Project Report:

Find Similar Hotels

by

Alexis Kyaw

Abstract

In this document we work out a method to find hotels that are similar, between two cities, comparing the venues that can be found in the neighbourhood of the hotel and in the actual hotel vicinity.

We do this by actually performing the analysis for a sample hotel in New York and we find the three most similar hotels in Toronto.

The motivation for this is to complement the existing hotel comparisons based on hotel features with a comparison based on the venues close to a hotel, so a hotel guest should experience a similar experience during his stay when leaving the hotel.

Table of Contents

1. Introduction

We are looking at a common almost "daily" problem city travellers have.

Consider you've been travelling to one city and liked the stay, especially how your hotel was located. Maybe you preferred a neighbourbood a bit outside the city center where it is more quiet and some parks to go for a walk, but still there were a number of restaurants and bars nearby so you could spend the evening right next to the hotel without going back late in the evening from the city center.

Now you want to travel to some city and not find a similar hotel by hotel standards (you don't need much more than bed to sleep and a shower), but find a hotel which is in a similar part of the city and has a similar environment in terms of venues around the hotel.

This the problem are taking a deeper look at in this analysis.

2. Methodology Overview

2.1 Problem Definition

As a (frequent) traveller to a specific city you've become used to spending your stay in a certain neighbourhood, regarding the neighbourhood itself and the venues nearby. So, when travelling somewhere else you want to find a hotel which has similar venues nearby and is in a similar neighbourhood.

So, given a hotel in one city (which also defines the neighbourhood) we want to find similar neighbourhoods in some other city and then find the hotels which have the most similar venues nearby.

The **question** we want to answer is:

Which N hotels (target hotels) in city X have the most similar environment compared to a given hotel (origin hotel) in city Y?

For this sample analysis we'll use a hotel in New York and we'll try to find a list of similar locations to stay in Toronto. We'll just list the hotels by similarity, so the number N of hotel is not fixed and will depend on the data and the origin hotel's location.

Business Relevance

The similarity of the environment of a hotel can help customers of online booking services like Booking.com to find not simply a similar hotel (which we do not look at in this analysis), but find a hotel which is similarly located, based on the venues in it's environment.

This can help improving hotel recommendation significantly because simple hotel recommendations based on collaborative filtering or using content-based recommendations are not aware more than a simple location rating for a hotel and even no information about the environment of the hotel at all. It's probably best to combine more than one recommendation algorithm to get the best result for the customer.

Exclusions

To get a good answer to our question we need to consider the overall general location of the hotel and it's closer vincinity. We do not consider travel time to and from the airport or similar, assuming we are doing this analysis for a longer stay, so travel to and from the hotel is not part of the comparison. We also ignore the proximity to monuments or museums for simplicity's sake.

2.2 Analytical Approach

General Idea

We solely base our comparison on similarity of the neighbourhood of the hotel, considering a walking distance of 1000m and the close proximity of the hotel, considering a radius of 250m.

Data Requirements

We need the venues in the neighbourhoods and around the hotels. These can be obtained using online APIs, which will require geocoordinates for the neighbourhoods and the hotels. We also need a list of neighbourhoods for the target city and potentially also for the origin city (from the origin citry we only need information about the neightbourhood of the hotel).

Modeling

We will transform the data about venues into one-hot encoded information about the neighbourhoods and the hotels. Then we can sum up the venues by type and calculate the mean across all neighbourhoods/hotels. Using this vector the proximity of two neighbourhoods or hotels can be calculated.

After getting the information for the neighbourhood of our origin hotel and all neighhourhoods of the target city we can calculate the similarity of the neighbourhoods using eucledian distance and pick the top 3 target neighbourhoods. Then we go through all hotels of these three neighbourhoods and get the venues in their immediate environment and do the same by calculating the distance to the data vector of the origin hotel. We choose the top 10 results as possible similar candidates target hotels.

Note: Due to the limitations of the free FourSquare API we'll be limited to a maximum of 100 venues per neighbourhood/hotel. This should be fine for the direct comparison of hotels, but it's definitely not really enough to compare two lively neighbourhoods in two big cities. Therefore we will choose a hotel in a neighbourhood with less than 100 venues as our origin hotel.

2.3 Data Sources

New York

Neighbourhood information

We use the data from this link [https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json] for a list of neighbourhoods and their geocoordinates

Venue information

To get information about the hotel's closeby venues and the information about the venus in a neighbourhood we use the FourSquare "explore" API.

Toronto

Neighbourhood information

To get neighbourhood information about Toronto we use the Wikipedia page "List of postal codes of Canada: M". This list has no geocoordinates yet.

Neighbourhood geocoordinates

For geocoordinates for the postal codes of the neighbourhoods in Toronto we use data from the following link [https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv]

Venue information

To get venue information for the neighbourhoods and the hotels we use the FourSquare API as well

2.4 Steps of the Analysis

1. The analysis will follow the following sequence:

2. Get data about origin city (New York) and choose origin hotel

3. Get/Save venue information about the neighbourhood the origin hotel is in

4. Get venue information about the direct environment of the origin hotel

5. Get data about target city neighbourhoods

6. Find the top 3 neighbourhoods in the target city that are most similar to the neighbourhood or the origin hotel

7. Get all hotels from these 3 neighbourhoods

8. Get the venues in immediately vicinity of each hotel in target city

9. Find the most similar hotels

2.5 Data Challenges

As part of the analysis we found the following challenges

- The number of venues reported by FourSquare varies a lot from city to city, their data doesn't seem to have the same quality in all cities
- FourSquare seems to report only a very small number of hotels

