

# 社群媒體分析第四次讀書會作業

指導老師:黃三益 教授

組別:第六組

組員:

N104020001 李采容

N104020002 廖英捷

N104020007 郭育雯

N104020008 游淑媛

N104020009 蔡雨臻

N104020010 盧貴聰

B096060020 黃湘安

M106020015 林猷盛

## 目錄

一、分析議題說明	3
二、工作流程設計	3
三、社會網路圖	5
四、LDA主題模型	12
五、Gephi軟體繪圖	17
六、結論	19

## 一、分析議題說明

- 主題:以PPT中包含「韓劇」的留言萃取與主題模型來建立社會網路圖和gephi圖
- 議題發想:

近年來，韓劇在全球掀起了一股熱潮。韓劇不僅在韓國國內獲得了廣泛的關注，也在全球各地贏得了廣大的粉絲群體。它們成功地跨越了語言和文化的障礙，成為全球觀眾喜愛的影視劇品牌。韓劇的全球成功得益於其獨特的敘事風格、精湛的演技和豐富的故事內容。然而，韓劇的成功不僅僅體現在收視率和商業價值上，它也在文化傳承和跨國影響方面發揮著重要作用。為了進一步研究韓劇在網路上引起的效應，本組將運用社群媒體分析課程所學，進行以下討論：

1. 網路上對於韓劇的相關主題分類有哪些？
2. 網路上對於韓劇的相關社會網路圖會如何呈現？
3. 網路上對於韓劇的相關gephi圖會如何呈現？

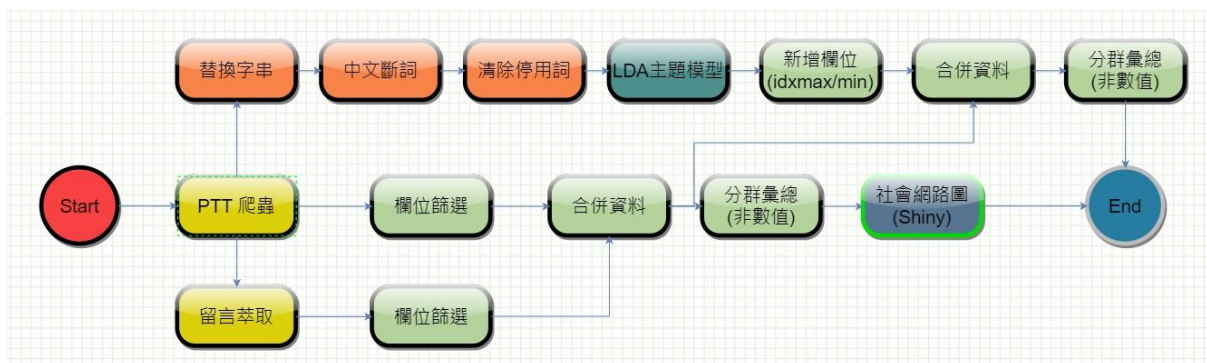
- 使用平台:文字探勘工作流程設計平台

## 二、工作流程設計

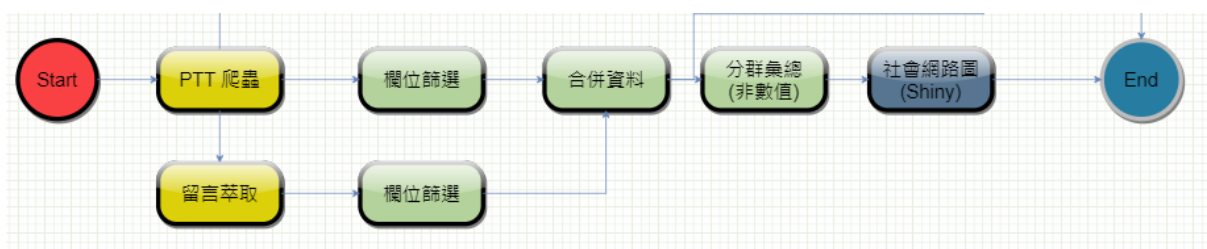
- 工作流程:013
- 軟體:Gephi
- 資料來源:PTT的KoreaDrama(韓劇)
- 分析期間:2023/4/1~2023/5/20
- 流程概述:
  1. 爬取PTT韓劇版今年4/1~5/20之內容，並進行留言萃取。
  2. 將原爬蟲資料與留言萃取資料篩選適當欄位，並進行合併。

3. 為了進行之後的社會網路圖，將2的合併資料進行分群彙總。
4. 將原爬蟲資料進行資料清理(替換字串、中文斷詞、清除停用字)。
5. 將清理之後的資料進行LDA主題模型。
6. 取權重最高的數值當作文章主題分類，並將分類的欄位新增至資料中。
7. 將LDA主題模型的資料與步驟2的資料進行合併。
8. 進行分群彙總，並將彙總之後的資料下載CSV檔案。
9. 將CSV檔案編輯Source和Target欄位，並按照LDA之資料，將原本的主題數字替換為文字(將0~3改為EP花絮、上線預告、心得、人物、四個討論主題)。
10. 將CSV檔匯入Gephi軟體進行繪圖。

● 總工作流程設計圖一覽：



### 三、社會網路圖



● PTT爬蟲(4)

1. 本組爬取2023/4/1至5/20的KoreaDrama(韓劇)資料

## 2. 共計抓取406筆資料

- 欄位篩選(13)

### 1. 保留system id、artPoster欄位

## 2. 擷取出貼文者

欄位篩選 ( 13 )

參數設定		Input - 4	任務結果
任務結果			
Show 10 entries		Search:	
system_id	artPoster		
1	AlsinGloro		
2	kawasau		
3	leione		
4	kakashi71		
5	ipipwrong		
6	kawasau		
7	Dodoro		
8	yihuan1122		
9	raininglight		
10	kawasau		
Showing 1 to 10 of 100 entries		Previous	1 2 3 4 5 ... 10 Next

## ● 留言萃取(11)

### 1. 萃取PTT留言紀錄

留言萃取 ( 11 )

參數設定	Input - 4	任務結果
保留本文內容		
否		
儲存更改		

### 2. 共抓取28564筆留言

留言萃取 ( 11 )

參數設定

Input - 4

任務結果

統計資訊

28564

總留言數量

73.24

平均單篇留言數

任務結果

Show

10

entries

Search:

system_id	comment_idx	cmtStatus	cmtPoster	cmtContent	cmtDate
1	1	推	hpzs	我也是前陣子開始看這部，剛開始看時要記太多人物融入不了劇	2023-04-01 00:13:00
1	2	→	hpzs	情，差點放棄，但越看越喜歡這種淡淡的、沒有勾心鬥角的朋	2023-04-01 00:13:00
1	3	→	hpzs	友、同事、病人之間的感情	2023-04-01 00:13:00
1	4	→	AlsinGloro	我本來習慣會看Wiki記人但太多人直接放棄了	2023-04-01 00:17:00
1	5	推	breakingdown	看完機智牢房不追這部第一集沒看完就棄了	2023-04-01 04:47:00
1	6	推	misod93	你那邊還記得買台換電嗎XD	2023-04-01 04:51:00
1	7	推	Smallgisp4	第二季你會失望的	2023-04-01 09:35:00

- 欄位篩選(14)

- 保留system\_id、comment\_idx、cmtPoster欄位，以保留貼文及其留言者的關聯資訊

欄位篩選 ( 14 )

參數設定 Input - 11 任務結果

選擇要保留的欄位(按住ctrl(Windows)或command(MAC)可以複選) \*

- system\_id
- comment\_idx
- cmtStatus
- cmtPoster
- cmtContent
- cmtDate

儲存更改

- 擷取出留言者

欄位篩選 ( 14 )

參數設定 Input - 11 任務結果

任務結果

Show 10 entries Search:

system_id	comment_idx	cmtPoster
1	1	hpzs
1	2	hpzs
1	3	hpzs
1	4	AlsinGloro
1	5	breakingdown
1	6	misod93
1	7	Smaligisp4
1	8	visrenee
1	9	visrenee
1	10	visrenee

Showing 1 to 10 of 100 entries Previous 1 2 3 4 5 ... 10 Next

- 合併資料(18)

- 將貼文者及留言者，透過system\_id關聯，進行資料合併

合併資料 ( 18 )

參數設定 Input - 13 Input - 14 任務結果

JOIN規則

新增規則 刪除規則

left_key	right_key
system_id	system_id
-----請選擇-----	-----請選擇-----

儲存更改

- 擷取出社會網路圖所需，貼文者與留言者間的關連及筆數

任務結果

Show 10 entries Search:

system_id	artPoster	comment_idx	cmtPoster
1	AlsinGloro	1	hpzs
1	AlsinGloro	2	hpzs
1	AlsinGloro	3	hpzs
1	AlsinGloro	4	AlsinGloro
1	AlsinGloro	5	breakingdown
1	AlsinGloro	6	misod93
1	AlsinGloro	7	Smallgisp4
1	AlsinGloro	8	visrenee
1	AlsinGloro	9	visrenee
1	AlsinGloro	10	visrenee

Showing 1 to 10 of 100 entries Previous 1 2 3 4 5 ... 10 Next

- 分群彙總(非數值)(21)

- 透過(artPoster,cmtPoster)作為鍵值，計算count(system\_id)，找出同一位貼文者與留言者間有多少則留言數

使用...欄位進行分群(按住ctrl(Windows)或command(MAC)可以複選)

system\_id  
artPoster  
comment\_idx  
cmtPoster

匯總函數

count  
numique  
min  
max  
first  
last

計算欄位(按住ctrl(Windows)或command(MAC)可以複選)

system\_id  
artPoster  
comment\_idx  
cmtPoster

儲存更改

- 取得網路圖間的關聯，特別一提的是，因貼文者會透過留言回復其他的留言者(如AblazeStar)，因此，可能形成節點(node)自己的邊(edge)



分群彙總 (非數值) ( 21 )

參數設定


Input - 18

任務結果

統計資訊

4495

群組數量



任務結果

Show10entries

Search:

artPoster	cmtPoster	system_id@count
AblazeStar	AblazeStar	3
AblazeStar	bown	1
AblazeStar	cashko	1
AblazeStar	eureka	2
AblazeStar	q750830	2
AblazeStar	raininglight	1
AblazeStar	yutan0802	1

全覽結果

點按下載完整CSV資料

點按下載完整data

點按下載完整json資料

PTT內, AblazeStar於5/13新增一則貼文

【板主:XDDDD555/wil183tw1】		[韓劇]		系列《KoreaDrama》	
[←]離開 [→]閱讀 [Ctrl-P]發表文章 [d]刪除 [z]精華區 [i]看板資訊/設定 [h]說明				人氣:206	
編號	日期	作者	文章標題		
≥ 1	3 5/13	AblazeStar	[問題] 黑暗榮耀的第一集與後面劇情矛盾		

※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 119.77.132.236 (臺灣)					
※ 文章網址: <a href="https://www.ptt.cc/bbs/KoreaDrama/M.1683910786.A.D56.html">https://www.ptt.cc/bbs/KoreaDrama/M.1683910786.A.D56.html</a>					
推	cashko:	第一集開頭是同珉腦內對仇人的假想吧			05/13 01:13
→	raininglight:	把名字打對 同珉			05/13 01:36
推	yutan0802:	第一集是腦中的想像			05/13 01:39
→	AblazeStar:	原來是想像的呈現 一開始還以為是倒敘法 謝謝解惑			05/13 02:06
→	q750830:	但我對這手法不是很喜歡, 只是引起觀眾的興趣但跟後續沒			05/13 03:14
→	q750830:	相關			05/13 03:14
→	eureka:	不覺得這段劇情只是為了引起觀眾興趣而已。被霸凌、被性騷			05/13 11:08
→	eureka:	擾過的人大都會一次又一次的想像、為了修復自己而反擊。			05/13 11:08
→	AblazeStar:	可以理解是想像的畫面 但前後的呼應不足			05/13 13:31
→	AblazeStar:	本來還以為後面劇情會演到回來這段			05/13 13:31
推	bown:	開頭很明顯是想像吧 涎鎮臉上有傷卻沒有痛感			05/14 02:00

