

資料科學與程式科學導論 Final Project

主題: Google map analysis

組員

B07901069 電機三 劉奇聖

B09902102 資工一 陳冠辰

B06901046 電機四 葉曜德

B06901187 電機四 韓承霖

程式碼

原本的 repository: <https://github.com/MortalHappiness/rlads2021Spring-Final>

繳交的 repository: <https://github.com/rlads2021/project-dhcl8881>

簡介

研究動機

我們常常使用 google map, 除了使用導航的功能之外, 有時也會參考上面的評分和評論幫助我們認識地點。我們隱約注意到地點的種類似乎會影響到評分、評論內容用詞似乎會隨著評分高低有所不同等現象。因此想藉由分析 google map 評論、評分、地標種類三者的關係驗證我們的猜想。

研究目標

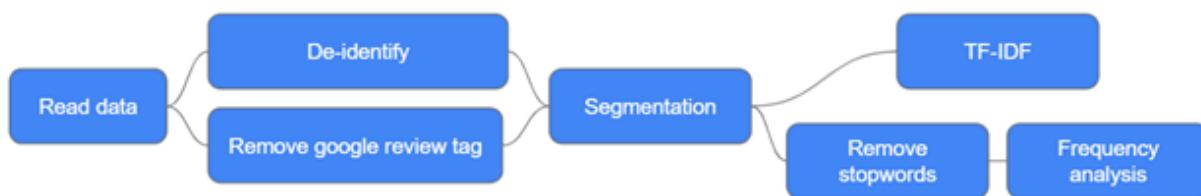
1. 不同類型的地標在 Google Map 上評分的分佈情形
2. 不同類型地標的評論重點詞
3. 評分高低的評論重點詞差異
4. 觀察 Google Map 上整體評論重點詞

方法

總流程



資料處理流程



資料取得

- 首先使用 Google Place API 裡的 Nearby search 抓出台大附近的各種地標資訊約 1000 筆，裡面包含名稱、評分、地標種類等資訊。
- 接著使用 Google Place API 裡的 Place Details Request 得到各地點在 Google map 上的網址。
- 使用 Python 的 Selenium 模擬瀏覽器捲動將各地點的評論爬取下來，以中文評論為主，每個地點最多爬取約 400 筆評論，總評論數量約 86000 筆，之所以使用 Python 是因為可以用 multi-thread 平行爬蟲，加快速度。

資料處理

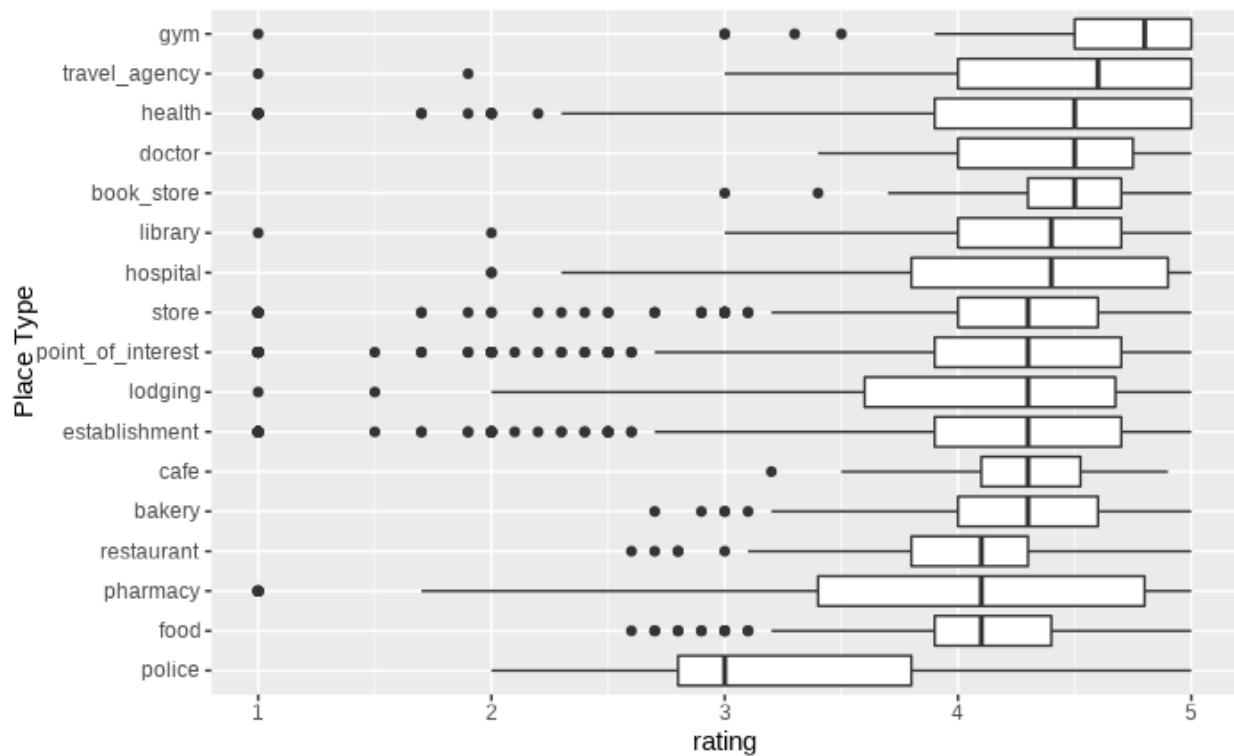
文字資料的部份經過斷詞後，我們在不同情境下使用兩種不同方式分析重點詞。第一種是 TF-IDF 分析，這種分析方法我們用來擷取一家店的特色評論。比如說一家 5 顆星星的店家，我們在分析重點詞的時候，不會想抓到「好吃」這種可能會出現在每一個五星店家的評論，相對的，我們會想知道這家有哪些其他店家少有的特色。第二種情況是詞頻分析，這種分析方法我們用來分析評論整體的指標，我們在乎哪些指標是經常被拿出來討論的，比如說在分析餐廳不同評分的重點詞時，前面「好吃」的例子就是我們希望能被擷取出來的指標。

