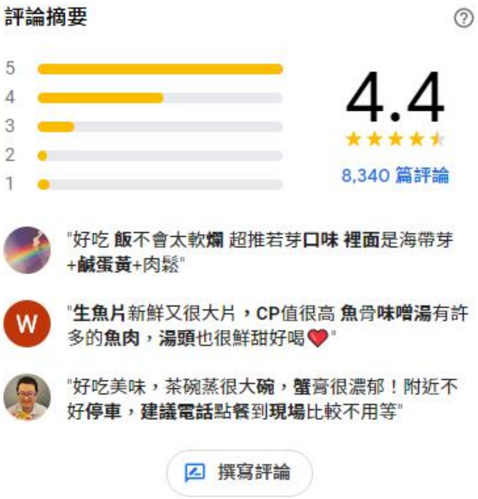
Undergraduate AI Capstone NYCU
Spr2022 Final Group Project

0816175 高靖 0816006 郭冠德 0816081 張以廉

Our project is about NLP. We try to collect reviews on google map & try to do some prediction of the reviews rating.



評論摘要

5
4
3
2
1

4.4

8,340 篇評論

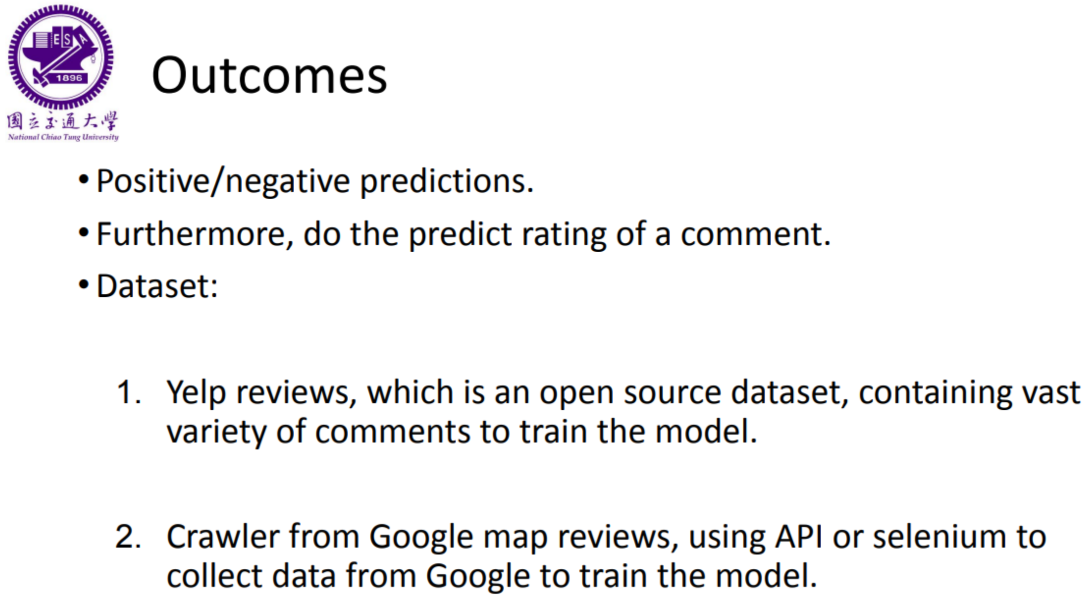
"好吃 飯不會太軟爛 超推若芽口味 裡面是海帶芽+鹹蛋黃+肉鬆"

"生魚片新鮮又很大片，CP值很高 魚骨味噌湯有許多的魚肉，湯頭也很鮮甜好喝❤"

"好吃美味，茶碗蒸很大碗，蟹膏很濃郁！附近不好停車，建議電話點餐到現場比較不用等"

撰寫評論

Our original slides of our short present:



Outcomes

- Positive/negative predictions.
- Furthermore, do the predict rating of a comment.
- Dataset:

1. Yelp reviews, which is an open source dataset, containing vast variety of comments to train the model.
2. Crawler from Google map reviews, using API or selenium to collect data from Google to train the model.

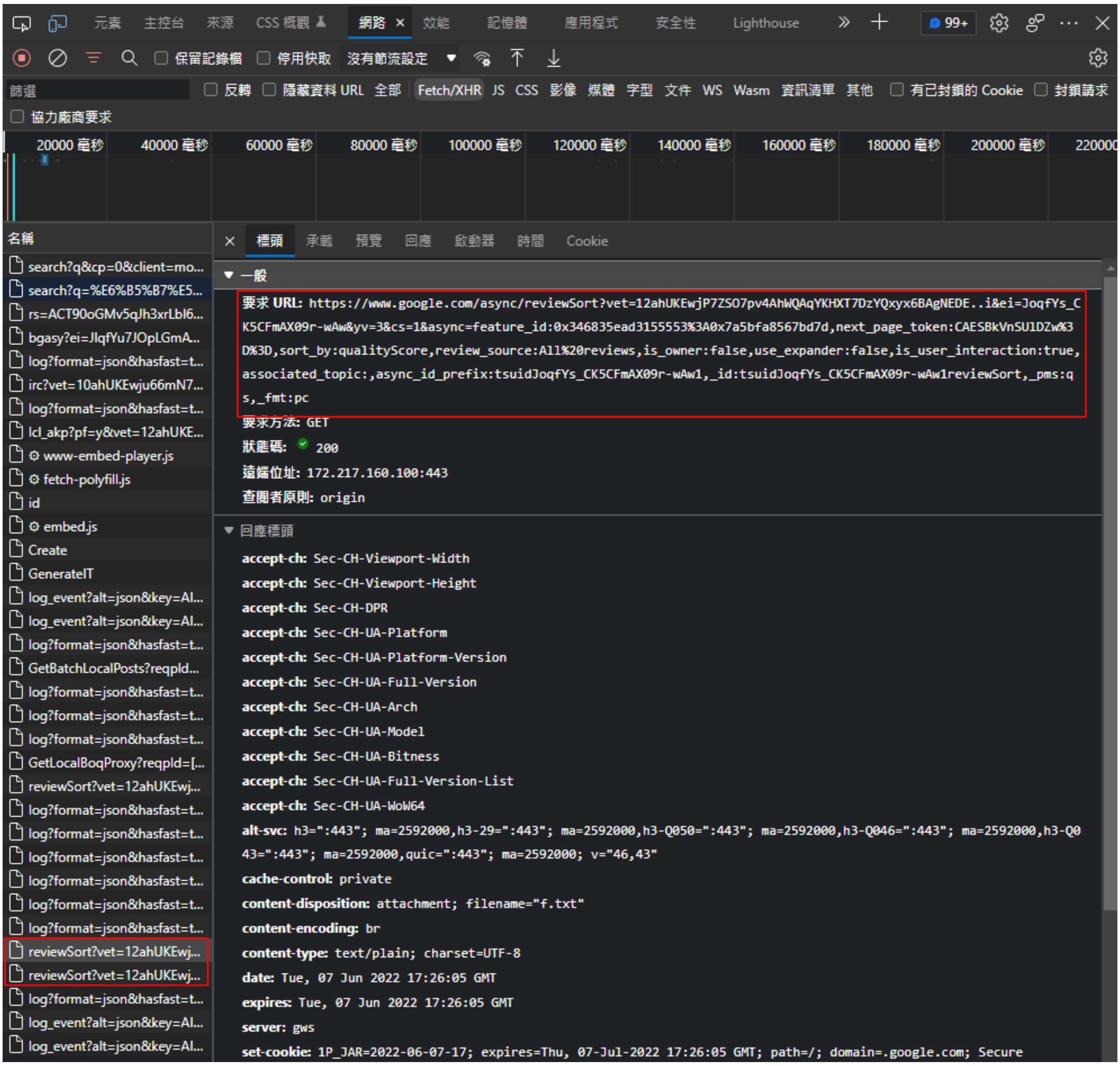
Data collection:

So, originally, we try to use some online open-source dataset. But in the end, we try to make it our own way. So instead to use some existing data on the Internet, we decide to do crawler on our own.

