

影片網址

<https://youtu.be/mDzT3AOg09s>





社群媒體分析 期中專案報告

日本旅遊

指導教授：黃三益

第十組 組員

N104020001 李采容

N104020007 郭育雯

N104020021 馮慧嬌

大綱

- 動機和目的
- 資料集
- 資料清理過程
- 資料分析過程
- 視覺化的分析結果與解釋
- 結論

動機和目的

- 動機：過去幾年由於疫情關係很長一段時間無法出國，而隨著近期疫情趨緩、國門開放，國人爭先恐後出國旅遊，旅行的話題隨之增加。

- 目的：希望藉此分析鄉民對到日本京都、東京和北海道旅行的看法。

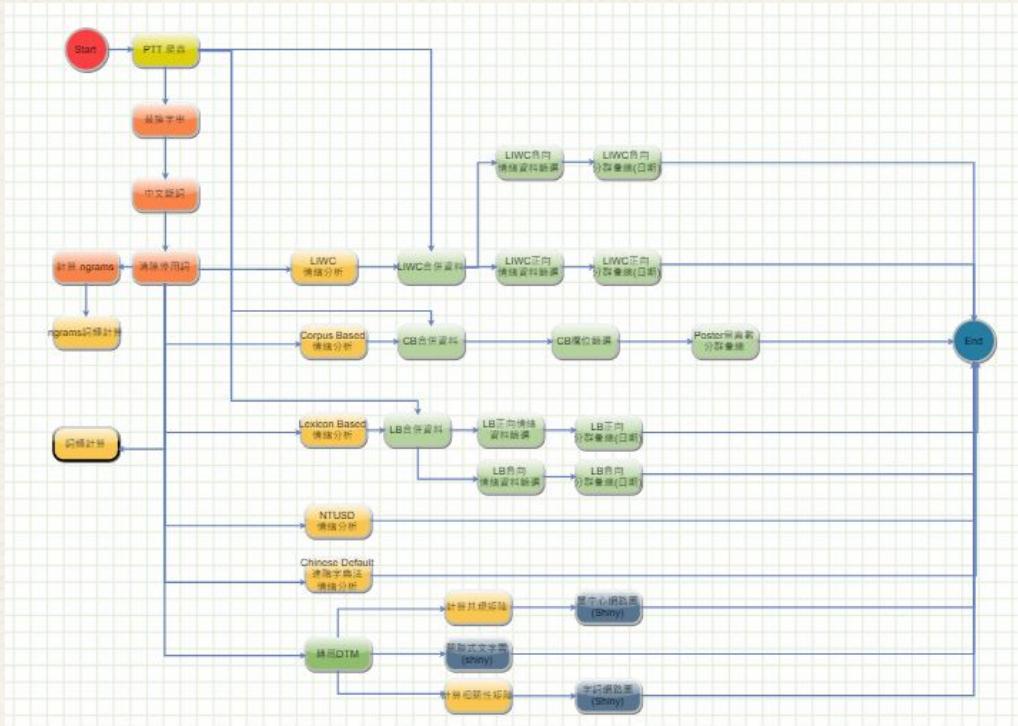
資料集

- 來源:PTT Japan_Travel 日旅版
- 關鍵字:京都、東京、北海道
- 日期:2023/1/1~2023/4/10

資料清理過程

Tarflow 平台

工作流程名稱:日本旅遊



整體流程圖

PTT爬蟲

PTT 爬蟲 (20)

參數設定

任務結果

選擇看板 *

- homemaker(家管)
- home_sale(房屋)
- Hsinchu(新竹)
- hypermall(賣場)
- Insurance(保險)
- iOS(iOS)
- Japan_Travel(日旅)

搜尋關鍵字 ⓘ

- 京都
- 東京
- 北海道

排除關鍵字 ⓘ

- 台灣
- 國旅

搜尋起始日期

2023/01/01

搜尋結束日期

2023/04/10

儲存更改

篩選PTT日旅版在2023/01/01~04/10期間有關「京都、東京、北海道」等關鍵字的內文，並排除「台灣、國旅」關鍵字，最後得到1440筆資料。

資料清理:替換字詞

替換字串 (2)

參數設定

Input - 20

任務結果

選擇處理欄位 *

artContent

選擇替換規則檔案 ①

-----請選擇-----

替換字串設定 ②

```
Re:([.*])>>
([.*])>>
\n\n\n>> 。
\n\n>> 。
\n>> ,
```

儲存更改

The screenshot shows a user interface for data cleaning, specifically for replacing characters or strings. At the top, there's a header with the title '資料清理:替換字詞'. Below it is a sub-header '替換字串 (2)'. The interface is divided into several sections: '參數設定' (Parameter Settings), 'Input - 20' (Input - 20), and '任務結果' (Task Result). In the '參數設定' section, there's a dropdown labeled '選擇處理欄位 *' containing the value 'artContent'. Another dropdown labeled '選擇替換規則檔案 ①' has the placeholder text '-----請選擇-----'. To the right, under '替換字串設定 ②', there's a code editor window containing the following regular expression code:

```
Re:([.*])>>
([.*])>>
\n\n\n>> 。
\n\n>> 。
\n>> ,
```

At the bottom right of the interface is a green button labeled '儲存更改' (Save Changes).

將內文部分字串替換，如斷行符號替換成標點符號，還有網址的替換。

資料清理:中文斷詞

中文斷詞 (3)

參數設定

Input - 2

任務結果

選擇處理欄位 *

result

定義詞彙 ⓘ

- 羽田機場 1000
- 成田機場 1000
- 關西機場 1000
- 新千歲機場 1000
- 御殿場 1000

選取字典 ⓘ

-----請選擇-----

儲存更改

This screenshot shows a user interface for a Chinese word segmentation tool. At the top, there's a header with the title '資料清理:中文斷詞' (Data Cleaning: Chinese Word Segmentation). Below the header, a sub-header reads '中文斷詞 (3)'. The interface is divided into several sections: '參數設定' (Parameter Settings) with a dropdown menu set to 'result'; 'Input - 2' which is currently empty; '任務結果' (Task Results) which also appears to be empty; '選擇處理欄位 *' (Select Processing Column) with a dropdown menu set to 'result'; a '定義詞彙' (Defined Vocabulary) section containing a list of five entries: '羽田機場 1000', '成田機場 1000', '關西機場 1000', '新千歲機場 1000', and '御殿場 1000'; '選取字典' (Select Dictionary) with a dropdown menu set to '-----請選擇-----'; and a large green '儲存更改' (Save Changes) button at the bottom.