- Venue categories cannot simply be compared between cities because of cultural difference, especially the restaurant types will not be the same if you travel to different countries and surely you a more interested in a similar number of restaurants and not in having exactly the same restaurant types around your hotel. We combined all restaurant types into a single category "restaurant" to overcome this
- Restaurants should probably be classified more coarsly like "Fast Food", "Bar with Food", "Full blown restaurant". We left this open, but it could have been included in the data cleanup steps
- There are some almost duplicate categories like Gym and Gym/Training Center. For productive use this would probably need further analysis
- It turns out that even though we have not so different cities (New York and Toronto) obviously the categories used to categorize venues are not overlapped as much as one might think. This might be due to cultural differences. To make neighbourhoods and hotel environments comparable we reduced the comparison to the categories found in both cities/neighbourhoods.
- FourSquare's 100 venue limit makes the free API not so useful for this analysis for places that are in environments where there are a lot of venues. We picked a neighbourhood with less than the 100 venues limit for this reason. This should allow to find a reasonable similar environment. The question remains if we should have excluded all neighbourhoods with 100 venues from the target city from the comparison as well (which we didn't)
- After running this jupyter notebook several times on different days and different times of the day I can see, that the result depend on the time of the day. This leaves the impression that FourSquare is not the right data source for the problem.

3 Analysis/Methodology

3.1 Get data about New York

In this section we'll get the information for our origin hotel and it's neighbourhood and take a look at the data to understand the structure of the information. For this we actually get the data for all neighbourhoods of our origin city and see the number of venues in each neighbourhoods and then pick an origin hotel as per the note above.

In a second step we'll get the neighbourhood information for our target city.

New York Neighbourhood Information

From our data source we obtained all the neighbourhood for New York, as a result we get a table of 306 neighbourhoods in 5 boroughs, of the following structure:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

New York Neighbourhood Venues

Then we load the venues for all of the neighbourhoods by calling the FourSquare API, resulting in a table of 20555 venues of the following structure:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Ripe Kitchen & Bar	40.898152	-73.838875	Caribbean Restaurant
2	Wakefield	40.894705	-73.847201	Ali's Roti Shop	40.894036	-73.856935	Caribbean Restaurant
3	Wakefield	40.894705	-73.847201	Jackie's West Indian Bakery	40.889283	-73.843310	Caribbean Restaurant
4	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop

3.2 Choose a hotel

Find the top 10 neighbourhoods of all the neighbourhoods that have hotels with almost 100 venues. Well choose our hotel from one of these.

Find Neighbourhoods With Hotels

By filtering the above venue data we obtain a list of all neighbourhoods that have hotels:

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Arrochar	1	1	1	1	1	1
Astoria Heights	2	2	2	2	2	2
Battery Park City	3	3	3	3	3	3
Blissville	1	1	1	1	1	1
Borough Park	1	1	1	1	1	1
Brooklyn Heights	2	2	2	2	2	2
Charleston	1	1	1	1	1	1
Chelsea	4	4	4	4	4	4
Chinatown	2	2	2	2	2	2
Civic Center	4	4	4	4	4	4
Clinton	5	5	5	5	5	5
Co-op City	1	1	1	1	1	1
Concord	1	1	1	1	1	1
Concourse	1	1	1	1	1	1

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Concourse Village	1	1	1	1	1	1
Corona	1	1	1	1	1	1
Downtown	1	1	1	1	1	1
Dumbo	1	1	1	1	1	1
East Elmhurst	3	3	3	3	3	3
Eastchester	1	1	1	1	1	1
Edenwald	1	1	1	1	1	1
Edgemere	1	1	1	1	1	1
Financial District	2	2	2	2	2	2
Flatiron	3	3	3	3	3	3
Fresh Meadows	3	3	3	3	3	3
Fulton Ferry	1	1	1	1	1	1
Gerritsen Beach	1	1	1	1	1	1
Gramercy	2	2	2	2	2	2
Greenpoint	1	1	1	1	1	1
Greenwich Village	2	2	2	2	2	2

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Hudson Yards	5	5	5	5	5	5
Hunters Point	1	1	1	1	1	1
Lincoln Square	2	2	2	2	2	2
Little Italy	3	3	3	3	3	3
Long Island City	7	7	7	7	7	7
Lower East Side	1	1	1	1	1	1
Malba	1	1	1	1	1	1
Manhattan Beach	1	1	1	1	1	1
Melrose	2	2	2	2	2	2
Midtown	4	4	4	4	4	4
Midtown South	5	5	5	5	5	5
Morris Park	1	1	1	1	1	1
Morrisania	1	1	1	1	1	1
Murray Hill	3	3	3	3	3	3
Noho	4	4	4	4	4	4
North Riverdale	1	1	1	1	1	1

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Ocean Hill	1	1	1	1	1	1
Park Hill	1	1	1	1	1	1
Pelham Bay	1	1	1	1	1	1
Queensbridge	11	11	11	11	11	11
Ravenswood	1	1	1	1	1	1
Red Hook	2	2	2	2	2	2
Rockaway Beach	2	2	2	2	2	2
Roosevelt Island	2	2	2	2	2	2
Rosebank	1	1	1	1	1	1
Sheepshead Bay	2	2	2	2	2	2
Shore Acres	1	1	1	1	1	1
Soho	3	3	3	3	3	3
South Ozone Park	3	3	3	3	3	3
Steinway	1	1	1	1	1	1
Sunnyside	1	1	1	1	1	1
Sunset Park	1	1	1	1	1	1

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Sutton Place	2	2	2	2	2	2
Travis	2	2	2	2	2	2
Tribeca	6	6	6	6	6	6
Turtle Bay	2	2	2	2	2	2
Upper East Side	3	3	3	3	3	3
Utopia	3	3	3	3	3	3
Vinegar Hill	1	1	1	1	1	1

Filter Venues by Neighbourhoods With Hotel and Count, Top 10

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Rockaway Beach	89	89	89	89	89	89
East Elmhurst	88	88	88	88	88	88
Melrose	88	88	88	88	88	88
Fresh Meadows	80	80	80	80	80	80
Astoria Heights	79	79	79	79	79	79
Pelham Bay	77	77	77	77	77	77
Co-op City	77	77	77	77	77	77

