

社群媒體分析

第一次讀書會報告_第十組

一、分析主題

台灣民眾對大陸美食的情緒分析

二、組員

N124320007 藍筱琦 N124320011 林紀吟 N124320012 鄭義璋 N124320017 郭良益

N124320019 吳邦齊 N124320020 陳軍弦 N124320024 何牧 N124320029 黃靖紋

三、分析工具

中山大學工作流程平台

工作流程名稱：第十組_第一次讀書會報告

四、緣起：

近年來，中國大陸美食在台灣掀起一陣風潮，藉由社群媒體的傳播更是挑起台灣民眾的味蕾，無論是螺螄粉、酸菜魚、肉夾饅等等五花八門的中國大陸美食在台灣隨處可見，在台灣夜市及路邊實體店都如雨後春筍般的誕生。以前對於中國大陸食品安全的考量問題較有疑慮，很多人甚至不敢嘗試從中國來的食品，隨著時代變遷兩岸互動往來頻繁，食安政策也逐步完善食安備受重視後，這幾年越來越多台灣商人從中國大陸引進各省美食來台灣擴展市場。

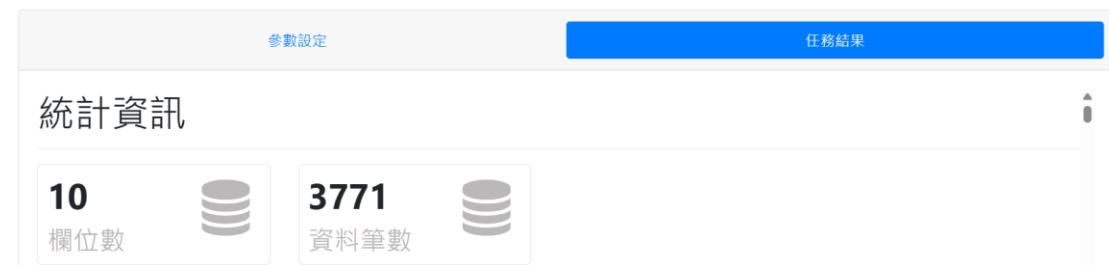
中國大陸美食在台灣市場回響頗受歡迎，我們本組一組員是廣西台商，在近年疫情期間將中國廣西的螺螄粉在台製作，成為第一支在台灣可公開販售的產品，加上本組周遭的親朋好友也都逐一推薦中國大陸進入台的美食，進而我們對於台灣民眾對大陸美食的情緒分析產生濃厚的興趣，想透過文字分析工具了解更多分析結果，於是本組決定以此題目開始著手進行。

