美食公道伯

Michael Chen

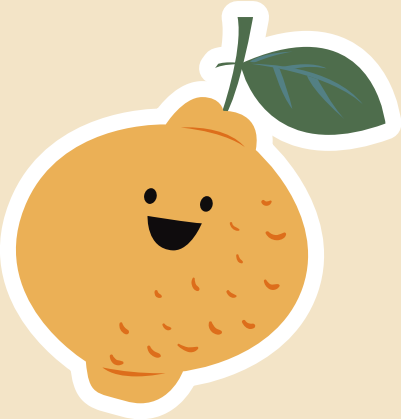


Table of contents

01 發想 &
資料探索

02 定義 Graph

03 資料處理

04 Algorithm

05 落地實現

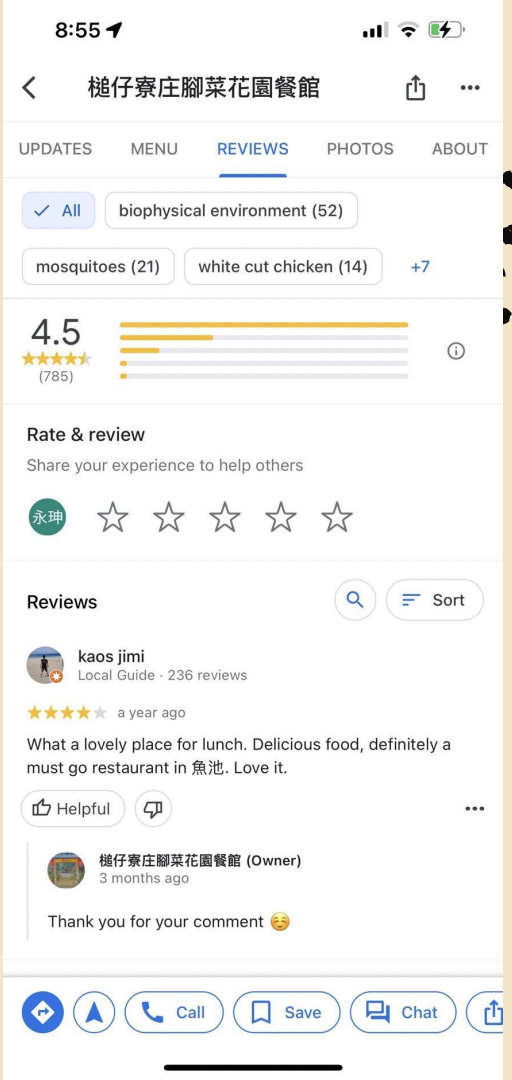
06 心得分享

01

發想＆資料探索

情境 & 提問

大家是否曾經與家人、朋友出遊，使用Google Map搜尋餐廳？除了參考其他網頁，在 Google Map 中哪些資訊會列入你的考量？



參考資訊

以下是我們團隊會參考的資訊:

星度、餐廳被留言數量、TopN留言、近期留言、留言照片、該篇留言讚數、留言者有沒有頭貼、留言內容、留言者聲量(留言次數、留言被按讚數)。

總結來說我們除了看他的評分高低以外也會關注這個評分是否真實。

真實度

我們參考的資訊大多與用戶留言相關(用戶與餐廳的互動), 那是否可以透過社群網絡來找出餐廳評分的真實度？

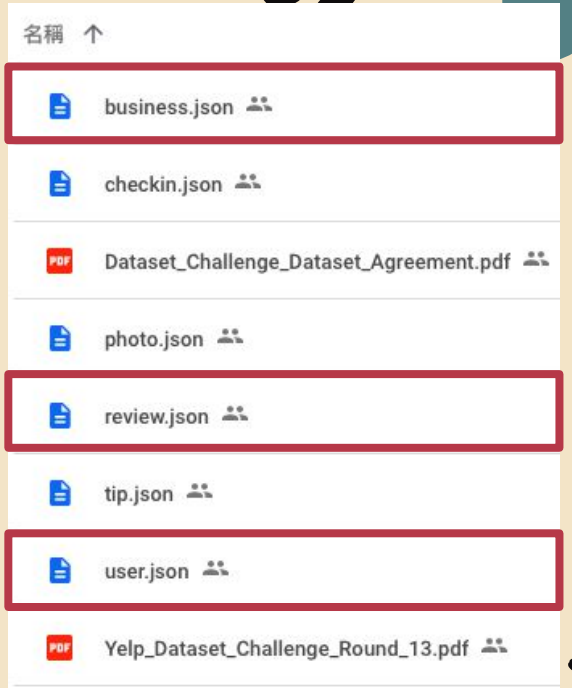
餐廳名稱	Stars	**Truth**
ABC bar	3	**1(?)**
CDE bar	5	**1.5(?)**
FGH bar	5	**6(?)**