原始碼運作說明

- 爬蟲:依序執行 crawler 資料夾裡的 crawl_places.py、crawl_urls.py、crawl_reviews.py
- 資料前處理:執行 Rscript/preprocessing.sh, 它會執行其他的 R script, 將資料處理好並存成 rds 的形式。
- 靜態圖片視覺化:執行 Rmds/index.Rmd 則可以看到所有的圖片

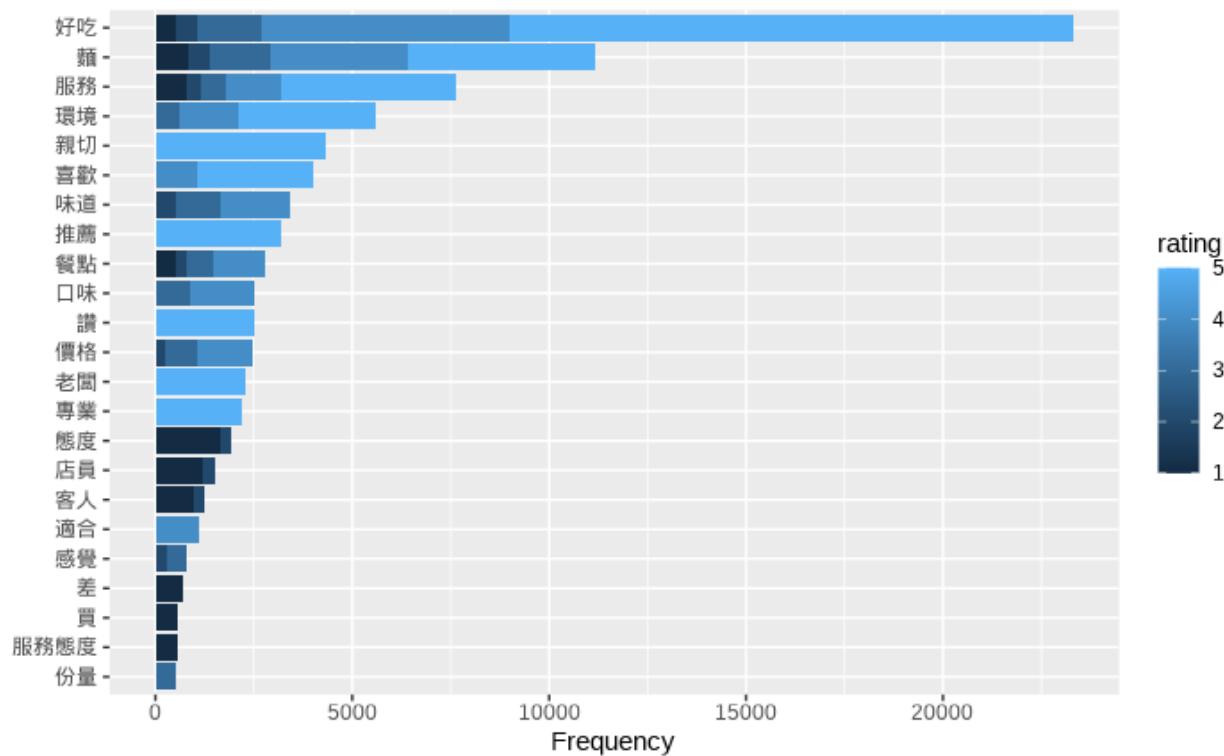
結果

地標類型的評分分佈



可以看出地點種類確實對評分有所影響，如警察局的評分明顯偏低，休閒場所如體育館、旅遊社等評分偏高。