五、資料範圍、來源、關鍵字

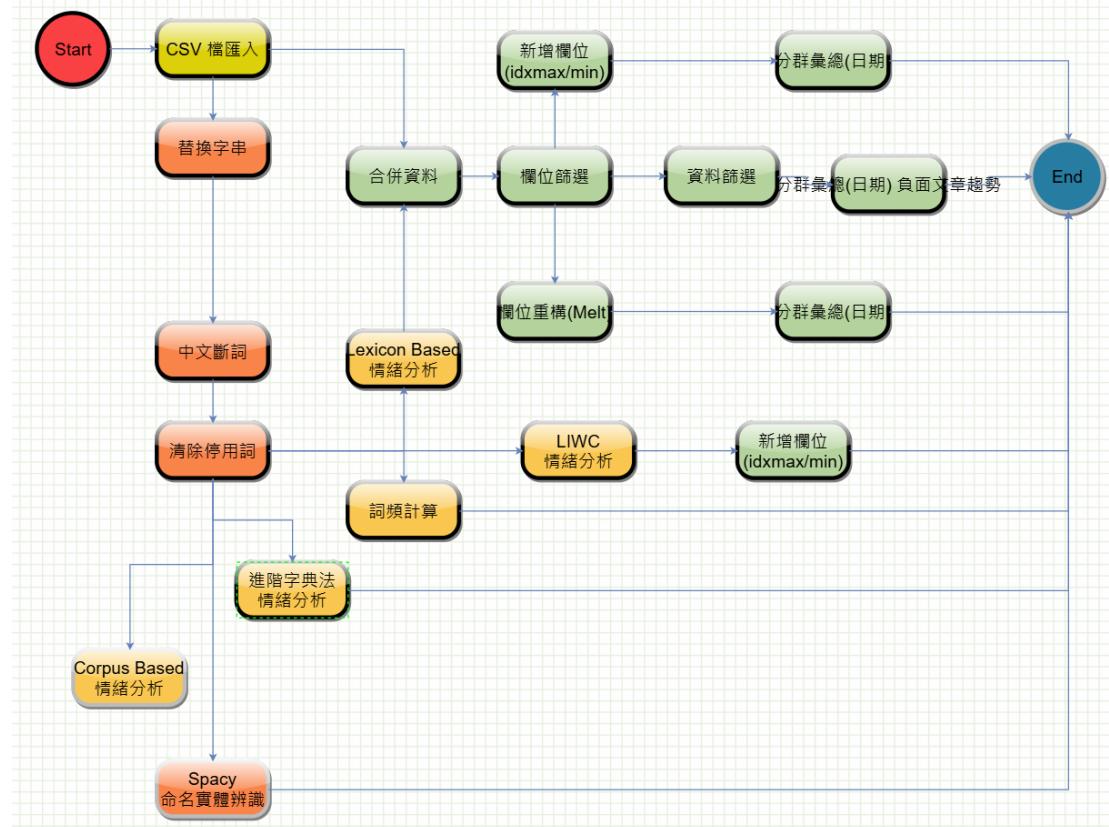
因 PTT 為純文字平台，因此原使用 PTT 抓取資料，但是發現 PTT 對於中國美食的討論文章較少，因此改為使用 Instagram 爬蟲。但由於 Instagram 爬蟲一次只能搜索 1000 筆，因此將 Instagram 文章分月份抓取後，合併為一筆，範圍為 2024/01~2024/12 月共一年。關鍵字使用“中華美食、爆紅美食、特色小吃、街邊小吃、川味、天津小吃、重慶小吃、上海小吃”。

六、爬蟲任務結果

CSV 檔匯入 (65)



七、系統流程



八、分析過程

8-1 替換字串

透過正規表達式將網址刪除 (`(https?:\/\/[^\\s]+)>>`)

替換字串 (7)

參數設定 Input - 65 任務結果

選擇處理欄位 * artContent

替換字串設定 `(https?:\/\/[^\\s]+)>>`

選擇替換規則檔案 -----請選擇-----

儲存更改

任務結果：取代了 60265 個字串

替換字串 (7)

參數設定 Input - 65 任務結果

統計資訊

60265 取代數量

8-2 中文斷詞

因搜尋目標為中國美食，因此尋找 30 個中國常見但台灣比較少見之中國美食加入權重 500

中文斷詞 (9)

參數設定 Input - 7 任務結果

選擇處理欄位 * result

定義詞彙 `北京烤鴨 500
酸菜魚 500
中國 500
水煮牛肉 500
螺蛳粉 500`

8-3 清除停用字

因後來詞頻計算時發現 IG PO 文很多無意義的 emoji(非正面也非負面)，因此後續又有回來

加入停用字清除，例如：

任務結果：清除 488544 個無用字元

■ 清除停用詞 (17)