We were trying to use Yelp dataset(Yelp Dataset) which is a famous Open dataset which includes personal ,educational, and some academic purpose and reviews & user data dataset. It has JSON & SQL form. However, we think do the Chinese part is more interesting. So we collect the google reviews of some famous store in Taiwan. So, how wee collect our data?

By developer tool, we can check the console of Internet/Fetch/XHR/Headers, to see how the google map backend call the backend api. So as we scroll down the reviews, we can check there's an request URL which is called repeatedly.

So we can cut in this point, just to get the reviews by calling the same URL.(next page)



名稱

search?q&cp=0&client=mo...
search?q=%E6%B5%B7%E5...
rs=ACT90oGMv5qJh3xrLbl6...
bgasy?ei=JlqfYu7JOpLGmA...
log?format=json&hasfast=t...
irc?vet=10ahUKEwju66mN7...
log?format=json&hasfast=t...
lcl_akp?pf=y&vet=12ahUKE...
⚙ www-embed-player.js
⚙ fetch-polyfill.js
id
⚙ embed.js
Create
GenerateIT
log_event?alt=json&key=AI...
log_event?alt=json&key=AI...
log?format=json&hasfast=t...
GetBatchLocalPosts?reqpId...
log?format=json&hasfast=t...
log?format=json&hasfast=t...
log?format=json&hasfast=t...
GetLocalBoqProxy?reqpId=[...
reviewSort?vet=12ahUKEwj...
log?format=json&hasfast=t...
log?format=json&hasfast=t...
log?format=json&hasfast=t...
log?format=json&hasfast=t...
log?format=json&hasfast=t...
log?format=json&hasfast=t...
reviewSort?vet=12ahUKEwj...
reviewSort?vet=12ahUKEwj...
log?format=json&hasfast=t...
log_event?alt=json&key=AI...
log_event?alt=json&key=AI...

▼ 一般

```
要求 URL: https://www.google.com/async/reviewSort?vet=12ahUKEwjP7ZSO7pv4AhWQAqYKHXT7DzYQxyx6BAgNEDE..i&ei=JoqfYs_CK5CFmAX09r-wAw&yv=3&cs=1&async=feature_id:0x346835ead3155553%3A0x7a5bfa8567bd7d,next_page_token:CAESBkVnSU1DZw%3D%3D,sort_by:qualityScore,review_source:All%20reviews,is_owner:false,use_expander:false,is_user_interaction:true,associated_topic:,async_id_prefix:tsuidJoqfYs_CK5CFmAX09r-wAw1,_id:tsuidJoqfYs_CK5CFmAX09r-wAw1reviewSort,_pms:qs,_fmt:pc
要求方法: GET
狀態碼: ● 200
遠端位址: 172.217.160.100:443
查閱者原則: origin
```

▼ 回應標頭

```
accept-ch: Sec-CH-Viewport-Width
accept-ch: Sec-CH-Viewport-Height
accept-ch: Sec-CH-DPR
accept-ch: Sec-CH-UA-Platform
accept-ch: Sec-CH-UA-Platform-Version
accept-ch: Sec-CH-UA-Full-Version
accept-ch: Sec-CH-UA-Arch
accept-ch: Sec-CH-UA-Model
accept-ch: Sec-CH-UA-Bitness
accept-ch: Sec-CH-UA-Full-Version-List
accept-ch: Sec-CH-UA-WoW64
alt-svc: h3=":443"; ma=2592000,h3-29=":443"; ma=2592000,h3-Q050=":443"; ma=2592000,h3-Q046=":443"; ma=2592000,h3-Q043=":443"; ma=2592000,quic=":443"; ma=2592000; v="46,43"
cache-control: private
content-disposition: attachment; filename="f.txt"
content-encoding: br
content-type: text/plain; charset=UTF-8
date: Tue, 07 Jun 2022 17:26:05 GMT
expires: Tue, 07 Jun 2022 17:26:05 GMT
server: gws
set-cookie: 1P_JAR=2022-06-07-17; expires=Thu, 07-Jul-2022 17:26:05 GMT; path=/; domain=.google.com; Secure
```

We can compare the 2 URL of the request, to do the loop crawler. However, I found that

URL1:https://www.google.com/async/reviewSort?vet=12ahUKEwjP7ZSO7pv4AhWQAqYKHXT7DzYQxyx6BAgNEDE..i&ei=JoqfYs_CK5CFmAX09r-wAw&yv=3&cs=1&async=feature_id:0x346835ead3155553%3A0x7a5bfa8567bd7d,next_page_token:CAESBkVnSUlDZw%3D%3D,sort_by:qualityScore,review_source:All%20reviews,is_owner:false,use_expander:false,is_user_interaction:true,associated_topic:,async_id_prefix:tsuidJoqfYs_CK5CFmAX09r-wAw1,_id:tsuidJoqfYs_CK5CFmAX09r-wAw1reviewSort,_pms:qs,_fmt:pc

URL2:https://www.google.com/async/reviewSort?vet=12ahUKEwjP7ZSO7pv4AhWQAqYKHXT7DzYQxyx6BAgNEDE..i&ei=JoqfYs_CK5CFmAX09r-wAw&yv=3&cs=1&async=feature_id:0x346835ead3155553%3A0x7a5bfa8567bd7d,next_page_token:CAESBkVnSUlGQQ%3D%3D,sort_by:qualityScore,review_source:All%20reviews,is_owner:false,use_expander:false,is_user_interaction:true,associated_topic:,async_id_prefix:tsuidJoqfYs_CK5CFmAX09r-wAw1,_id:tsuidJoqfYs_CK5CFmAX09r-wAw1reviewSort,_pms:qs,_fmt:pc

