The Battle of Neighborhoods - New York City and Toronto

Saubhik Bagchi

July 21, 2019

1. Introduction

In today's world, it's very common for people to relocate from one city to other across the world for education, business, jobs and other personal reasons. Someone getting a better job offer from another city may decide to relocate however she/he would like to find out a neighborhood in the new city where similar amenities are available so that she/he can feel comfortable after relocating to the new place.

If someone likes the comfort of quality Italian food in her neighborhood restaurants, she would probably look for a neighborhood in the new city where there are plenty of good Italian restaurants around her. Same goes for other amenities like coffee shops, school, gym, amusement park, grocery stores, super markets etc.

1.1. Business Problem

Let me define a business problem for this project. Debbie lives in Bronx, Riverdale, New York and she loves her neighborhood. She got a great job offer from Toronto, and she decided to relocate to Toronto in 2 weeks' time to take up the new opportunity. Debbie wants to find out a neighborhood in Toronto where she would get similar amenities available around her that she gets in Bronx. Our problem statement is to explore the neighborhoods of Toronto and see which of them closely matches with Bronx and provides similar amenities. Based on our analysis we can recommend Debbie which Toronto neighborhoods should be the top choice for her!

1.2. Target Audience

Target audience for this project is anyone like Debbie who is planning to relocate to another city and trying to find out the top neighborhoods in the new city that will provide him/her similar amenities that the neighborhood in their current city provides where they currently live.

2. Data Acquisition and Cleaning

2.1. Data Sources

For this project, the following data sources will be used:

(1) Location data of New York City to be downloaded as a json file from url https://geo.nyu.edu/catalog/nyu_2451_34572. From this json file, only 'Borough', 'Neighborhood', 'Latitude' and 'Longitude' will be extracted.

	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

Figure 1

(2) Neighborhood data of Toronto City will be scrapped from the Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. From the scrapped data, 'Postal code', 'Borough' and 'Neighborhood' of Toronto will be extracted

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Figure 2

(3) Location data will be downloaded using the link http://cocl.us/Geospatial_data and extract the 'Latitude' and 'Longitude' for Toronto neighborhoods.

(4) Foursquare API will be used to retrieve geo-location information about various types of venues for each neighborhood in New York City and Toronto City and compare them based on similar categories of venues.

Note: Data from Foursquare API changes over time.

Venues information of New York City is presented in Figure-3. Shape of data is: (306, 473), which indicates New York City has 469 categories of venue.

	Borough	Neighborhood	Latitude	Longitude	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	Airport Terminal	American Restaurant
0	Staten Island	St. George	40.644982	-74.079353	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.013514
1	Staten Island	New Brighton	40.640615	-74.087017	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.023256
2	Staten Island	Stapleton	40.626928	-74.077902	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
3	Staten Island	Rosebank	40.615305	-74.069805	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
4	Staten Island	West Brighton	40.631879	-74.107182	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.016667

Figure 3

Venues information of Toronto is presented in Figure-4. Shape of data is: (103, 338), which indicates Toronto has 333 categories of venue.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	American Restaurant
0	M8V	Etobicoke	Humber Bay Shores,Mimico South,New Toronto	43.605647	-79.501321	0.0	0.000000	0.0	0.0	0.0	0.0	0.052632
1	M8W	Etobicoke	Alderwood,Long Branch	43.602414	-79.543484	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000
2	M8X	Etobicoke	The Kingsway,Montgomery Road,Old Mill North	43.653654	-79.506944	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000
3	M8Y	Etobicoke	Humber Bay,King's Mill Park,Kingsway Park Sout...	43.636258	-79.498509	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000
4	M8Z	Etobicoke	Kingsway Park South West,Mimico NW,The Queensw...	43.628841	-79.520999	0.0	0.017544	0.0	0.0	0.0	0.0	0.035088

Figure 4

(5) Finally, exploratory data analysis and machine learning algorithms will be used to identify top neighborhoods in Toronto which resembles Bronx, so that we can help Debbie with her search for the right place to settle when she relocates to Toronto.

