Use location data to look for the best coffee shop candidates

Phan Hai Quang

October 2, 2019

1. Introduction

1.1. Background

Nha Trang is a beautiful city of Vietnam and has a lot of visitors from all over the world. As apart of Vietnam culture, there are many coffee shops. There are around ~400 coffee shops around this small city so that it is high competition for a person to open a new location. Therefore, it is advantageous for people who could analyze the coffee shops density to find out the locations that do not have so many similar shops in an area. The location should not too far from the center of the city. There are also many successful shops available, so that we could analyze the similarity between the candidate locations with such successful coffee shop to narrow down candidates to find the best suitable places.

1.2. Problem

Data that might contribute to determining the best candidates might include the current available coffee shop locations (coordinates of latitude and longitude) so that we could look for the low density areas. Data might also include the nearby venues that might greatly contribute to the coffee shop venue so that not only density ratio but also the important venues nearby are important of the data.

1.3. Interest

The location plays an important role in the entertainment business such as coffee shops where families and friends are there to meet others. So that people who plan to open a new coffee shops are interest in this result.

2. Data acquisition and cleaning

2.1. Data sources

The location data could be extracted from 2 big data sources that are Google Geocoder and Foursquare. We will use Google Geo service to decode location (lat/lng) to get the address as well as from address to location (lat/lng). And we will use Foursquare to search for venues such as hotels, schools, shops,... that are nearby the location.

2.2. Data cleaning

Because we are interested in areas that are around the center of the city, we will discard the information that are more than 6km from the center of the city. The information from Google

Geocoding server contains editable information of location name so that we also need to clean it up to keep the addresses in the same format.

Second, we will get venues from Foursquare for about 500m around the location. There will be some areas that have a lot of coffee shops already. We are interested in coffee shop density so that we will ignore other categories. In this scope of the report, we are interested in “coffee” category from Foursquare. But in practical, we might need to check other categories that might have coffee service such as restaurants, fast food services,..

Thirth, we will discard these areas that have more than 1 coffee shops in 500m. We also do not care so many small venues because such information is bias. People might not input to system if their venue is too small.

Finally, we will discard very expensive location that cost so much on hiring fee. So that we will need to discard such locations from the data.

2.3. Feature selection

After data cleaning, there were about 10 candidate locations that met our criteria. We will add a success coffee shops (high rate coffee shop for example. However, in this report, I am interest in a known coffee shop that I am interested in). Then, we could group locations by Kmean and pick up the cluster that contains the interest coffee shop.

To group locations, we will pick all venues from Foursquare. These information are decoded as one-hot to features. It is summarized in this list:

'American Restaurant', 'Art Gallery', 'Asian Restaurant', 'BBQ Joint', 'Bakery', 'Bar', 'Bistro', 'Breakfast Spot', 'Café', 'Coffee Shop', 'Comfort Food Restaurant', 'Convenience Store', 'Dessert Shop', 'Distillery', 'French Restaurant', 'Greek Restaurant', 'Gym', 'Hookah Bar', 'Hostel', 'Hotel', 'Ice Cream Shop', 'Indian Restaurant', 'Italian Restaurant', 'Japanese Restaurant', 'Jewelry Store', 'Juice Bar', 'Lounge', 'Market', 'Modern European Restaurant', 'Motel', 'Pizza Place', 'Plaza', 'Pop-Up Shop', 'Pub', 'Resort', 'Russian Restaurant', 'Seafood Restaurant', 'Smoothie Shop', 'Snack Place', 'Soccer Field', 'Soup Place', 'Spa', 'Spanish Restaurant', 'Sporting Goods Shop', 'Sports Bar', 'Steakhouse', 'Sushi Restaurant', 'Taco Place', 'Thai Restaurant', 'Tourist Information Center', 'Vietnamese Restaurant'.

3. Exploratory Data Analysis

3.1. Coffee shop density around the center of the city

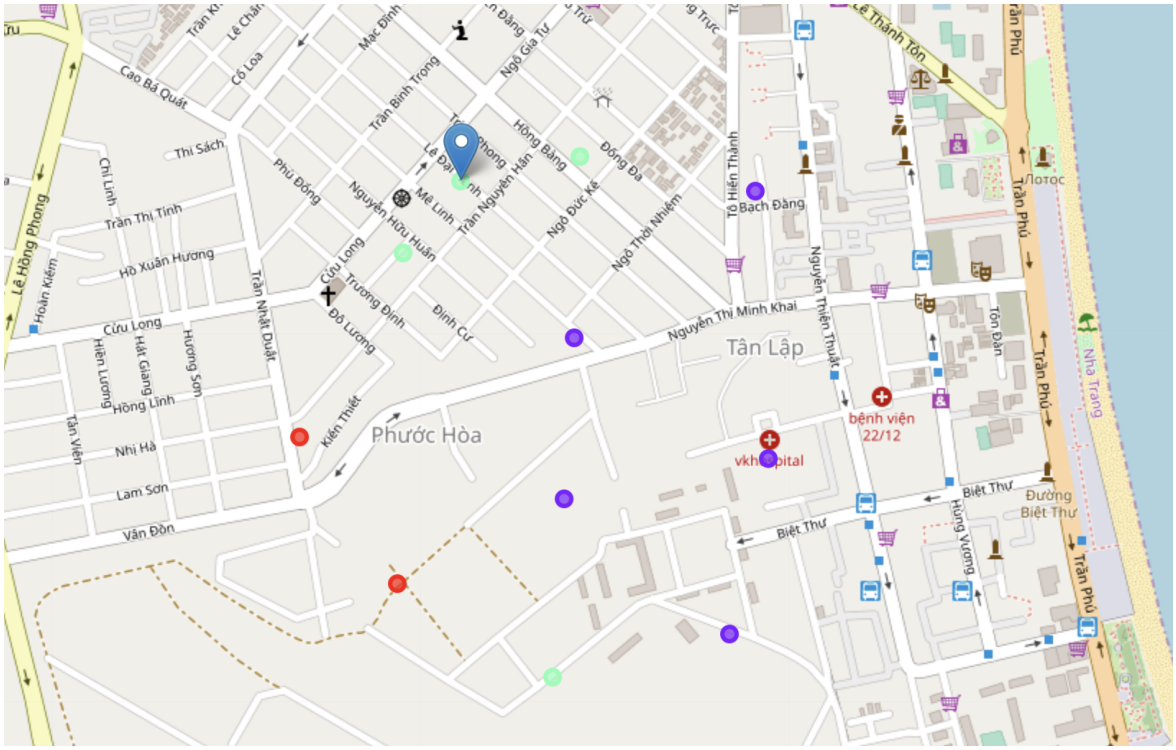
Nha Trang is beach city. It has very long beach so there is no coffee shop on the east. Moreover, there are many hotels along the beach so that it coffee shops are high density at the top and bottom of the center of the city. The below image looks correct.

