Further analysis in Kyoto & Tokyo Restaurant Reviews Dataset

組員：許博閔、白佳灝

Introduction

從網路上可以搜尋到很多種對於餐廳評論的評論，像是Yelp、TripAdvisor以及Tabelog等…。因此，我們可以發現，對於相同的餐廳在不同的網頁，也會有不同的評論。假如今天我們想對的特定的餐廳有了解，並上網搜尋，發現很多時候，對於同一家餐廳的評論常常會有一些出入，那哪一個網頁評論是值得我們相信的呢？

以下我們將舉一個例子，請參考下圖。我們發現在Tabelog上，網友給予「草喰 なかひがし」這家餐廳的評價是京都的第一名，滿分5分下得到4.69分。但是，在TripAdvisor上，「草喰 なかひがし」的分數似乎就沒有那麼高，只有4.5分(滿分亦為5分)，而在評論的排名裡也並非是京都第一好吃的餐廳。

右圖是草喰 なかひがし在Tabelog上的評論，其總分得到4.96且為京都第一好吃餐廳。
左圖是草喰 なかひがし在TripAdvisor的評論，其總分得到4.5。

所以有那些因素會影響排名呢？來自不同國家的人會不會因為口味不同而對同一家日本餐廳有著不同的評論？或又是一些人們刻意製造假評論呢？(像是之前選舉熱潮，挺韓國瑜的粉絲對支持陳其邁的肉粽店狂刷負評)而我們又該如何辨別這些評論？使得我們可以有一個公正的態度看這家店。

OUR GOAL

我們希望當我們完成這份報告時可以回答以下問題。

哪家網站的評論是我們可以相信的？

首先，我們認為一家好的餐廳，它在不同的網頁皆有著相當不錯的評論，若結果如此，則在這幾家的網頁上即沒有存在假評論。而在這次報告中，我們嘗試去尋找一個函數可以轉換R餐廳在A網站的評論分數到B網站的評論分數，若R餐廳在A和B的網頁上都有著不錯的表現，則說它應該為一家不錯的餐廳。

例如：在Tablelog上，餐廳R得到3.5分，經過轉換的函數後，我們得到餐廳R在TripAdvisor應該會得到4分左右。若此時發現真實的TripAdvisor上顯示R餐廳的分數若遠大於4或著遠小於4，那對於R餐廳在TripAdvisor的評論可能存在一些錯誤。

Data collection

我們分別蒐集了京都與東京餐廳在Tablelog以及TripAdvisor的資料。Tablelog抓了前1200家餐廳，因此我們把所有餐廳的資料都抓下來；另外，對於每個城市，在TripAdvisor則有上萬家的餐廳，因此我們抓了依排名前100個page的餐廳(大約3000多家)做為我們的資料。以下分別說明對於此兩個網站，我們截取的詳細內容。

[註]:附上抓取資料的網址。
Tablelog : https://tabelog.com/en/kyoto/rstLst/?SrtT=rt
https://tabelog.com/en/tokyo/rstLst/?SrtT=rt
TripAdvisor : https://www.tripadvisor.com/Restaurants-g298564-Kyoto_Kyoto_Prefecture_Kinki.html
https://www.tripadvisor.com/Restaurants-g298184-Tokyo_Tokyo_Prefecture_Kanto.html

Tablelog

一開始，我們嘗試使用課程提供的PART A dataset「Kyoto Restaurant Reviews Dataset」。然而，我們發現抓取此資料的作者似乎有加入一些條件在擷取此資料，因為在原網站上total rating最高為4.69，但是在此dataset上它的total rating似乎都是3點多，請參考下圖。

JapaneseName	Station	FirstCategory	SecondCategory	DinnerPrice	LunchPrice	TotalRating	DinnerRating
オールデイダイニング　ラジョウ	Kyoto	Buffet style	Cafe	￥4000～￥4999	￥2000～￥2999	3.39	3.2
ステックフリット ガスパール ザンザン	Karasuma	Bistro	Steak	￥3000～￥3999	￥1000～￥1999	3.18	3.06
和馬	Sanjo	Izakaya (Tavern)	Japanese food (other)	￥3000～￥3999	NA	3.28	3.28
お好み焼き 鉄板焼き 三喜	Tambaguchi	Okonomiyaki	Izakaya (Tavern)	￥3000～￥3999	NA	3.14	3.14