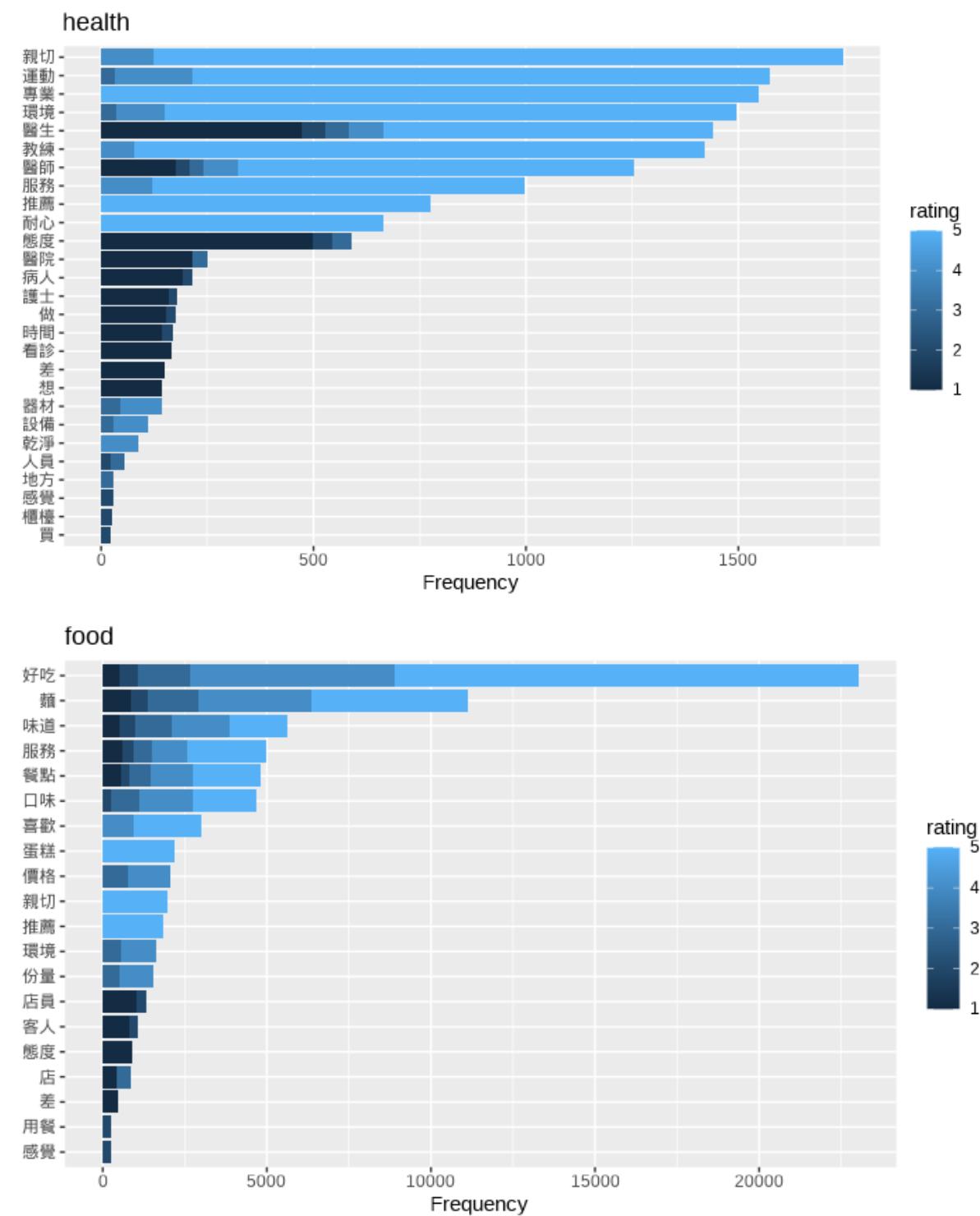
整體 Google Map 評論的重點詞

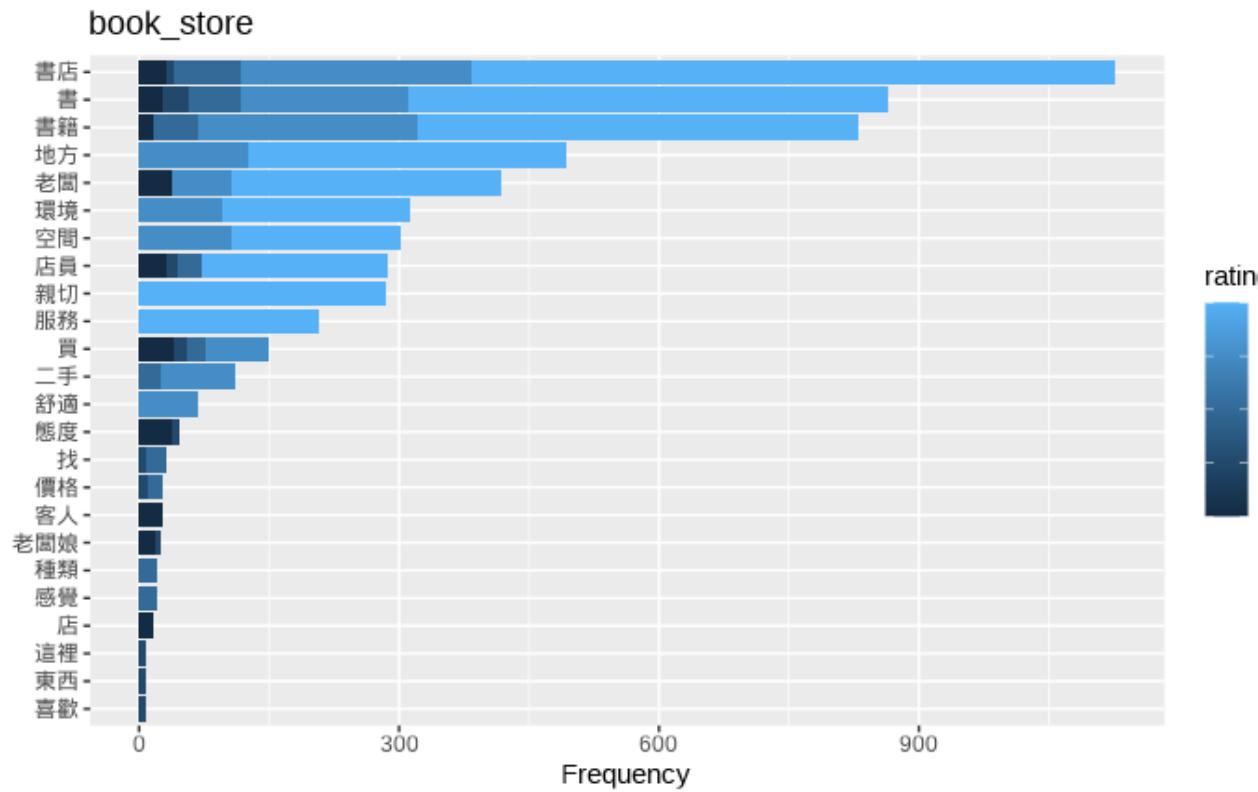
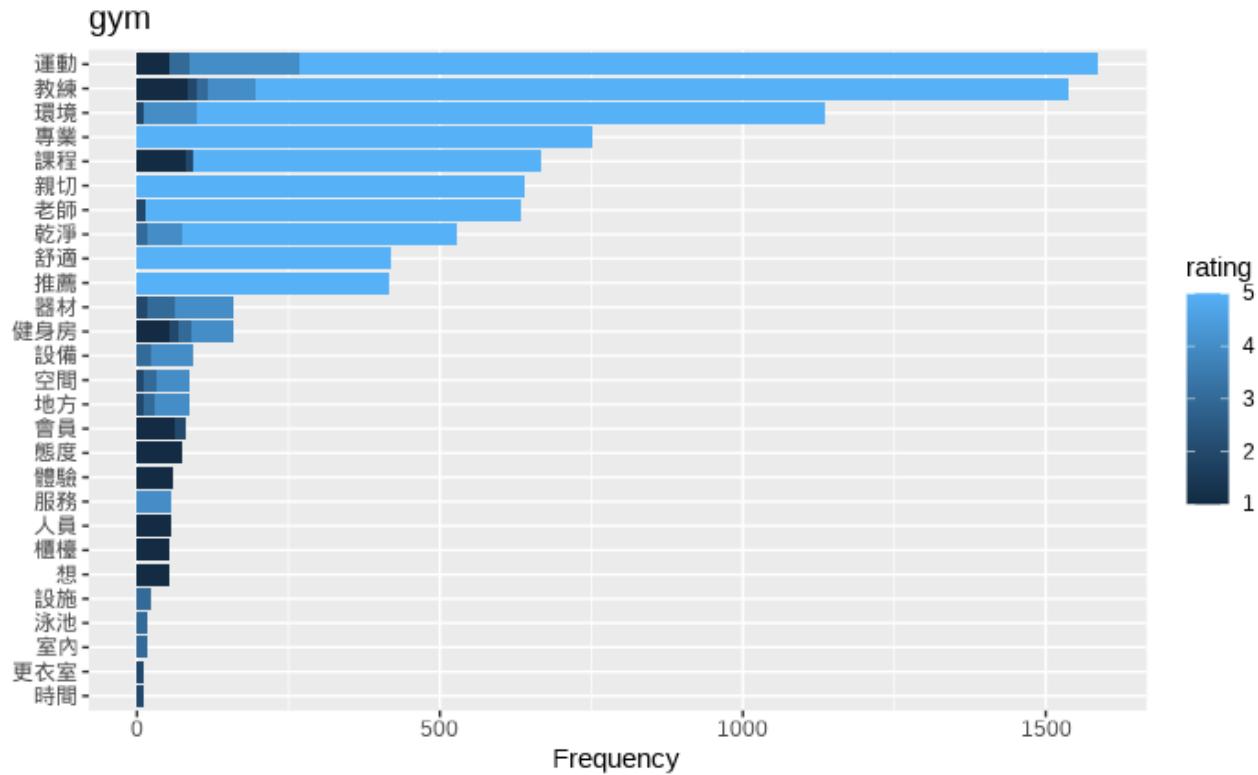


推測是因為資料裡餐廳的地點過多，導致結果裡面多是與吃有關的詞語，如「好吃」、「餐點」、「麵」等。

不同類別地標評論重點詞

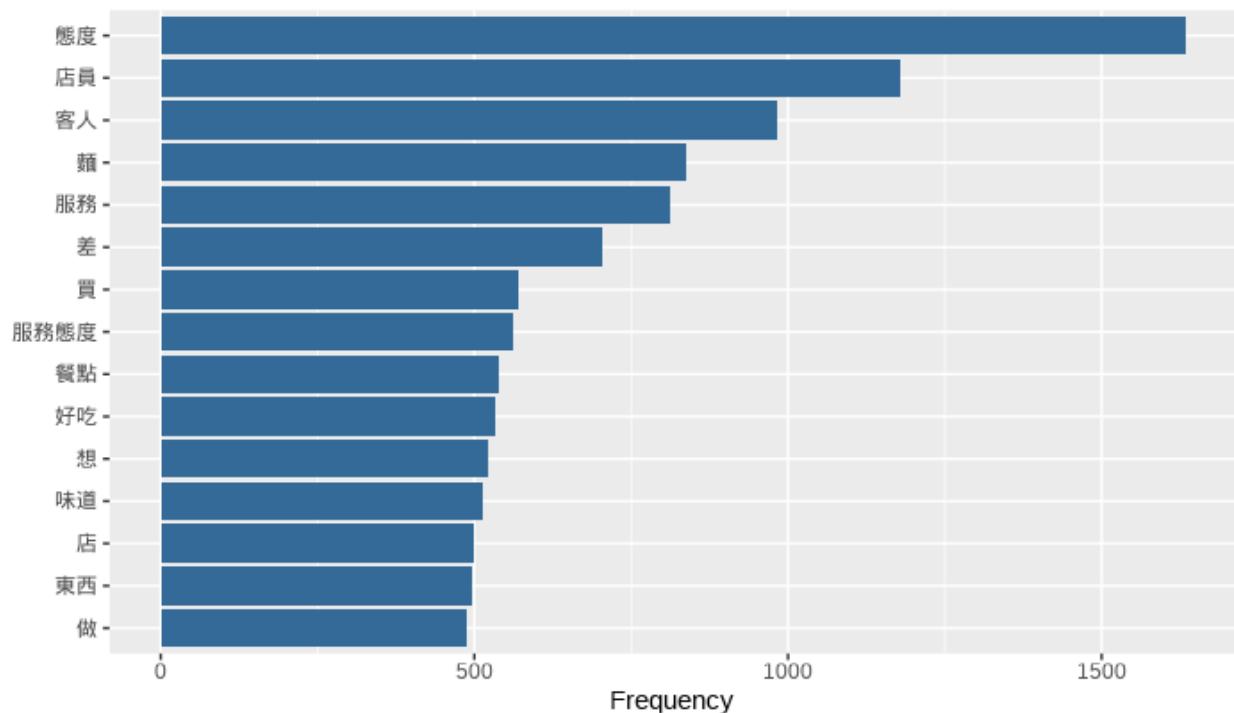
有鑑於分析整體 google map 的評論重點詞會受到地點類別的影響，我們對每個類別做重點詞分析，的確可以看出符合該地點類型容易聯想到的詞語，如下面 4 張圖所示，分別分析了 health, food, gym, book_store 這四種底點的重點詞。