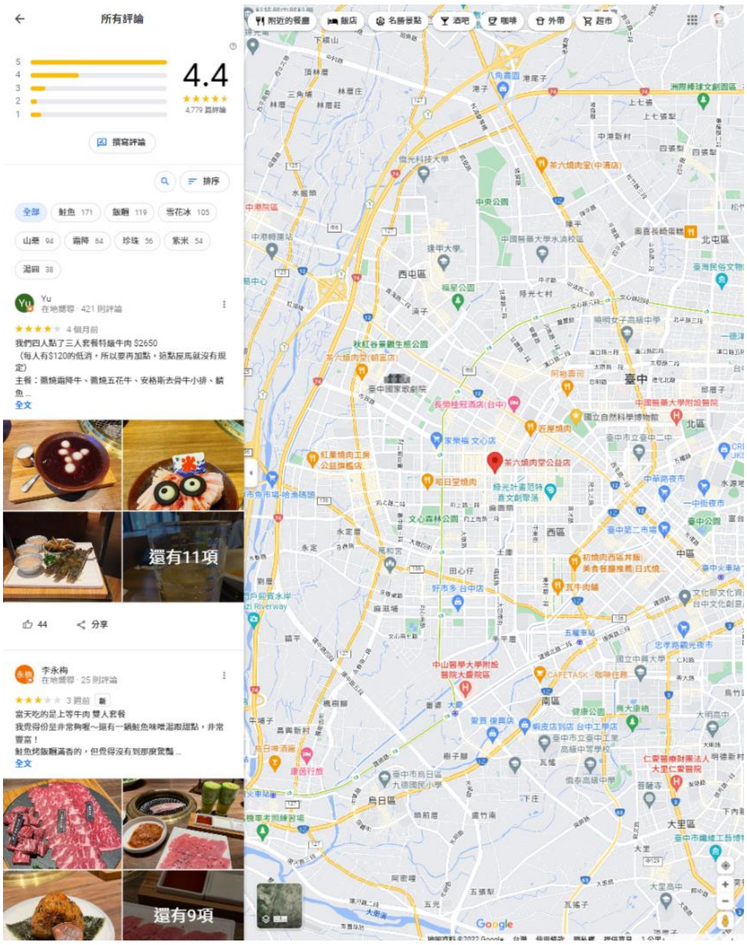
The crucial part is about the param: **next_page_token.**

Rest part are the same. So here comes the question, how can we decide find the pattern of the URL? Because the token seems to be random. After some research, I found that it do have regular pattern is that, **each store start with DZw, GQQ,IZw,LQQ**……. So the token seems no rules, but can we get the next page token before do the crawler? The answer is negative, we still need to find the token manually and, then, apply to every store.

So, it may be not a ‘smart’ way to collect way to collect data but a ‘not bad’ way, if you do have time to do so. For us, we want to find more ‘intelligent’ way that not to do any manual work. So here’s work around we find:

Instead of using the reviews on web page of the google store

We change to using the map application interface, because the api called here is different from web interface.



所有評論

5
4
3
2
1

4.4
4,779 篇評論

撰寫評論

排序

全部 鮭魚 171 飯糰 119 雪花冰 105 山藥 94 霜降 64 珍珠 56 紫米 54 湯圓 38

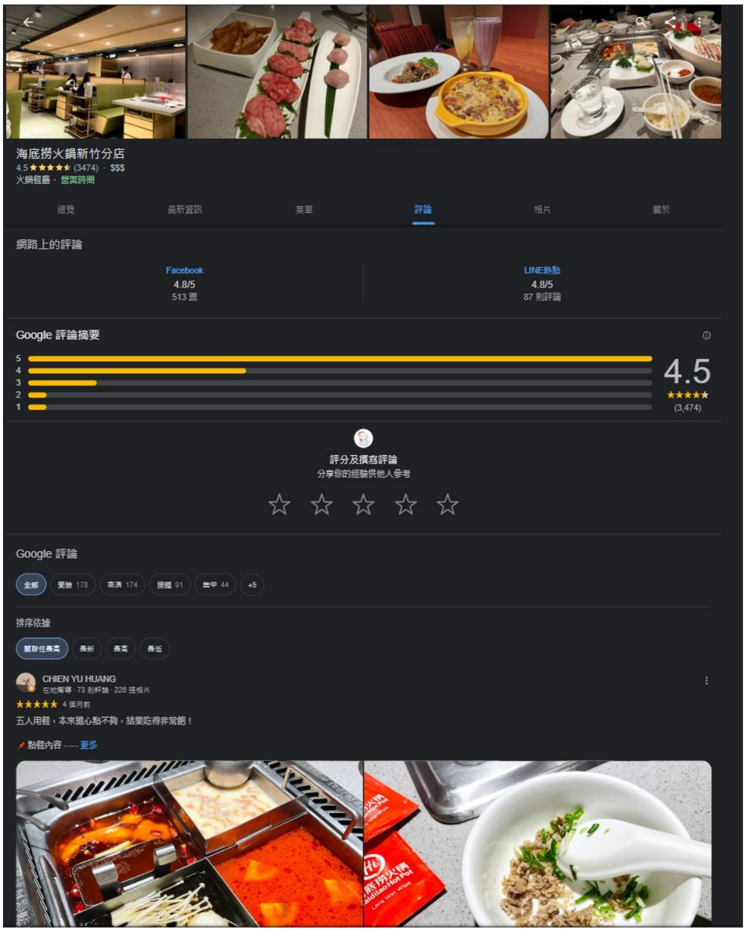
Yu
在地嚮導 · 421 則評論
★★★★ 4 個月前
我們四人點了三人套餐特級牛肉 $2650
（每人有$120的低消，所以要再加點，菇點屋馬就沒有規定）
主餐：醬燒霜降牛、醬燒五花牛、安格斯去骨牛小排、鮪魚 ...
全文

還有11項

44 分享

李永梅
在地嚮導 · 25 則評論
★★★ 3 週前 新
當天吃的是上等牛肉 雙人套餐
我覺得份量非常夠喔～還有一鍋鮭魚味噌湯跟甜點，非常豐富！
鮭魚烤飯糰滿香的，但覺得沒有到那麼驚豔 ...
全文

還有9項



海底撈火鍋新竹分店
4.5 ★★★★★ (3474) · $$$
火鍋餐廳 · 營業時間

總覽 最新資訊 菜單 評論 相片 關於

網路上的評論

Facebook
4.8/5
513 票

LINE熱點
4.8/5
87 則評論

Google 評論摘要

5
4
3
2
1

4.5
★★★★★
(3,474)

評分及撰寫評論
分享你的經驗供他人參考
☆ ☆ ☆ ☆ ☆

Google 評論

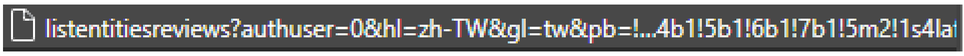
全部 獎勵 178 表演 174 撈麵 91 蝦滑 44 +5

排序依據

關聯性最高 最新 最高 最低

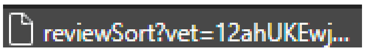
CHIEN YU HUANG
在地嚮導 · 73 則評論 · 226 張相片
★★★★★ 4 個月前
五人用餐，本來擔心點不夠，結果吃得非常飽！
點餐內容 ...... 更多