3.2. Low density area

Because north-west is expensive locations, so that we move our target to south-west about 1km. As a result, we could find out ~10 low density areas

3.3. Add interest coffee shop and clustering

Then, by adding a successful coffee shop (or, an interest coffee shop), we make 3 groups. The “marker” in below image is the interest coffee shop and it labels as green color.



There are 3 other locations that are also marked as green. Their information is shown in below table. Please be noticed that the item 10th is the interest coffee shop that we added to candidate group before clustering.

	Longitude	Latitude	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	109.187359	12.239194	56 Nguyễn Hữu Huân, Tân Lập, Thành phố Nha Tra...	2	Café	Hotel	Vietnamese Restaurant	Seafood Restaurant	Asian Restaurant
1	109.189593	12.233002	Biệt Thự, Phước Hoà, Thành phố Nha Trang, Khán...	2	Hotel	Café	Vietnamese Restaurant	Seafood Restaurant	Hostel
9	109.190000	12.240601	75 Hồng Bàng, Tân Lập, Thành phố Nha Trang, Kh...	2	Café	Hotel	Asian Restaurant	Spa	Convenience Store
10	109.188217	12.240251	40 Le Dai Hanh, Nha Trang, Viet Nam	2	Café	Vietnamese Restaurant	Hotel	Seafood Restaurant	Asian Restaurant

3.4. The study from result

As shown in the above table, the common venue nearby successful coffee shop are Hotel and Restaurant. So, it is easy to understand that the 1st most common venue is Coffee (it is Cafe in Vietnamese). So, it seems that these 2 venues (Hotel and Restaurant) is important venues to consider the place.

4. Cluster Modeling

There are 2 models. First model is to find low density area to find candidates. Second model is to clustering candidates after adding an interest coffee shop.

4.1. KMeans model to find low density area of coffee shops

4.1.1. Problems

The new coffee shop should not be too far from the center of the city. And it should not be in areas that have too much coffee shops around. In other words, we have to determine low density area, we have to know where coffee shops are in the city and their location through the center of the city.

4.1.2. Solution to the problems

To resolve the problems, we firstly determine the center of the city. The center of the city might not be the “middle location” of the map, it should be the most popular location that everyone should know and have a lot of traffic come to/from. As a Nha Trang citizen, I know that it is Thap Tram Huong that is popular building near the beach.

From this point, we could get a list of coffee shops around 6km from the center of the city. Foursquare API returns these information.

By creating a grid around 6km around the center of the city, we can find the number of coffee shops within each area grid to make heatmap as well as identify empty clusters.

4.2. KMeans model to find group of interest coffee shop

4.2.1. Problems

After getting list of low density area, we need to determine the “quality” of each venue. If we already know a good candidate (a successful coffee shop), we try to find out the venues that are the most similarity with that good place.

4.2.2. Solution to the problems

To find the similar venues of the good place, we append the successful coffee shop to candidate list. This good point could be get from some online service that reviews with high rate of quality, focusing on position factor. However, our stakeholders had a successful coffee shop (it means the service quality is the same with the new one), we could use this venue.

By getting all venues around the location candidates, we clean it by making one-hot to a KMeans model to group candidates to 3 groups. Then, we look for the group that contains the added coffee shop.

5. Conclusions

In this study, I analyzed the density of coffee shops around Nha Trang city and find out areas with low coffee shop density. I used Google and Foursquare services to get address, location (lat/lng), nearby venues and filter interest features that are number of coffee shops in the area, venues nearby and their categories. I also realized that the successful coffee shop is nearby Hotel and Restaurant and good location candidates are 3 locations that are grouped in the same cluster with interest coffee shop.

6. Future directions

In this study, I still use some manual information for the decision such as the expensive positions. Moreover, the traffic to location candidates might be hard. Moreover, the extract venues are individually considered, there might be interactions with other factors. These practical information are needed to be considered during final review to decide the best location to open new coffee shop.