## ● 社會網路圖(Shiny)(23)

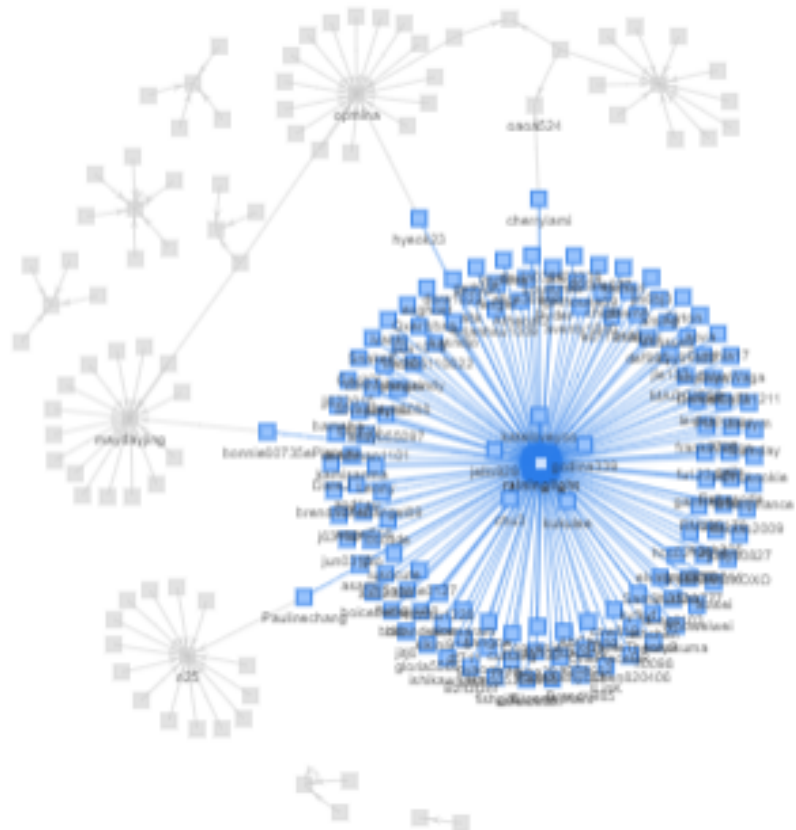
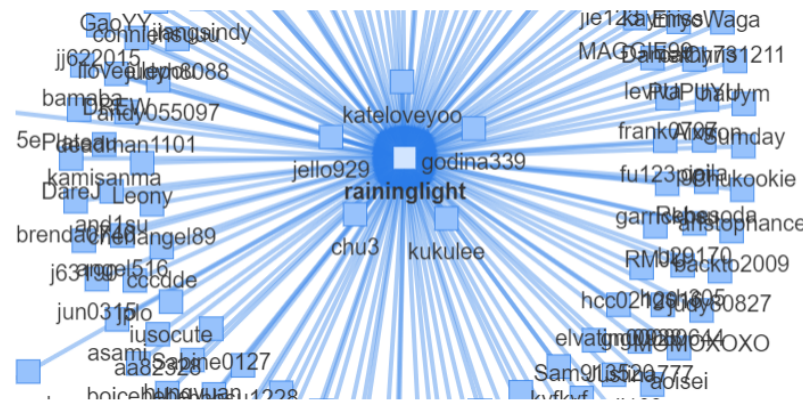
- 繪製社會網路圖, 由貼文者(artPoster)連結至留言者(cmtPoster), 期間加上權重(留言總數, system\_id@count)

社會網路圖 (Shiny) ( 23 )

參數設定		Input - 21	任務結果
節點權位 (來源) *		節點權位 (目標) *	
artPoster		cmtPoster	
連結權位 *		system_id@count	
儲存更改			