調整詞彙權重，將特定關鍵字給予較高的權重，如花見小路、伏見稻荷、廣域周遊卷等，以利斷詞分析。

資料清理:清除停用詞

清除停用詞 (4)

參數設定	Input - 3	任務結果	
語言 *	Chinese	使用預設停止詞 是	
是否清除單字元 ⓘ	是	是否轉為小寫英文 是	
清除英文字母 *	是	清除數字 *	是
清除換行符號 *	是	清除特殊標點符號 *	是
清除html tag *	是	自定義停止詞	物品 幫忙 位置 確認 網誌

儲存更改

將斷詞後的結果進行清除停用詞，其中包含預設、英文字母、數字，另自定義停用詞，如物品、幫忙、位置、確認等不影響分析結果的詞彙。

資料分析過程

Tarflow 平台

計算ngrams

計算 ngrams (33)

參數設定 Input - 4 任務結果

165305 ngram數量

任務結果

Show 10 entries Search:

system_id	result
1	[大阪 京都, 京都 首次 首次 必訪, 必訪 住宿, 住宿 七條, 七條 鄰近, 鄰近 京板, 京板 電車, 電車 七條, 七條 七條, 七條 河原町, 河原町 飯店, 飯店 還沒訂, 還沒訂 關西機場, 關西機場 京都, 京都 飯店, 飯店 走路, 走路 放完, 放完 四條, 四條 河原町, 河原町 鴨川, 鴨川 納涼, 納涼 床看, 床看 夜景, 夜景 燒肉, 燒肉 回七條, 回七條 祇園, 祇園 四條, 四條 早去, 早去 市場, 市場 市場, 市場 四條, 四條 搭至, 搭至 淸水, 淸水 下車, 下車 清水, 清水 清水寺, 清水寺 二年, 二年 寧寧, 寧寧 之道, 之道 八坂, 八坂 神社, 神社 祇園, 祇園 花見小路, 花見小路 祇園, 祇園 居酒屋, 居酒屋 四號, 四號 回家, 回家 四條, 四條 河原町, 河原町 七條, 七條 河原町, 河原町 伏見, 伏見 奈良, 奈良 一早, 一早 京阪, 京阪 電車, 電車 伏見, 伏見 稲荷, 稲荷 稲荷, 稲荷 穏居, 穏居 解決, 解決 伏見, 伏見 奈良, 奈良 轉搭, 轉搭 奈良, 奈良 循環, 循環 春日, 春日 神社, 神社 本殿, 本殿 春日, 春日 神社, 神社 奈良, 奈良 公園, 公園 志津香, 志津香 東大寺, 東大寺 福興, 福興 東向, 東向 商店街, 商店街 東福寺, 東福寺 京阪, 京阪 電車, 電車 七條, 七條 大阪, 大阪 天守閣, 天守閣 心齋橋, 心齋橋 逛街, 逛街 天守閣, 天守閣 京板, 京板 電車, 電車 七條, 七條 滿橋, 滿橋 步徒, 步徒 天守閣, 天守閣 心齋橋, 心齋橋 道頓堀, 道頓堀 地鐵, 地鐵 緑地, 緑地 森之宮, 森之宮 心齋橋, ...]
2	[徵求 吃吃, 吃吃 旅伴, 旅伴 這趟, 這趟 臨時, 臨時 號跟 環球, 環球 天都還, 天都還 彈性, 彈性 標題, 標題 大阪, 大阪 住宿, 住宿 市區, 市區 偏好, 偏好 日料, 日料 甜食, 甜食 關西, 關西 地區, 地區 歡迎, 歡迎 討論, 討論 大阪, 大阪 照應, 照應 曾經, 曾經 日本, 日本 留學, 留學 一年, 一年 出差, 出差 閃過, 閃過 京都, 京都 還算]

全螢幕瀏覽 點我下載完整CSV資料 點我下載完整Rdata 點我下載完整json資料

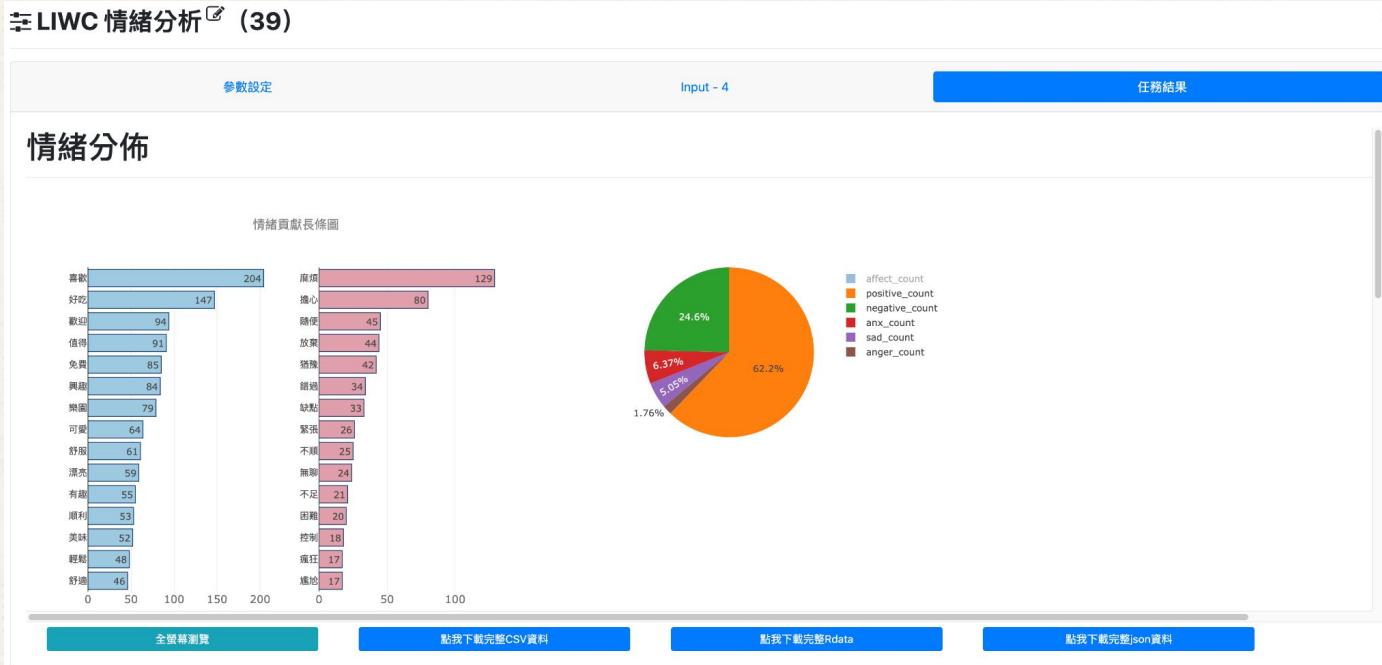
清除停用字後，計算ngrams，採用 $n=2$ 找出兩個較常共同出現的詞彙，並反覆調整斷詞。

