Recommendation of New Venues to a Commercial Plaza

Qin Zhang

July 19, 2019

1. Introduction

1.1 Business problem

A client recently bought a commercial plaza in Lanham, Maryland. Most of the units in the plaza have already been under contract, but there is one empty unit. He wants to find a venue to rent out the empty unit as soon as possible, preferably for long-term. In this data science project, a new venue (category) or a few venue candidates (categories) will be recommended for him as his potential future tenant(s).

1.2 Interest

This problem could be interesting to landlords of commercial properties or developers of a new community/neighborhood. It could also be interesting to chain stores that are looking for new locations.

2. Data

2.1 Data Source

The client's property is in Lanham, a census-designated place in Prince George's (PG) county, Maryland. The analysis in this project is based on venue information in PG county, which is obtained from Foursquare location exploration. First, all the names of cities/towns/census-designated places in PG county are scraped from a Wikipedia page link, and the latitude and longitude for each city/town are obtained using Geocoder Python package. Then Foursquare API (application programming interface) is used to explore venues for each city/town in PG county, with a radius of 1200 meters. Information of venues that are close to the client's property (within a radius of 1000 meters) is also obtained from Foursquare.

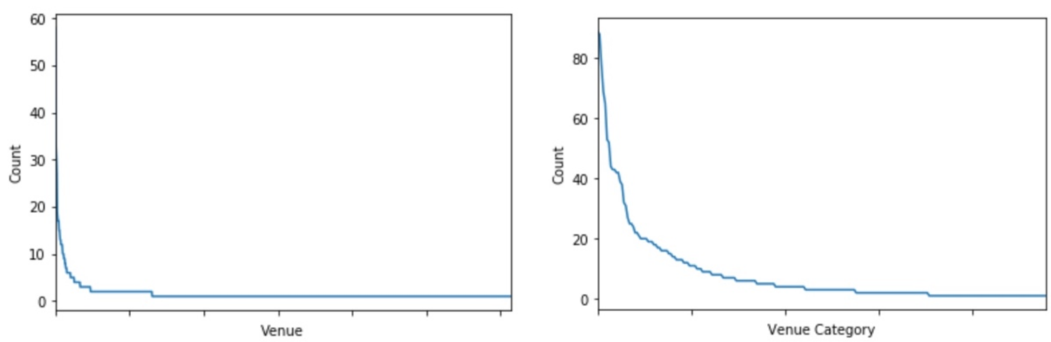
2.2 Data Preprocessing

Data from Wikipedia and Geocoder are put into a table with columns of "Town", "Latitude", and "Longitude", Table 1. There are a couple of places where the latitude and longitude are not available. After dropping them, there are 79 places in PG county that are visualized on the map. It is then observed on map that some places are very close to each other, which will cause a problem when using Foursquare API: some venues are generated multiple times when exploring the closely located neighbor towns and will affect the accuracy of the recommender system. A simple way to reduce the error in this scale (79 places) is to remove the towns that are too close to others. After dropping them, there are 72 places on the map.

Table 1. Head of location data for cities/towns in PG County, Maryland

	Town	Latitude	Longitude
0	Bowie, MD	38.942966	-76.731234
1	College Park, MD	38.980666	-76.936919
2	District Heights, MD	38.857613	-76.889417
3	Glenarden, MD	38.929278	-76.861639
4	Greenbelt, MD	38.993733	-76.883066

Foursquare API is then used to explore the venues at the 72 places and put the data into a table. There are 1228 venues (searched on 7/18/2019) in 240 venue categories, many of which do not frequently appear (Fig. 1). In this project, only "popular" venues (chained venues) and venue categories are considered to reduce the noise. To recommend a specific chained venue, venues shown in more than 5 locations are considered, Table 2(a). To recommend a venue category, in which independent venue can also be included, venue categories shown in more than 10 locations are considered, Table 2(b).



(a) (b)

Fig.1 (a) Venue and (b) Venue Category count in PG county have "long tail" in distribution. Only a small fraction of the venues/venue categories have many locations, which are considered "popular".

Table 2. Head of (a)popular venues and (b)popular venue categories in PG County, Maryland

(a)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Bowie, MD	38.942966	-76.731234	Five Guys	38.943590	-76.734322	Burger Joint
2	Bowie, MD	38.942966	-76.731234	Five Below	38.941971	-76.736021	Miscellaneous Shop
6	Bowie, MD	38.942966	-76.731234	Bath & Body Works	38.944056	-76.734953	Cosmetics Shop
8	Bowie, MD	38.942966	-76.731234	Panera Bread	38.944654	-76.734430	Bakery
10	Bowie, MD	38.942966	-76.731234	LA Fitness	38.943978	-76.736636	Gym / Fitness Center