2. 可以看到有形成數個聚落，包含一個最大的聚落，其中心點為發文

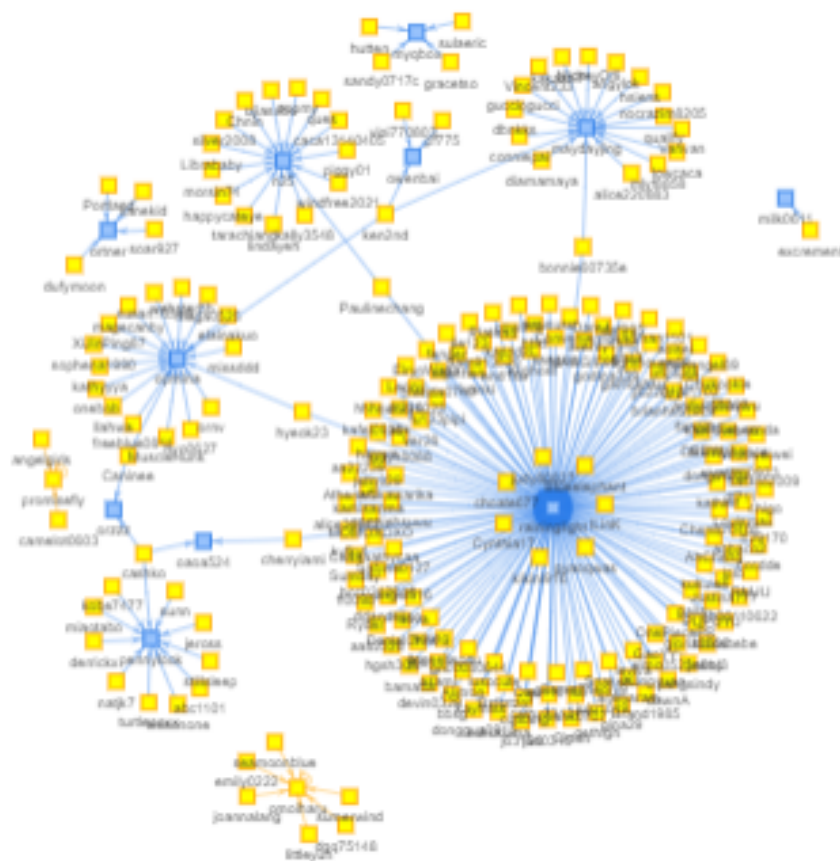
ID:raininglight



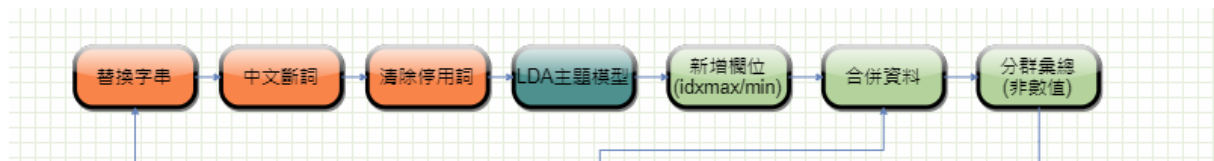
PTT內，可以看到raininglight貼文非常的活絡

[板主:XDDDD555/wil183tw1]			[韓劇]		系列《KoreaDrama》	
←離開 [→]閱讀 [Ctrl-P]發表文章 [d]刪除 [z]精華區 [i]看板資訊/設定 [h]說明						
編號	日期	作者	文章標題	人氣:203		
297	爆	5/06	raininglight	[閒聊]	Netflix《魔幻之音》EP04~06 討論(END)	
298	爆	5/06	raininglight	[LIVE]	第58屆百想藝術大賞 紅毯、頒獎典禮	
299	爆	5/06	raininglight	[LIVE]	第58屆百想藝術大賞 頒獎典禮(Part2)	
300	爆	5/06	raininglight	[情報]	第58屆百想藝術大賞 得獎名單	
301	M27	7/28	raininglight	[徵文]	綠葉你最紅-芮秀貞(藝秀晶)	
302	+2012	2/28	raininglight	[閒聊]	咖啡之約/要喝一杯咖啡嗎?	
303	+37	2/28	raininglight	[情報]	3月新劇 JTBC 離婚律師申晟瀚/神聖的離婚	
304	+爆	3/04	raininglight	[閒聊]	離婚律師申晟瀚/神聖的離婚 EP01-02	
305	爆	3/11	raininglight	[閒聊]	離婚律師申晟瀚/神聖的離婚 EP03-04	
306	+爆	3/18	raininglight	[閒聊]	離婚律師申晟瀚/神聖的離婚 EP05	
307	+爆	3/19	raininglight	[閒聊]	離婚律師申晟瀚/神聖的離婚 EP06	
308	+爆	3/25	raininglight	[閒聊]	離婚律師申晟瀚/神聖的離婚 EP07-08	
309	+爆	4/01	raininglight	[閒聊]	離婚律師申晟瀚/神聖的離婚 EP09-10	
310	+爆	4/08	raininglight	[閒聊]	離婚律師申晟瀚/神聖的離婚 EP11	
311	M26	4/09	raininglight	[徵文]	2022 明星韓劇-飛起來吧,蝴蝶	
312	+爆	4/09	raininglight	[閒聊]	離婚律師申晟瀚/神聖的離婚 EP12(END)	
313	+19	4/24	raininglight	[情報]	4月新劇 ENA 紙之月	
314	+38	4/26	raininglight	[情報]	第59屆百想藝術大賞(人氣賞結果、頒獎典	
315	+爆	4/28	raininglight	[LIVE]	第59屆百想藝術大賞 紅毯、頒獎典禮	
316	+爆	4/28	raininglight	[情報]	第59屆百想藝術大賞 得獎名單	

3. 我們也可以透過顏色標記, 區分來源(藍色)節點與目標(黃色)節點



## 四、LDA主題模型



### 1. LDA主題模型

設定主題數為4個(原本設定為5個主題, 但發現有兩個主題較為類似, 故將兩者合併);  
迭代次數為50次; 保留關鍵字20個。

設定詞彙頻率下限為40, 該詞彙最少要出現在20篇文章中。

設定詞彙頻率上限為0.5, 在所有文章中高於0.5者將排除計算。

#### ≡ LDA主題模型 ( 35 )

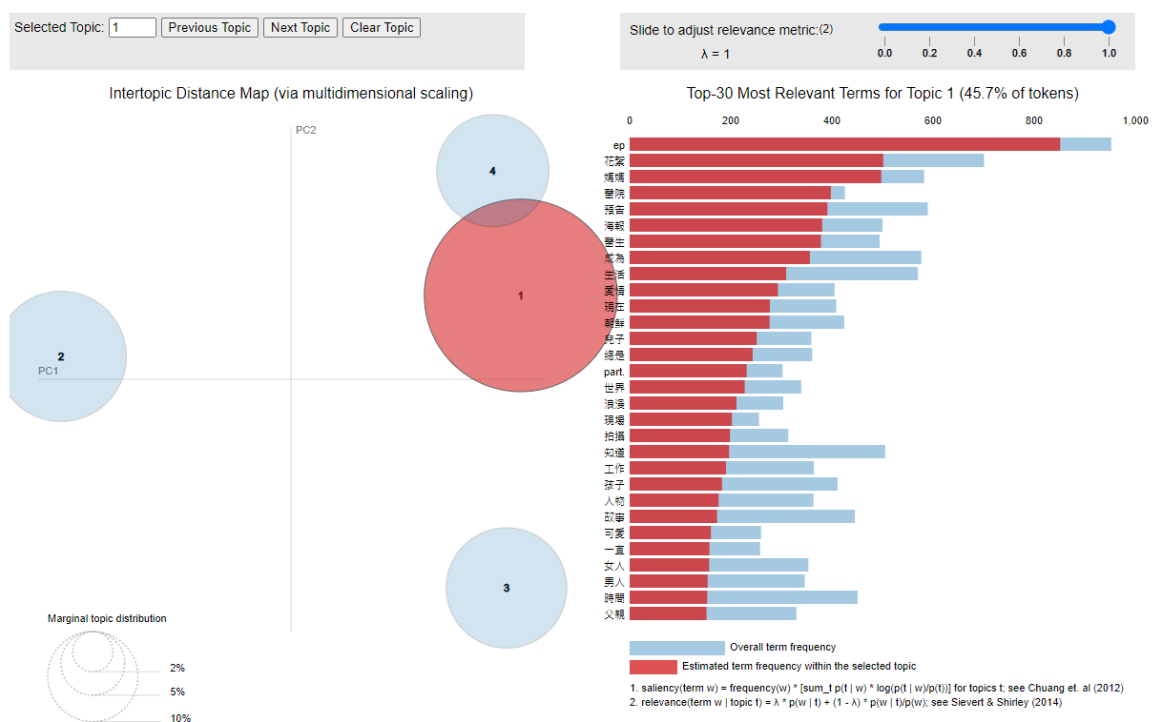
參數設定	Input - 30	任務結果
目標欄位 *	result	
主題數 *	4	
詞彙頻率下限 ⓘ	40	
alpha	預設為主題數/50	
chucksize ⓘ	預設為2000	
是否輸出字典	是	
迭代次數	50	
主題保留關鍵字數量	20	
詞彙頻率上限 ⓘ	0.5	
Beta	預設為0.1	
update_every ⓘ	1	

