Capstone - The Battle of Neighborhoods

Peiyu Xiao

Introduction

Toronto is a vital city that the technology industry is booming which attracting countless young people for seeking opportunities. The data shows that as the third largest city for tech talent in North America, Toronto had a net inflow of 117,000 people in 2018. Large numbers of migrant workers have also brought a huge market for housing rentals. However, mostly when people move to a new city, they have to decide where to live in a short time. The process for housing agent may not transparent enough which may cause dissatisfaction. Last but not least, renters may have demand that are even unrealized by themselves. This project introduces an extremely convenient solution in location choosing of housing rentals - it is so transparent that enable real estate companies communicate smoothly to clients with a list of venue categories, and clients can simply tick on the list, which is of great value to Toronto real estate companies.

Data

Data Source

In this project I will use the data of Toronto shows below to look at the algorithm of the project with building a recommendation system. The data used in this project are:

1. Postal Code information of Toronto, categorized in Borough and Neighbourhood from Wikipedia:(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M);
2. Geospatial data of Boroughs (http://cocl.us/Geospatial_data);
3. Venues and relevant data supported by Foursquare API;
4. Random user data for examining the model.

Data Cleaning & Preparation

Preparing and Obtaining the Data. First, scrap the information about boroughs and neighborhoods in Toronto from the following Wikipedia page: wiki = 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'. Pandas' command "read_HTML" to read HTML tables into a list of DataFrame objects and remove cells with a borough that is "Not assigned". Then substitute unnamed neighborhoods for the name of relevant borough. After that, group the table by the postal code.

Out[5]:

	PostalCode	Borough	Neighborhood
0	M1B	Scarborough	Malvern, Rouge
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

The next step is to add the geographical coordinates (Latitude and Longitude) of each postal code to our data. Some prepared date from http://cocl.us/Geospatial_data will be used, but it is possible to use geocoder as well.

Out[7]:

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

In the prototype, use the Foursquare API to get the venues for each neighbourhood, after that I limit the amount of venues per neighbourhood to 100 and the range from the centre of the neighbourhood to 500 m. With this API I can get all the venues for each neighbourhood and group them for each neighbourhood.

Out[17]:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Agincourt	4	4	4	4	4	4
Alderwood, Long Branch	8	8	8	8	8	8
Bathurst Manor, Wilson Heights, Downsview North	23	23	23	23	23	23
Bayview Village	4	4	4	4	4	4
Bedford Park, Lawrence Manor East	27	27	27	27	27	27
Berczy Park	57	57	57	57	57	57
Birch Cliff, Cliffside West	4	4	4	4	4	4
Brockton, Parkdale Village, Exhibition Place	25	25	25	25	25	25
Business reply mail Processing Centre, South Central Letter Processing Plant Toronto	19	19	19	19	19	19
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara,						

Get a new table with the neighbourhood as the index and percentage of each category available in that neighbourhood applying OneHot encode in the categories and the mean for the amount venues for each category.

Out[19]:

	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

[20]:
```
toronto_onehot.shape
```
Out[20]: (2156, 272)

Methodology

In this project I will create a data based on a recommendation system that allows a user to choose the neighbourhood, which fit the best to his interests. There are used Toronto's data and random generated for user data to test the system.

1. Model examination
   In this project, we randomly choose a user from *Foursquare.com*. To generate a random user, get a list of all categories available in the city. After that, select a random number from 1 to 10 to represent the number of categories selected by the user. Then, from the list of categories, sample the same amount obtaining the list of categories that our user will have interest in. Create a table with the categories as the columns and one row, where the values are 1 if the user has that category in his list and 0 for vice versa. This will result in a user profile that will be used in the recommendation system.
2. Recommendation system
   Now, to make a recommendation system. Compare our user profile to the table with the neighbourhoods and the mean of value for the amount of venues of each category in it. So, I will multiply both matrix and apply a sum for each row. As the result I get a new matrix with the neighbourhoods and the score for each one of them. The higher the score is, the better the neighbourhood matches the user's interests.