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Ocean Hill	75	75	75	75	75	75
Blissville	72	72	72	72	72	72
Concourse	68	68	68	68	68	68

Now Find Hotels in These Neighbourhoods

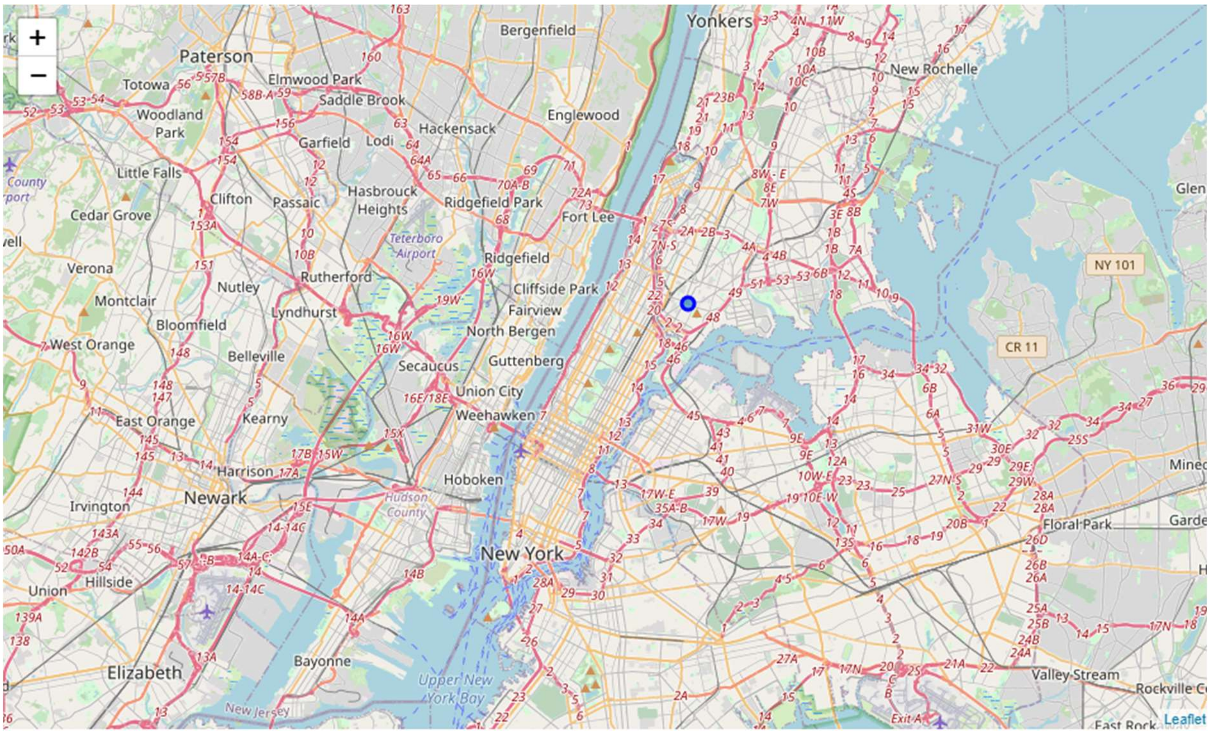
Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Co-op City	40.874294	-73.829939	Ramada by Wyndham Bronx	40.879865	-73.831504	Hotel
Melrose	40.819754	-73.909422	Opera House Hotel	40.815250	-73.916090	Hotel
Melrose	40.819754	-73.909422	Days Inn Bronx-Yankee Stadium	40.827092	-73.912007	Hotel
Pelham Bay	40.850641	-73.832074	Residence Inn by Marriott New York The Bronx a...	40.850038	-73.842574	Hotel
Concourse	40.834284	-73.915589	Days Inn Bronx-Yankee Stadium	40.827092	-73.912007	Hotel
Ocean Hill	40.678403	-73.913068	Days Inn	40.674713	-73.905807	Hotel
East Elmhurst	40.764073	-73.867041	ibis Styles New York LaGuardia Airport	40.770173	-73.869321	Hotel

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
East Elmhurst	40.764073	-73.867041	New York LaGuardia Airport Marriott	40.769107	-73.867732	Hotel
East Elmhurst	40.764073	-73.867041	Aloft New York LaGuardia Airport	40.770412	-73.870143	Hotel
Fresh Meadows	40.734394	-73.782713	Wyndham Garden Fresh Meadows	40.739418	-73.787829	Hotel

We Choose The Second Hotel

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Melrose	40.819754	-73.909422	Opera House Hotel	40.81525	-73.91609	Hotel

We've Chosen Opera House Hotel, Which Is Located Here



3.3 Get Venues Close to the Origin Hotel

We accessed FourSqaure to get the venues with 250m of the chosen origin hotel. We received 88 venues close to the hotel. The top categories of venues were

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
Restaurant	7	7	7	7	7	7
Kids Store	3	3	3	3	3	3
Department Store	2	2	2	2	2	2
Fried Chicken Joint	2	2	2	2	2	2
Mobile Phone Shop	2	2	2	2	2	2
Bank	1	1	1	1	1	1

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
Donut Shop	1	1	1	1	1	1
Financial or Legal Service	1	1	1	1	1	1
Gym	1	1	1	1	1	1
Hotel	1	1	1	1	1	1
Sandwich Place	1	1	1	1	1	1
Shoe Store	1	1	1	1	1	1
Shop & Service	1	1	1	1	1	1
Supermarket	1	1	1	1	1	1
Video Game Store	1	1	1	1	1	1

3.4 Get Neighbourhood Information for Target City Toronto

We Load the Neighbourhoods From Wikipedia

We get a table starting like that:

	PostalCode	Borough	NeighbourHood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village

	PostalCode	Borough	NeighbourHood
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park	Ontario Provincial Government
5	M9A	Etobicoke	Islington Avenue
6	M1B	Scarborough	Malvern, Rouge
7	M3B	North York	Don Mills)North
8	M4B	East York	Parkview Hill, Woodbine Gardens
9	M5B	Downtown Toronto	Garden District, Ryerson

Combine With Spatial Data

Using the data file Geospatial_Coordinates.csv from [https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv] (provided by IBM), we get the following table of neighbourhoods with geocoordinates which we can use to access the FourSquare API based on the geocoordinates

	PostalCode	Borough	NeighbourHood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242

	PostalCode	Borough	NeighbourHood	Latitude	Longitude
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills)North	43.745906	-79.352188
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937

Get the Venues for Toronto

After accessing the FourSquare API for the neighbourhoods in Toronto we get 2132 venues.