資料搜索

我們嘗試在 yelp 資料集中找到以下資料:

餐廳被留言數量、TopN留言、近期留言、留言的照片、該篇留言讚數、留言者有沒有頭貼、留言內容、留言者聲量(留言次數、留言被按讚數)。

使用了review、business、user資料集



誰反映了真實度？

餐廳被留言數量	
Top N留言	
近期留言	
該篇留言讚數	
留言者質量(留言次數、留言被按讚數)	

資料篩選

基於Top N留言、近期留言、留言者質量,
我們篩選了什麼資料？

1. Las Vegas 當地屬於 restaurants & bars 、有營業的店家(運算考量)
2. 2018 年有 elite 資格的用戶
3. 20180101 以後的評論

篩選結果

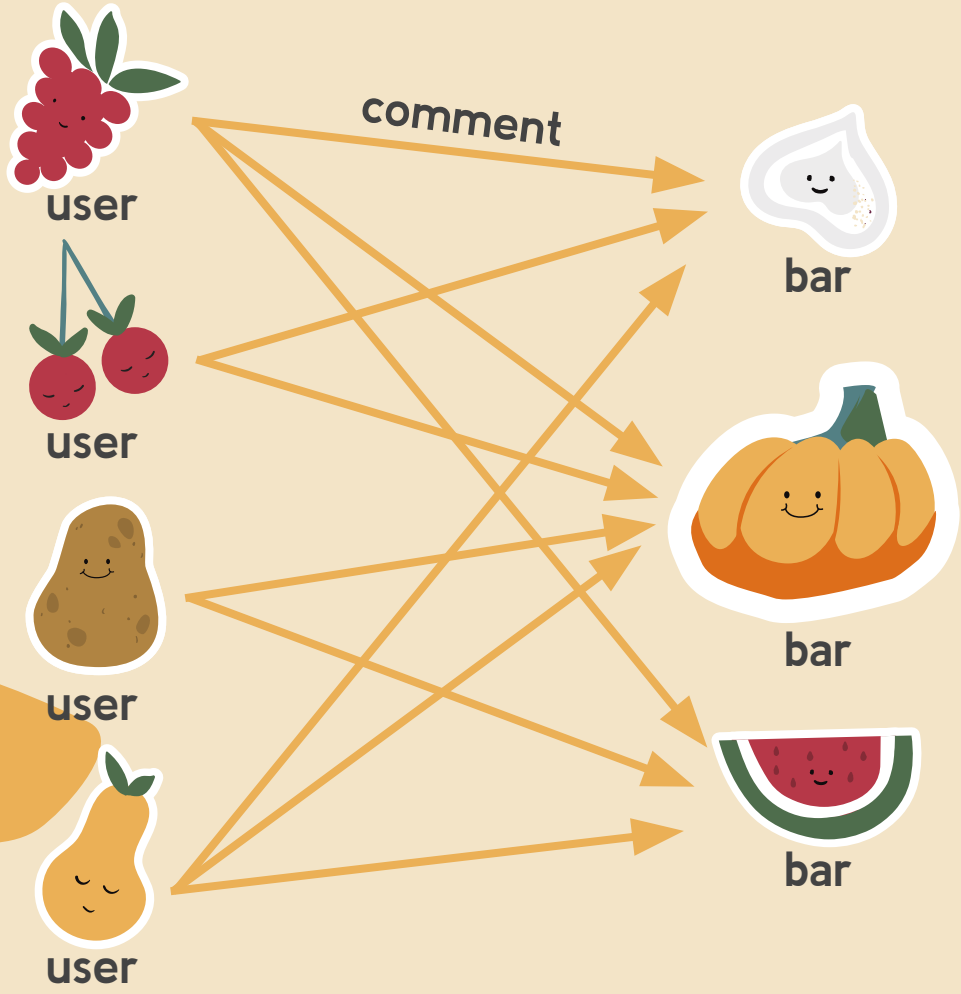
	review_id	user_id	business_id	stars	useful	funny	cool	text	date
0	C0onMemq7n2bMx3fMQXF9w	1RQaXb0xSLsUzwJ0u4uZ-w	gOOfBSBZlffCkQ7dr7cpdw	4	0	0	0	Went in on a Friday since I was strolling arou...	2018/4/27 18:53
1	V_Wd_Ybjvr47o3utbEa9dg	qXcTj0AyR6BF5wQpzRSYaQ	H5Y6o9H4rtxbHJFgLqI6Fw	3	0	0	0	Came to have dinner and they were pretty busy....	2018/2/11 3:30
2	g2s2D_gDIdt1SHA9_03DHQ	qEjm0_ivRn8WIfJipeFAgg	XsSgv3vBOyOBXn3Co8EVIg	5	1	1	1	I love Q Karaoke so I decided to eat at QBistr...	2018/5/28 21:25
3	T0HsOKAAuHEzHXq00baJUQ	kdDTqKBbfNZeKEbWAcqZWQ	pH0BLkL4cbxKzu471VZnuA	4	0	0	0	Great service at the bar for happy hour and in...	2018/8/14 4:17
4	H9iVzGZklifXAHBSanqWOw	lPlLcdNG426qOopHH4EwkQ	isw3cS3hOKdKeBgi3lF3-A	4	0	0	0	I went on national burger day ha ha so I had a...	2018/5/29 2:12

