi 加權度中心圖再次的驗證本組在度中心性圖上的結果是相同的，顯示主要活躍用戶的影響力：



(3) 介數中心性，參數設定為 52271.80320686704-544497.9500715317 進行篩選，Gephi 介數中心性可看出某些回文者具影響力，其回文會誘發其他回文者的

議題回覆：



七、結論

(一) LDA 運用在主題分類中是一個很良好的工具，本次將資料來源限縮在日旅版，內容以十個主題區分，日本最頻繁的旅遊地點就屬東京、北海道及大阪關西，而季節使然櫻花也佔據一個版面，這其實對旅遊業者是一個很好的參考，包含應用於行程規畫，亦或機酒的預定等。

(二) 由於旅遊的特性，住宿交通是去任何地方都會有關聯的，也因此某些主題有重疊的情況。

(三) 有最多回文的文章 PO 文者，在社群媒體分析可以運用來找出意見領袖，但是在本次的文章中，是一篇分享婚禮的 PO 文者，因此對意見領袖的意義似乎不大，因為大家回文亦僅為祝福為主，可能需要將時間拉長，對找出意見領袖的機率能有所提升。

(四) Gephi 的功能性很強，但由於時間限制，無法嘗試更多設定與圖檔繪製，或許多方嘗試後能找出更適合圖型並且用來分析說明。