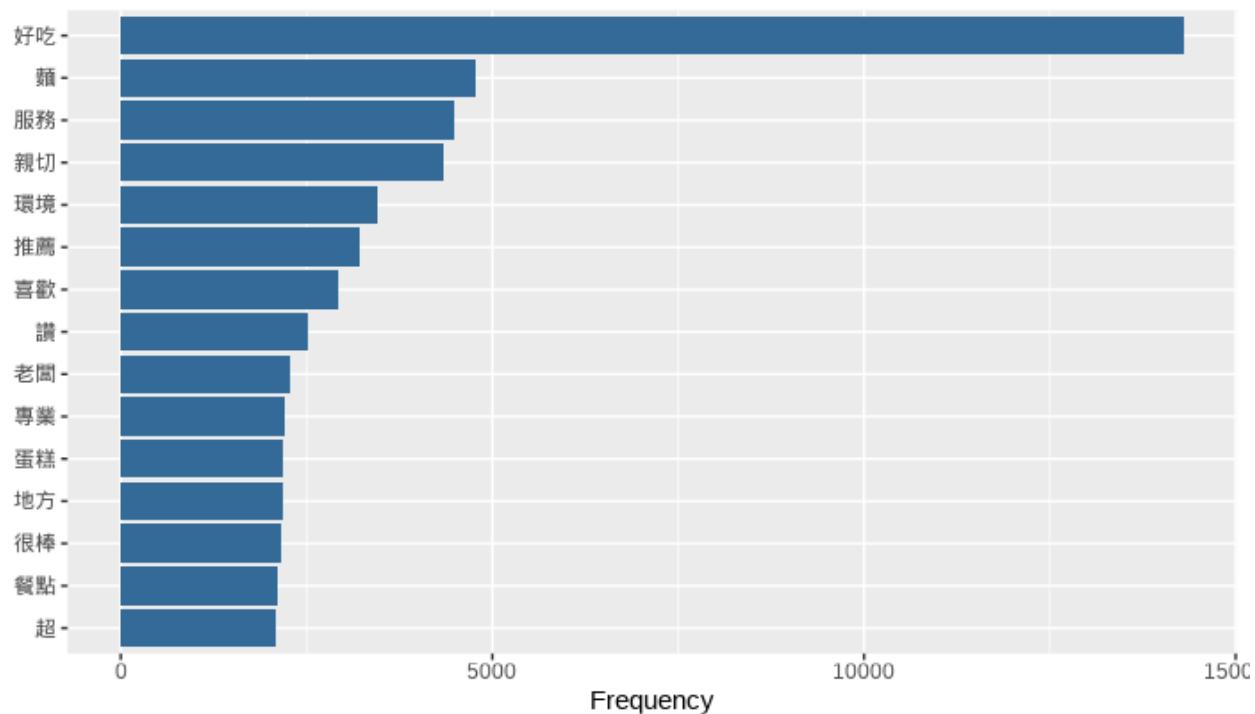


不同評分的評論重點詞

Rating = 1



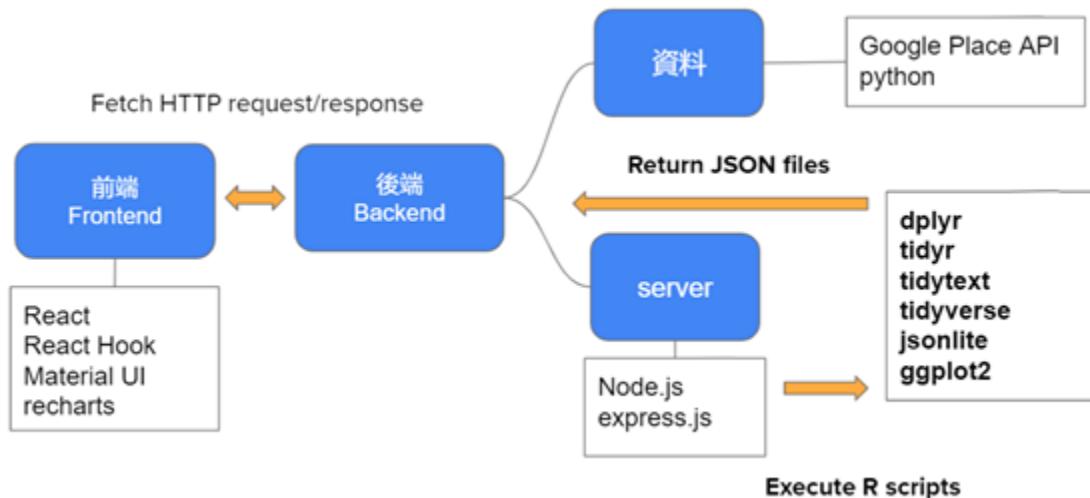
Rating = 5



上面挑選了 1 顆星和 5 顆星的評論來分析重點詞，可以看出 1 顆星的重點詞包含「態度」、「差」等，5 顆星的評論重點詞包含「好吃」、「親切」等。

套件、API與架構

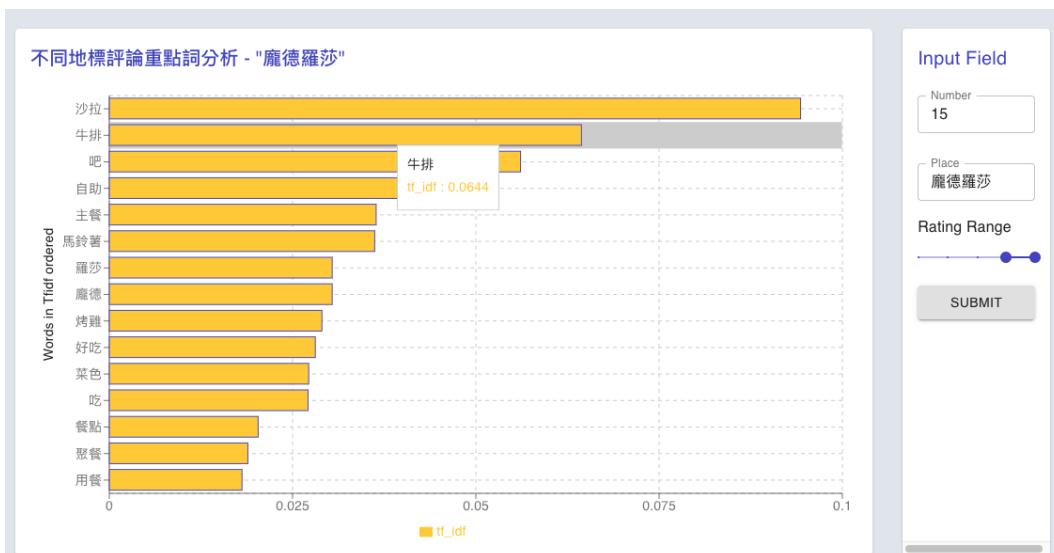
除了靜態視覺化之外，我們還寫了動態網頁來呈現我們的結果，以下是我們的網頁的架構圖。



App 功能介紹

1. 地標評論重點詞分析：

- 使用者輸入：
 - Number : 想要觀察的重點詞數量(長條圖的數目)
 - Place : 輸入想要搜尋的地標名稱(目前可支援台大附近的地標)
 - Rating Range : 輸入想要觀察的rating的間距, 範圍為1~5
- 輸出結果：根據該地標名稱，在輸入的rating範圍中所出現的評論重點詞分佈