誰反映了真實度？

餐廳被留言數量	
Top N留言	✓
近期留言	✓
該篇留言讚數	
留言者質量(留言次數、留言被按讚數)	✗

02

定義 Graph



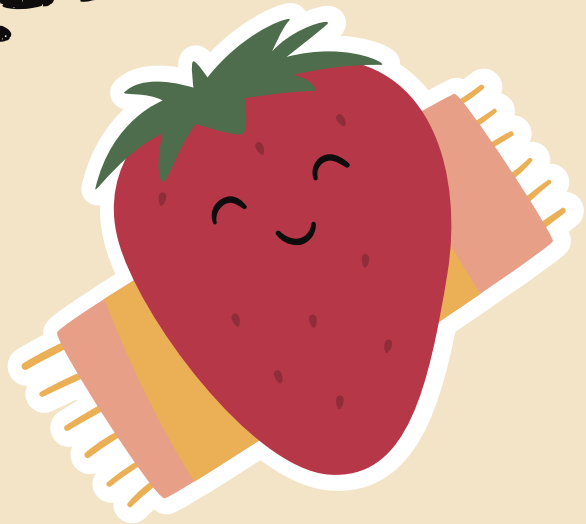
節點：菁英用戶、餐廳＆酒吧

關係：用戶「評論」餐聽

權重：useful、funny、cool 相加

誰反映了真實度？

餐廳被留言數量	
Top N留言	✓
近期留言	✓
該篇留言讚數	✓
留言者質量(留言次數、留言被按讚數)	✗



03

資料處理

use python pandas

1. 以這張表當主鍵, 去合併user及business資料集

	review_id	user_id	business_id	stars	useful	funny	cool	text	date
0	C0onMemq7n2bMx3fMQXF9w	1RQaXb0xSLsUzwJ0u4uZ-w	gOOfBSBZlffCkQ7dr7cpdw	4	0	0	0	Went in on a Friday since I was strolling arou...	2018/4/27 18:53
1	V_Wd_Ybjvr47o3utbEa9dg	qXcTj0AyR6BF5wQpzRSYaQ	H5Y6o9H4rtxbHJFgLqI6Fw	3	0	0	0	Came to have dinner and they were pretty busy....	2018/2/11 3:30
2	g2s2D_gDIdt1SHA9_03DHQ	qEjm0_ivRn8WlfJipeFAgg	XsSgv3vBOyOBXn3Co8EVIg	5	1	1	1	I love Q Karaoke so I decided to eat at QBistr...	2018/5/28 21:25
3	T0HsOKAAuHEzHXq00baJUQ	kdDTqKBbfNZeKEbWAcqZWQ	pH0BLkL4cbxKzu471VZnuA	4	0	0	0	Great service at the bar for happy hour and in...	2018/8/14 4:17
4	H9iVzGZklifXAHBSanqWOw	IPlLcdNG426qOopHH4EwkQ	isw3cS3hOKdKeBgi3lF3-A	4	0	0	0	I went on national burger day ha ha so I had a...	2018/5/29 2:12

2. 將重複評論的列合併。

3. 篩選要使用的欄位：business_name、 user_name、stars、useful、funny、cool

4. 將 useful、funny、cool相加當作權重(weight)

結果

	business_name	user_name	stars	weight
0	Chica	Melody	4	0
1	Chili's	Mayito	3	0
2	Q Bistro	Crystal	5	3
3	SUSHISAMBA - Las Vegas	Lulu	4	0
4	Eureka!	Theresa	4	0
7	SUSHISAMBA - Las Vegas	Juliet	4	0
8	Sammy's Beach Bar & Grill	Rodrigo	3	2
9	Hawthorn Grill	George	5	3
10	Battista's Hole In the Wall	Jhordan	3	7

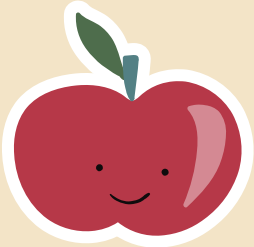
餐廳數量：663

評論數：9997

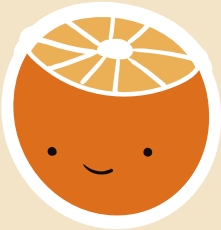
用戶數：1298

04 Algorithm

Algorithm



PageRank



Birank

PageRank – define method

「PageRank」是 Google 用於「評等網頁重要性」的一種方式。

	Google	Our case
數量假設	一個網頁收到的其他網頁指向的 **入鏈結(in-link)越多**，說明該網頁越重要。	一間餐廳**被評論的數量越多**，說明該家餐廳評分**越趨近於真實**。
質量假設	當一個**質量高的網頁指向(out-link)的網頁也質量也高。**	**評論餐廳的人質量越高**，說明這家餐廳的**評分質量也越高**。

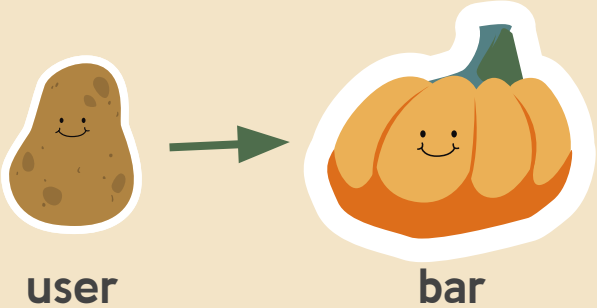
誰反映了真實度？

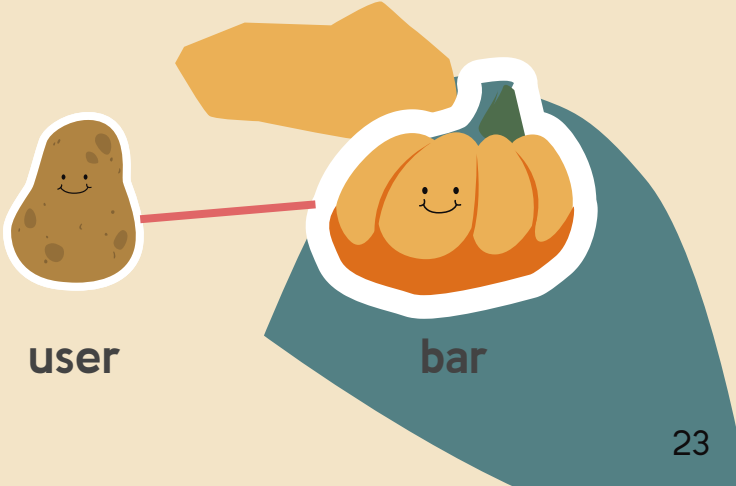
餐廳被留言數量	✓
Top N留言	✓
近期留言	✓
該篇留言讚數	✓
留言者質量(留言次數、留言被按讚數)	✗