此為Kyoto Restaurant Reviews Dataset，紅色框住的地方為total rating。

因此，我們認為此dataset並不夠完整，因此我們重新截取Tabelog上的資料，請參考下圖。如圖所示，除了total rating外，我們還考慮了一些細部的rating，像是dishes、service、atmosphere 以及drink。

此為Tabelog上草喰 なかひがし這家餐廳，網友評分的詳細資料。每一家店，我們都抓取紅色框住的部分。

TripAdvisor

在TripAdvisor的網站上，除了total rating外，我們同樣也考慮了詳細排名(如Food, Service, Value, Atmospher)做為我們在蒐集data上的重點。然而，我們發現有很多的餐廳它的total rating幾乎不是5，即是4.5。因此，我們也考慮了組成total rating的評分細節(包含Excellent, very good, average, poor, terrible)，亦即每個分數項目的人數組成，以下舉草喰 なかひがし這家店為例，請參考下圖。

此為TripAdvisor上草喰 なかひがし這家餐廳，網友評分的詳細資料，紅色框框住的部分為我們抓取每一家店的詳細資料。另外，在這邊我們還考慮了組成total rating的評分細節，此例中評分的人數比例分別是Excellent 54%, very good 26%, average 14%, poor 4%, terrible 2%。

資料擷取實作部分

主要是使用python套件requests和beautifulsoup4讀取網址。關於讀取細節的部分(例如，餐廳名稱、排名、食物總類...等)，主要考慮html中的class以及id來辨別，例如在Tableog中，尋找English name時，就得先進入id位置在rd-header__rst-name-main之裡面的text1部分，以此類推尋找其他需要的feature。然而，並非所有的內容都可以靠前述方法讀取，有些需要使用到json套件，如在TripAdvisor讀取total rating時，得先找到type中的application/ld+json，以及其script，最後才使用json.load()的方式讀取其值。

Tools

Python 3.6.5
numpy 1.14.3
Pandas 0.23.0
SequenceMatcher
beautifulsoup4 4.6.0
requests 2.18.4
json 0.1.1