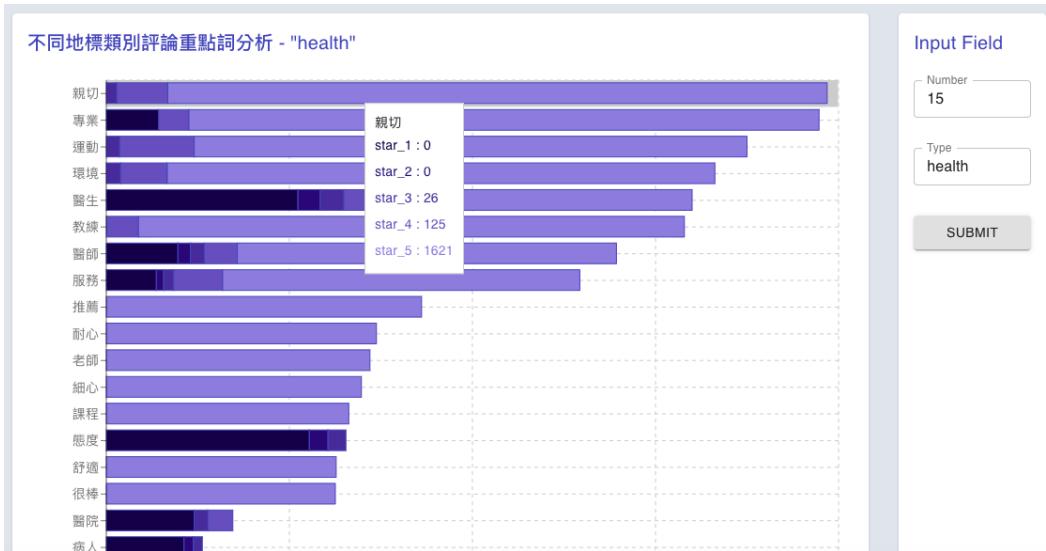
2. 類別評論重點詞分析:

a. 使用者輸入:

i. Number : 想要觀察的重點詞數量(長條圖的數目)

ii. Type : 輸入想要觀察的地標類型 (Ex: health, gym, food, ...)

b. 輸出結果:根據該地標類型, 分析與該類型有關的所有評論中, 不同rating下該重點詞的分佈情形



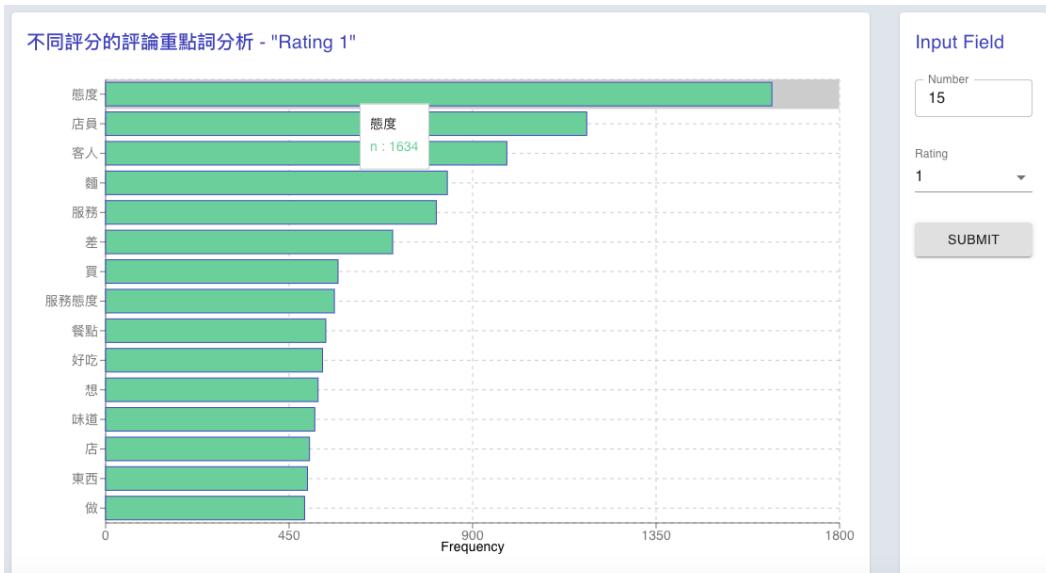
3. 評分評論重點詞分析:

a. 使用者輸入:

i. Number : 想要觀察的重點詞數量(長條圖的數目)

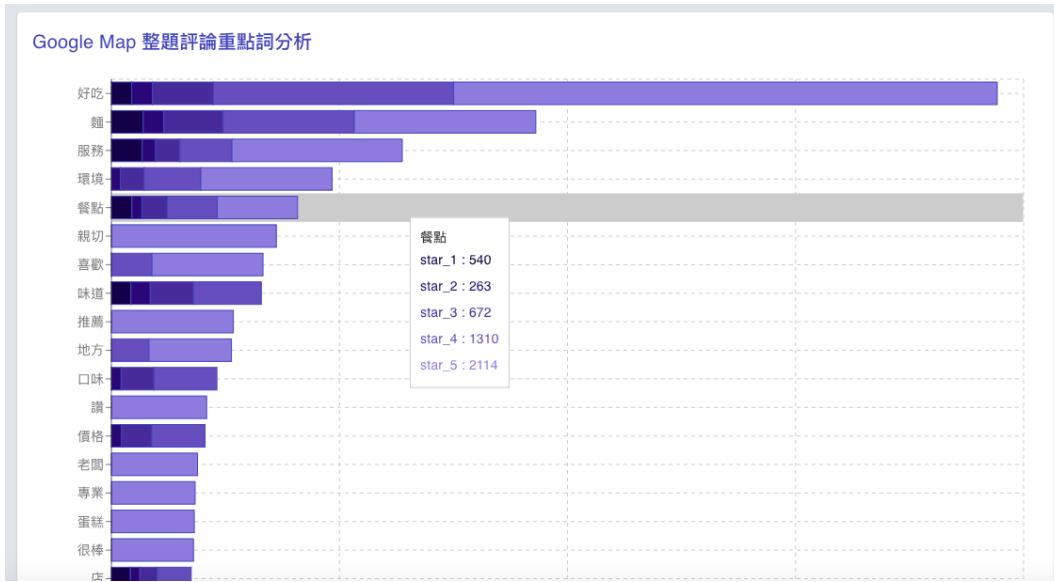
ii. Rating : 想要查詢的特定rating

b. 輸出結果:根據輸入的rating, 分析在該rating下的所有評論中, 評論重點詞的分佈情形



4. Google Map 整體評論重點詞分析:

- a. 輸出結果: 分析Google Map整體的評論中, 在各個rating下出現的各個評論重點時的分佈頻率情形並排序出來

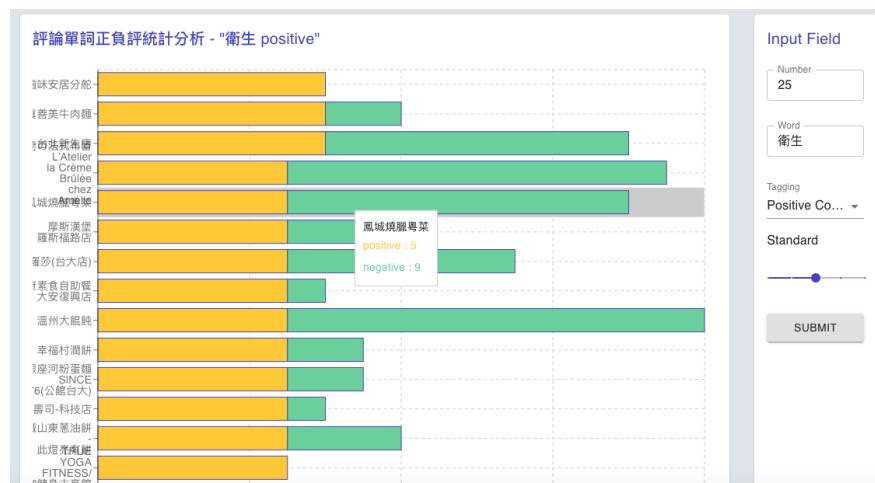


5. 評論單詞正負評統計:

- a. 使用者輸入:

- Number: 想要觀察的地標數量(長條圖的數目)
- Word: 輸入想要在評論中查找的重點詞 (Ex: 衛生、燈光、食物...)
- Tagging: 輸入想要觀察的評論狀況, 例如選擇Positive Tagging, 便會依照這個字詞出現在正評中的數量進行排序, 著重觀察出現在正評的比例; Negative就是出現在負評的情形, 而Total是以正負評加起來的和作排序
- Standard: 自行設定正負評的分界, 例如數字3便是將rating 1-3設定為負評, 而rating 4-5設定為正評

- b. 輸出結果: 依照輸入的字詞, 可以看出該字詞在不同地標的評論中出現在正評或是負評的比例狀況



討論與貢獻

經過了以上的分析，我們驗證了 google map 評論、評分、地標種類三者之間的確有關係：不同的地標類型一定程度上影響了評分、不同地標種類的評論重點詞都不太一樣、不同評分的評論重點詞也可某種程度上看出正面或負面的語氣。我們藉由分析實際資料得出以上的結果，而不是隨意猜測，希望藉由我們的專題可以讓其他跟我們有相同疑惑的人可以解惑。

附錄

組員分工

| 姓名 | 負責內容 | 投入程度 (1 ~ 6) |
|-----|---|-----------------|
| 劉奇聖 | 網頁爬蟲、程式碼整理、資料前處理、不同評分的評論分析、不同種類的評論分析、投影片製作、前端畫面微調 | 6 |
| 陳冠辰 | 資料前處理、資料分析、作圖、投影片與報告 | 5 |
| 葉曜德 | 資料後處理、網頁前後端架設、部分資料分析與作圖 | 5 |
| 韓承霖 | 整理投影片與報告 | 3 |