3. Exploratory Data Analysis

3.1. Data Sources

New York City and Toronto has different number of venues. New York has 469 and Toronto has 333 venue categories. There are some common venue categories in both cities. We'll find the common venue categories and keep only these common categories to find similar neighborhoods. Based on our analysis of data from both datasets, we can identify that

```
Number of common venue categories in both data are       : 301
Number of different venue categories in New York City are: 168
Number of different venue categories in Toronto are      : 32
```

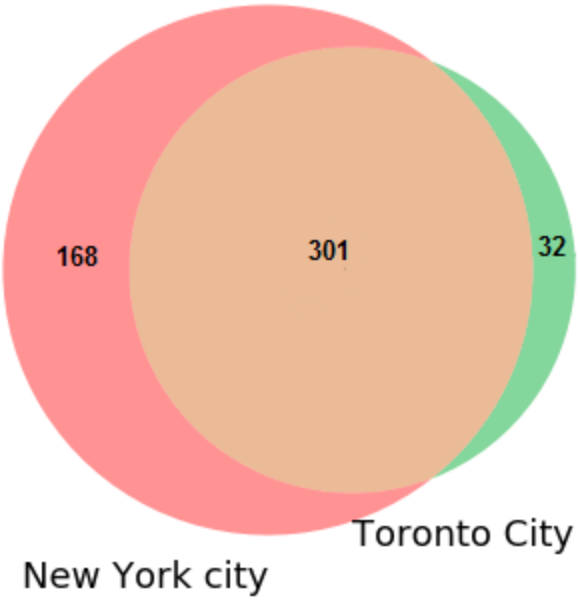

Figure 5

4. Find most similar location in another city

To find the most similar city, Cosine Similarity Machine Learning recommender engine will be used. For more details on Cosaine Similarity, please refer to section 5.

As mentioned in the business problem statement of this project, Debbie's current location is Bronx, Riverdale, New York and we want to find the top 7 similar boroughs in Toronto city.

As the first step, we find the index of the current city in New York City data (in df1) and corresponding latitude and longitude.

Next, we get all common venue categories of Bronx and calculate matrix product with Toronto City common venue data. This will give quantitative measure of similarity of each borough of Toronto city to Bronx. Sort in descending order and select top 7 similar boroughs.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Lounge	America Restaura
84	M9L	North York	Humber Summit	43.756303	-79.565963	0.0	0.0	0.0	0.0	0.0	0.00000
65	M2L	North York	Silver Hills,York Mills	43.757490	-79.374714	0.0	0.0	0.0	0.0	0.0	0.00000
6	M9B	Etobicoke	Cloverdale,Islington,Martin Grove,Princess Gar...	43.650943	-79.554724	0.0	0.0	0.0	0.0	0.0	0.05555
76	M3L	North York	Downsview West	43.739015	-79.506944	0.0	0.0	0.0	0.0	0.0	0.00000
87	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	0.0	0.0	0.0	0.0	0.0	0.00000
83	M6L	North York	Downsview,North Park,Upwood Park	43.713756	-79.490074	0.0	0.0	0.0	0.0	0.0	0.00000
94	M1M	Scarborough	Cliffcrest,Cliffside,Scarborough Village West	43.716316	-79.239476	0.0	0.0	0.0	0.0	0.0	0.00000

Figure 6

Finally, using folium we can visualize the top 7 boroughs of Toronto which are similar to Bronx