3.5 Find Similar Neighbourhoods in Toronto for the Origin Hotel Neighbourhood

The Dimensions of Our Comparison Space

Looking at the data we have so far, we get the following result:

```
There 2132 venues in Toronto
There are 218 uniques categories in Toronto.
There are 483 uniques categories in New York.
There are 41 uniques categories in Origin Hotel Hood.
```

Obviously there are a lot of categories we don't need

Therefore we remove everything from Toronto that doesn't has a category that exists in New York and also remove everything from our New York venues that doesn't have a category that exists in Toronto. This way we'll get the "dimensions" in which neighbourhoods or environments can be compared.

Category Cleanup

But first we will also combine all restaurants into a single "Restaurant Category". This will make the restaurants as a criterion more useful.

Why do we do this? Well, if you compare different cities in different countries there is always a big difference in restaurant culture. If you compare the venues on the restaurant type granularity you will not get a good similarity because of course you'll find as many French restaurants as you'll find American restaurants in New York.

Probably it would be possible to be a bit more fine granular e.g. distinguish between "Restaurant" and "Fast Food", but for simplicity's sake I'll leave it at a single category of restaurants. However, we will consolidate "Gym" and "Gym/Fitness Center" into a single "Gym" category.

So, first let's combine all types of restaurants into a single category and merge the two types of gym categories

After cleaning up the categories we get:

```
There are 218 uniques categories in Toronto.
There are 483 uniques categories in New York.
There are 34 uniques categories in Origin Hotel Hood.
There are 15 uniques categories in Origin Hotel Vicinity.
```

Filter categories in Toronto and New York

Now we remove all venues from Toronto that are in categories we don't find in our origin hotel hood and all venues in our origin hotel hood and origin hotel vicinity that have categories that don't exist in Toronto

As a result we get:

```
There are 29 uniques categories in Toronto.
There are 29 uniques categories in Origin Hotel Hood.
```

The remaining 29 categories are

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
Art Gallery	13	13	13	13	13	13
Bakery	48	48	48	48	48	48
Bank	28	28	28	28	28	28
Bus Line	4	4	4	4	4	4
Bus Station	2	2	2	2	2	2
Clothing Store	34	34	34	34	34	34

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
Convenience Store	12	12	12	12	12	12
Discount Store	11	11	11	11	11	11
Donut Shop	4	4	4	4	4	4
Fish Market	5	5	5	5	5	5
Food Truck	2	2	2	2	2	2
Fried Chicken Joint	12	12	12	12	12	12
Grocery Store	28	28	28	28	28	28
Gym	46	46	46	46	46	46
Hotel	35	35	35	35	35	35
Ice Cream Shop	10	10	10	10	10	10
Kids Store	1	1	1	1	1	1
Martial Arts School	1	1	1	1	1	1
Mobile Phone Shop	5	5	5	5	5	5
Office	7	7	7	7	7	7
Park	53	53	53	53	53	53
Pharmacy	23	23	23	23	23	23

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
Pizza Place	47	47	47	47	47	47
Restaurant	548	548	548	548	548	548
Sandwich Place	42	42	42	42	42	42
Shopping Mall	8	8	8	8	8	8
Supermarket	8	8	8	8	8	8
Supplement Shop	2	2	2	2	2	2
Video Game Store	3	3	3	3	3	3

Convert to One-Hot Encoding and Sum up

After converting to One-Hot Encoding and summing up we get the following table like this for the origin hotel

Neighborhood	Melrose
Art Gallery	2
Bakery	1
Bank	1
Bus Line	1
Bus Station	4
Clothing Store	1
Convenience Store	1
Discount Store	2
Donut Shop	6
Fish Market	1
Food Truck	1
Fried Chicken Joint	4
Grocery Store	5
Gym	3
Hotel	2
Ice Cream Shop	1
Kids Store	3
Martial Arts School	1
Mobile Phone Shop	3
Office	1
Park	2
Pharmacy	2
Pizza Place	5
Restaurant	20
Sandwich Place	4
Shopping Mall	1
Supermarket	3
Supplement Shop	1
Video Game Store	1