The left hand side is



listentitiesreviews?authuser=0&hl=zh-TW&gl=tw&pb=!...4b1!5b1!6b1!7b1!5m2!1s4la

While the right hand side is

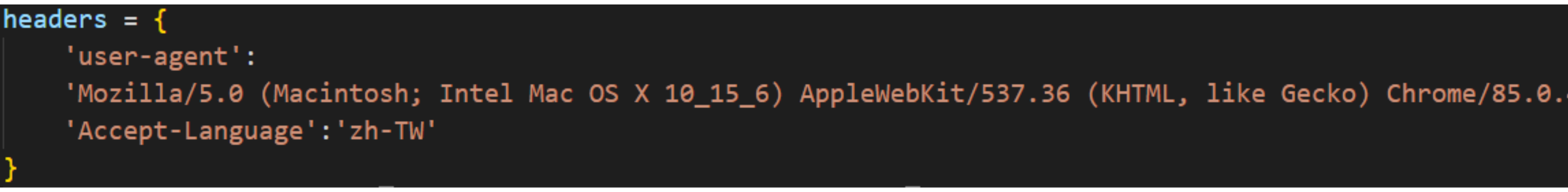


reviewSort?vet=12ahUKEwj...

So the map version has rules to follow to do the crawler, by iterating one of the params i, which loop as we scroll the page.

So next part is about how to do the crawler.

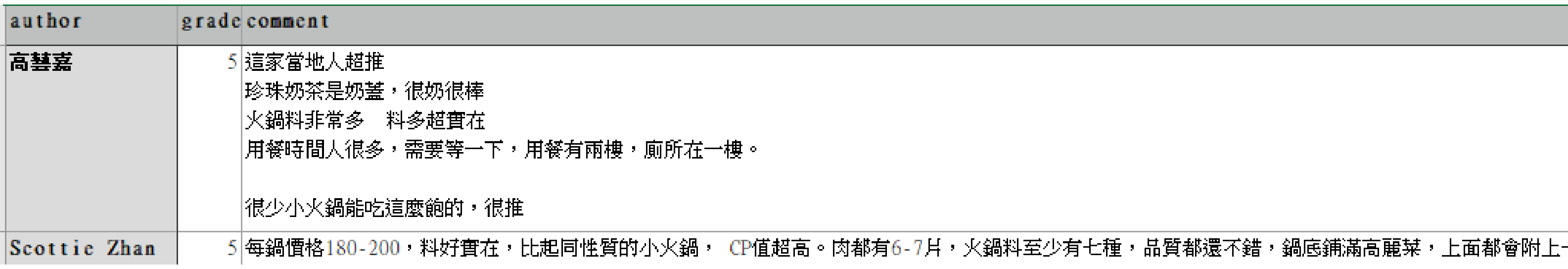
Initially, we try to use selenium/webdriver, but we finally decide to call the api of backend. And, google's anti-crawler still block my continuous request in about 1 min. So, to avoid such circumstance, I add the header(language for not to get reviews other than Chinese, or we would receive English review)



```
headers = {
    'user-agent':
    'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/85.0.
    'Accept-Language':'zh-TW'
}
```

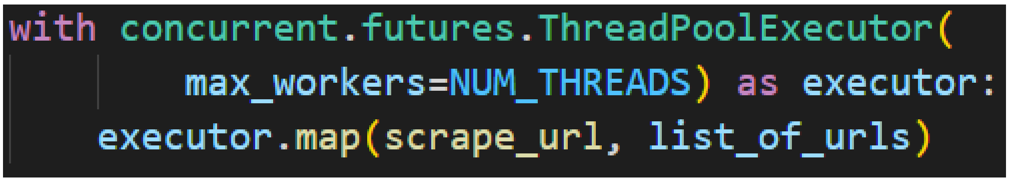
Hope not be blocked again, however, it's of no use. So I use VPN, try not to request from same ip for too many times. But it still of no use, because our demand is so big that these method can't prevent we from being blocked. So, we look up for another third party tool. Scraper API.

It can automatically change from millions of proxy server, and do the retry also. So I use the tool and do implementation in python. (all codes will be on github also).



author	grade	comment
高慧嘉	5	這家當地人超推 珍珠奶茶是奶蓋，很奶很棒 火鍋料非常多　料多超實在 用餐時間人很多，需要等一下，用餐有兩樓，廁所在一樓。 很少小火鍋能吃這麼飽的，很推
Scottie Zhan	5	每鍋價格180-200，料好實在，比起同性質的小火鍋， CP值超高。肉都有6-7片，火鍋料至少有七種，品質都還不錯，鍋底鋪滿高麗菜，上面都會附上

So I use the tool to get data, and here comes the last question, It’s too slow. Only 2500 reviews takes few hours to craw. So I change the way we craw to multi-thread. Use the python module, `import concurrent.futures`,we create all URL as a list, and view them as a pool, then the workers will execute the crawler from pool once they’re free.



```
with concurrent.futures.ThreadPoolExecutor(
        max_workers=NUM_THREADS) as executor:
    executor.map(scrape_url, list_of_urls)
```

After crawling, export the data as excel.

So, that’s about the whole process of the data collection.

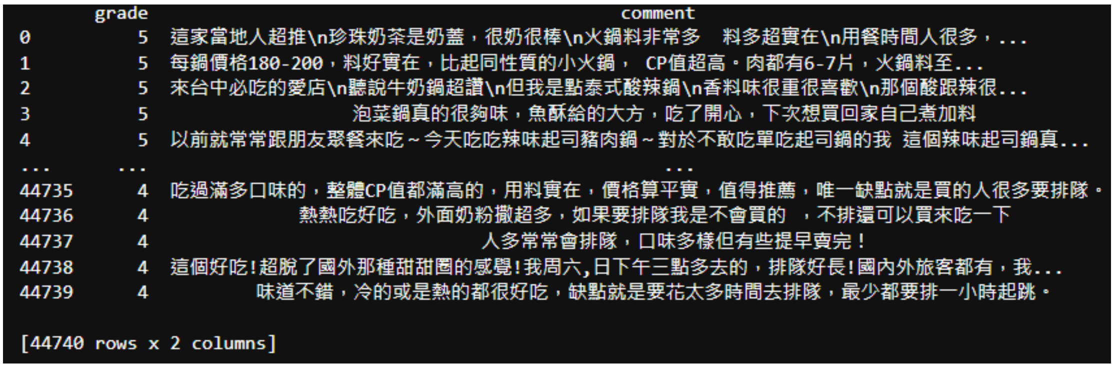
NLP part:

For data analysis, preprocessing, I use 3 famous tool in Machine Learning & in NLP field. The Pandas & matplotlib & numpy.