PageRank – define method

➔ 調整成無向圖

註:計算完 PageRank 分數後再將用戶節點的分數移除。



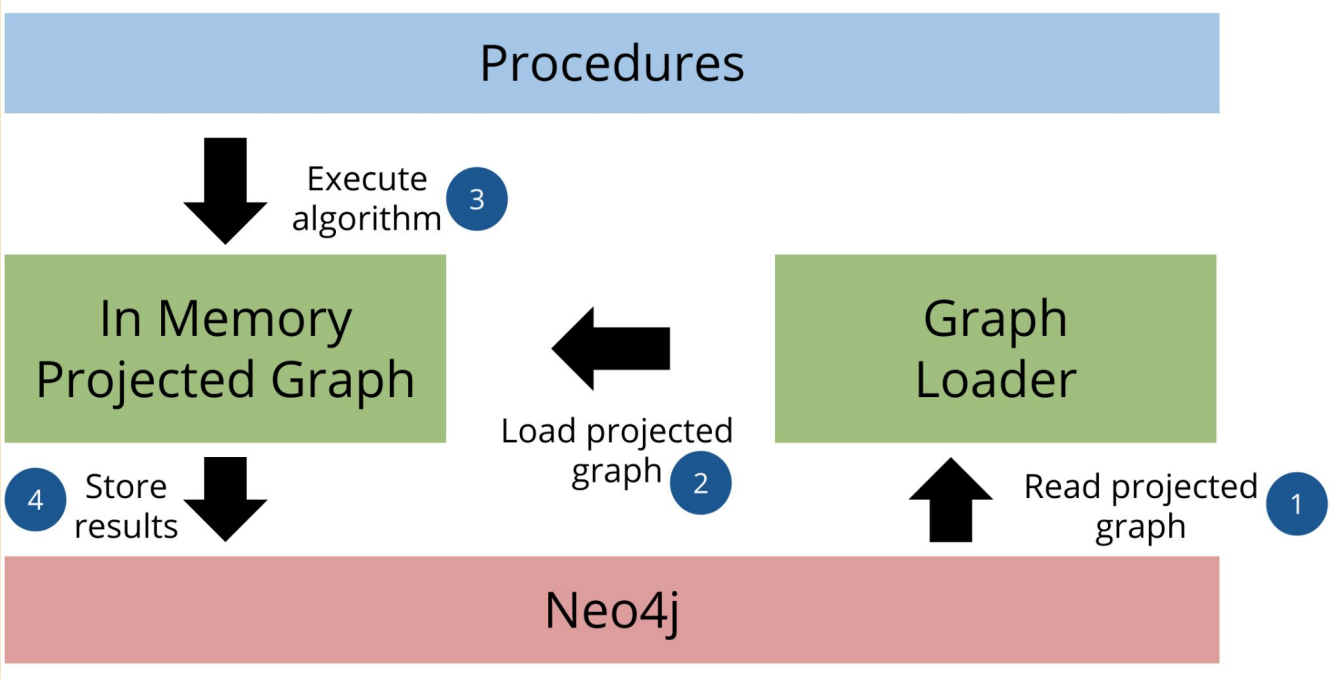


PageRank – graph

user(3) business(38)

PageRank – gds info

Graph Data Science library



PageRank – code_add_graph

```
CALL gds.graph.create(
  'myGraph',
  ['user', 'business'],
  {
    Related: {
      orientation: 'UNDIRECTED',
      properties: 'weight'
    }
  }
)
```

PageRank – code_PageRank

```
CALL gds.pageRank.stream('myGraph') YIELD nodeId, score
WITH gds.util.asNode(nodeId) AS n, score
MATCH (n)-[r:Related]-()
RETURN n.Name AS name, score, count(r) AS interactions
ORDER BY score DESC
```

Table | Text | Code

"name"	"score"	"interactions"
"Momofuku Las Vegas"	25.641347145149098	204
"HEXX kitchen + bar"	16.85194591643035	136
"Lotus of Siam"	15.666573188885478	132
"Sapporo Revolving Sushi"	12.962212150692016	124
"Gordon Ramsay Pub & Grill"	12.777973644927615	102
"Holsteins"	12.272574440683506	103
"SkinnyFATS"	11.958860096266198	108
"Black Tap"	11.931775801865307	105
"Yard House"	11.036061905182281	98
"Beauty & Essex"	10.984403766069727	89
"The Peppermill Restaurant & Fireside Lounge"	10.752671837482556	90