詞頻計算



清除停用字後進行詞頻計算，停止詞設定如：地址、交易、徵求等與旅遊不相關的字詞，利於後續分析。

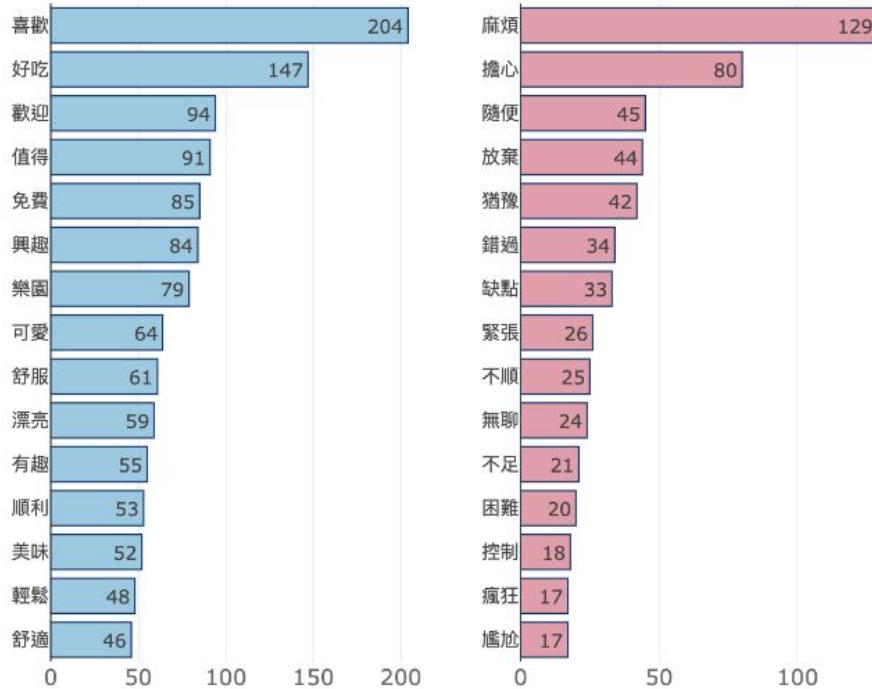
LIWC情緒分析



將清除停用字後的結果進行LIWC情緒分析，結果正向情緒大於負向情緒，正向情緒達48.8%，負向情緒為32.1%。

LIWC情緒分析

情緒貢獻長條圖



前15的正負情緒詞

LIWC合併資料及篩選

The screenshot shows a software interface for filtering LIWC emotional data. The title bar reads "LIWC負向 情緒資料篩選 (116)" with a note "參數有做更動，建議重新執行". The main area has tabs for "參數設定" (selected), "Input - 7", and "任務結果". Under "參數設定", there is a section for "條件式" with a note "必填". A text input field contains the condition "\$sentiment_value < 0". At the bottom is a green button labeled "儲存更改".

將LIWC情緒分析後的結果與爬蟲資料進行合併，並利用情緒值區分出整體正向文章641筆及負向文章310筆。

LIWC正向分群彙總

LIWC正向 分群彙總(日期) (8)		
參數設定	Input - 6	任務結果
<h2>任務結果</h2>		
Show 10 entries		Search: <input type="text"/>
artDate	: positive_count	
2023-03-16		86
2023-03-18		82
2023-03-07		78
2023-03-20		76
2023-03-12		75
2023-02-02		66
2023-02-21		66
2023-02-13		63
2023-03-09		61
2023-02-07		60
Showing 1 to 10 of 96 entries		
Previous 1 2 3 4 5 ... 10 Next		
全螢幕瀏覽	點我下載完整CSV資料	點我下載完整Rdata
		點我下載完整json資料

依照日期彙總每天正向文章的正向情緒字彙總數，前五名皆集中在三月。

LIWC負向分群彙總

LIWC負向 分群彙總(日期) (52)

參數設定 Input - 116 任務結果

任務結果

Show 10 entries Search:

artDate	: negative_count
2023-04-03	17
2023-03-15	16
2023-03-05	15
2023-02-15	14
2023-02-19	14
2023-03-20	14
2023-02-17	13
2023-02-21	13
2023-03-09	13
2023-03-17	13

Showing 1 to 10 of 95 entries Previous 1 2 3 4 5 ... 10 Next

[全螢幕瀏覽](#) [點我下載完整CSV資料](#) [點我下載完整Rdata](#) [點我下載完整json資料](#)

依照日期彙總每天負向文章的負向情緒字彙總數，彙總每日不超過20個字彙，且分佈月份不均。

Corpus Based 情緒分析

Corpus Based 情緒分析 (37)

參數設定 Input - 4 任務結果

統計資訊

924	負向情緒數	
516	正向情緒數	

任務結果

Show 10 entries Search:

system_id	sentiment_value
1	0.999
2	0.000
3	0.160
4	1.000
5	0.000
6	0.450
7	1.000

[全螢幕瀏覽](#) [點我下載完整CSV資料](#) [點我下載完整Rdata](#) [點我下載完整json資料](#)

將清除停用字後的結果進行 Corpus Based 情緒分析，結果負向情緒文章數大於正向情緒，負向情緒文章共有 924 篇，正向情緒文章共有 516 篇。

Corpus 合併資料及篩選

CB欄位篩選 (79) 參數有做更動，建議重新執行

任務結果

system_id	artDate	artPoster	sentiment_value
1	2023-01-01 04:31:03	jjane068tw	0.999
2	2023-01-01 13:33:46	william81413	0.000
3	2023-01-01 16:40:35	vuosu	0.160
4	2023-01-01 16:44:12	hidein	1.000
5	2023-01-01 18:15:56	SteveVai	0.000
6	2023-01-01 22:40:09	jenny0904	0.450
7	2023-01-01 23:11:44	JacksonWu	1.000
8	2023-01-01 23:50:18	notion0125	0.208
9	2023-01-02 01:31:46	Sherlock56	0.000
10	2023-01-02 10:57:25	snoopy63	0.329

Showing 1 to 10 of 100 entries

Previous [1](#) [2](#) [3](#) [4](#) [5](#) ... [10](#) Next

[全螢幕瀏覽](#) [點我下載完整CSV資料](#) [點我下載完整Rdata](#) [點我下載完整json資料](#)

將Corpus Based 情緒分析後的結果與爬蟲資料進行合併，並保留系統編號、發文日期、發文者、情緒分數。

發文者發文數分群彙總

Poster留言數 分群彙總 (83)

參數設定 Input - 79 任務結果

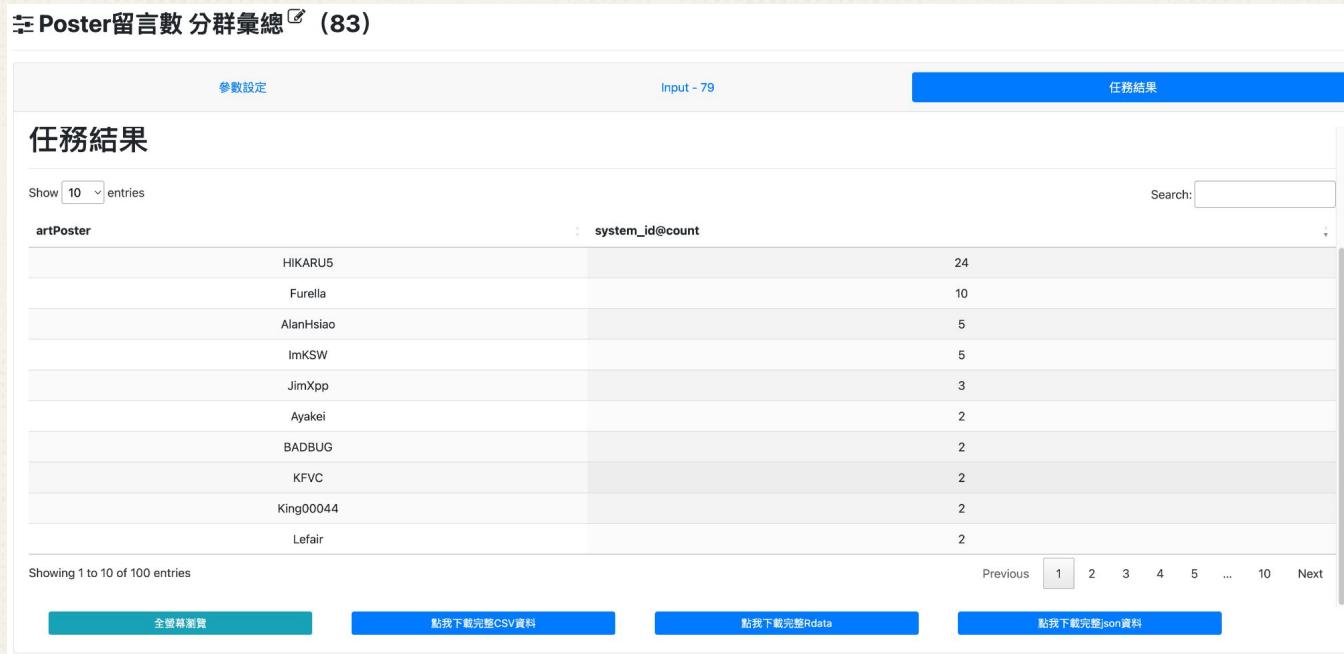
任務結果

Show 10 entries Search:

artPoster	system_id@count
HIKARU5	24
Furella	10
AlanHsiao	5
ImKSW	5
JimXpp	3
Ayakei	2
BADBUD	2
KFVC	2
King00044	2
Lefair	2

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

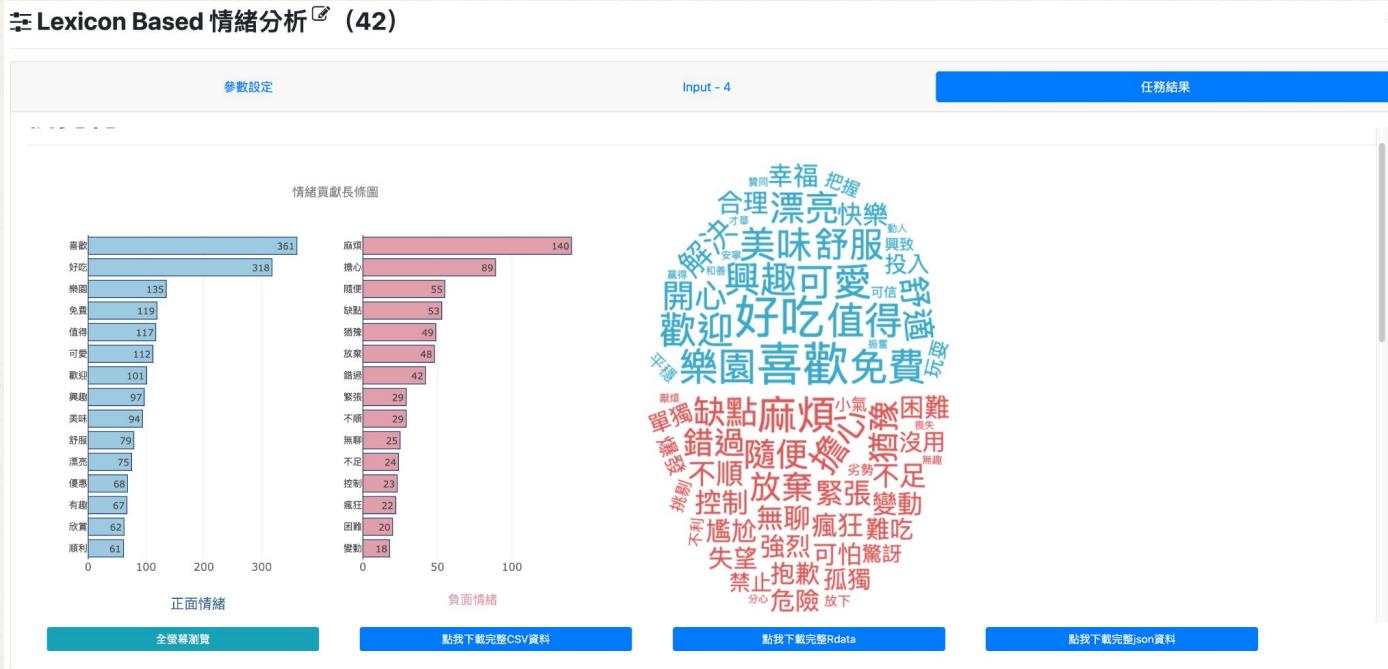
全螢幕瀏覽 點我下載完整CSV資料 點我下載完整Rdata 點我下載完整json資料



The screenshot shows a data analysis interface with a title bar 'Poster留言數 分群彙總 (83)'. Below it is a navigation bar with tabs: '參數設定' (Parameter Settings), 'Input - 79', and '任務結果' (Task Result). The main area is titled '任務結果' (Task Result) and displays a table of poster counts. The table has two columns: 'artPoster' and 'system_id@count'. The data shows 10 rows of results, with HIKARU5 having the highest count of 24. At the bottom, there are links for '全螢幕瀏覽' (Full Screen View), '點我下載完整CSV資料' (Download Complete CSV Data), '點我下載完整Rdata' (Download Complete Rdata), and '點我下載完整json資料' (Download Complete json Data). There are also page navigation controls for 'Previous' and 'Next'.

利用發文者分群，找出每個發文者發文的數量，其中第一名共發表24篇文章。

Lexicon Based 情緒分析



將清除停用字後的結果進行Lexicon Based 情緒分析，結果正向情緒大於負向情緒，正向詞彙前五有喜歡、好吃、樂園、免費、值得，負向詞彙前五有麻煩、擔心、隨便、缺點、猶豫。

Lexicon 合併資料及篩選

The screenshot shows a user interface for data filtering. At the top, it says "LB正向情緒 資料篩選 (45)". Below this, there are three tabs: "參數設定" (selected), "Input - 44", and "任務結果". Under "參數設定", there is a section titled "條件式 * ⓘ" containing the condition "\$sentiment_value >0". At the bottom of the screen, there is a green button labeled "儲存更改".

將情緒分析後的結果與爬蟲資料進行合併，並利用情緒值區分出整體正向文章717筆及負向文章210筆。

Lexicon正向分群彙總

LB正向 分群彙總(日期) (49)

參數設定 Input - 45 任務結果

任務結果

Show 10 entries Search:

artDate	sentiment_value
2023-03-16	106
2023-02-21	87
2023-03-07	87
2023-03-12	84
2023-02-23	83
2023-03-18	83
2023-03-20	75
2023-02-13	73
2023-03-11	72
2023-02-02	69

Showing 1 to 10 of 96 entries Previous 1 2 3 4 5 ... 10 Next

全螢幕瀏覽 點我下載完整CSV資料 點我下載完整Rdata 點我下載完整json資料

依照日期彙總每天正向文章的正向情緒總分，前十名多數集中在二三月。

Lexicon負向分群彙總

LB負向 分群彙總(日期) (54)

參數設定 Input - 120 任務結果

任務結果

Show 10 entries Search:

artDate	sentiment_value
2023-02-07	-14
2023-02-16	-12
2023-03-09	-10
2023-02-08	-9
2023-01-12	-8
2023-01-22	-8
2023-03-19	-8
2023-03-27	-8
2023-04-03	-8
2023-03-13	-7

Showing 1 to 10 of 86 entries Previous [1](#) [2](#) [3](#) [4](#) [5](#) ... [9](#) Next

[全螢幕瀏覽](#) [點我下載完整CSV資料](#) [點我下載完整Rdata](#) [點我下載完整json資料](#)