So, first, let’s import our data using pandas.read_excel()

pandas Matplotlib NumPy

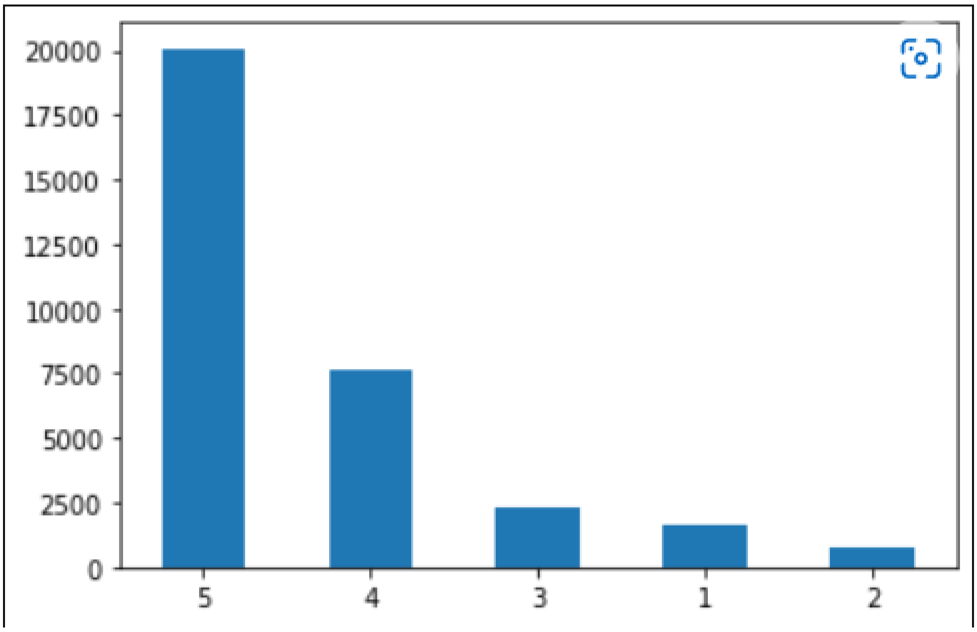
So the data is import as dataFrame like this:



```
       grade                                            comment
0          5  這家當地人超推\n珍珠奶茶是奶蓋，很奶很棒\n火鍋料非常多  料多超實在\n用餐時間人很多，...
1          5  每鍋價格180-200，料好實在，比起同性質的小火鍋， CP值超高。肉都有6-7片，火鍋料至...
2          5  來台中必吃的愛店\n聽說牛奶鍋超讚\n但我是點泰式酸辣鍋\n香料味很重很喜歡\n那個酸跟辣很...
3          5                        泡菜鍋真的很夠味，魚酥給的大方，吃了開心，下次想買回家自己煮加料
4          5  以前就常常跟朋友聚餐來吃～今天吃吃辣味起司豬肉鍋～對於不敢吃單吃起司鍋的我 這個辣味起司鍋真...
...      ...                                                ...
44735      4  吃過滿多口味的，整體CP值都滿高的，用料實在，價格算平實，值得推薦，唯一缺點就是買的人很多要排隊。
44736      4             熱熱吃好吃，外面奶粉撒超多，如果要排隊我是不會買的 ，不排還可以買來吃一下
44737      4                                  人多常常會排隊，口味多樣但有些提早賣完！
44738      4  這個好吃!超脫了國外那種甜甜圈的感覺!我周六,日下午三點多去的，排隊好長!國內外旅客都有，我...
44739      4          味道不錯，冷的或是熱的都很好吃，缺點就是要花太多時間去排隊，最少都要排一小時起跳。

[44740 rows x 2 columns]
```

Let’s using matplotlib to see how our data distribution:



There’s huge difference between amounts of 5 starts & others,

Actually it’s normal, because real trend on Google maps is that

Taiwanese seem love to give 5 stars on revies.

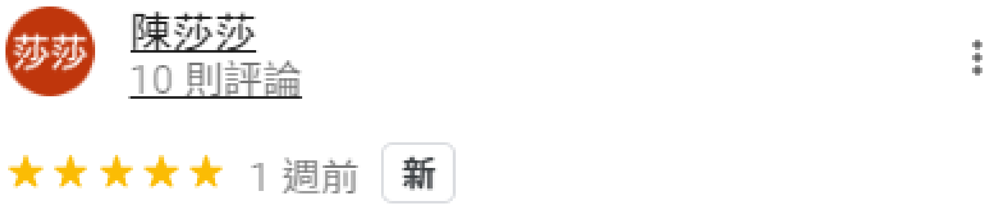
Let’s hand over it in next version, let’s move on first.

Next step, we’re going to deal with the row without comment.

Because some people have no comments but only stars like

Sasa Chen:

(this step is address reviews below)



莎莎

陳莎莎

10 則評論

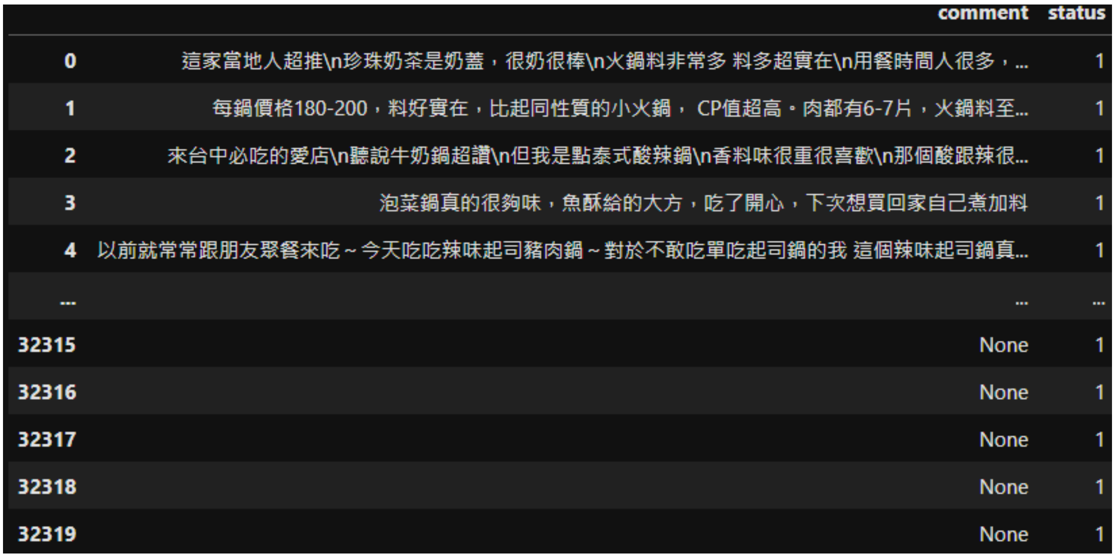
★★★★★ 1 週前 新

So our crawler will only get None value in excel. So we replace ‘None’ to Nan, then use pandas.dropna() to get rid of those empty value. And after dealing with the some basic operation. We can start to train our model.

Let’s see positive/negative analysis first, So we change the reviews into 2 categories, rating >3 are positive, others are negative. And we choose to use bert as our model, which is a bidirectional encoder from transformers. So using pyTorch & CUDA, I build the model on my own environment. Finishing the fine tune part of bert is to set the label as int, and use f1score, confusion matrix to present the outcome. I use sklearn’s package.