The methodology used

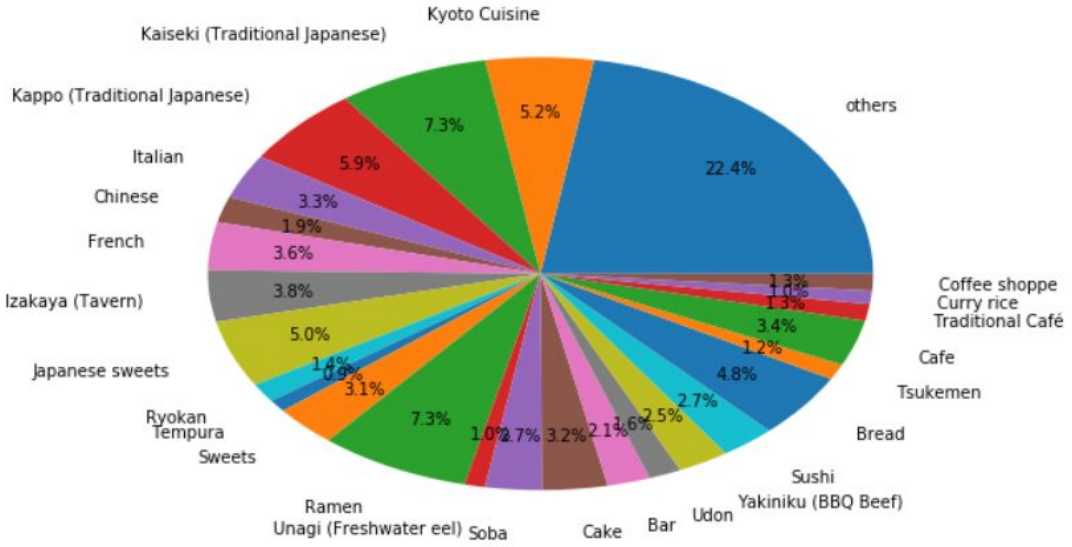
Clustering, Linear transformation

Our implementation and finding

1.Tabelog data : 總共有1200筆資料，從網站上最高分的餐廳到第1200名。

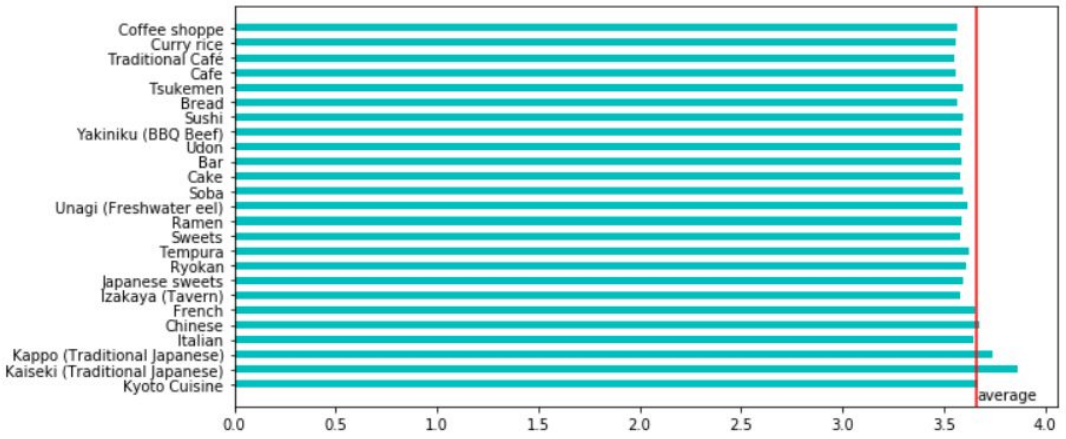
平均分數是3.625，標準差是0.029

(1) 用食物的種類來分群，將餐廳總數小於10的種類歸類到others，得到的結果如下

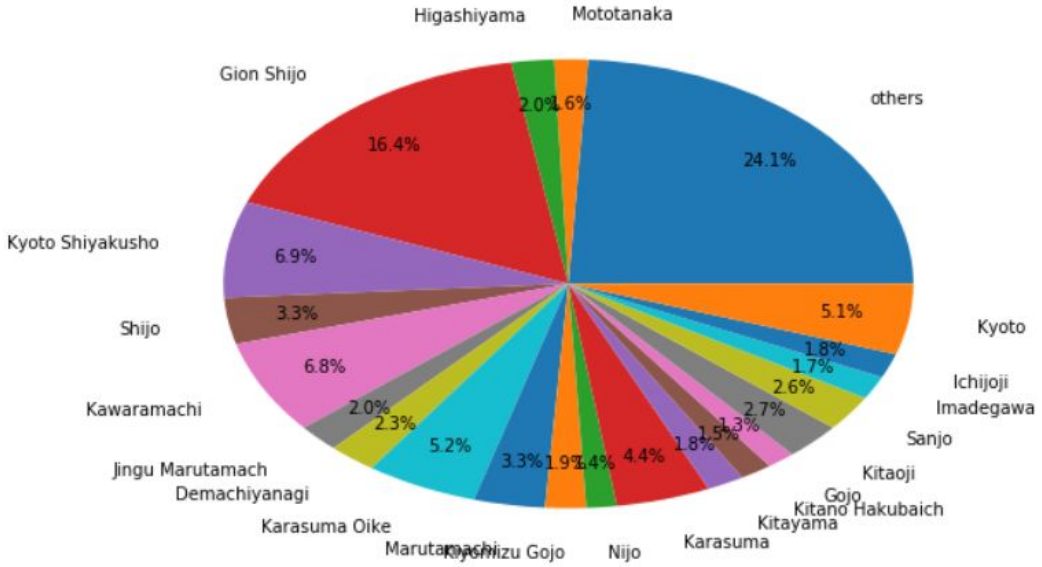


可以看到圖中佔比最多的二類是都佔7.3%的Kaiseki(懷是料理)和Ramen(拉麵)