And a table like this for Toronto

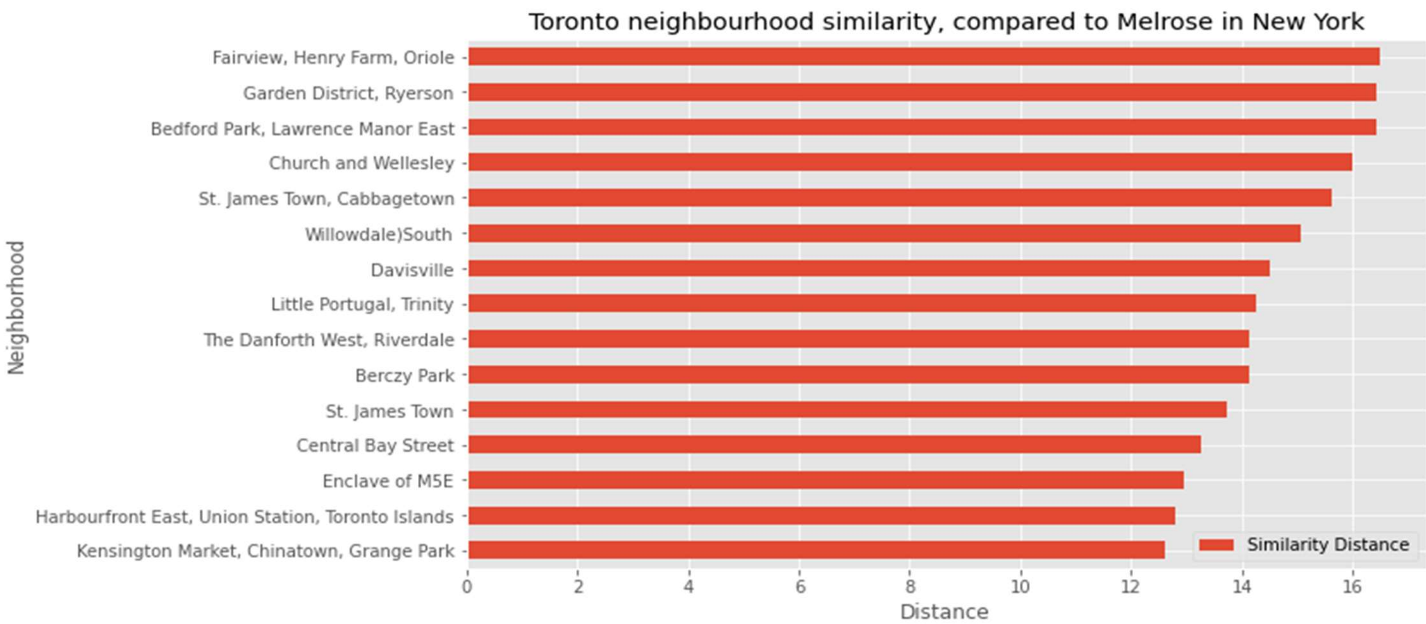
Neighborhood	Agincourt	Alderwood, Long Branch	Bathurst Manor, Wilson Heights, Downsview North	Bayview Village	Bedford Park, Lawrence Manor East	Berczy Park	Brockton, Parkdale Village, Exhibition Place	Caledonia-Fairbanks	Cedarbrae	Central Bay Street	Christie	Church and Wellesley	Clarks Corners, Tam O'Shanter, Sullivan
Art Gallery	0	0	0	0	0	1	0	0	0	0	0	0	0
Bakery	0	0	0	0	0	3	1	0	1	0	0	0	0
Bank	0	0	2	1	0	0	0	0	1	0	0	0	1
Bus Line	0	0	0	0	0	0	0	0	0	0	0	0	0
Bus Station	0	0	0	0	0	0	0	0	0	0	0	0	0
Clothing Store	1	0	0	0	0	1	0	0	0	0	0	1	0
Convenience Store	0	0	0	0	0	0	1	0	0	0	0	0	0
Discount Store	0	0	0	0	0	0	0	0	0	1	0	0	0
Donut Shop	0	0	0	0	0	0	0	0	0	1	0	0	0
Fish Market	0	0	0	0	0	1	0	0	0	0	0	0	0
Food Truck	0	0	0	0	0	0	0	0	0	0	0	0	0
Fried Chicken Joint	0	0	1	0	0	0	0	0	1	0	0	0	1
Grocery Store	0	0	1	0	1	1	1	0	0	0	4	1	0
Gym	0	1	0	0	0	0	1	0	0	1	0	0	0
Hotel	0	0	0	0	0	1	0	0	0	1	0	2	0
Ice Cream Shop	0	0	1	0	0	0	0	0	0	1	0	1	0
Kids Store	0	0	0	0	0	0	0	0	0	0	0	0	0
Martial Arts School	0	0	0	0	0	0	0	0	0	0	0	1	0
Mobile Phone Shop	0	0	1	0	0	0	0	0	0	0	0	0	0
Office	0	0	0	0	0	0	1	0	0	1	0	0	0
Park	0	0	1	0	0	1	0	1	0	1	2	1	0
Pharmacy	0	1	1	0	1	2	0	0	0	0	0	0	1
Pizza Place	0	2	1	0	1	0	0	0	0	0	0	1	2
Restaurant	1	0	4	2	10	15	2	0	3	24	2	29	5
Sandwich Place	0	1	1	0	2	1	0	0	0	4	0	0	0
Shopping Mall	0	0	1	0	0	0	0	0	0	0	0	0	0
Supermarket	0	0	1	0	0	0	0	0	0	0	0	0	0
Supplement Shop	0	0	0	0	0	0	0	0	0	0	0	0	0
Video Game Store	0	0	0	0	0	0	0	0	0	0	0	0	0

Combine the One-Hot Encoded Tables And Calculate Euclidian Distances

After combining the two tables above and calculating the Euclidian distances we get the following similarities/distances for the top 10

Neighborhood	Similarity Distance
Kensington Market, Chinatown, Grange Park	12.609520
Harbourfront East, Union Station, Toronto Islands	12.806248
Enclave of M5E	12.961481
Central Bay Street	13.266499
St. James Town	13.711309
Berczy Park	14.142136
The Danforth West, Riverdale	14.142136
Little Portugal, Trinity	14.247807
Davisville	14.491377
Willowdale) South	15.066519
St. James Town, Cabbagetown	15.620499
Church and Wellesley	16.000000
Bedford Park, Lawrence Manor East	16.431677
Garden District, Ryerson	16.431677
Fairview, Henry Farm, Oriole	16.492423

Or, as a bar chart:

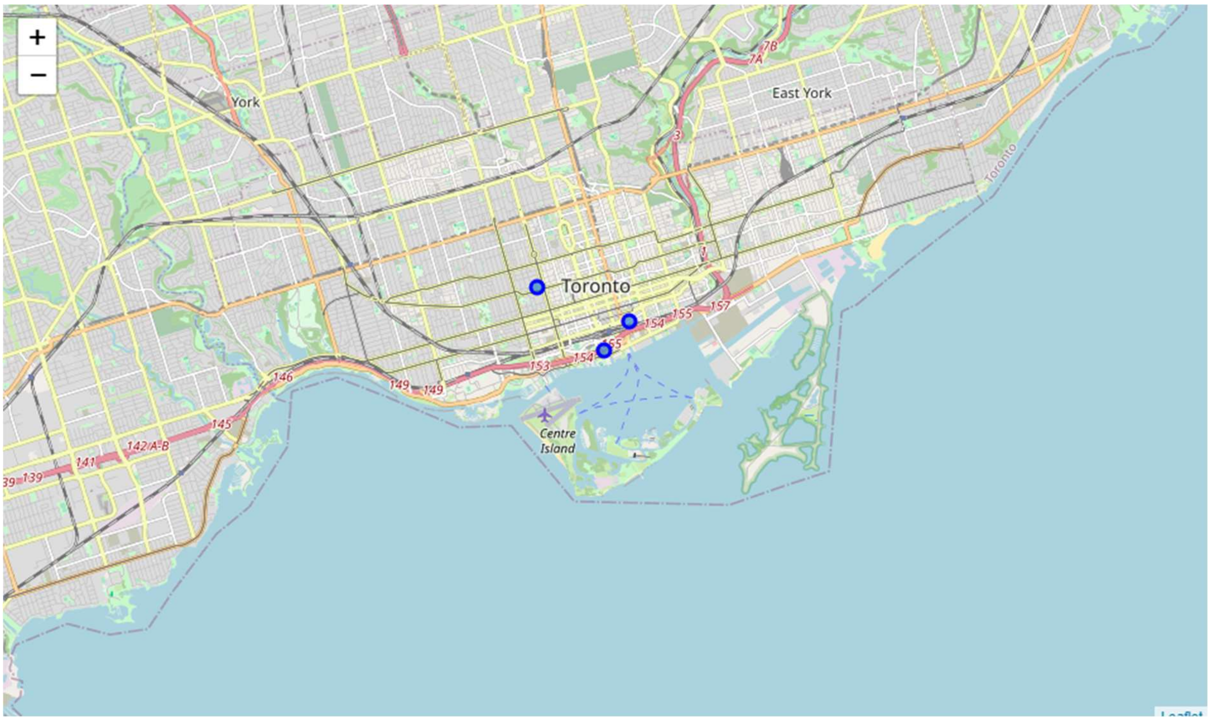


Choose The Top 3 Neighbourhoods

Of those neighbourhoods we take the 3 most similar neighbourhoods:

	PostalCode	Borough	NeighbourHood	Latitude	Longitude
36	M5J	Downtown Toronto	Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752
84	M5T	Downtown Toronto	Kensington Market, Chinatown, Grange Park	43.653206	-79.400049
92	M5W	Downtown Toronto Stn A	Enclave of M5E	43.646435	-79.374846

In the map they are located here:



3.6 Find Most Similar Hotels in Top 3 Neighbourhoods

Get the Hotels in the Top 3 Neighbourhoods

This results in a table like this:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752	Delta Hotels by Marriott Toronto	43.642882	-79.383949
Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752	Le Germain Hotel	43.643125	-79.380918
Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752	Radisson Admiral Hotel Toronto-Harbourfront	43.638765	-79.385871
Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752	The Westin Harbour Castle, Toronto	43.641211	-79.375749
Enclave of M5E	43.646435	-79.374846	The Omni King Edward Hotel	43.649191	-79.376006
Enclave of M5E	43.646435	-79.374846	Cosmopolitan Toronto Centre Hotel & Spa	43.649064	-79.377598

Get Venues Close to the Hotels in the Top 3 Neighbourhoods, Clean up Categories, One Hot Encode

Again we use FourSquare to get venues, this time in 250m from the hotels in the top 3 neighbourhoods in Toronto. Again we cleanup the categories as we did for the neighbourhoods.

For the origin hotel we get

Then we do one-hot encoding, sum up, and as a result we get:

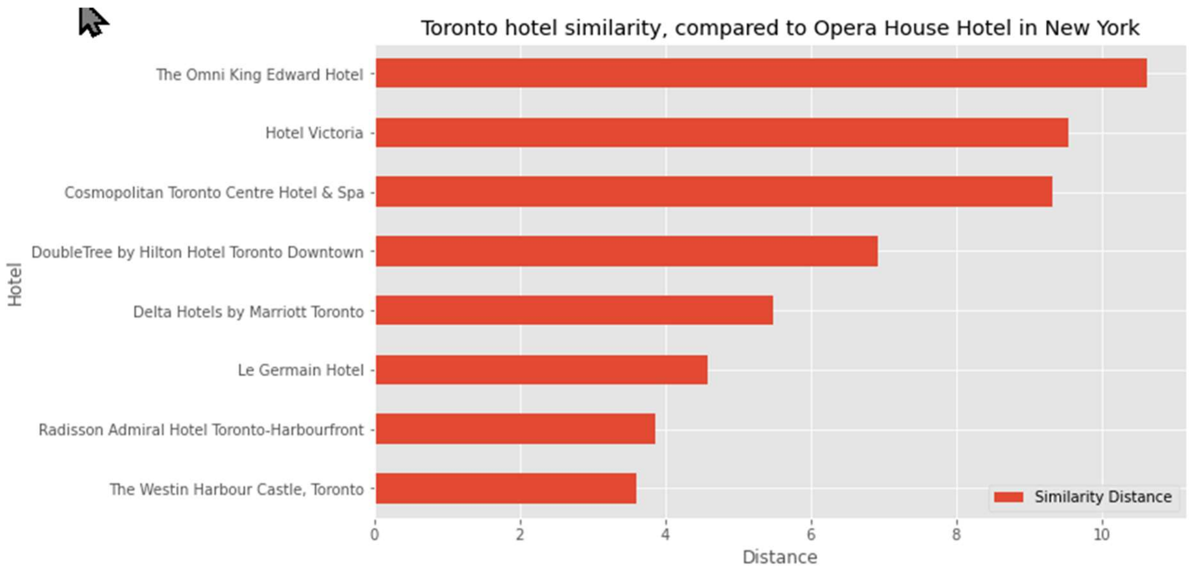
	Neighborhood	Bank	Department Store	Fried Chicken Joint	Gym	Hotel	Restaurant	Sandwich Place	Supermarket
0	Opera House Hotel	1	2	2	1	1	7	1	1
0	Cosmopolitan Toronto Centre Hotel & Spa	0	0	0	3	4	15	1	0
1	Delta Hotels by Marriott Toronto	0	0	0	1	3	11	0	1
2	DoubleTree by Hilton Hotel Toronto Downtown	0	0	0	0	1	1	0	0
3	Hotel Victoria	1	1	0	1	3	16	1	0
4	Le Germain Hotel	1	1	1	0	2	11	0	1
5	Radisson Admiral Hotel Toronto-Harbourfront	0	0	0	0	1	5	1	0
6	The Omni King Edward Hotel	0	1	0	2	3	17	0	0
7	The Westin Harbour Castle, Toronto	1	0	0	1	1	9	1	0