Simple Transformers

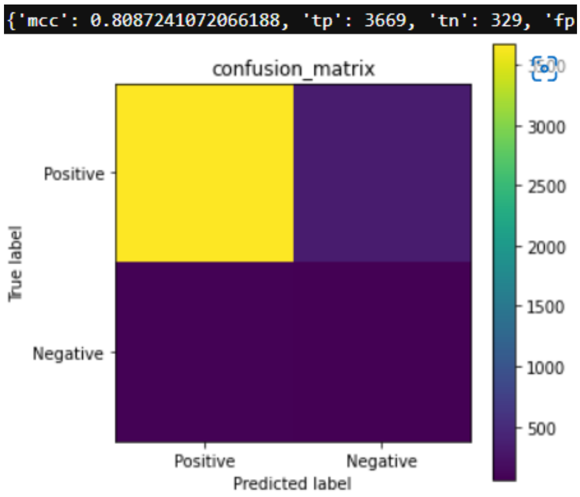
Transformers
Documentation



	comment	status
0	這家當地人超推\n珍珠奶茶是奶蓋，很奶很棒\n火鍋料非常多 料多超實在\n用餐時間人很多，...	1
1	每鍋價格180-200，料好實在，比起同性質的小火鍋， CP值超高。肉都有6-7片，火鍋料至...	1
2	來台中必吃的愛店\n聽說牛奶鍋超讚\n但我是點泰式酸辣鍋\n香料味很重很喜歡\n那個酸跟辣很...	1
3	泡菜鍋真的很夠味，魚酥給的大方，吃了開心，下次想買回家自己煮加料	1
4	以前就常常跟朋友聚餐來吃～今天吃吃辣味起司豬肉鍋～對於不敢吃單吃起司鍋的我 這個辣味起司鍋真...	1
...	...	...
32315	None	1
32316	None	1
32317	None	1
32318	None	1
32319	None	1

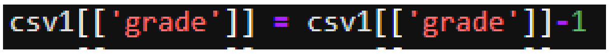
So the original data would look like this after doing 2-label.

Using train_test_split in sklearn.model_selection, I use test data Size of 4200. And most of data are positive, which is quite normal and fulfill original trend. So let’s do 5-label version. With the same processing before, we change the None to Nan then remove them from data. And minus 1 from original score, Cause the bert’s label need to start from 0, instead of 1.



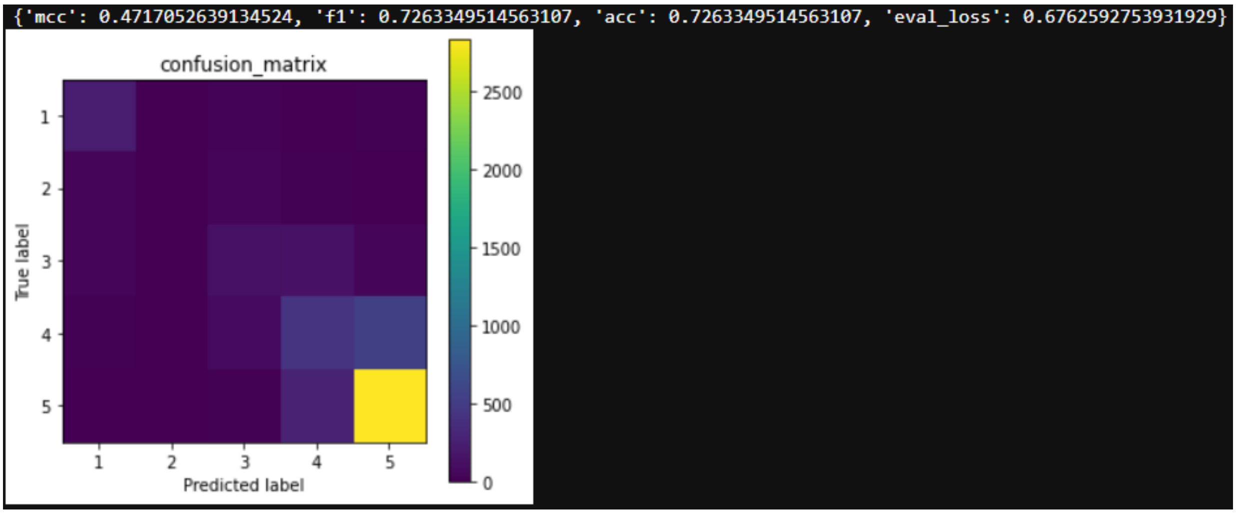
```
{'mcc': 0.8087241072066188, 'tp': 3669, 'tn': 329, 'fp
```

2-label version



```
csv1[['grade']] = csv1[['grade']]-1
```

Here’s the outcome of 5-label:

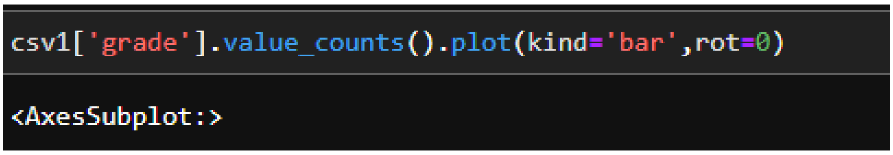


```
{'mcc': 0.4717052639134524, 'f1': 0.7263349514563107, 'acc': 0.7263349514563107, 'eval_loss': 0.6762592753931929}
```

5-label version

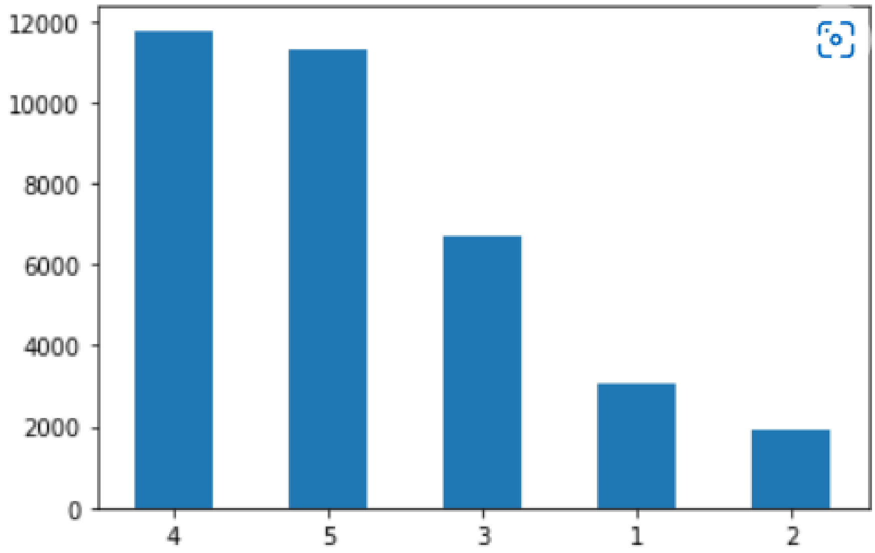
We can see that most of the data still appears in label 5-stars.