依照日期彙總每天負向文章的負向情緒總分，其中排名分布不均。

NTUSD 情緒分析



將清除停用字後的結果進行NTUSD情緒分析，結果正向情緒大於負向情緒，正向詞彙前五有喜歡、好吃、經驗、便宜、說明，負向詞彙前五有麻煩、擔心、沒辦法、可惜、沒想到，其結果和Lexicon based有些微差異。

進階字典法情緒分析



將清除停用字後的結果進行進階字典法情緒分析，結果正面情緒大於負向情緒，正面詞彙前五有喜歡、好吃、歡迎、值得、免費，負向詞彙前五有麻煩、擔心、隨便、放棄、猶豫。

計算共現矩陣

計算共現矩陣 (110)

參數設定 Input - 60 任務結果

熱圖

熱圖

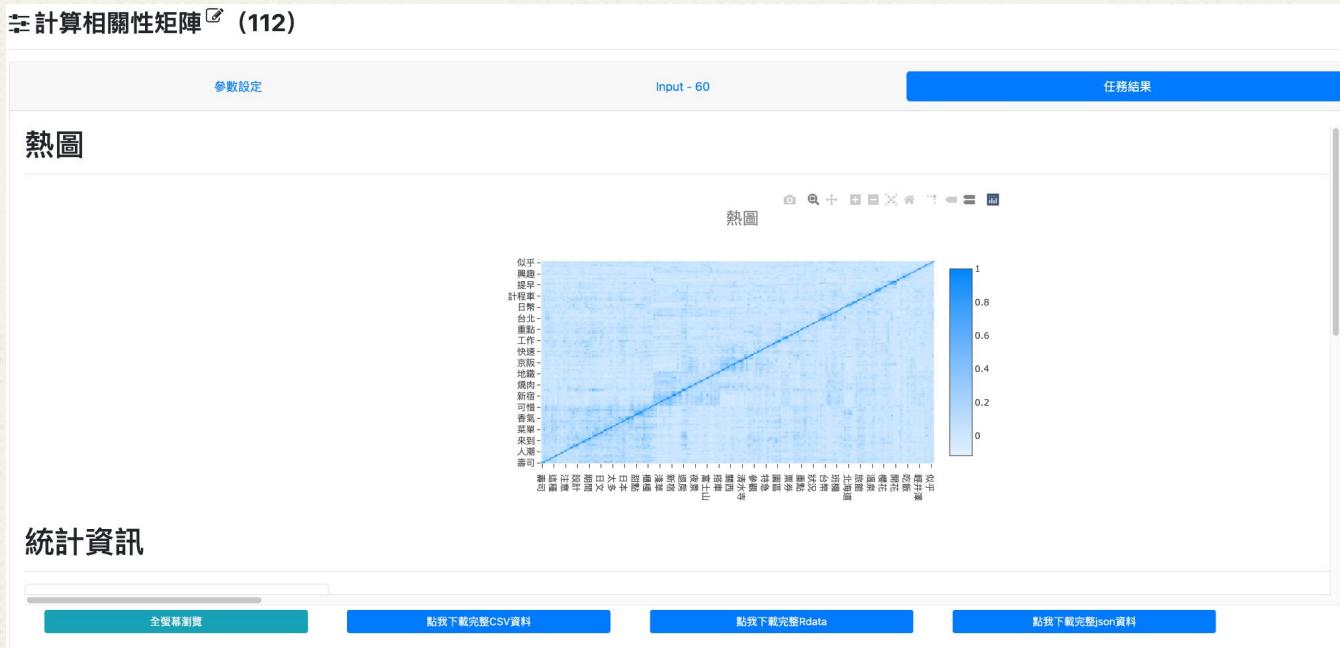
The heatmap displays the co-occurrence matrix for 60 input items. The x-axis and y-axis both list the same 60 items, with labels rotated vertically. A color scale bar on the right indicates frequency from 0 to 3500, with major ticks at 0, 500, 1000, 1500, 2000, 2500, 3000, and 3500. The matrix shows high density along the diagonal and varying levels of cross-correlation between different items.

統計資訊

全螢幕瀏覽 點我下載完整CSV資料 點我下載完整Rdata 點我下載完整json資料

將資料轉為DTM後計算共現矩陣，用於之後視覺化分析。

計算相關性矩陣



將資料轉為DTM後計算相關性矩陣，用於之後視覺化分析。

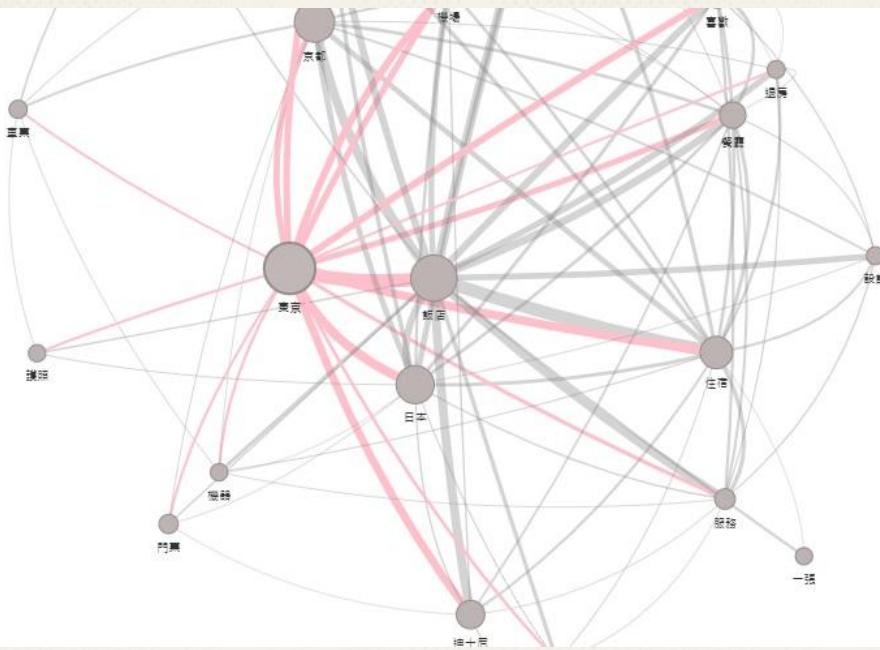
視覺化的分析結果與解釋

Tarflow 平台 + 視覺化儀表板(日本旅遊)

關聯式文字雲

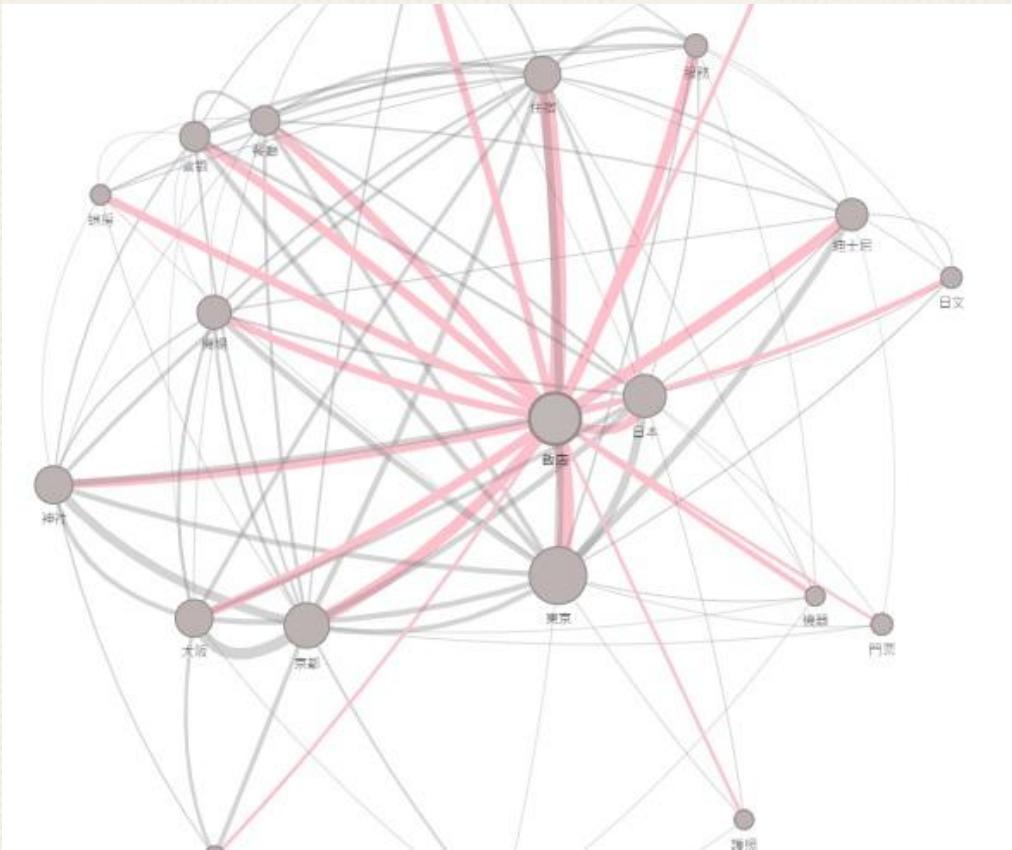
在轉為DTM後的關聯式文字雲可以看到，東京與成田機場、晴空塔、上野等有關聯；京都跟大阪、關西、奈良等地區常一起討論；住宿與飯店服務等話題也具有關聯性。