(b)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Bowie, MD	38.942966	-76.731234	Five Guys	38.943590	-76.734322	Burger Joint
4	Bowie, MD	38.942966	-76.731234	Red Robin Gourmet Burgers and Brews	38.944233	-76.732983	Burger Joint
5	Bowie, MD	38.942966	-76.731234	California Tortilla	38.941848	-76.734349	Mexican Restaurant
6	Bowie, MD	38.942966	-76.731234	Bath & Body Works	38.944056	-76.734953	Cosmetics Shop
8	Bowie, MD	38.942966	-76.731234	Panera Bread	38.944654	-76.734430	Bakery

3. Methodology

3.1 Method

Item based collaborative filtering is used for the recommender system. First, find the patterns of existing venues/venue categories near the client's commercial plaza. Second, build a venue/venue category-venue/venue category matrix to determine relationships between pairs of venues/venue categories using k-nearest-neighbors (knn) model: i.e. wherever there is the home depot, there is highly possible a pizza hut or wherever there is movie theatre, there is highly possible a fast food. Thirdly, exam the matrix and match the data of the client's commercial plaza to recommend new venues/venue categories.

3.2 Existing venues/venue categories near the client's commercial plaza

Using Foursquare API, twenty-six venues are found near the client's commercial plaza (searched on 7/18/2019), Fig. 2. There are only seven "popular" venues: *Chipotle Mexican Grill*, *Ledo Pizza*, *Dollar Tree*, *AT&T*, *7-Eleven*, *Shell*, and *McDonald's*. And there are 17 "popular" venue categories: *Café*, *Sandwich Place*, *Mexican Restaurant*, *Spa*, *Pizza Place*, *Discount Store*, *Department Store*, *Mobile Phone Shop*, and so on.

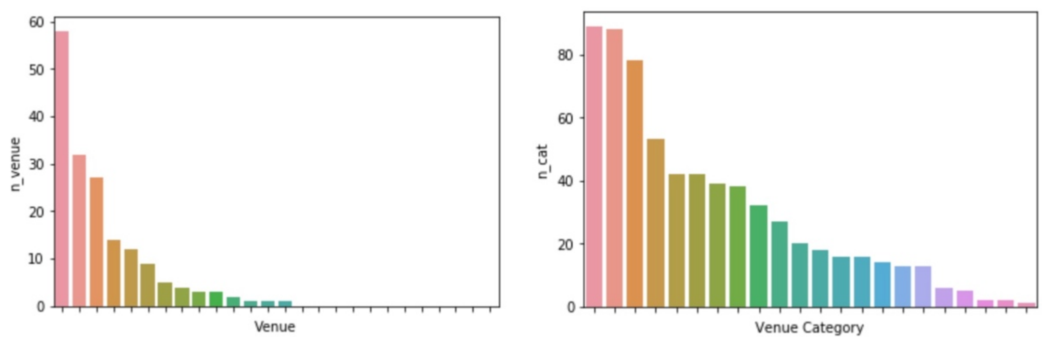


Fig.2 (a) The count of nearby venues in PG county has "long tail" in count distribution. Only 7 out of the 26 venues are considered "popular" (have 5 or more locations in PG county), while (b) most of the nearby venue categories are "popular" (have 10 or more locations in PG county).

3.3 Relationship between pairs of venues/venue categories

Relationships between pairs of venues/venue categories are determined by k-nearest-neighbors (knn) model. Knn relies on item feature similarity. The input of the knn model is an m x n array, where m is the numbers of venues/venue categories and n is the number of cities/towns. The dataframes shown in Table 2 are transformed to the dataframes with venues/venue categories as index, cities/towns (neighborhoods) as columns, and venue/venue category densities as content, Table 3.

When knn makes inference about a venue/venue category, it calculates the "distance" between the target venue/venue category and all the venues/venue categories in the database, and return the top k nearest venues/venue categories as the closest venues/venue categories. So, we expect to learn something like wherever there is the home depot, there is highly possible a pizza hut or wherever there is movie theatre, there is highly possible a fast food.

Table 3. Head of densities of (a)popular venues and (b)popular venue categories in PG County, Maryland

(a)

Neighborhood	Adelphi, MD	Andrews AFB, MD	Beltsville, MD	Berwyn Heights, MD	Bladensburg, MD	Bowie, MD	Calverton, MD	Camp Springs, MD	Cheverly, MD	Chillum, MD	...	Silver Hill, MD	South Laurel, MD	Suitland, MD
7-Eleven	0.285714	0.0	0.045455	0.071429	0.333333	0.000000	0.0	0.166667	1.0	0.153846	...	0.105263	0.0	0.166667
ALDI	0.000000	0.0	0.045455	0.000000	0.000000	0.000000	0.0	0.166667	0.0	0.000000	...	0.000000	0.0	0.000000
AT&T	0.000000	0.0	0.000000	0.000000	0.000000	0.043478	0.0	0.000000	0.0	0.000000	...	0.000000	0.0	0.000000
America's Best Contacts & Eyeglasses	0.000000	0.0	0.000000	0.000000	0.000000	0.043478	0.0	0.000000	0.0	0.076923	...	0.000000	0.0	0.000000
America's Best Wings	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.2	0.000000	0.0	0.000000	...	0.052632	0.0	0.000000