So to get a more accurate outcome, we try to make the numbers of 5-stars not exceed others too much. So we remove 10000 reviews of 5-stars. And the data distribution now:



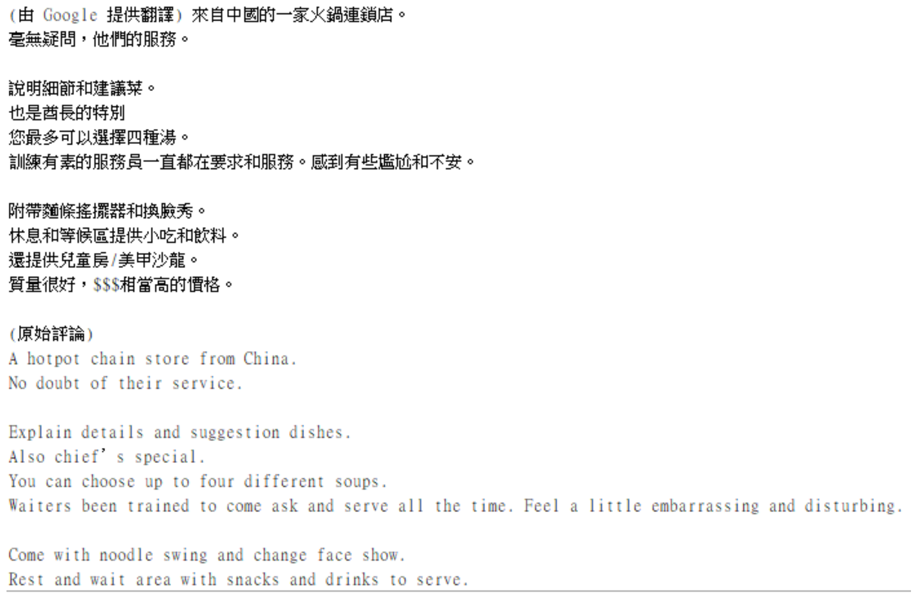
```
csv1['grade'].value_counts().plot(kind='bar',rot=0)
```

```
<AxesSubplot:>
```



Also, we remove the reviews that contains translation.

Because some foreigner might comment in language other than Chinese. Like the example of a comment of 海底撈:



(由 Google 提供翻譯) 來自中國的一家火鍋連鎖店。
毫無疑問，他們的服務。

說明細節和建議菜。
也是酋長的特別
您最多可以選擇四種湯。
訓練有素的服務員一直都在要求和服務。感到有些尷尬和不安。

附帶麵條搖擺器和換臉秀。
休息和等候區提供小吃和飲料。
還提供兒童房/美甲沙龍。
質量很好，$$$相當高的價格。

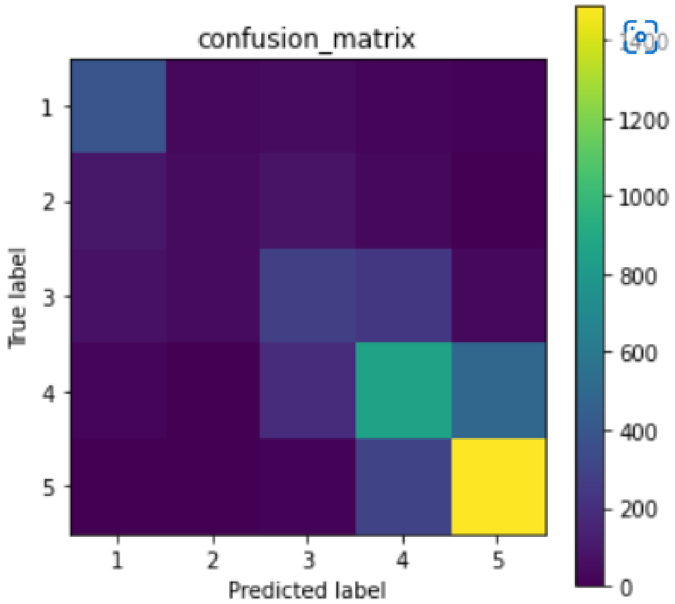
(原始評論)
A hotpot chain store from China.
No doubt of their service.

Explain details and suggestion dishes.
Also chief’ s special.
You can choose up to four different soups.
Waiters been trained to come ask and serve all the time. Feel a little embarrassing and disturbing.

Come with noodle swing and change face show.
Rest and wait area with snacks and drinks to serve.

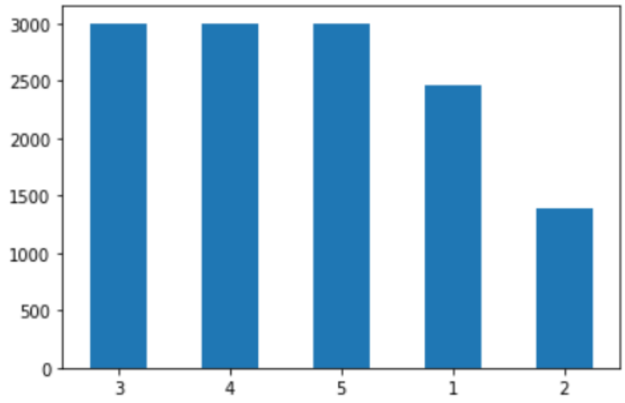
So if comment contains ‘Google 提供翻譯’, we remove it.

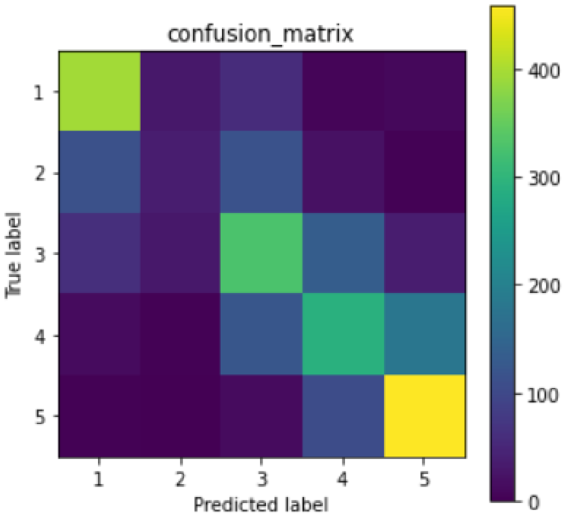
And here’s the consequence:



We can see the diagonal is more clear, which means our model is getting better after the label become more balanced.

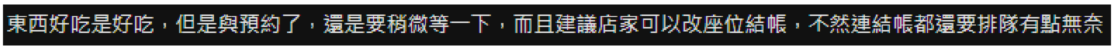
So we change the data distribution again, make data more balanced. And here’s our final outcome:

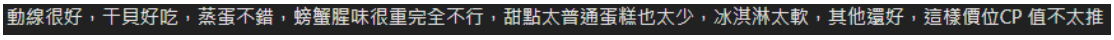


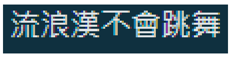


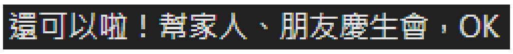
We can see the diagonal is clear but a little bit divergent.