Calculate the Distances

We calculate the Euclidian distances and get the following result:

Hotel	Similarity Distance
The Westin Harbour Castle, Toronto	3.605551
Radisson Admiral Hotel Toronto-Harbourfront	3.872983
Le Germain Hotel	4.582576
Delta Hotels by Marriott Toronto	5.477226
DoubleTree by Hilton Hotel Toronto Downtown	6.928203
Cosmopolitan Toronto Centre Hotel & Spa	9.327379
Hotel Victoria	9.539392
The Omni King Edward Hotel	10.630146

Or, as a bar chart:



4 Result and Summary

Our origin hotel was the

- Opera House Hotel in Neighbourhood Melrose

And the most similar neighbourhoods in Toronto for Melrose are

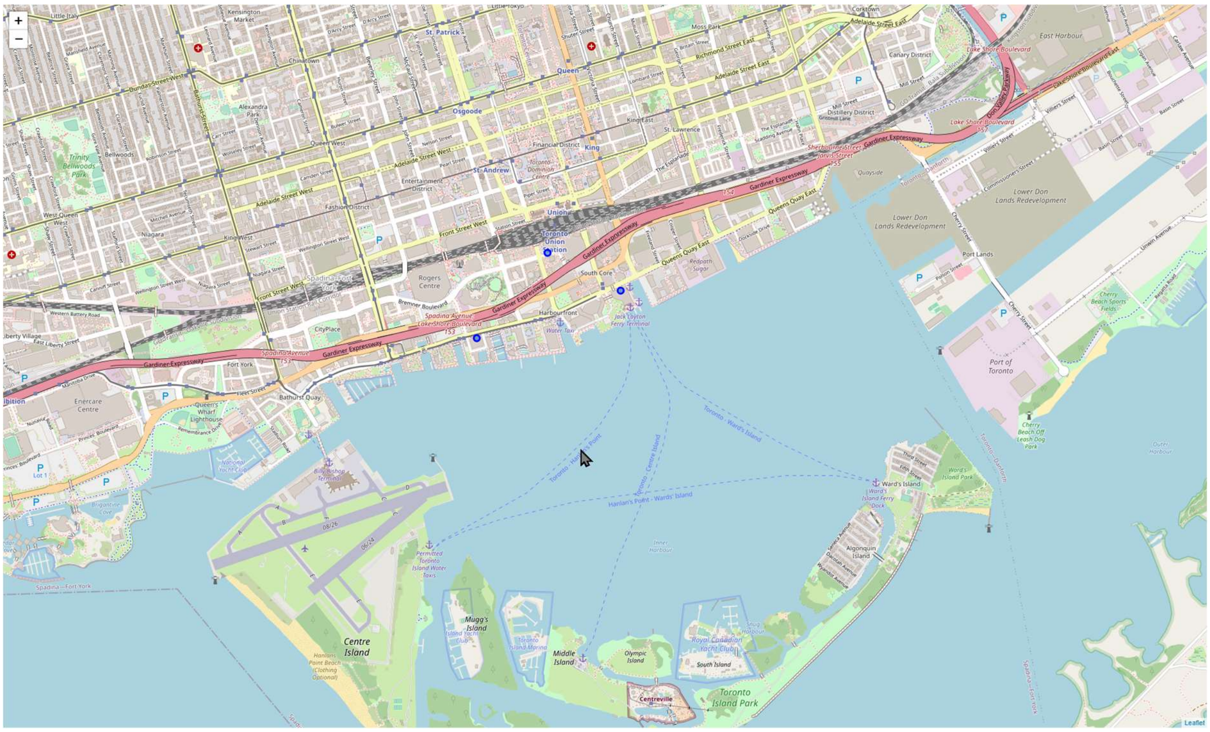
- Kensington Market, Chinatown, Grange Park
- Harbourfront East, Union Station, Toronto Islands

