Tzu-I Lee
February, 20, 2016

SI 601 Individual Project

The relationship between restaurants' rating and the numbers of check-in.

Motivation

I would like to propose a project to investigate the relationship between the rating of a restaurant and the numbers of check-in of the restaurant in the social media website. To be specific, I'd like to utilize the dataset from Yelp as well as the check-in data from Instagram. It is interesting to understand whether the rating of the restaurant will influence people's willing of check-in. Sometimes when we see our friends checked in at certain restaurant, we will think that probably the restaurant is a good place to go. Therefore, I want to investigate the correlation between the rating and the numbers of check-in in terms of restaurants.

Data Sources

The data sources are Yelp academic dataset and Instagram API.

Yelp:

I downloaded the the challenge dataset from Yelp (https://www.yelp.com/dataset_challenge). I only utilized the business dataset which contained the location, review_counts, rating and city of a business. The business dataset includes not only the data of a restaurant, but also other businesses. The file is a json file which contains 77444 data. The format of the file is a json object, as a result, I use python to load json file.

Instagram API:

The url of Instagram API is: https://www.instagram.com/developer/. Instagram provides API for python (https://github.com/lnstagram/python-instagram). At first, I used the location_search function to attain location_id of certain location based on the longitude and latitude data from Yelp. Then, I used the location_recent_media function to retrieve the numbers of check-in in certain location by passing location_id as parameter. To use Instagram API, I have to first install the api and then import simplejson and InstagramAPI in the python file.

Data Manipulation Methods

Process:

When using Instagram API, we have to firstly enter a location and then we can search the numbers of check-in of that location. As a result, I firstly parse data from Yelp to retrieve the longitude and latitude of a certain place and then input the coordinate to Instagram API to retrieve the check- in data. More detailed description of the manipulation methods are as follows.

Yelp:

1. Minimize data: The data from Yelp are too many, as a result, I only parse data from one state by filtering out business from other states.
2. Parse only restaurant data: The business object contains categories data. Since I only focus on restaurants, as a result I only get data with restaurants categories.
3. Review counts: I think if the review counts of a restaurant are too small. The rating of a restaurant might be biased, as a result, I only include data with review counts larger than 10.
4. Store data: I stored a restaurant's name, rating, review counts, longitude and latitude.

Instagram API:

1. Search for location id in Instagram: I used the longitude and latitude retrieved from yelp to search for the location id in Instagram.
2. Validate that the locations from Yelp and Instagram are the same businesses: It is possible that places with same coordinate are not the same businesses. As a result, I use regular expression to match the name of the restaurant from yelp with the name from Instagram. Because the name of a certain restaurant might not 100% match the name from Instagram, but the names will be similar. As a result, I split the name string from Yelp and then check if there is a word match the name of Instagram. If not matched, I'll exclude the data.
3. Numbers of check-in: I then use *api.location_recent_media* function to retrieve the recent check-in in certain location.
4. Store data: I stored the data from Yelp and the numbers of check-in from Instagram and output a csv file.
5. Final data: Name, Rating, Review_counts, Check-in.

Analysis and Visualization

Methods:
I use Pearson product-moment correlation coefficient to analyze the relationship between the numbers of check-in and rating. Also, I think that review counts might be correlated with the rating. Because if a restaurant is very famous, it is possible that more people will go to that restaurant and add review on Yelp.

In order to examine the correlation, I also draw scatter plot to see whether the correlation is due to outliers or not.

Results:
I retrieved 360 valid data for this project (see Table). The result showed that there is no significant correlation between check-in and rating (see Figure). However, the review counts and the rating is modestly correlated (r= 0.136).

Discussion:
The results showed that there is no relationship between the numbers of check-in and rating. Possible explanation for the results is that probably for people who likes to check-in, they will check-in in wherever they go. Another explanation is that maybe the reason for people to check-in is for the discounts, but not related to the overall rating of a restaurant.

Table:

Correlations

		Stars	Reviews	Checkin
Stars	Pearson Correlation	1	.136**	-.059
	Sig. (2-tailed)		.010	.262
	N	360	360	360
Reviews	Pearson Correlation	.136**	1	.041
	Sig. (2-tailed)	.010		.438
	N	360	360	360
Checkin	Pearson Correlation	-.059	.041	1
	Sig. (2-tailed)	.262	.438	
	N	360	360	360

**. Correlation is significant at the 0.01 level (2-tailed).

Tzu-I Lee
February, 20, 2016

Figure: