Deyao Kong, Rui Lin
DSCC 440
10/27/2021
Prof. Jiebo Luo

What is the Best Location for Your Next Business?

Introduction

Yelp provides people with a comprehensive knowledge for different restaurants. Customers now can rate different businesses through the app, and based on the information given from the Yelp App, other people get to choose their favourite restaurants nearby. Most previous works mainly focus on how people will choose different types of restaurants, whereas we want to analyze the dataset in an opposite direction.

Location is the most important feature for each business man to consider before starting one. Based on different build-in features of different restaurants, the main purpose of our project is to find the optimal location for a certain restaurant to start.

Using the Yelp public dataset, we will acquire related attribute lists for restaurant's features including price range, amenities, etc to label different areas. Our proposed recommendation algorithm will return the top 10 recommended locations for a certain business, and we will use the ground truth from the dataset to test the accuracy.

Problem Statement:

What is the best location to start a new business?

Data Acquisition

In our study, we will use the Yelp Open Dataset (https://www.yelp.com/dataset). The downloaded dataset contains multiple json files, the only one we need is bussiness.json, which gives us some information about all businesses on yelp.

The following table contains a brief description of the json file.

bussiness.json

Attribute Name	Description
business_id	string, 22 character unique string business id
name	string, the business's name
city	string, the city

state	string, 2 character state code, if applicable
latitude	float, latitude
longitude	float, longitude
stars	float, star rating, rounded to half-stars
review_count	integer, number of reviews
is_open	integer, 0 or 1 for closed or open, respectively
attributes	object, business attributes to values. note: some attribute values might be objects
categories	an array of strings of business categories
hours	an object of key day to value hours, hours are using a 24hr clock

Data Cleaning and Preprocessing:

We want to first partition the whole dataset into different subsets based on the business's location (per city). We will use the top five cities with the largest number of tuples as our samples.

City Name	State Name	Count
Portland	OR	18196
Orlando	FL	10635
Atlanta	GA	12611
Vancouver	BC	10299
Austin	TX	22412

From the original .json file, we have generated 5 new .csv files, each containing tuples in 5 cities listed above. We will then clean the dataset.

These are attributes we will need for the dataset: Business_id, Name, City, State, Latitude, Longitude, Stars, Is_open.

We will also choose valid attributes for classification based on its frequency and how they are related to our goal.

Attributes: We are going to use One-Hot encoder to transfer the categorical data into binary attributes for later calculations.

RestaurantsTableService	WiFi	RestaurantsReservations
WheelchairAccessible	RestaurantsPriceRange2	Ambience
HasTV	Alcohol	RestaurantsTakeOut
NoiseLevel	RestaurantsAttire	RestaurantsDelivery
GoodForKids	Music	CoatCheck

Categories: We only want to study restaurants' distribution, so we eliminate all other kinds of business.

After cleaning, the stats of current datasets is:

City Name	State Name	Count
Portland	OR	5729
Orlando	FL	3748
Atlanta	GA	4179
Vancouver	BC	4275
Austin	TX	4956

Methodologies:

We will first use clustering to label our restaurants into multiple classes (refer to actual location in selected city). The main method used here is k-means clustering since we are mainly dealing with binary attributes. We will use both Euclidean distance measure and cosine dissimilarity to determine the similarity between each data point, and we will compare their permanence later in the analysis part. We may also utilize a density-based clustering method such as DBSCAN or OPTICS for comparison. During the clustering, a k-fold cross validation process will be implemented to evaluate the accuracy of both methods. After labeling the dataset, we will use ridge regression (or other recommendation algorithms such as neural networks) to recommend new business with a list of top 10 locations, comparing the result with the ground truth from the original Yelp dataset.

Main method used:

K-means
DBSCAN/OPTICS
K-fold cross validation
Ridge Regression (or other recommendation algorithms)

Related Work:

The optimal placement problem is also addressed by several articles and papers. Previous work also introduced some alternative methods to achieve the same goal. For example, *Where to Place Your Next Restaurant? Optimal Restaurant Placement via Leveraging User-Generated Reviews*[4] had utilized a density-based clustering method (OPTICS) to increase the accuracy by clustering the geometrical data. In addition, several other regression models were proposed to generate the recommendation list such as SVR and GBRT. We will also consider utilizing such methods to see the permanence difference.

References

[1]
A. K. P, S. S. G, P. K. R. Maddikunta, T. R. Gadekallu, A. Al-Ahmari, and M. H. Abidi, “Location Based Business Recommendation Using Spatial Demand,” *Sustainability*, vol. 12, no. 10, p. 4124, May 2020.
[2]
R. Deshpande, *How to Choose the Ideal Site for Designing Your Restaurant Using Data Science*, Aug 10, 2020. Accessed on: Oct 27, 2021. [Online].
Available:
https://medium.com/swlh/how-to-choose-the-ideal-site-for-designing-your-restaurant-using-data-science-2cbfb9853f93
[3]
Y.Ding, *Restaurant Location and Planning Study Using Machine Learning*, Oct 25, 2020. Accessed on: Oct 27, 2021. [Online]
Avaliable:
https://www.linkedin.com/pulse/restaurant-location-planning-study-using-machine-learning-yimu-ding/
[4]
F. Wang, L. Chen, en W. Pan, “Where to Place Your Next Restaurant? Optimal Restaurant Placement via Leveraging User-Generated Reviews”, 11 2016.
[5]
A. S. M. Tayeen, A. Mtibaa, en S. Misra, “Location, Location, Location! Quantifying the True Impact of Location on Business Reviews Using a Yelp Dataset”, in Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Vancouver, British Columbia, Canada, 2019, bll 1081–1088.