(b)

Neighborhood	Accokeek, MD	Adelphi, MD	Andrews AFB, MD	Aquasco, MD	Baden, MD	Beltsville, MD	Berwyn Heights, MD	Bladensburg, MD	Bowie, MD	Brandywine, MD	...	Springdale, MD	Suitland, MD	Summerfield, MD
American Restaurant	0.00	0.0	0.166667	0.0	0.333333	0.023810	0.023256	0.000000	0.016949	0.0	...	0.0	0.0	0.04
BBQ Joint	0.25	0.0	0.000000	0.0	0.000000	0.000000	0.023256	0.000000	0.000000	0.5	...	0.0	0.0	0.00
Bakery	0.00	0.0	0.000000	0.0	0.000000	0.071429	0.000000	0.043478	0.033898	0.0	...	0.0	0.0	0.00
Bank	0.00	0.0	0.000000	0.0	0.000000	0.071429	0.046512	0.000000	0.033898	0.0	...	0.0	0.0	0.00
Breakfast Spot	0.00	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.016949	0.0	...	0.0	0.0	0.00

In this study, three inferences are made by the knn model: from a venue to venues, from a venue category to venue categories, and from a venue to venue categories. Tests are made with well-known venue and venue categories. First, "Starbucks" was tested as the target venue, the model returned the closest venues as "*Starbucks, Chipotle Mexican Grill, T-Mobil, Safeway, CVS pharmacy, AT&T, ALDI, Capital One Bank, Target*". The test is good, because the closest venue should be itself, and we know that some Safeway have built-in Starbucks in the area. Then "Hotel" and "Gym/Fitness Center" were tested as the target venue categories, the model returned the closest venue categories as "*Hotel (itself), Restaurant, Gas Station, Convenience Store, Italian Restaurant, American Restaurant, Coffee Shop, Pizza Place, Mobile Phone Shop, Breakfast Spot, Gym/Fitness Center*" and "*Gym/Fitness Center (itself), Burger Joint, Donut Shop, Mobile Phone Shop, Mexican Restaurant, Bank*" respectively. It makes sense since hotels usually have built-in restaurants, coffee, convenience store and gym, and gym is often seen surrounded by many restaurants. Again, "Starbucks" was tested and inferred to the closest venue categories to be "*Starbucks (itself), Coffee Shop (its own category), Mexican Restaurant, Gym/Fitness Center, Bank, Mobile Phone Shop, Pharmacy, Italian Restaurant, Café, Restaurant*", which is consistent with its closest venues. Recommendations can be made from all three inferences.

4. Results

Now, the knn model will list the closest venues/categories to the venues/categories near the client's commercial plaza. Comparing the closest venues/categories and the existing venues/categories, it is expected to see many venues/categories coincide, the venues that are on the closest venue/category list while not on the existing venue/category list are the ones we want to recommend to the client.

First the nearby venues were used to infer the closest venues, and *CVS* and *Redbox* were returned on the top list while they were not in the existing venue list. However, when I double checked with google map, I found there already are *CVS* and *redbox* right next to the commercial plaza that Foursquare didn't include. Missing information may affect the accuracy of the results, see Discussion section. The good thing is that it, in a way, proves our recommender system working. Looking at other venues on the closest venue list, maybe another fast food place (Subway, Burger King, Wendy's, Dunkin's...) would be recommended.

Then the closest venue categories were inferred by the nearby venue categories and venues respectively. By comparing the results and the existing categories, a *Chinese Restaurant* would be recommended.

5. Discussion

This simple recommender is working. We would recommend *Chinese Restaurant* and *Fast Food Restaurant (Subway, Burger King, Wendy's, Dunkin's...)* to the client. But there are a few problems that may affect the accuracy: First, the venue information is not complete or accurate. The search is within a certain distance of each town center, therefore it doesn't cover the whole area of the county. Furthermore, some venue information from Foursquare is missing or incorrect. For example, some existing venues are not found, or some venues are assigned to multiple venue categories. Secondly, there are duplicated venues, because they belong to multiple towns and are generated multiple times during the search. A better data cleaning can help solve the problem. There are also limitations to the approach: Only popular venues (chained) or venue categories can be recommended. Less known venues, papa mama stores or new venues are ignored. Also, it is hard to recommend an existing venue category even the area may need more venues of the same category. Ratings of the venues are not considered for now, which could be helpful for better recommendations.

6. Conclusion

In this study, item based collaborative filter method was used to build a recommender system to recommend which venues/venue categories would be added to a client's commercial plaza. Data is obtained from Wikipedia and Foursquare API, and a k-nearest-neighbors model was built to analyze the relationship between venues/venue categories in the county where the client's commercial plaza is located. The model was tested with well-known venues and venue categories, and then applied to the client's commercial plaza to make recommendations. This recommender system can be useful to help landlords/developers make decisions or get inspiring ideas.