### 統計資訊

80 字數	4 主題數	-0.919 主題連貫性 (UMass)	-3.238 主題連貫性(PMI)
0.328 主題連貫性(Cv)	288.33 混淆度		

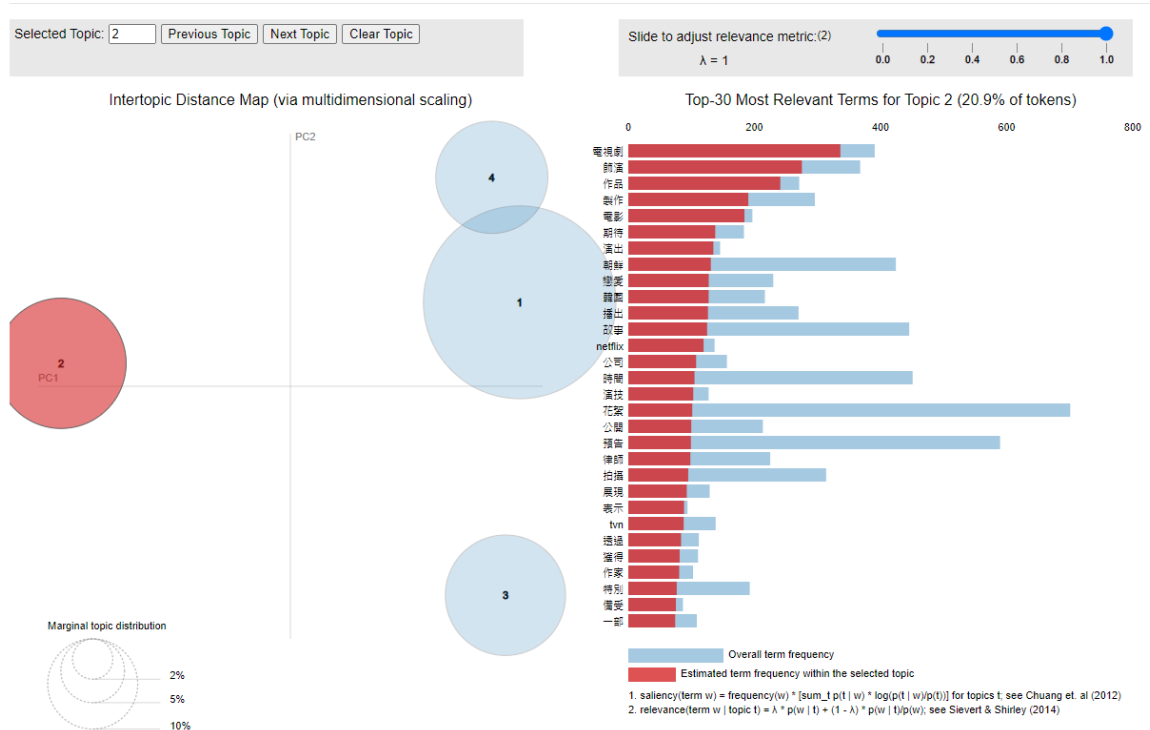
第一個主題(45.7%), 前三大詞彙: ep> 花絮> 媽媽

## LDA Vis



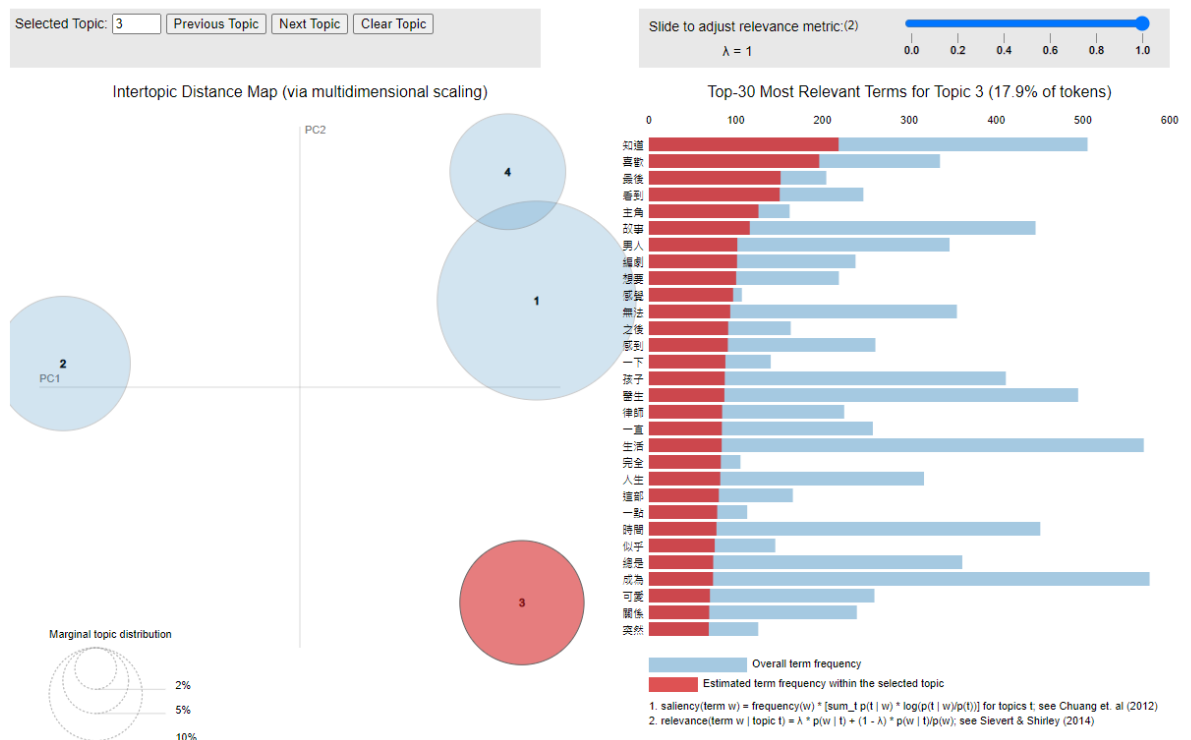
第二個主題(20.9%), 前三大詞彙: 電視劇>飾演>作品

## LDA Vis



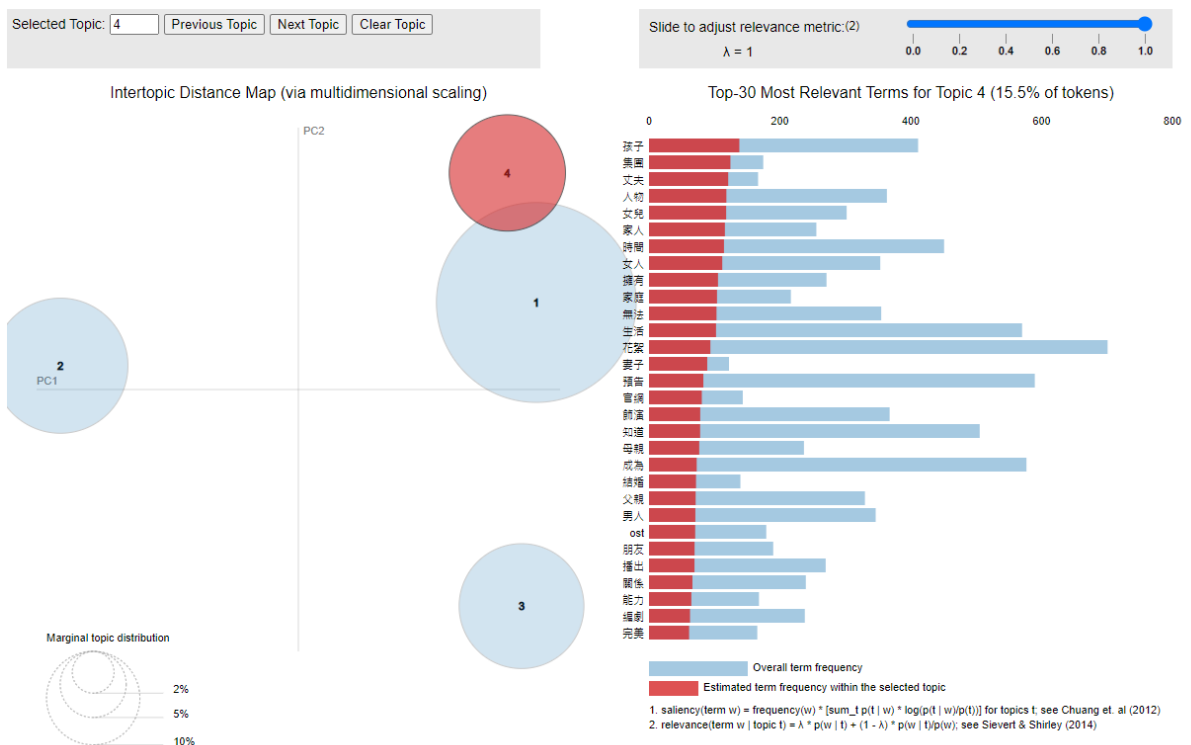
第三個主題(17.9%), 前三大詞彙: 知道>喜歡>最後

## LDA Vis



第四個主題(15.5%), 前三大詞彙: 孩子>集團>丈夫

## LDA Vis



### 2. 新增欄位(idxmax/min)(37)

利用函數「max」, 將LDA分數計算每篇文本的最大值, 定義為新欄位「topic」

## 新增欄位 (idxmax/min) ( 37 )

參數設定

Input - 35

任務結果

匯總函數 \* ⓘ  
max

計算欄位(按住ctrl(Windows)或command(MAC)可以複選) \*  
system\_id  
0  
1  
2  
3

新增的欄位名稱 \*  
topic

利用新增欄位「topic」來儲存主題編號

## 任務結果

Show 10 entries Search:

system_id	0	1	2	3	topic
1	0.000000	0.000000	0.000000	0.993569	3
2	0.000000	0.000000	0.000000	0.994191	3
3	0.000000	0.131175	0.000000	0.866833	3