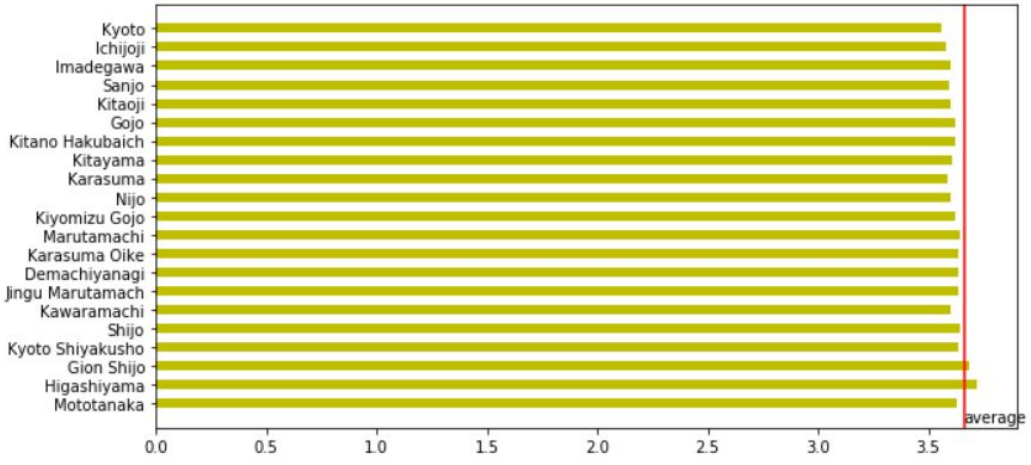
用食物種類分群後，分別計算各種類在rating上表現，得到的結果是：懷石料理的平均分數高達3.88分，第二高的是Kappo(割烹)，兩者皆是日本傳統料理。其餘種類幾乎都比平均分數低，結果如下圖



(2)用餐聽得位置來分群，將餐廳總數小於15的種類歸類到others，得到的結果如下



可以看出最多餐廳位於Gion Shijo。分別計算不同位置在rating上表現，結果只有只有兩個地區的高於總平均rating，分別是Higashiyama和Gion Shijo，結果如下圖。Gion Shijo的狀應該可以理解成競爭越激烈的地方，平均rating分數也會偏高。



(3)比較

a. 若和東京的資料相比，同樣是前1200名，東京的平均負數高達3.89分，比京都的的3.625高出不少

	平均	標準差
東京	3.89	0.053
京都	3.625	0.029

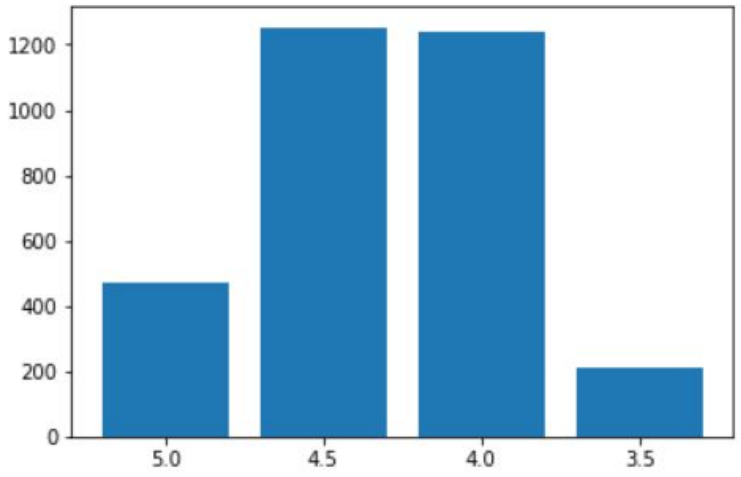
b. 和Kaggle作者得到的結果相比，最大的差異在：kaggle作者的dataset中最多的餐廳種類是Izakaya(居酒屋)。而前1200名的dataset則是懷石料理的餐廳總數最大，可以理解成居酒屋雖然很常見，但要得到很高的rating卻不容易。