統計資訊

8-4 詞頻計算

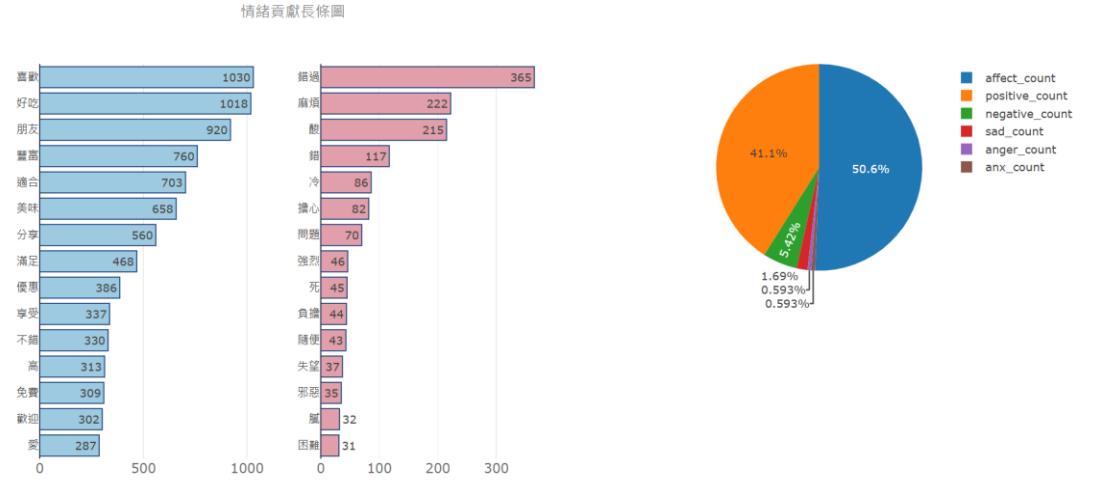
因結果有出現一、二等無意義字元，因此加入自定義停止詞



8-5-1 LIWC 情緒分析

分析後發現負面情緒中有一個“油膩”，但回去看原始來源幾乎都是“不油膩”，因此將“

不油膩”加入斷詞比重，讓系統重新斷詞



任務結果：最大正向情緒 32，最大負向情緒 14

統計資訊

32
最大正向情緒

0
最小正向情緒

14
最大負向情緒

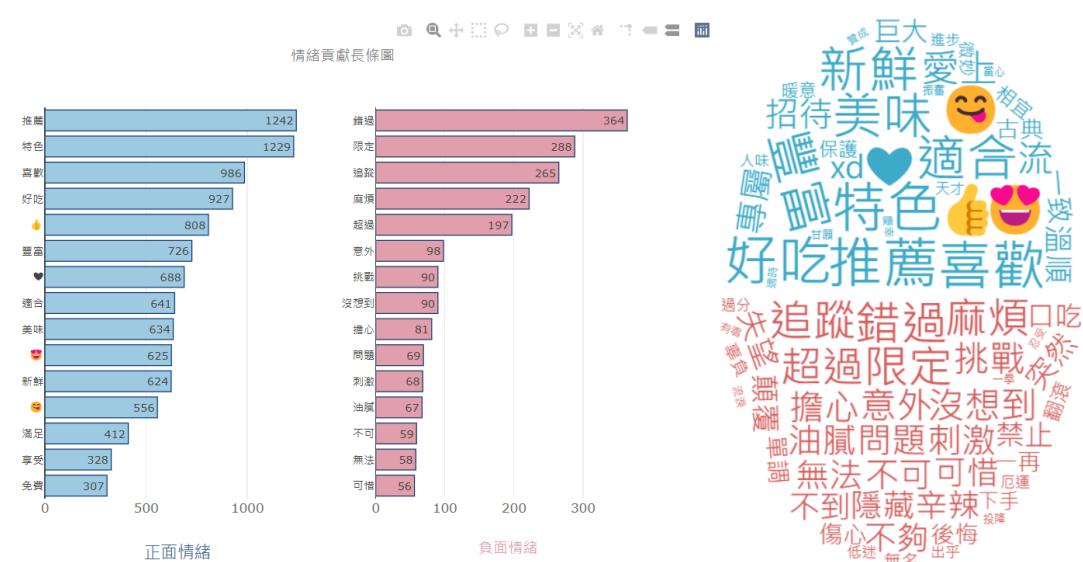
0
最小負向情緒

8-5-2 LIWC 後的新增欄位(idxmin/max)