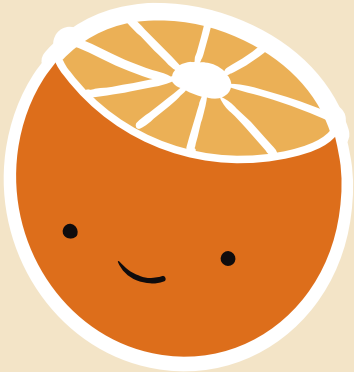
PageRank – remove_user

	name	score
0	Momofuku Las Vegas	25.641347
1	HEXX kitchen + bar	16.851946
2	Lotus of Siam	15.666573
3	Sapporo Revolving Sushi	12.962212
4	Gordon Ramsay Pub & Grill	12.777974
...	...	...
659	Holo Holo	0.237453
660	Agua El Manantial	0.237397

Birank – method intro

Bipartite networks are commonly reduced to unipartite networks for further analysis, such as calculating node centrality (e.g. PageRank,see Figure 1(c)).

To overcome the issues of one-mode projection, we present BiRank, an R and Python package that performs PageRank on bipartite networks directly.



Birank – package

BrianAronson / birankr Public

Watch 2 | Fork 6 | Star 22

Code | Issues | Pull requests | Actions | Projects 2 | Wiki | Security | Insights

master | birankr / birankpy / README.md | Go to file | ···

yy minor typo fixes ✓ | Latest commit b2d428f on 7 Feb 2020 | History

2 contributors

186 lines (135 sloc) | 5.06 KB | Raw | Blame

API documentation of birankpy

PageRank

```
pagerank(adj, d=0.85, max_iter=200, tol=0.0001, verbose=False)
```

Birank – data

	business_name	user_name	stars	weight
0	Chica	Melody	4	0
1	Chili's	Mayito	3	0
2	Q Bistro	Crystal	5	3
3	SUSHISAMBA - Las Vegas	Lulu	4	0
4	Eureka!	Theresa	4	0
7	SUSHISAMBA - Las Vegas	Juliet	4	0
8	Sammy's Beach Bar & Grill	Rodrigo	3	2
9	Hawthorn Grill	George	5	3
10	Battista's Hole In the Wall	Jhordan	3	7

Birank – code

To performing BiRank on this bipartite network, just:

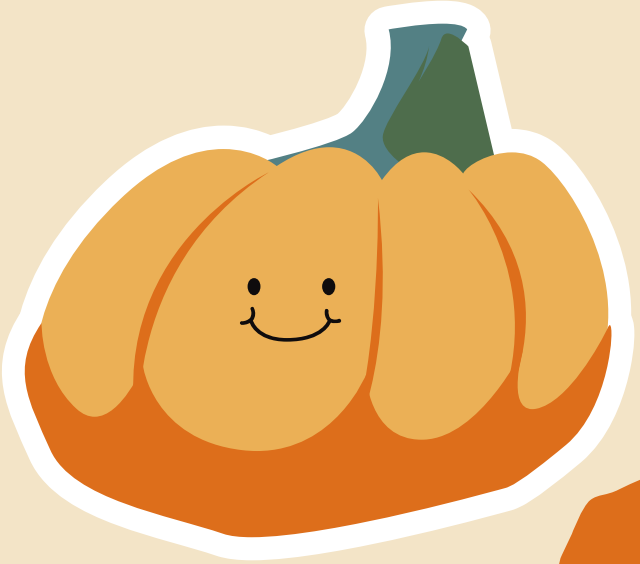
```
import birankpy

bn = birankpy.BipartiteNetwork()

bn.set_edgelist(df, top_col='business_name', bottom_col='user_name',

weight_col=weight)

top_birank_df, bottom_birank_df = bn.generate_birank()
```

Birank – output

	business_name	business_name_birank
0	Holsteins	3.352381e-01
1	Malena's Yogurt Plus	3.101665e-01
2	Blue Martini Lounge	3.219785e-02
3	Rí Rá Irish Pub	2.012090e-02
4	Lazy Dog Restaurant & Bar	1.937041e-02
...	...	...
658	Al-Hayat Hookah Lounge	2.869219e-07
659	Agua El Manantial	2.869219e-07
660	Margaritaville Casino	2.869219e-07
661	Oliva Minimart & Cafeteria	2.869219e-07
662	OTORO Robata Grill & Sushi	2.869219e-07