4	0.000000	0.079760	0.000000	0.917631	3
5	0.000000	0.983240	0.000000	0.000000	1
6	0.522355	0.000000	0.477192	0.000000	0
7	0.997901	0.000000	0.000000	0.000000	0
8	0.000000	0.216362	0.783073	0.000000	2
9	0.331351	0.032264	0.304889	0.331497	3
10	0.019500	0.000000	0.000000	0.979587	3

Showing 1 to 10 of 100 entries Previous 1 2 3 4 5 ... 10 Next

## 3. 合併資料

將原始PTT爬蟲結果與LDA所計算結果，利用system\_id關聯，進行資料合併

## 合併資料 ( 42 )

參數設定

Input - 18

Input - 37

任務結果

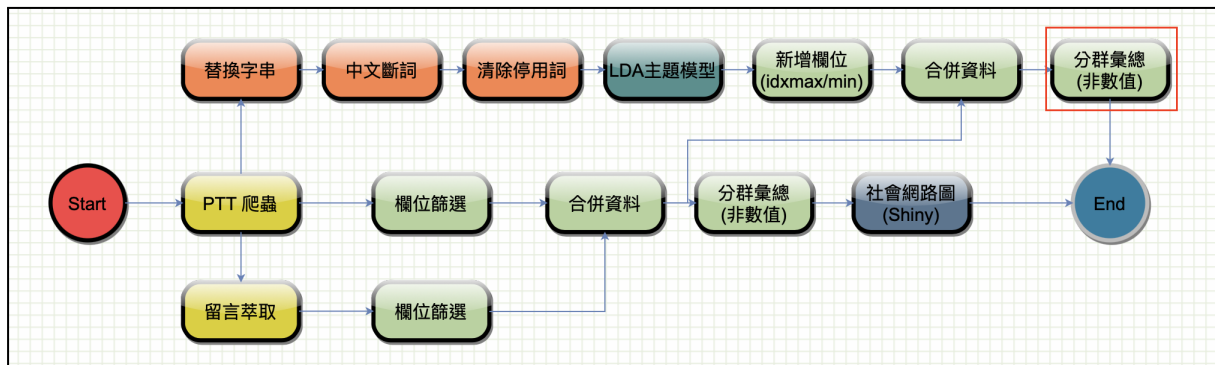
JOIN規則  
新增規則 刪除規則

left\_key  
system\_id  
-----請選擇-----

right\_key  
system\_id  
-----請選擇-----

合併之後下載CSV檔案，用於之後的Gephi繪圖軟體。

## 五、Gephi繪圖



1. 利用分群彙總(39)篩選出的LDA主題模型資料與發文—留言者的次數, 匯出csv檔後, 增加主題的大略分類以利辨識, 包含「人物」、「上線預告」、「心得」、「EP預告」四大項目。
2. 投入Gephi繪製人物關係圖, 共有2289個Nodes 與4682個Edges。
3. 將system\_id@count複製到Weight欄位。

