文字探勘專題

2019

嚴選台北地區五星飯店之評論

高子淇、賴永琪、劉萱

目录 Contents

01

動機與目的

動機與目的

01 動機

隨著時間累積大量的顧客評論也成為重要的研究資料，利用文字探勘的方式，找出有用的關鍵字來推薦使用者最佳訂房選擇，以縮短使用者訂房時間。

02 介紹

由於技術上的困難，在嘗試過後，選擇以評分與評論之間的關係做成分類。

03 目的

了解消費者對於旅館的評論與評分之間的關聯，進而能使網站平台，飯店了解消費者的需求，最終能使消費者更佳便利的瀏覽訂房網站。

目录 Contents

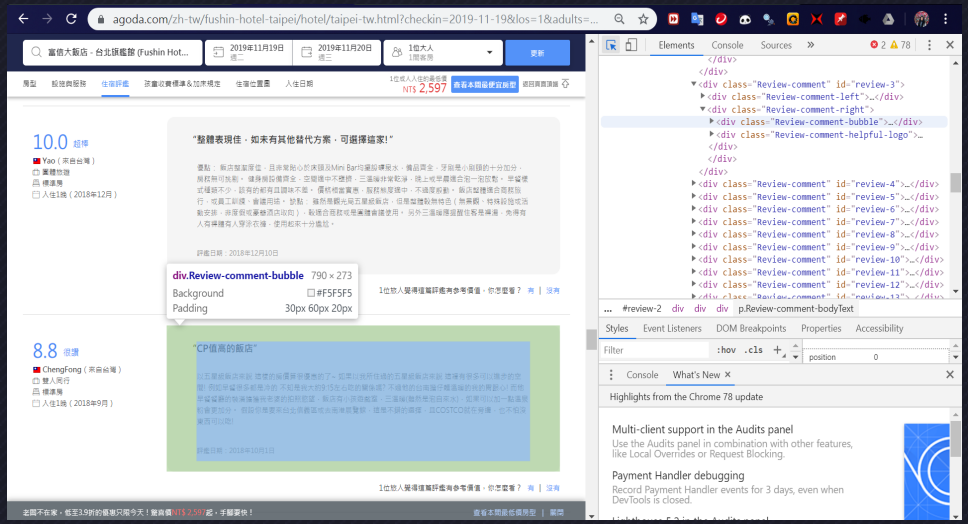
02

資料收集

資料收集

01 Agoda網站爬取

使用(https://www.agoda.com/zh-tw/)提取台北地區五星級飯店的顧客評鑑



資料收集

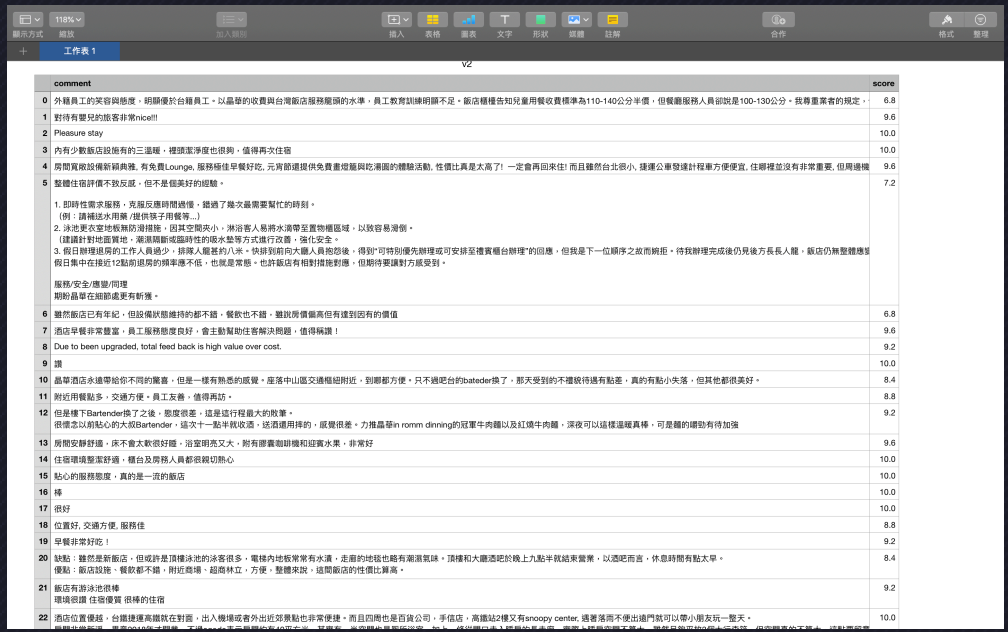
02 使用方法：selenium

直接運行在瀏覽器中，通過一系列命令來模擬操作，可以將這些命令轉化成實際的HTTP請求在瀏覽器中運行

In [1]:
```
import requests
from bs4 import BeautifulSoup
import selenium
from selenium import webdriver
from selenium.common.exceptions import NoSuchElementException
from selenium.webdriver.common.keys import Keys
import time
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.common.by import By
from selenium.webdriver.support import expected_conditions as ec
import pandas as pd
```