使用彙總函數 max，計算 positive_count 和 negative_count 哪者較大，並將結果輸出為 result 欄位。

8-6-1 Lexicon Based 情緒分析

發現負面情緒最多的是錯過，然而美食文章中大多都是“不能錯過”之類的正面情緒，因此改採用可以移除情緒詞的進階字典法情緒分析



統計結果：最大正向情緒 41，最大負向情緒 30

統計資訊



8-6-2 進階字典法 情緒分析

使用 NTUSD 字典，根據結果新增正面詞彙”愛吃、xd、邪惡、罪惡、❤️、👍、🎉、😊、

上癮”，並將上述的“錯過”加入移除情緒詞。

進階字典法情緒分析 (84)

參數設定

選取字典：NTUSD

移除情緒詞：錯過

定義正面詞彙：

- 愛吃
- xd
- 邪惡
- 罪惡
- ❤️
- 👍
- 🎉
- 😊

是否使用否定詞：是

定義負面詞彙：

- 以旅仔特體語彙
- 錯過
- 失敗
- 彷彿

是否使用加強詞：是

任務結果：最大正向情緒 41，最大負向情緒 30

視覺化



8-7 合併資料

將情緒分析結果與 csv 檔使用 system_id 合併

JOIN規則

任務一欄位	任務二欄位
system_id	system_id
----- 請選擇 -----	----- 請選擇 -----

8-8 欄位篩選

只保留 system_id、artDate、artPoster、commentcount、likecount、positive_count、

negative_count、sentiment_value 等欄位

參數設定

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

system_id
dataSource
artUrl
artDate
artContent
artPoster

任務結果

任務結果

system_id	artDate	artPoster	commentcount	likecount	positive_count	negative_count	sentiment_value
1	2024-01-22 14:48:48	HKFoodie ❤ Eat Hard, Play Hard 🍕	38	341.0	18	1	17.00000
2	2024-01-22 09:09:29	寶寶！露米淇的電愛不釋懷人科班	79		3	4	-1.00000
3	2024-01-22 12:28:25	肉食 台中台北 新竹 雲林 美食旅遊分享	18		7	0	7.00000
4	2024-01-22 03:46:30	台北市兩岸處 - 路是裡的人	0		8	0	8.00000
5	2024-01-23 04:16:16	基層360	0	4.0	26	3	23.00000
6	2024-01-16 14:00:27	新新聞 Weekend Weekly	0	293.0	18	4	14.00000
7	2024-01-16 10:43:00	HK Foodie 🌟揭露國職賽好評	5	250.0	8	0	8.00000
8	2024-01-16 23:43:29	Elaine elaine elaine	8		10	0	10.00000
9	2024-01-16 02:47:09	吃貨阿嬤女子 🇹🇼 台北美食 新北美食	20	575.0	7	0	7.00000
10	2024-01-15 10:00:00	吃貨阿嬤女子 中華 桃園 新竹 台中美食旅遊	12	116.0	8	1	7.29300

Showing 1 to 10 of 100 entries

Previous 1 2 3 4 5 ... 10 Next

8-9-1 新增欄位(idxmax/min)後，根據月份分群彙總

先使用 max 函數去比較 positive_count 和 negative_count 兩者何者大，將結果放入 senti_class 欄位中

參數設定

匯總函數 * :