05

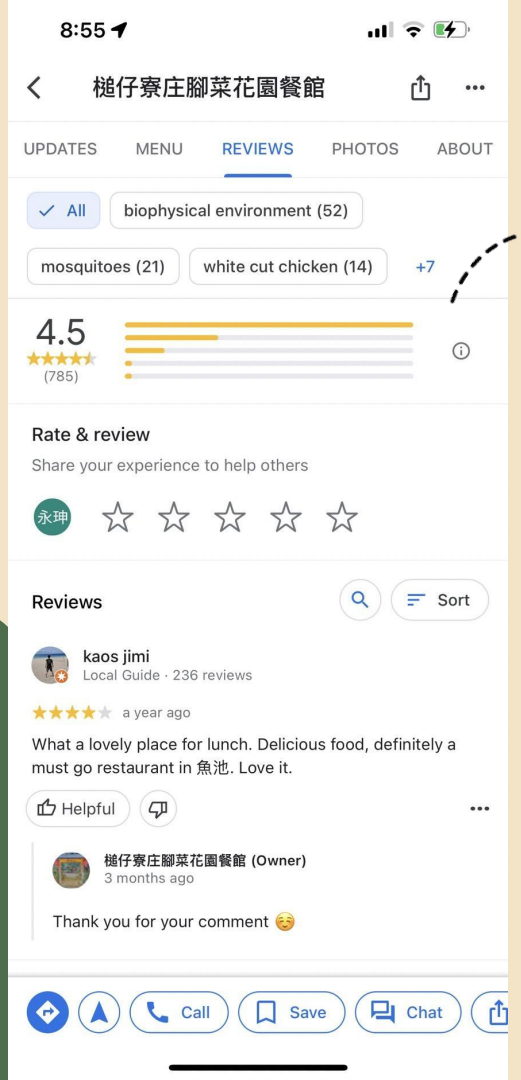
落地實現

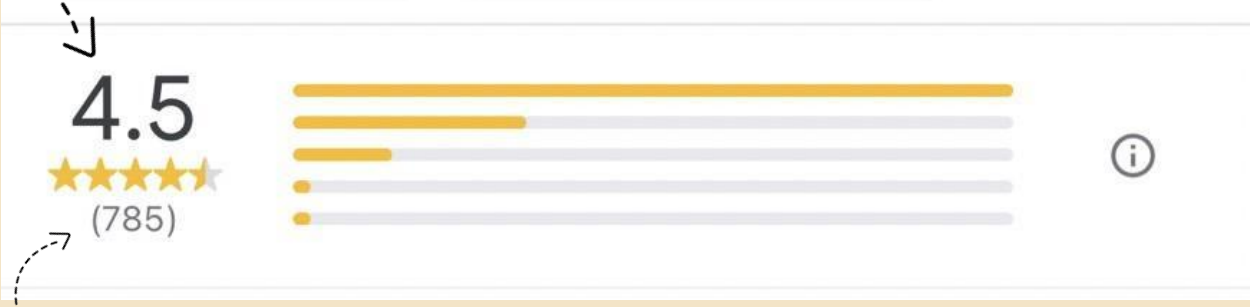


➔ 將透過 PageRank 計算得到的排序轉換為 PR 值

	business_name	business_name_birank	rank	pr
0	Holsteins	3.352381e-01	1	99
1	Malena's Yogurt Plus	3.101665e-01	2	99
2	Blue Martini Lounge	3.219785e-02	3	99
3	Rí Rá Irish Pub	2.012090e-02	4	99
4	Lazy Dog Restaurant & Bar	1.937041e-02	5	99
...	...	...	...	...
658	Al-Hayat Hookah Lounge	2.869219e-07	659	0
659	Agua El Manantial	2.869219e-07	660	0

```python
output['rank'] = output.index + 1

N = len(output)
def pr(s):
    pr_val = int((N - s)*100 / N)
    return pr_val

output['pr'] = output['rank'].apply(pr)
```





PR 98

➔ 相較於 Google Map 本身提出的評分數量更能反映客觀的真實度

PR 分數的疑慮

惡性競爭

無法爭取絕對分數的高分
-> 攻擊比自己高分的店家

使用者判讀

找到兩家 PR 不高的餐廳
-> 感覺 star 都是假的

商家流失

平台上一定有 PR 低的店家
-> 不願待於己不利的平台