資料來源

03 爬下飯店評論及評分將兩項製程資料表做分類分析



v2

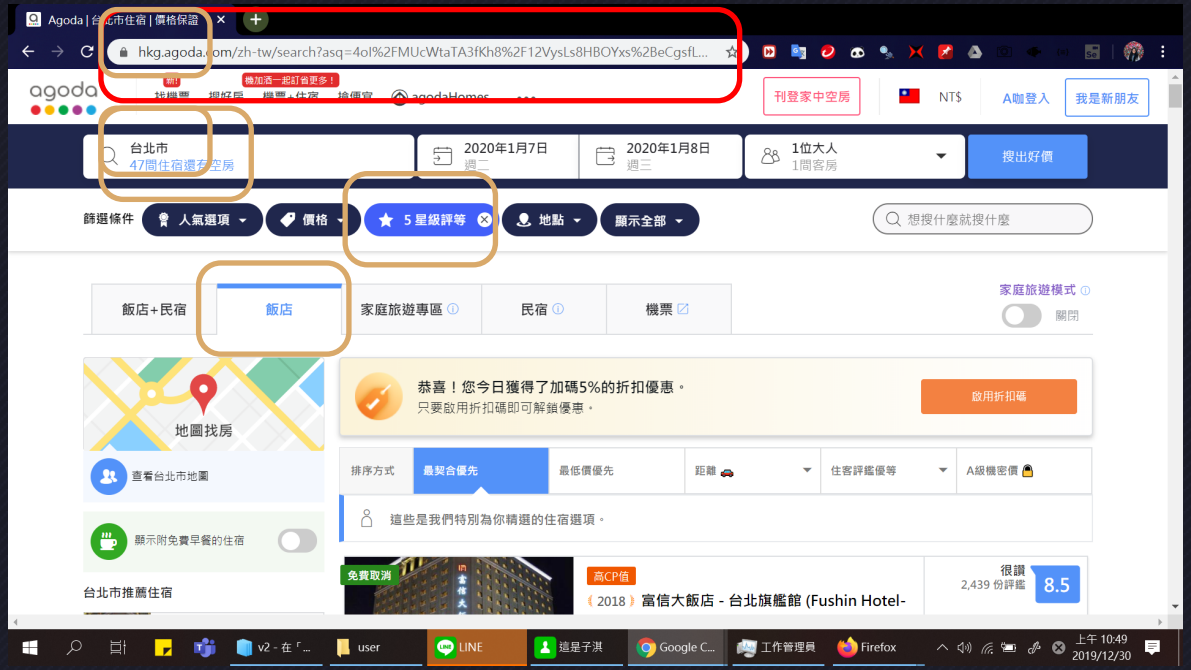
	comment	score
0	外籍員工的笑容與態度，明顯優於台籍員工。以晶華的收費與台灣飯店服務層頭的水準，員工教育訓練明顯不足。飯店櫃檯告知兒童用餐收費標準為110-140公分半價，但餐廳服務人員卻說是100-130公分。我尊重業者的規定，	6.8
1	對待有嬰兒的旅客非常nice!!!	9.6
2	Pleasure stay	10.0
3	內有少數飯店設施有的三溫暖，裡頭潔淨度也很夠，值得再次住宿	10.0
4	房間寬敞設備新穎典雅, 有免費Lounge, 服務極佳早餐好吃, 元宵節還提供免費畫燈籠與吃湯圓的體驗活動, 性價比真是太高了! 一定會再回來住! 而且雖然台北很小, 捷運公車發達計程車方便便宜, 住得裡並沒有非常重要, 但周邊機	9.6
5	整體住宿評價不致反感，但不是個美好的經驗。 1. 即時性需求服務，克服反應時間過慢，錯過了幾次最需要幫忙的時刻。 （例：請補送水用筷 /提供筷子用餐等...） 2. 泳池更衣室地板無防滑措施，因其空間夾小，淋浴客人易將水滴帶至置物櫃區域，以致容易滑倒。 （建議針對地面質地，潮濕隔斷或臨時性的吸水墊等方式進行改善，強化安全。） 3. 假日辦理退房的工作人員過少，排隊人龍甚約八米，快排到前向大廳人員抱怨後，得到"可特別優先辦理或可安排至禮賓櫃台辦理"的回應，但我是下一位順序之故而婉拒。待我辦理完成後仍見後方長長人龍，飯店仍無整體應變	7.2
	假日集中在接近12點前退房的頻率應不低，也就是常態。也許飯店有相對措施對應，但期待要讓對方感受到。 服務/安全/應變/同理 期盼晶華在細節處更有斬獲。	
6	雖然飯店已有年紀，但設備狀態維持的都不錯，餐飲也不錯，雖說房價偏高但有達到因有的價值	6.8
7	酒店早餐非常豐富，員工服務態度良好，會主動幫助住客解決問題，值得稱讚！	9.6
8	Due to been upgraded, total feed back is high value over cost.	9.2
9	讚	10.0
10	晶華酒店永遠帶給你不同的驚喜，但是一樣有熟悉的感覺。座落中山區交通樞紐附近，到哪都方便。只不過吧台的bateder換了，那天受到的不禮貌待遇有點差，真的有點小失落，但其他都很美好。	8.4
11	附近用餐點多，交通方便，員工友善，值得再訪。	8.8
12	但是樓下Bartender換了之後，態度很差，這是這行程最大的敗筆。 很懷念以前點心的大叔Bartender，這次十一點半就收酒，送酒連用拌的，感覺很差。力推晶華in romm dinning的冠軍牛肉麵以及紅燒牛肉麵，深夜可以這樣溫暖真棒，可是麵的嚼勁有待加強	9.2
13	房間安靜舒適，床不會太軟很好睡，浴室明亮又大，附有膠囊咖啡機和迎賓水果，非常好	9.6
14	住宿環境整潔舒適，櫃台及周務人員都很親切熱心	10.0
15	點心的服務態度，真的是一流的飯店	10.0
16	棒	10.0
17	很好	10.0
18	位置好, 交通方便, 服務佳	8.8
19	早餐非常好吃！	9.2
20	缺點：雖然是新飯店，但或許是頂樓泳池的泳客很多，電梯內地板常常有水漬，走廊的地毯也略有潮濕氣味，頂樓和大廳酒吧於晚上九點半就結束營業，以酒吧而言，休息時間有點太早。 優點：飯店設施、餐飲都不錯，附近商場、超商林立，方便，整體來說，這間飯店的性價比算高。	8.4
21	飯店有游泳池很棒 環境很讚 住宿優質 很棒的住宿	9.2
22	酒店位置優越，台鐵捷運高鐵就在對面，出入機場或者外出近郊景點也非常便捷。而且四周也是百貨公司，手信店，高鐵站2樓又有snoopy center, 邊著落雨不便出境門就可以帶小朋友玩一整天。	10.0