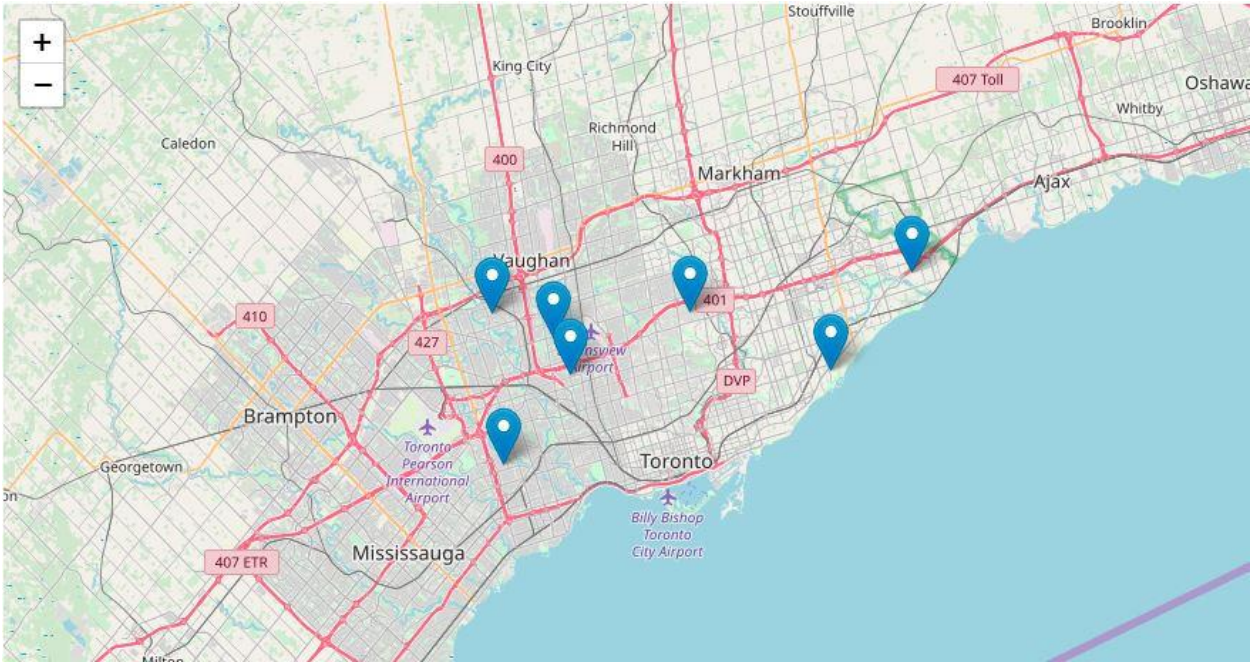

Figure 7

5. Additional Information on Cosine Similarity

Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity.

5.1. Introduction

A commonly used approach to match similar documents is based on counting the maximum number of common words between the documents.

But this approach has an inherent flaw. That is, as the size of the document increases, the number of common words tend to increase even if the documents talk about different topics.

The cosine similarity helps overcome this fundamental flaw in the ‘count-the-common-words’ or Euclidean distance approach.

5.2. What is Cosine Similarity and why is it advantageous?

Cosine similarity is a metric used to determine how similar the documents are irrespective of their size.

Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. In this context, the two vectors I am talking about are arrays containing the word counts of two documents.

As a similarity metric, how does cosine similarity differ from the number of common words?

When plotted on a multi-dimensional space, where each dimension corresponds to a word in the document, the cosine similarity captures the orientation (the angle) of the documents and not the magnitude. If you want the magnitude, compute the Euclidean distance instead.

The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance because of the size (like, the word ‘cricket’ appeared 50 times in one document and 10 times in another) they could still have a smaller angle between them. Smaller the angle, higher the similarity.

6. Conclusion

We started this project with a problem statement for Debbie. She lives in Bronx, Riverdale, New York and she loves her neighborhood. She got a great job offer from Toronto, and she decided to relocate to Toronto in 2 weeks’ time to take up the new opportunity. Debbie wants to find out a neighborhood in Toronto where she would get similar amenities available around her that she gets in Bronx.

We retrieved location information using Foursquare API, used exploratory data analysis and Cosine Similarity algorithm to identify top 7 neighborhood of Toronto which closely resembles Bronx. Now Debbie can get over the battle of neighborhood and can confidently choose a Toronto neighborhood that will be identical to Bronx!

7. Reference

https://en.wikipedia.org/wiki/Cosine_similarity
https://www.machinelearningplus.com/nlp/cosine-similarity/