單中心網路圖



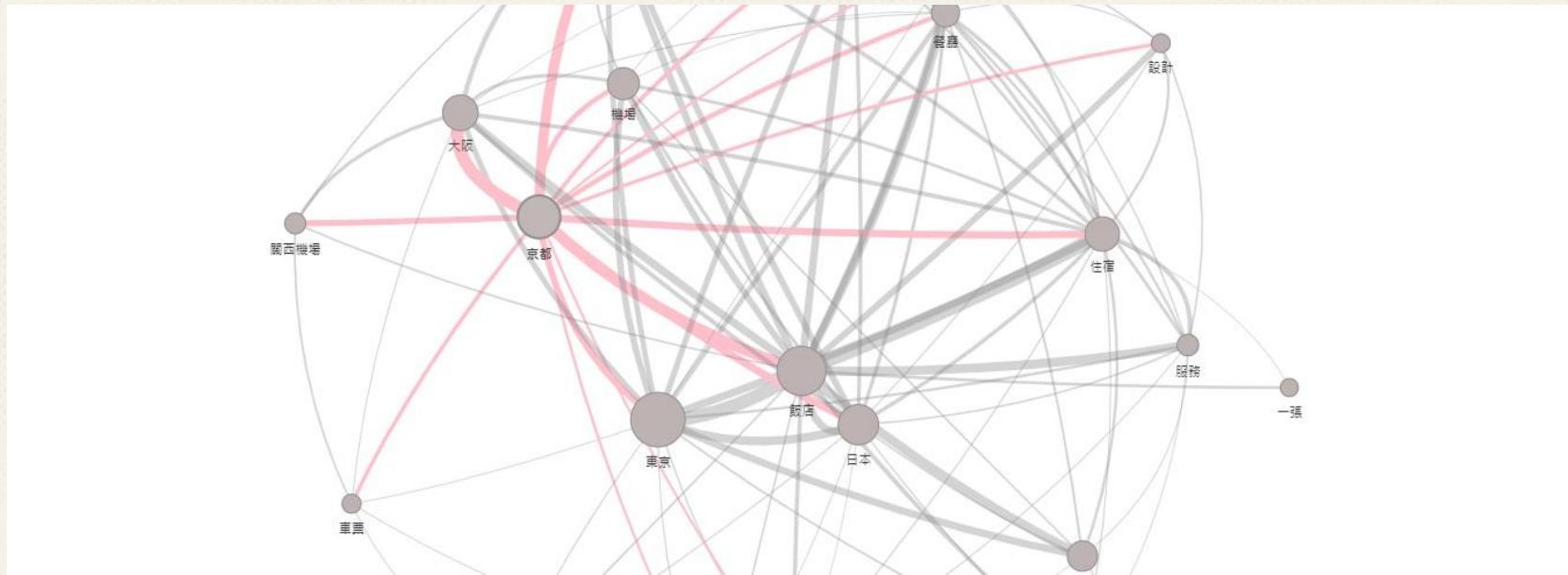
東京主題常與迪士尼、護照、日文、日本和飯店等主題一起討論。

單中心網路圖



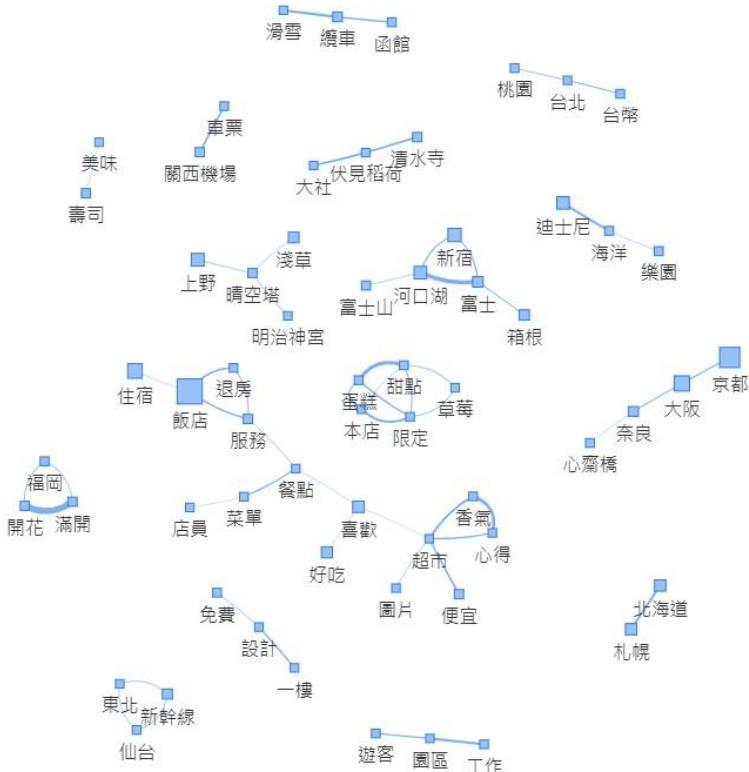
由圖可看出，飯店主題常與機場、餐廳、服務和各日本地點一起做討論。

單中心網路圖



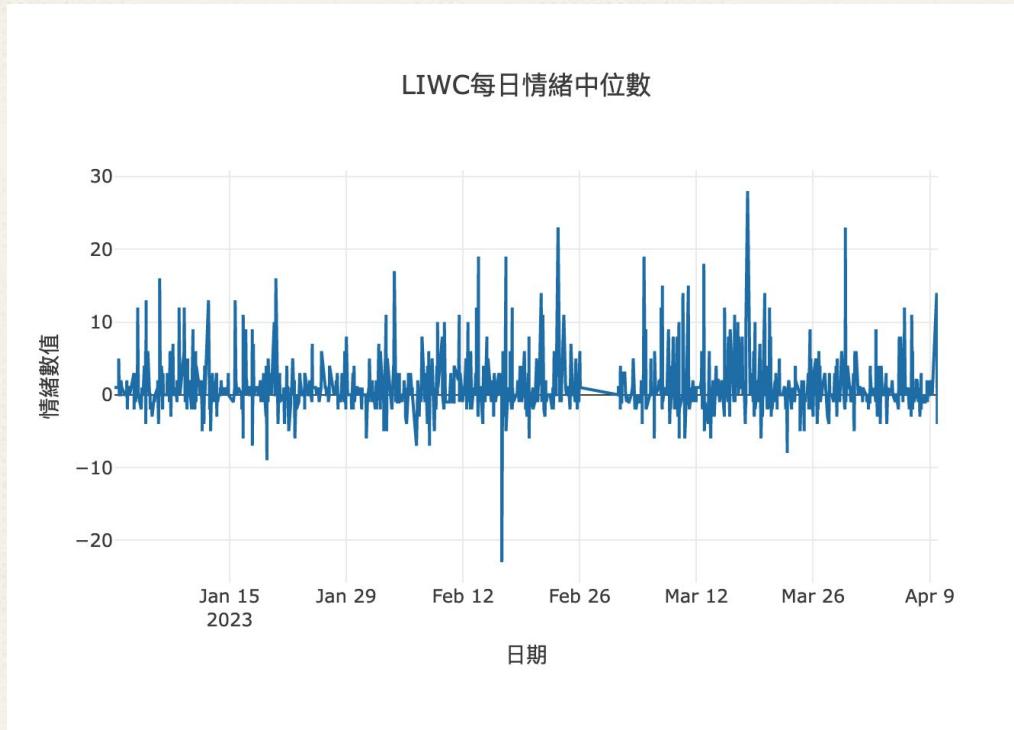
京都主題常與大阪、神社、關西機場、車票等主題一起討論。

字詞網路圖



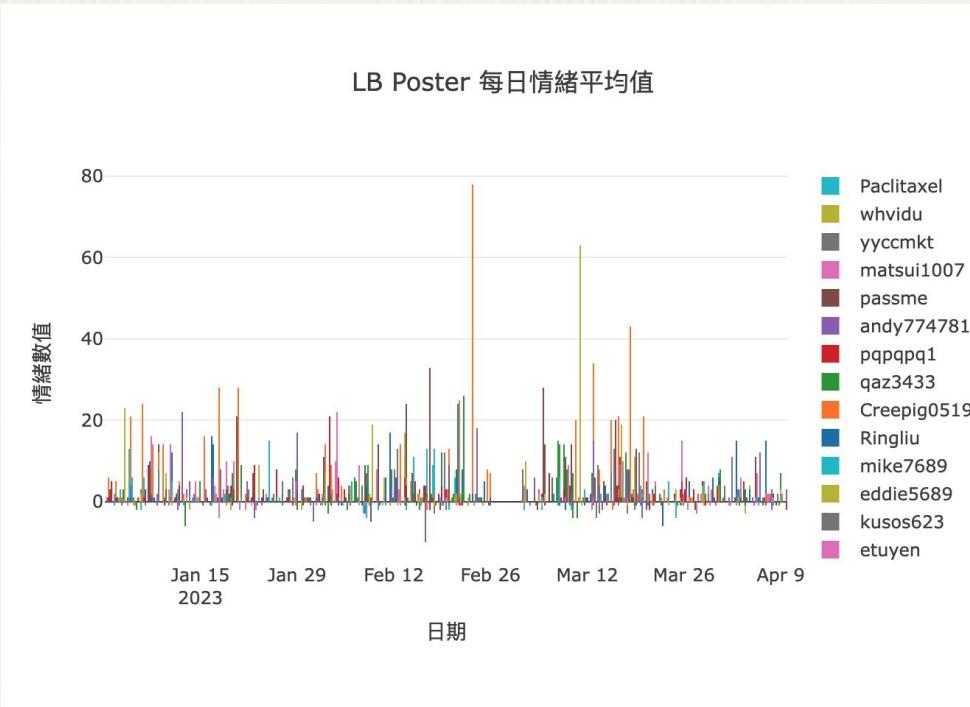
字詞網路圖可以看到，飯店與住宿、退房的相關性高，另外還有草莓甜點、蛋糕等相關討論。

每日情緒中位數



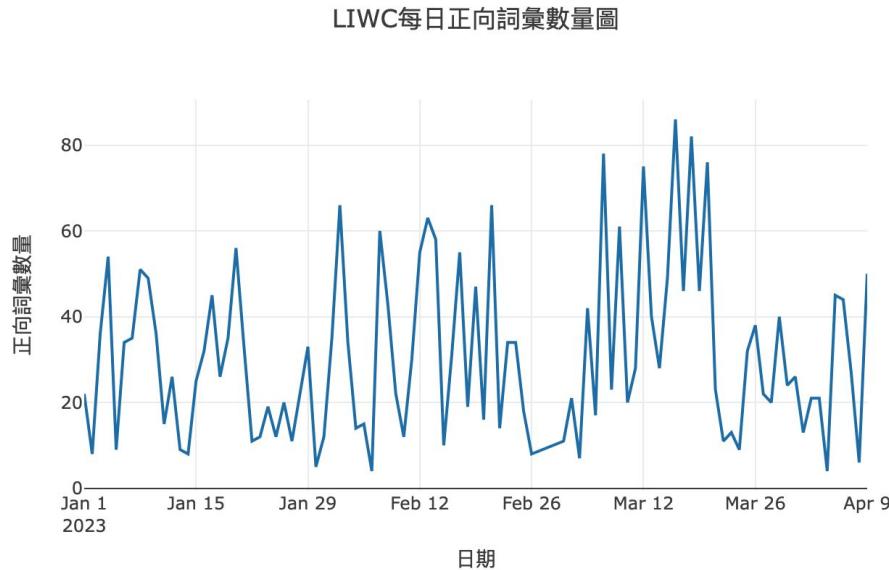
利用LIWC情緒分析的結果依照日期取情緒中位數，以3/28最高(數值28)，以2/16最低(數值-23)。