2.TripAdvisor data：總共有3000筆，並沒有排名先後之分

平均分數是4.31，標準差是0.166

(1) 實際上TripAdvisor 的評分是分區間的，3000筆資料中total_rating只有5.0、4.5、4.0、和

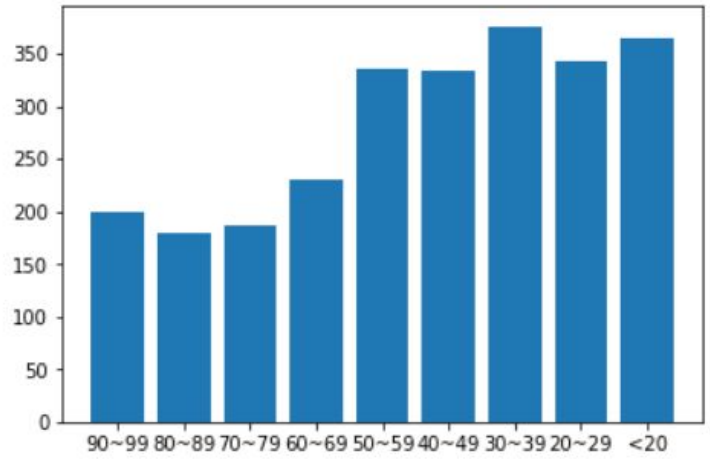
3.5分，實際分布如下:



從圖中可以看出大部分的餐廳分數都集中在4.5和4.0這兩個區間，和Tabelog的結果相比，TripAdvisor確實比較容易得到高分。

(2)TripAdvisor的細部評分

如果比較細部的評分，大約有2500家餐廳有細部的評分中，細部的評分共分成5個等級，從最好的excellent到最差的是terrible。基本上評分都集中在前2好的等第，較少人會給出很差的評價，這似乎也驗證了顧客通常都不會給出太差的評價。下圖是一家餐廳得到excellent評價佔全部評價的比例，通常>85%的餐廳，在total rating會得到5分



3.How to convert rating

原本想要用regression、decision tree或其他上課學到的方法來找出合適的轉換方法，可是這樣都會遇到一個同樣的問題，就是必須同時取得一家餐廳在兩個的評價網站的資料，但這其實是個繁雜的過程，因為有些餐廳可能就只出現在Tabelog中，或是語言或是翻譯上的差異，使得同一家餐廳在Tabelog和TripAdvisor中的名字略有不同，造成困難。

後來在presentation的時候，有同學提到了可以分別計算兩個評分網站的平均rating和標準差，就可以對分數做平移，當作是一種轉換的方法，這樣也可以避免需要同一家餐廳在不同評分網站的分數。我們覺得這是個好方法，所以後續就這麼做。

因此先列出Tabelog和TripAdvisor在rating上的比較，如下:

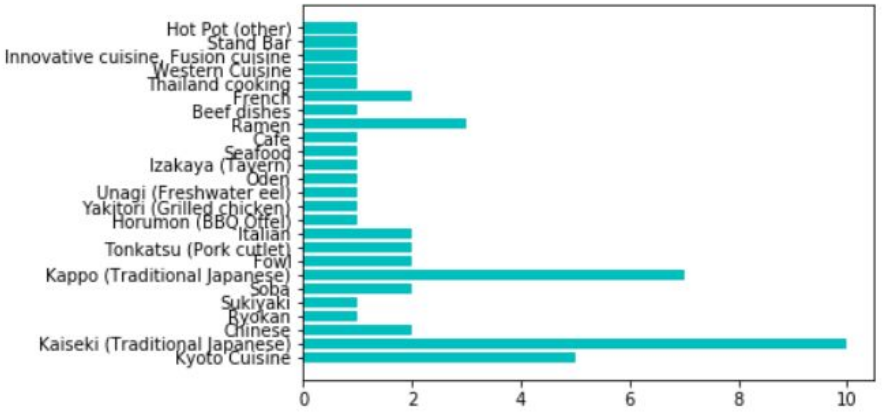
	平均	標準差
Tabelog	3.625	0.029
TripAdvisor	4.31	0.166