目录 Contents

03

模型與分析步驟

分析步驟



台北市

五星級

飯店

技術說明

```
browser = webdriver.Firefox()
browser.get("https://www.agoda.com/zh-tw/search?city=4951&languageId=20&userId=fbea6c98-bc10-4d2c-a45f-38aa37091ae0&sessionId=qi
#while len(soup.select(".btn pagination2__next"))>0:
browser.execute_script("window.scrollTo(0,336)")
browser.execute_script("window.scrollTo(0,354.6666564941406)")
browser.execute_script("window.scrollTo(0,469.3333435058594)")
browser.execute_script("window.scrollTo(0,696.6666870117188)")
browser.execute_script("window.scrollTo(0,813.3333129882812)")
browser.execute_script("window.scrollTo(0,928.6666870117188)")
browser.execute_script("window.scrollTo(0,1046.6666259765625)")
browser.execute_script("window.scrollTo(0,1160.6666259765625)")
browser.execute_script("window.scrollTo(0,1277.3333740234375)")
browser.execute_script("window.scrollTo(0,1396.6666259765625)")
browser.execute_script("window.scrollTo(0,1508.6666259765625)")
browser.execute_script("window.scrollTo(0,1814)")
browser.execute_script("window.scrollTo(0,1858)")
browser.execute_script("window.scrollTo(0,2134.666748046875)")
browser.execute_script("window.scrollTo(0,2204.666748046875)")
browser.execute_script("window.scrollTo(0,2440.666748046875)")
browser.execute_script("window.scrollTo(0,2553.333251953125)")
browser.execute_script("window.scrollTo(0,2985.333251953125)")
browser.execute_script("window.scrollTo(0,3047.333251953125)")
browser.execute_script("window.scrollTo(0,3364)")
browser.execute_script("window.scrollTo(0,3667.333251953125)")
browser.execute_script("window.scrollTo(0,3831.333251953125)")
browser.execute_script("window.scrollTo(0,4114.66650390625)")
browser.execute_script("window.scrollTo(0,4524)")
browser.execute_script("window.scrollTo(0,4576.66650390625)")
browser.execute_script("window.scrollTo(0,5190.66650390625)")
browser.execute_script("window.scrollTo(0,5340.66650390625)")
browser.execute_script("window.scrollTo(0,5727.33349609375)")
```