Result

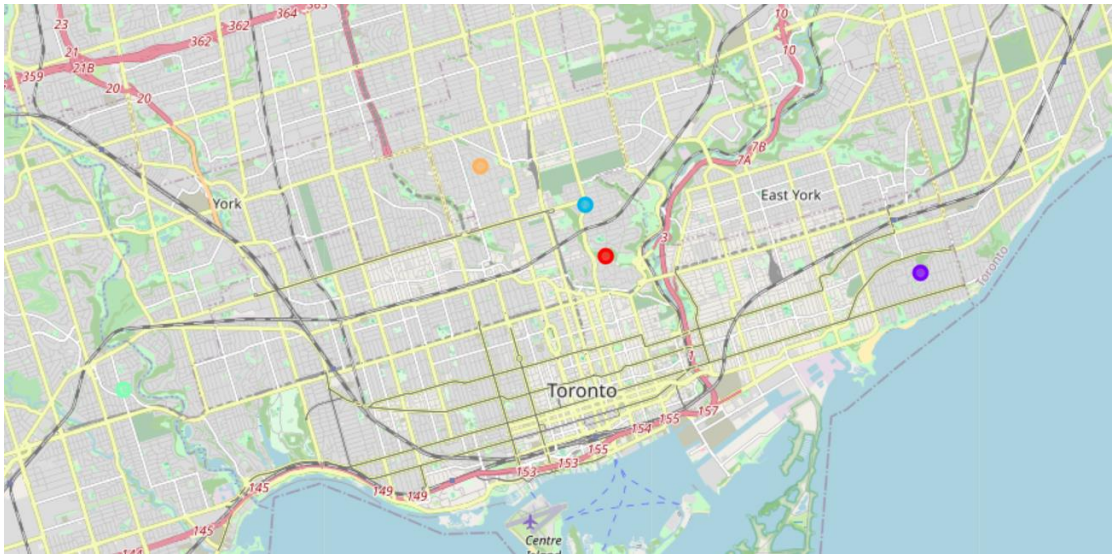
In this project, our user chose the next categories (Figure 1):

```
Out[26]: ['Health Food Store',
          'Liquor Store',
          'Pool',
          'Recording Studio',
          'Creperie',
          'Trail',
          'Art Gallery',
          'Brewery',
          'Board Shop']
```

With this user, we got the following score (descending ranked) (Table 3) and a map (Figure 2) with the best 5 areas, which could fit to our user.

Out[35]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Score
0	M4E	East Toronto	The Beaches	43.676357	-79.293031	0.500000
1	M4T	Central Toronto	Moore Park, Summerhill East	43.689574	-79.383160	0.333333
2	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944	0.333333
3	M5P	Central Toronto	Forest Hill North & West, Forest Hill Road Park	43.696948	-79.411307	0.250000
4	M4W	Downtown Toronto	Rosedale	43.679563	-79.377529	0.250000



Discussion

From this result, we can see that the best neighbourhood for our user is “The Beaches, East Toronto”. From the Table 3 we can see that the 2nd and the 3rd neighbourhoods have the same score of 0.33, while the 4th and 5th neighbourhoods have the same score of 0.25, There is a stable and significant gradient division between three different scores. A reasonable explanation is that for a large emerging city, urban functional zoning needs to adequately diverse.

In the process of checking the number of venues in each neighbourhood, it is found that 6 of the neighborhoods reach the limit of 100 venue. These 6 neighbourhoods are listed below:

First Canadian Place, Underground city
Garden District, Ryerson
Harbourfront East, Union Station, Toronto Islands
Richmond, Adelaide, King
Toronto Dominion Centre, Design Exchange
Commerce Court, Victoria Hotel

In terms of whether the limitation has an impact on validity of result, it is shown that these neighbourhoods account for a relatively small share of the total. If I compare this list with final result, none of these 6 is in result. Another explanation is that although there is an error in exact number of venues, it does shows a trend that there is a wide variety of venues. As a result, it has little impact on validity. However, in this project, the random user has not pick venues among either of these six neighbourhoods, which is a probability event.

Conclusion

This is a sample content-based recommendation system that still need to be improved. Since this recommendation system to an extent rely on the data source from Foursquare.com, the data quality related to venues and categories, as well as user profile greatly depends on the data supplier, which may cause objective problems that are difficult to examine.

References

[1] Scoring Tech Talent in North America 2020 – CBRE: <https://www.cbre.us/research-and-reports/Scoring-Tech-Talent-in-North-America-2020>

[2] List of postal codes of Canada - Wikipedia:_ <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

[3] Geospatial data: <http://cocl.us/Geospatial_data>

[4] Forsquare API

[5] Google Map