And start-2 seems the most bad, so let’s check the original data:

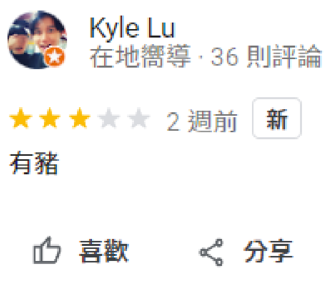

東西好吃是好吃，但是與預約了，還是要稍微等一下，而且建議店家可以改座位結帳，不然連結帳都還要排隊有點無奈

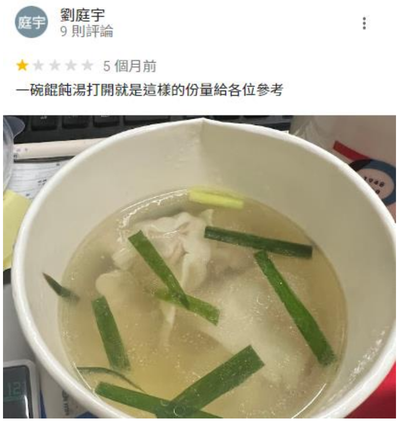

動線很好，干貝好吃，蒸蛋不錯，螃蟹腥味很重完全不行，甜點太普通蛋糕也太少，冰淇淋太軟，其他還好，這樣價位CP 值不太推

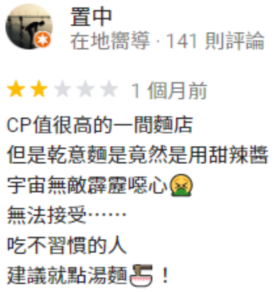

流浪漢不會跳舞


還可以啦！幫家人、朋友慶生會，OK

All of these comments are not ‘that’ bad I think. But they’re only worth of 2 stars. Basically, the comment is too subjective & someone will comment in strange viewpoint to affect our prediction.


Kyle Lu
在地嚮導 · 36 則評論
★★★☆☆ 2 週前 新
有豬


劉庭宇
9 則評論
★☆☆☆☆ 5 個月前
一碗餛飩湯打開就是這樣的份量給各位參考

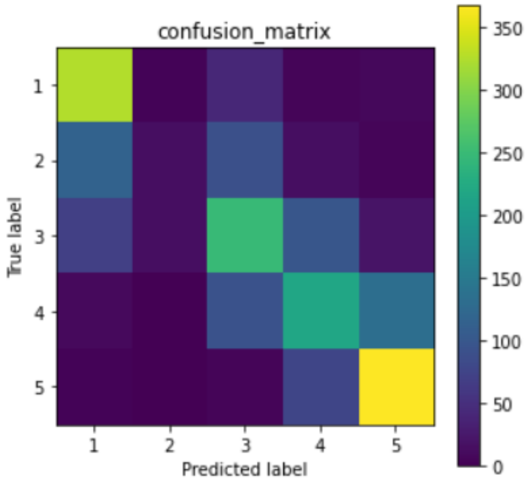

置中
在地嚮導 · 141 則評論
★★☆☆☆ 1 個月前
CP值很高的一間麵店
但是乾意麵是竟然是用甜辣醬
宇宙無敵霹靂噁心🤮
無法接受……
吃不習慣的人
建議就點湯麵🍜！

So comments seem to be neutral can be low stars also, which affect our model.

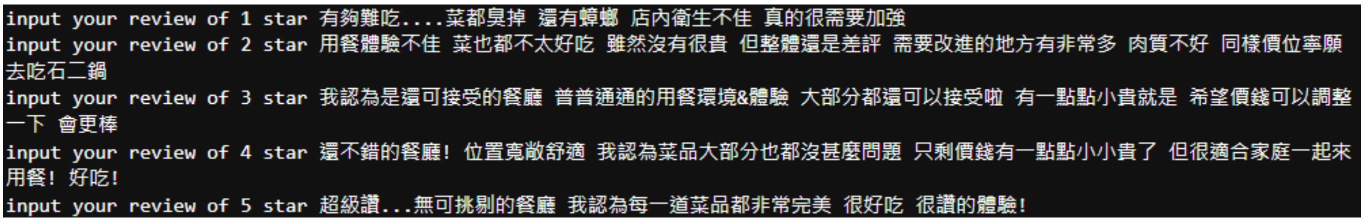
To do it more precise, I filter the comment that’s too short like: ‘難吃’,’還不錯’,’流浪漢不會跳舞’, only leave the comment whose length is more than 15.

And the outcome looks much the same as the previous one.

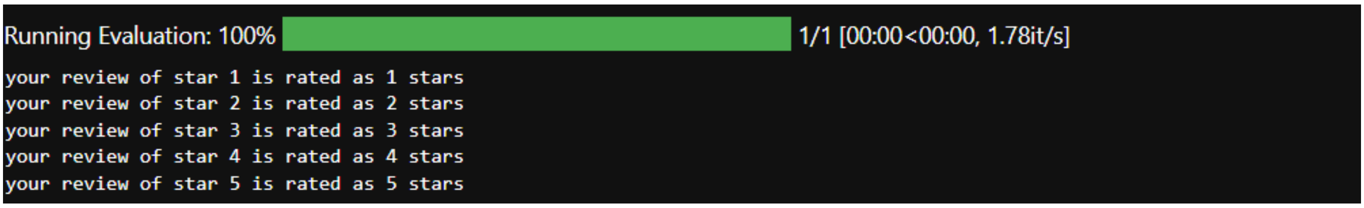
So I think the most of the reason is that reviews are subjective, and we can analysis in such way to get most of the result correct(70% of the majority people).



Last part, I do a manually input reviews, to check the outcome by your own.



```
input your review of 1 star 有夠難吃....菜都臭掉 還有蟑螂 店內衛生不佳 真的很需要加強
input your review of 2 star 用餐體驗不佳 菜也都不太好吃 雖然沒有很貴 但整體還是差評 需要改進的地方有非常多 肉質不好 同樣價位寧願
去吃石二鍋
input your review of 3 star 我認為是還可接受的餐廳 普普通通的用餐環境&體驗 大部分都還可以接受啦 有一點點小貴就是 希望價錢可以調整
一下 會更棒
input your review of 4 star 還不錯的餐廳! 位置寬敞舒適 我認為菜品大部分也都沒甚麼問題 只剩價錢有一點點小小貴了 但很適合家庭一起來
用餐! 好吃!
input your review of 5 star 超級讚...無可挑剔的餐廳 我認為每一道菜品都非常完美 很好吃 很讚的體驗!
```



```
Running Evaluation: 100%                                        1/1 [00:00<00:00, 1.78it/s]
your review of star 1 is rated as 1 stars
your review of star 2 is rated as 2 stars
your review of star 3 is rated as 3 stars
your review of star 4 is rated as 4 stars
your review of star 5 is rated as 5 stars
```

The reviews are all from my own opinion when I share comment in real condition, and the outcome is good!

So that’s all about our project. In the process, most of time spending on the model, tons of error and fix the crawler bug.

It’s a long journey & We learn a lot in the project.

Hope you enjoy our final project!

Group 23

Cute Bert 😊 (easy to learn, hard to utilize & debug, my environment was a chaos after the project)