- Enclave of M5E

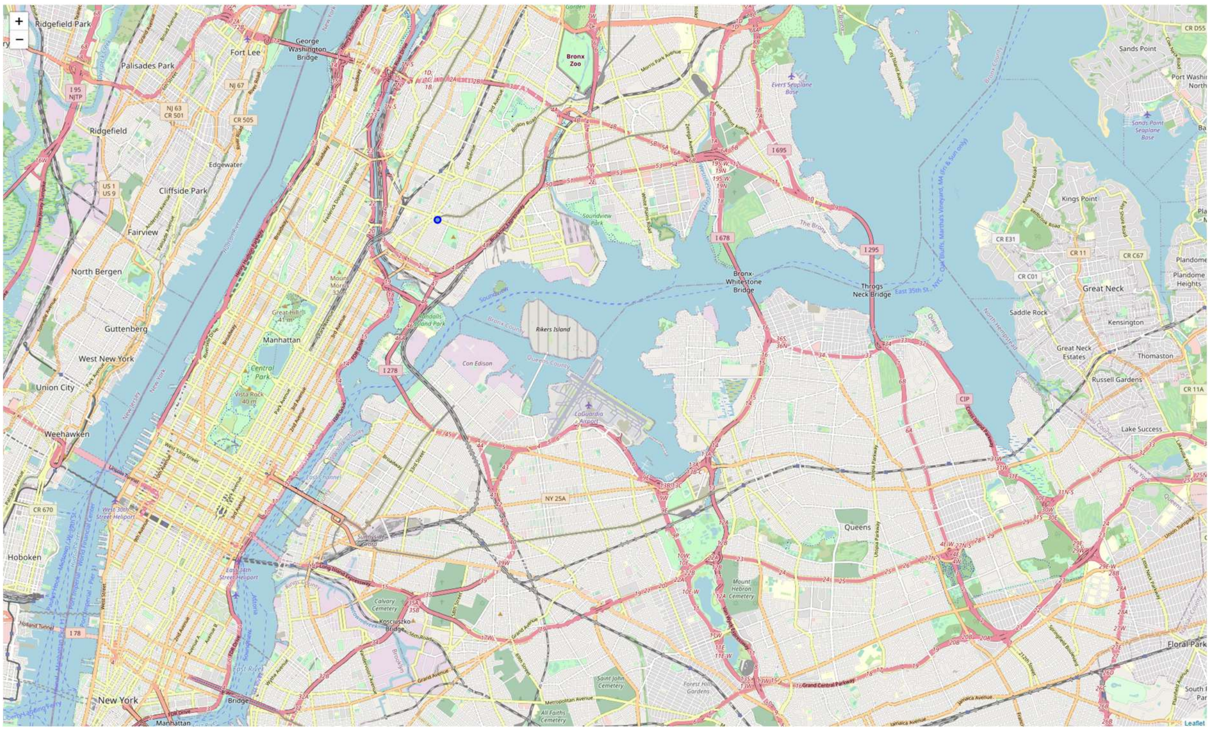
The most similar hotels found in these three neighbourhoods are

- The Westin Harbour Castle, Toronto
- Radisson Admiral Hotel Toronto-Harbourfront
- Le Germain Hotel

In the map these hotel are located here:



Whereas the hotel in New York was located here:



These three hotels would be our recommendation to the customer if he/she wants the most similar place in Toronto

Our approach lead us directly to a clear and understandable recommendation for our customers. This approach can be used in general for all hotel comparison/recommendation situations, with some restriction as discussed in the "Discussion" section below.

5. Discussion

FourSquare oddities

While doing this analysis I reloaded the venues from FourSquare several times. While the overall number of results remained more or less the same, the results were so different that the neighbourhoods with less than 100 venues in New York completely changed from run to run and I had to add the filter to only include neighbourhoods with hotels because on some runs there were no hotels in the resulting neighbourhoods. Which is another oddity - the number of hotels is not anywhere close to real numbers. I was actually forced to choose a different hotel in New York to get sufficient hotels in Toronto in the similar neighbourhoods to make this analysis worthy to show.

Therefore I've resorted to saving the data, to have reproducible results. However this also means that this approach is unusable for a real analysis beyond this Coursera Capstone, where I would say this is one of the key finding of the analysis. Reloading the data freshly from FourSquare can be triggered by setting the variable "refetch" to True.

Is Euclidean the correct distance measure to calculate the similarity?

There are different way how the distance between hotel neighbourhoods could be calculated. The euclidean distance works fine and gives truly similar results. However, one might pose the question if actually "better" results should appear closer, meaning that if the environment of a hotel has more venues of some type the distance should be equal to a hotel that has the same number of venues of that type (Example: if there are more restaurants around a hotel it might be better to at least not increase the distance because of that).

A second possibility would be to use just zero and one for "has/has not venue of same type" which would account better for the variety of venues around the hotel, not giving larger weight to venues with higher numbers.

Problems comparing two cities - different categories

It turns out more complicated than expected to compare lists of venues from different cities because categorization seems to be very different from city to city. After compacting the categories in the hotel neighbourhoods we ended up with only 8 categories left to make up our comparison space for the hotels and when taking a closer look some those should have been added to the restaurant category as well. Neighbourhood comparison still had 29 categories which seems to be ok. Probably FourSquare is not the best data source for this comparison.

Is this similarity what the customer is actually looking for?

The discussion of the eucildean distance brings up if this similarity is really what the customer would be looking for. This algorithm really finds a similar hotel, it doesn't find a potentially "better" place. For this analysis I'm simply assuming that this is what is desired. If someone picked a hotel in a quiet area the comparison will find a hotel in a quiet area, not one that is a busy environment because there are more places around. This could be adjusted by calculating the similarity differently.

Normalization/Calculation of mean() instead of using sum()

For this analysis I changed from taking the mean of the one-hot vales to taking the sums because this results in really similar to the original hotel. But calculation the mean() the differences between environments are nivellated and the distances between locations don't really reflect true similarity.

Where is machine learning in this analysis?

This analysis doesn't use any machine learning, it is not needed for this analysis. We could probably also cluster the hotels in the target city using K-Means and then find the best matching cluster using K-Nearest-Neighbour with some tuning. However, we'd still just end up with a cluster, so we'd still need to determine the best match within the cluster.

6. Conclusion(s)

- We were able to achieve our goal: we developed an approach to finding a similar hotel based on the surroundings of the hotel in a different city
- We should look for a more complete and consistent data source, FourSquare doesn't provide complete data, the data seems to vary, hotels don't seem to be well covered and the venue categorization is not really consistent between cities
- Not all data science requires machine learning
- I think we have solved an interesting problem for hotel booking companies like Booking.com. They also have hotel comfort similarity and a much more complete hotel database, so with this data the algorithm could be executed with much better result quality, just because there would be much more information about hotels but also because we could add a side condition of similar hotel quality as well to make the search for a similar hotel a really useful feature