技術說明

```
browser.execute_script("window.scrollTo(0,14500)")
soup = BeautifulSoup(browser.page_source)
http=[]
for ele in soup.select(".PropertyCardItem.ssr-search-result a"):
    http.append("https://www.agoda.com"+ele.get("href"))
browser.close()
comment=[]
score=[]
for i in http:
    url=i
    browser = webdriver.Firefox()
    browser.get(url)
    browser.execute_script("window.scrollTo(0, document.body.scrollHeight)")
    time.sleep(5)
    if browser.find_element_by_css_selector("span.Searchbox__searchButton__text"):
        browser.find_element_by_css_selector("span.Searchbox__searchButton__text").click()
        soup = BeautifulSoup(browser.page_source)

        for ele in soup.select(".Review-comment"):
            for ele1 in ele.select(".Review-comment-leftScore"):

                for ele2 in ele.select(".Review-comment-bodyText"):
                    if len(ele1.text)*len(ele2.text)!=0:

                        score.append(ele1.text)
                        comment.append(ele2.text)

    browser.close()
```

模型與分析步驟

用評分將評論分成三類

大於 9 為一類，在7、8區間呈另一類，不足者即為最後一類

分類完後，將分成訓練集與測試集

利用課堂上的分類方法

模型與分析步驟

01 評論與評分之高低關係

02 問題出現

03 模型訓練集有.95準確率

04 模型測試集有.55的準確率

```
evaluation on train data
confusion matrix:
[[242  10   9]
 [  0 188   0]
 [  0   0  51]]
accuracy:
0.962
------------------------------
evaluation on test data
confusion matrix:
[[230 113  42]
 [ 35  18   4]
 [  1   2   1]]
accuracy:
0.5582959641255605
```

目录 Contents

04

實驗與分析結果

實驗

01 過程中有使用jieba進行斷詞

02 然而，有過多不必要的單字使真正提及優缺點反而被掩蓋。

	support	itemsets
0	0.65	(\n)
1	0.50	(也)
2	0.55	(很)
3	0.50	(早餐)
4	0.60	(是)
5	0.65	(有)
6	0.90	(的)
7	0.55	(都)
8	0.65	(飯店)
9	0.50	(，)
10	0.55	(有, \n)
11	0.60	(的, \n)
12	0.50	(\n, 飯店)
13	0.50	(的, 也)
14	0.55	(的, 很)
15	0.60	(是, 的)
16	0.60	(的, 有)
17	0.55	(都, 的)
18	0.65	(的, 飯店)
19	0.50	(的, 有, \n)
20	0.50	(的, \n, 飯店)

實驗困難

1.selenium動態網站自動爬取下一頁，只能到第二頁，第三頁不能再繼續

2.訂房網站會出現沒有飯店名稱、詳細內容的推薦，解決方案if沒有詳細內容（程式碼）就跳過。

3.爬下來的comment 包含許多unicode string

模型與分析步驟

01 更多資料時測試集的準確率下降

02 問題又出現

03 模型訓練集有.95準確率

04 模型測試集有.48的準確率

```
evaluation on train data
confusion matrix:
[[370  14  12]
 [  1 242   2]
 [  0   0  59]]
accuracy:
0.9585714285714285
--------------------------------------------------
evaluation on test data
confusion matrix:
[[496 391 134]
 [ 48  62  16]
 [  0   0   0]]
accuracy:
0.4864864864864865
```

分析結果

只有近五成的準確率，可以再更探討更多因素，可以再了解情緒字詞。

目录 Contents

05

結論

結論

單純只看評論與評分的分類可能準確度較低，還需搭配情緒字詞或是個人對於分數的滿意程度上感覺的不同來深入了解

文字探勘專題計畫書

2019

感謝一路有你