max

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

artDate
artPoster
commentcount
likecount
positive_count
negative_count

新增的欄位名稱 *

senti_class

計算結果

system_id	artDate	artPoster	commentcount	likecount	positive_count	negative_count	sentiment_value	senti_class
1	2024-01-22 14:48:48	HKFoodie ❤ Eat Hard, Play Hard 🍕	38	341.0	18	1	17.00000	positive_count
2	2024-01-22 09:09:29	寶寶！露米淇的電愛不釋懷人科班	79		3	4	-1.00000	negative_count
3	2024-01-22 12:28:25	肉食 台中台北 新竹 雲林 美食旅遊分享	18		7	0	7.00000	positive_count
4	2024-01-22 03:46:30	台北市兩岸處 - 路是裡的人	0		8	0	8.00000	positive_count
5	2024-01-23 04:16:16	基層360	0	4.0	26	3	23.00000	positive_count
6	2024-01-16 14:00:27	新新聞 Weekend Weekly	0	293.0	18	4	14.00000	positive_count
7	2024-01-16 10:43:00	HK Foodie 🌟揭露國職賽好評	5	250.0	8	0	8.00000	positive_count
8	2024-01-16 23:43:29	Elaine elaine elaine	8		10	0	10.00000	positive_count
9	2024-01-16 02:47:09	吃貨阿嬤女子 🇹🇼 台北美食 新北美食	20	575.0	7	0	7.00000	positive_count
10	2024-01-15 10:00:00	吃貨阿嬤女子 中華 桃園 新竹 台中美食旅遊	12	116.0	8	1	7.29300	positive_count

Showing 1 to 10 of 100 entries

Previous 1 2 3 4 5 ... 10 Next

再由 senti_class 計算每月分分群的 system_id 數目(使用 count 函數)

選擇日期範圍 *

選擇日期類型 *

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

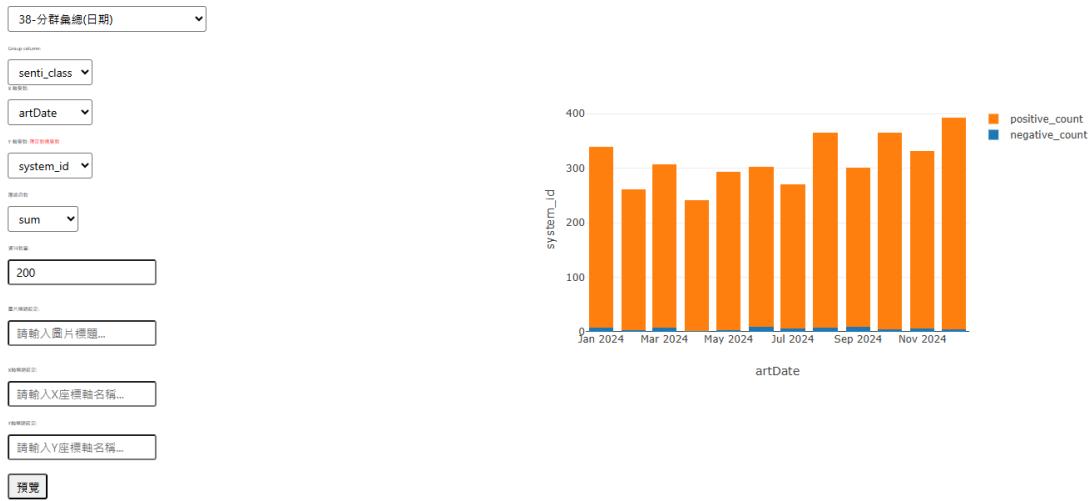
日期格式 *

匯總函數 *

保留欄位(按住ctrl(Windows)或command(MAC)可以複選) *

總共分 24 羣(2024/01~12 月份，正向負向分開計算共 24 羣)

以視覺化儀表板繪製，可以看出大陸美食有緩緩增加之趨勢，且正面聲量皆遠大於負面聲量



後來根據這個結果思考後，因為食記的一個特性，通常可能傾向會說好話避免招黑粉抨擊，或是被店主告，因此就算是覺得不好吃，也是會放一些好吃的因素來平衡，因此改成一個事不過三的精神，先以資料篩選計算 negative_count > 3 的數量，再以分群匯總方式繪製折線圖