因次將Tabelog的rating，平移增加0.7分，理想上應該不會和TripAdvisor的結果差太多。

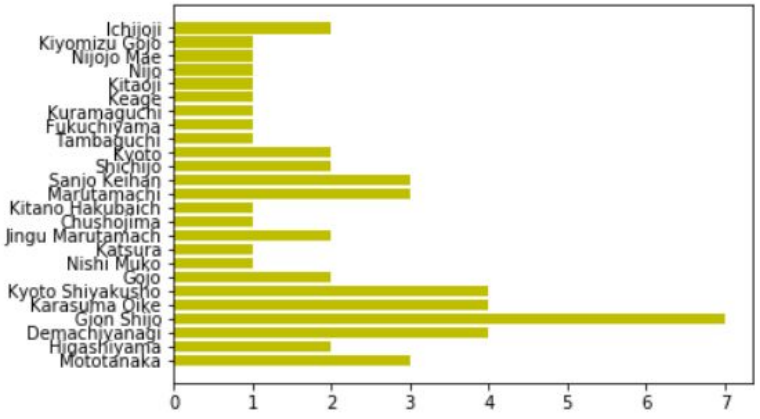
順帶一提，因為翻譯的問題，兩個網站的英文拼音可能會略有不同，因此用到了SequenceMatcher，用來衡量兩個string的相近程度，兩者越相似分數越接近1。把我們蒐集到的資料拿去比對後，總共找到了227家餐廳同時有在Tabelog和TripAdvisor的dataset內。

將這227餐廳的rating做平移後，總共有52家餐廳的rating不符合預期，佔23%，其中有23家餐廳在Tablog的分數偏高，而剩下的29家餐廳在Tablog的分數偏低。其實有接近八成的餐廳rating都在預期之內，代表這個轉換方法其實不差。

對這52家餐廳分析，以食物種類來看時，最常出現的兩類，正好就是Tablog中平均分數的前二名: Kaiseki(懷石料理) 和 Kappo(割烹)，可見對TripAdvisor的使用者來說，或許有些日本傳統料理不合他們的胃口，或是沒有達到他們的預期。相反的，和懷石料理並列最多餐廳的Ramen(拉麵)，就比較少發生rating不合預期的狀況。分析結果如下圖:



如果改用地區來做分類的話，其實並沒有哪個地區特別突出，Gino shijo會偏多應該是因為該地區的餐廳數量本來就比較多。分析結果如下圖:



如果把overated和underated的餐廳分開來做分析，其實得到的結果和前面差不多，會發生rating不符合預期的狀況比較常發生在懷石料理和割烹上，而和餐廳的位置沒甚麼關係。

The contribution of our work

我們分析了Tablog和TripAdvisor的資料，並試著對兩者的rating做轉換，得到了以下結果:

1.Tablog的資料中，懷石料理和割烹的rating特別突出
2.TripAdvisor的使用者，通常不會給出太差的評價，rating都偏高
3.用平移的方法，對兩網站的rating做轉換，發現大部分的餐廳在兩網站上的rating不會有太大的差異，代表這個轉換方法其實不賴，而我們也可以選擇相信網站的評論
4.分數有落差的餐廳，最常發生的正好是懷石料理和割烹，意味著Tablog和TripAdvisor的使用者之間，可能對這兩類的餐廳的要求或評價有落差
5.基本上，餐廳的位置對rating沒甚麼影響

Conclusion

無論是Tablog還是TripAdvisor，應該都是有口碑的評價網站了，大家都可以依自己的喜好或規劃，找到自己想要吃得餐廳，不需要太害怕踩雷。