來源	目標	類型	Id	Weight	topic	system_id@count
AblazeStar	AblazeStar	有向性	0	3.0	人物	3
AblazeStar	bown	有向性	1	1.0	人物	1
AblazeStar	cashko	有向性	2	1.0	人物	1
AblazeStar	eureka	有向性	3	2.0	人物	2
AblazeStar	q750830	有向性	4	2.0	人物	2
AblazeStar	raininglight	有向性	5	1.0	人物	1
AblazeStar	yutan0802	有向性	6	1.0	人物	1
AirOctopus	AirOctopus	有向性	7	5.0	上線預告	5
AirOctopus	AirOctopus	有向性	8	19.0	人物	19
AirOctopus	Barbarian123	有向性	9	1.0	人物	1
AirOctopus	Domobear	有向性	10	3.0	人物	3
AirOctopus	EELLSP	有向性	11	2.0	人物	2
AirOctopus	Howard61313	有向性	12	1.0	上線預告	1
AirOctopus	Howard61313	有向性	13	2.0	人物	2
AirOctopus	LaBoLa	有向性	14	1.0	人物	1
AirOctopus	Sabo5566	有向性	15	1.0	人物	1
AirOctopus	XAlOQ	有向性	16	3.0	人物	3
AirOctopus	XDDDD555	有向性	17	1.0	上線預告	1
AirOctopus	angylok	有向性	18	1.0	人物	1
AirOctopus	aseaeel	有向性	19	2.0	人物	2
AirOctopus	bonnenuit123	有向性	20	2.0	人物	2
AirOctopus	bonnie60735e	有向性	21	1.0	人物	1
AirOctopus	camelot0603	有向性	22	1.0	人物	1
AirOctopus	candyrain821	有向性	23	5.0	上線預告	5
AirOctopus	candyrain821	有向性	24	10.0	人物	10
AirOctopus	cashko	有向性	25	1.0	上線預告	1
AirOctopus	cashko	有向性	26	4.0	人物	4
AirOctopus	cherryiami	有向性	27	1.0	上線預告	1
AirOctopus	cherryiami	有向性	28	2.0	人物	2
AirOctopus	felicie	有向性	29	2.0	人物	2
AirOctopus	flkjsi	有向性	30	1.0	人物	1

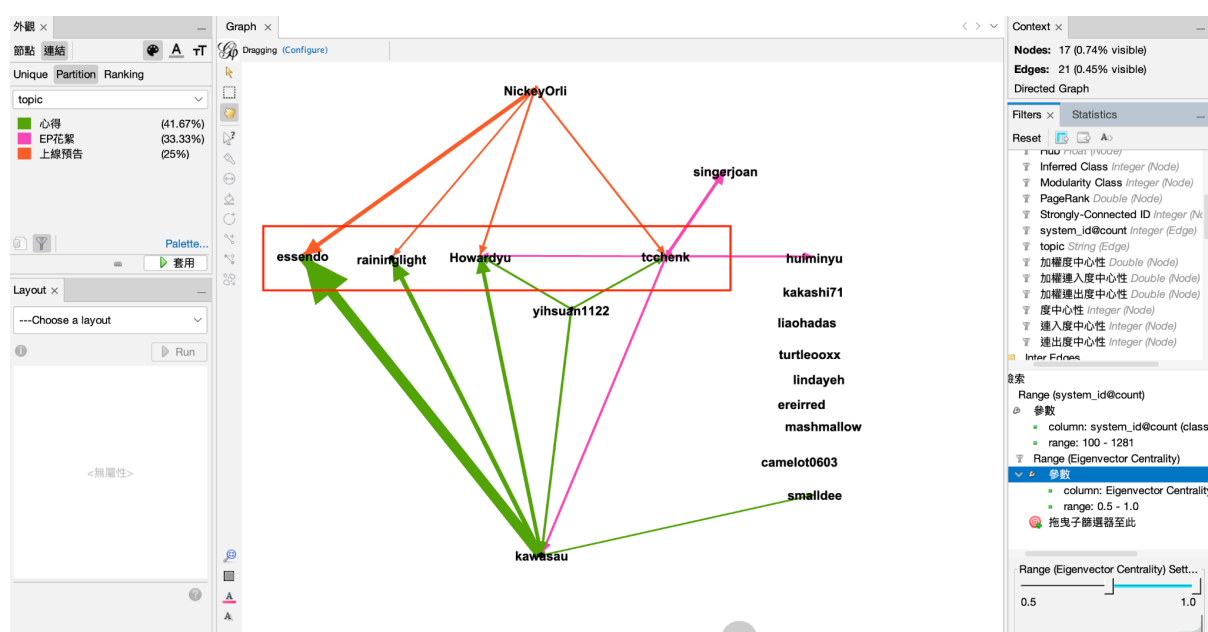
4. 利用分析功能計算各項數值, 包含平均度中心性 degree centrality、平均權重程度Avg. weighted Degree、PageRank Centrality、HIITS Score等, 數值呈現於