8-9-2 資料篩選 negative_count > 3

條件式 *

negative_count > 3

篩除 3349 筆數據，留下 422 筆

統計資訊

3349
移除數量

422
保留數量

任務結果

system_id	artDate	artPoster	commentcount	likecount	positive_count	negative_count	sentiment_value
2	2024-01-22 09:09:29	英羅日暮米的兩♥不喜歡輸人對吧	79	3	5	-2.00000	
6	2024-01-16 14:00:27	新聞用 Weekend Weekly	0	293.0	18	4	14.00000
15	2024-01-15 04:44:05	大大的世界小小的新魚	583	663.0	12	7	5.86435
37	2024-01-18 10:00:31	飛鷺台灣 台東虎頭岩六福村	0	77.0	9	4	5.26370
69	2024-01-24 11:45:39	◎◎◎ 宜蘭W美食 🌸 台北 新竹 苗栗 金台萬	2	21	6	15.86435	
85	2024-01-25 11:05:14	阿華劉威甲 11月中 東北 頭痛	7	7	5	-2.29300	
116	2024-01-29 12:13:47	雙魚逆流魚露豆漿	21	331.0	35	6	29.29300
123	2024-01-04 05:22:54	林鈞一煎什麼肉 台北美食 蘿蔴及蕃薯圓圓餅	9	114.0	6	4	2.00000
126	2024-01-01 05:00:20	Giphy動圖提供空空的	1	875.0	22	5	17.57135
127	2024-01-10 13:19:55	火影忍者 台灣藝能 美食 ◉ 大家都三三併排坐	38	13	6	7.00000	

Showing 1 to 10 of 100 entries

Previous 1 2 3 4 5 ... 10 Next

再以月分 分群匯總

52-分群彙總(日期) 負面文章趨勢 (52)

參數設定

Input - 49

任務記錄

選擇日期欄位 *

選擇日期類型 *

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

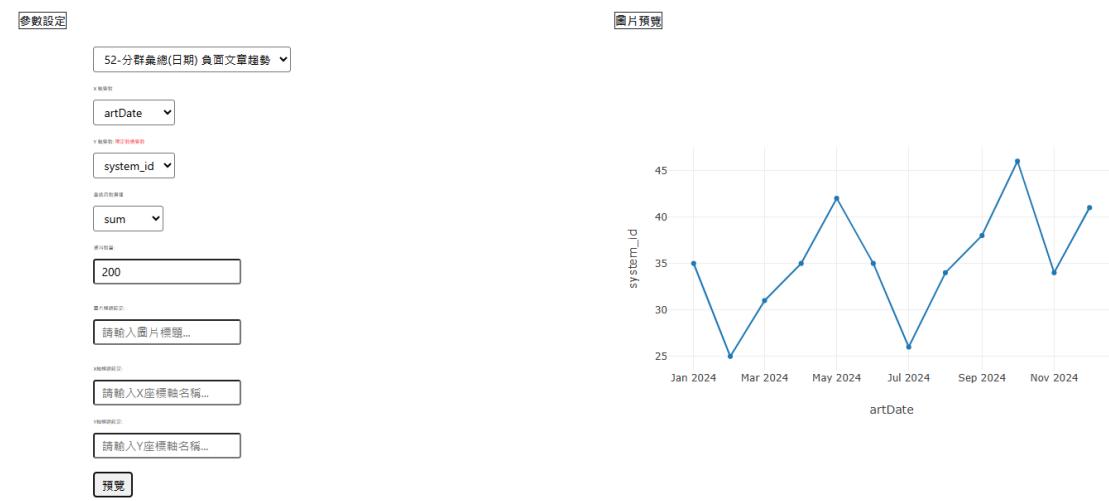
選擇欄位(按住ctrl(Windows)或command(MAC)可以複選) *

日期格式 *

匯總函數 *

保留欄位(按住ctrl(Windows)或command(MAC)可以複選) *

最後使用視覺化，繪製折線圖



可以看出負面的聲量，就較上方純粹只有正負比大小的文章還要多一點了。