發文者每日情緒平均值



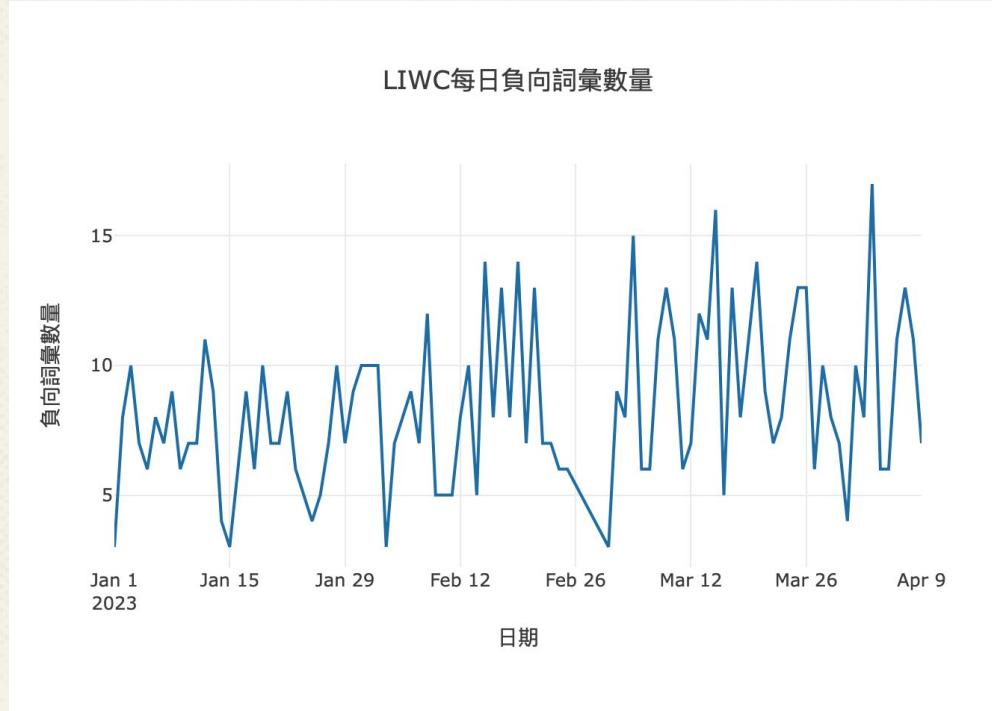
利用Lexicon Based情緒分析的結果依照日期和發文者進行平均。由圖中可看出多數留言者對於日本旅遊討論的情緒值為正向。

每日正向文章正向詞彙總數



利用LIWC情緒分析的結果依照日期彙總正向詞彙數量。最高峰為3/16(四)。

每日負向文章負向詞彙總數



利用LIWC情緒分析的結果依照日期彙總負向詞彙數量。最高峰為3/17(五)，推測是因3/20(一)為輕症免隔離的開始日期，造成網路上的旅遊話題較高。

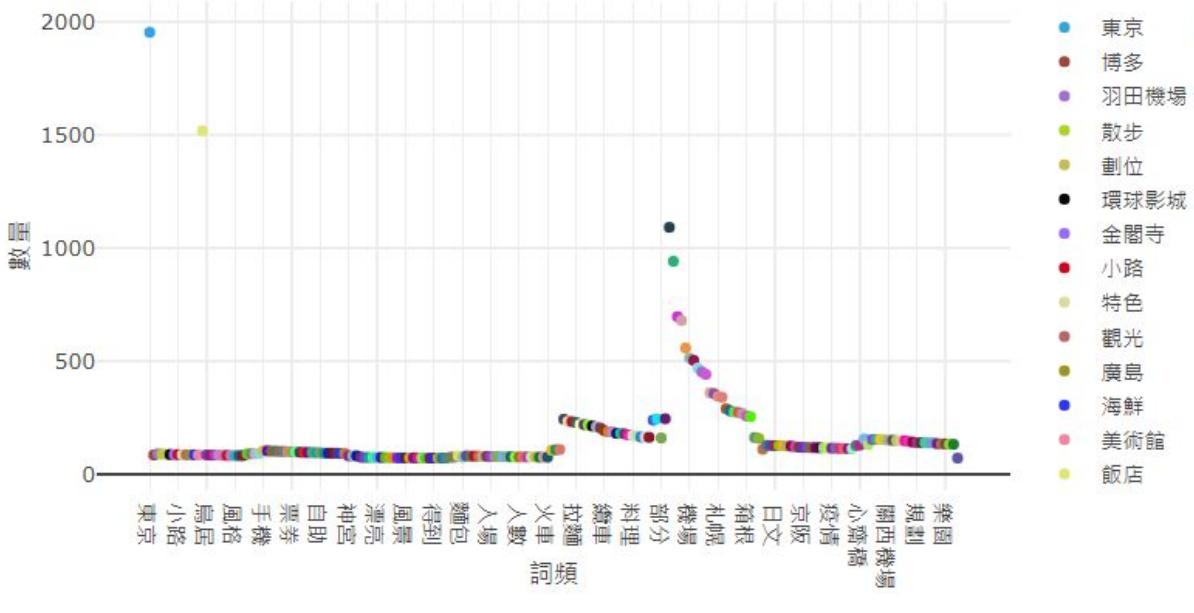
詞頻計算文字雲



詞頻計算中產生的文字雲，會依據討論出現次數的多寡決定文字大小，討論度越高則文字字體越大，圖中的東京、京都、大阪等地區都是討論度較高的日本旅遊地點。

詞頻計算數量散佈圖

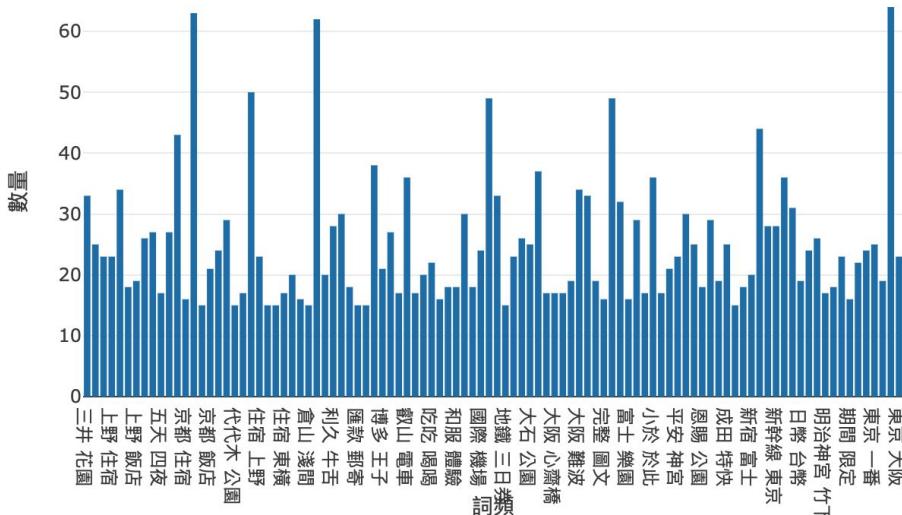
詞頻計算數量散佈圖



詞頻計算中的每個點代表一個詞彙，圖示出現數量越高則表示討論的頻率越高。圖中的”東京”出現1954次為討論度最高的詞彙。

ngrams詞頻計算長條圖

ngrams詞頻計算彙總長條圖



利用ngrams詞頻計算最常共同出現的兩個字，其中以東京地鐵、京都大阪、八坂神社最常共同出現。

結論

- 整體而言，日本旅遊的討論度除Corpus Based情緒分析外，多數都偏正向情緒。
- 由ngrams詞頻計算結果，可推測東京地區相關的旅遊議題的討論度最高。
- 由詞頻計算結果來看，討論度最高的議題前三名分別為東京、飯店與京都。
- 在三月疫情政策鬆綁之後，網路上討論度明顯增加。
- 依據上述各項情緒分析結果中發現，多數留言者的情緒皆為正向，由此亦推論其對於該地區旅遊多採正面評價。

THANKS!