Edges的數據庫中。將用於後續的篩選功能。

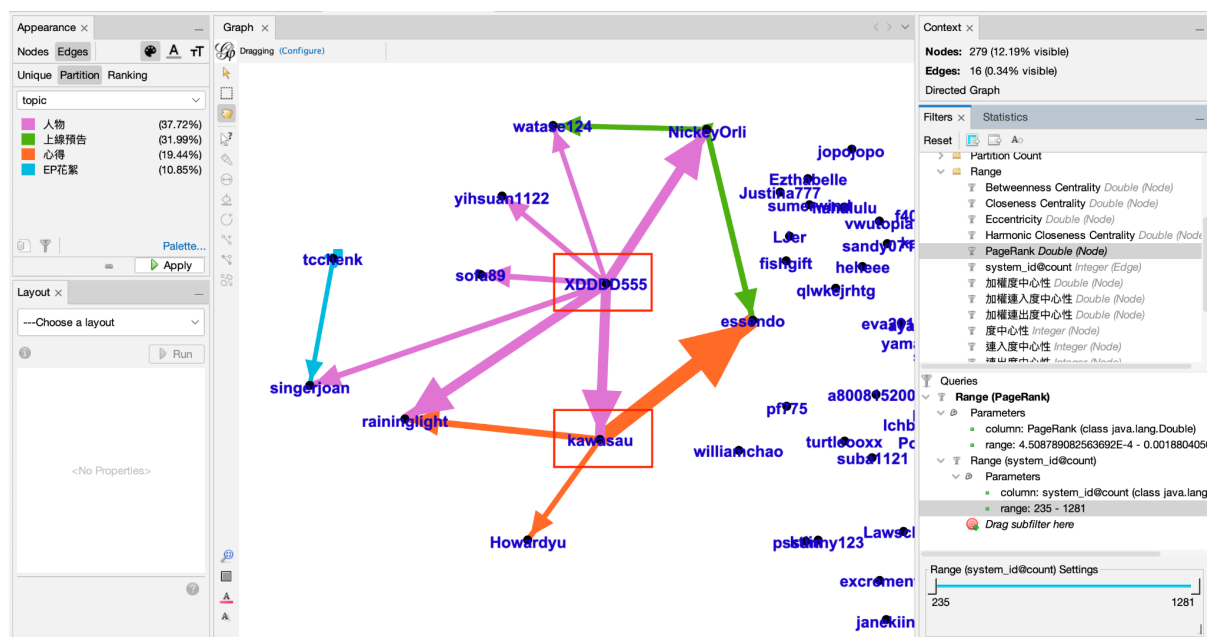
Id	Label	Interval	Authority	Hub	PageRank	Component ID	Strongly-Connected ID	Modularity Class	Inferred Class	Clustering Coefficient	Eigenvector Centrality
a0943780806	a0943780806		0.012092	0.0	0.000421	0	1528	0	0	0.0	0.063087
a100881	a100881		0.000176	0.0	0.000451	0	208	3	0	0.005011	
a1230215	a1230215		0.00202	0.0	0.000422	0	1753	0	0	0.000638	
a214567	a214567		0.011823	0.0	0.00042	0	1291	4	0	0.0	0.048631
a2526520	a2526520		0.014243	0.0	0.00042	0	915	3	0	0.0	0.062446
a36915077	a36915077		0.012092	0.0	0.000421	0	1527	0	0	0.0	0.063087
a4715646	a4715646		0.000479	0.0	0.000428	0	1860	5	0	0.0	0.000638
A48R018	A48R018		0.001991	0.0	0.000429	0	438	7	0	0.0	0.038944
a50202	a50202		0.048629	0.0	0.000436	0	435	3	0	0.666667	0.205444
a526jane	a526jane		0.000301	0.0	0.00043	0	1215	2	0	0.0	0.005223
a62312000	a62312000		0.014243	0.0	0.00042	0	914	3	0	0.0	0.062446
a68549271	a68549271		0.011823	0.0	0.00042	0	1290	4	0	0.0	0.048631
a7662888	a7662888		0.001991	0.0	0.000429	0	434	7	0	0.0	0.038944
a7700800	a7700800		0.014243	0.0	0.00042	0	913	3	0	0.0	0.062446
a8008152000	a8008152000		0.003381	0.0	0.000459	0	364	0	0	0.333333	0.020469
a86680280	a86680280		0.026117	0.0	0.00042	0	912	3	0	0.5	0.063272
a86851247	a86851247		0.028147	0.0	0.000436	0	393	6	0	0.583333	0.093793

5. 利用篩選功能依照留言加總次數100~1281, 以及Eigenvector centrality為0.5~1的留言者ID萃取出來, 並以顏色分類主題類別。可發現箭頭指向的發文者ID十分集中, 回溯自原始資料可發現這些帳號的發文多為韓劇資訊分享, 包含#新聞、#情報等, 因此推測會吸引有觀看相關戲劇的網友留言討論。



6. 再度利用篩選功能將PageRank範圍設定為大於4.5E-4, 以及system\_id@count大於235, 繪製圖如下。可發現帳號名「XDDD555」的留言者指向許多ID, 回溯至原始資料可發現其留言頻率與次數皆高, 除了分享心得與看法外, 也會提醒其他留言者注意版規, 因此與上述的頻繁發文者互動甚多。另外, 帳號名「kawasau」也是與他者互動頻繁, 回溯資料發現其不僅會發布#情報、#新聞等戲劇資訊, 也會參與韓劇版的徵文, 以及熱烈的在留言板中發表心得, 使得其與其他帳號的連結

程度極高。



## 六、結論

本次讀書會以韓劇為主題進行社會網路圖與LDA主題模型探究，進而將主題關聯至 Gephi繪圖，目的以此了解出在2023/4/1 ~ 2023/5/20之間在PPT韓劇版上的留言討論度及主要核心發文者意向。

首先，我們針對留言萃取，比對留言者與其發文者之間的節點關係，在社會網路圖，我們發現貼文者raininglight為最大聚落發展出關係，以此研判該位發文者為韓劇版的主要核心人物，常針對相關韓劇議題進行發文，引發其他使用者討論度並同時也會積極回應其發文留言者的評論，使得互動性越發熱絡。其次則為kateloveyoo、jello929、chu3、kukulee、godina339等貼文者，亦為韓劇版上常見貼文及留言者。

接下來，我們針對發文篇數裡的所有留言數內容，進行主題模型的探究，歸納出EP花絮、上線預告、心得、人等四大模型，主題間相符程度排行: Topic 1 (45.7%) > Topic 2 (20.9%) > Topic 3 (17.9%) > Topic 4 (15.5%)，由上述主題模型結果，以主題連貫性的頻率來看，各主題前三大詞彙:「主題1: ep > 花絮 > 媽媽」、「主題2: 電視劇 > 飾演 > 作品」、「主題3: 知道 > 喜歡 > 最後」、「主題4: 孩子 > 集團 > 丈夫」。故可推斷常在韓劇版

上引發討論度的韓劇故事背景，可能為遭老公不倫背叛的妻子、角色關係間有財閥集團、爭奪小孩扶養權等議題做劇情設定。最後看完戲劇的回覆留言者也會針對幕後花絮與其真實飾演的演員角色其他詮釋的作品進行關聯性的討論，也可能透過熱烈迴響進一步喜歡該部韓劇演員。經調查此期間相符的韓劇為「車真淑醫生」與「離婚律師申晟瀚」、「壞媽媽」，三部韓劇的故事角色背景設定與上述LDA主題模型相符，可佐證其LDA模型結果的準確度吻合性。

進而，我們想透過Gephi工具協助繪圖以視覺化呈現彼此間關係緊密程度，依照LDA模型計算點邊之間的權重，主要以page rank做排名顯示，用來回溯其回覆貼文者與發文者的關係，可佐證另有帳號ID為「XDDD555」 & 「kawasau」也是活躍於韓劇版具有影響力的指標性人物。