IBM DATA SCINCE PROFESSIONAL CERTIFICATE

APPLIED DATA SCIENCE CAPSTONE PROJECT

PART II

AUGUST 16, 2021

THE BATTLE OF NEIGHBORHOODS

RISHI SAXENA

INTRODUCTION

Travel between the United States and Canada has commenced once again with the borders officially opening. In the year 2019, approximately 36 million travelers crossed the U.S.-Canadian border. It is expected that these numbers would compound in the coming years to overcome the 2020 bottleneck. With the market finally finding a footing, experts predict that new businesses would flourish and travel between the two countries would soar higher than ever. As a consequence of business, international travel for employment and immigration should increase as well.

One of the key challenging elements of this travel is identifying neighborhoods that suit one's preferences. Finding suitable accommodation can be an up-hill battle for most travelers. With the rising costs of real estate and start contrast in the structure of available amenities – making a wrong decision in accommodation planning has the potential to be catastrophic.

This project aims to provide a high-level solution to this problem. By employing the use of Machine Learning Classification models, this project endeavors to compare neighborhoods between two prominent cities and identify ideal locations for given specifications. By providing a way to statistically evaluate these similarities and identify similar neighborhoods between these two cities, this venture aims to simplify relocation efforts between the United States and Canada.

In the future, this project has the potential to expand and support relocators between multiple different cities in North America. This expansion may one day even support relocators between different continents. Additionally, this could also be a handy tool for small and large businesses alike – to assess whether a neighborhood or locality is suitable for their business. The possibilities are endless.

It should be noted that this program only offers guidelines based on publicly available data and should not be solely used for making decisions. Consequently, the accuracy of this program is a direct function of the accuracy of the source data.

DATA

The most important element of a machine learning model is the quality of the training data. As the famous saying goes, *"Garbage in, garbage out!"*. This is particularly true in the field of machine learning and data science. If the data used to train the model is not diverse enough, or lacks the desired accuracy, the model would underfit and unreservedly fail in delivering factual results. Consequently, if the train data is excessive, the model would overfit the data and be unusable to predict anything outside of the train dataset.

While this program can be easily tailored to work on any combination of two cities across North America, provided the spatial data is available, the two cities used in this initial demonstration of the program are New York City and Toronto. In order to create this model, the data should be able to offer a detailed overview of the venues and amenities for every neighborhood in the city. A key player that offers a free to use, publicly available online tool for this service is FourSquare. FourSquare is an American technology company that provides a location tracking platform for a several businesses and consumer products.

The FourSquare API provides a detailed overlook of the amenities and services available at a given location. However, this location has be provided in the form of spatial coordinates, i.e., latitudes and longitudes. This implies that the program would need another source of data that can provide the coordinate for every neighborhood in the cities of interest. Effort is put into ensuring that the spatial data used for this program is accurate and reliable. The two primary sources for data are –

1. New York University – Spatial Data Repository (https://geo.nyu.edu/)
2. Wikipedia (https://wikipedia.org)

Unfortunately, this is the bottleneck for this program. If the coordinate data for a city is unavailable, the program cannot provide the necessary inputs to FourSquare API and fetch the results. Another liability of this program is that publicly available, free data would always have its limitations. On the bright side, FourSquare is a reliable and accurate partner that makes this process viable, despite all challenges.

METHODOLOGY

This program is created using python programming language. One of the key advantages of python is the vast variety of supporting libraries available for free that simplify the code tremendously. In addition to the FourSquare API discussed above, this program also uses NumPy, Pandas, BeautifulSoup and SciKit-Learn libraries. Furthermore, data visualization libraries such as Matplotlib and Folium are also used to generate intuitive visualizations that illustrate the results.

Recall that this program compares the neighborhoods of two cities – New York City and Toronto. In order for the data to be useable by the FourSquare API, it must be supplied with the names of the boroughs and neighborhoods along with the corresponding geospatial data. In order to fetch this neighborhood data and use the FourSquare API, the geospatial data needs to be available. This data is obtained from the sources mentioned above. More specifically, the borough and neighborhood data for New York City is obtained from here.

Unfortunately, obtaining the relevant data for the city of Toronto is not so straightforward. Unlike New York City, Toronto does not have a publicly accessible spatial data repository. Thus, the data available on Wikipedia is used. Wikipedia does not provide a ready-to-use data in the format that is required for the FourSquare API. However, it does provide two separate datasets –

1. A table containing all the postal codes within and around the city of Toronto along with their corresponding Boroughs and Neighborhoods. This data is available here.
2. Another table containing the geo-spatial coordinate data for every postal code within and around the city of Toronto. This data is available here.

By combining these two raw datasets, the desired data can be obtained. However, this data first needs to be scrapped from the internet. One powerful tool that achieves this is Beautiful Soup. Beautiful Soup is a python library that can be used for web scrapping and html parsing. With the help of Beautiful Soup and the Pandas python library, these raw datasets can be processed and normalized quite easily.

The process of normalization eliminates the entries that do not have the necessary data. If the borough is not assigned, the rows are dropped. However, if the borough is available but the neighborhood is missing, the borough name is used as the name of the neighborhood. Furthermore, if two or more rows contain the same postal code and borough, the neighborhoods for all these rows are merged. They are separated by a comma.

Once this normalization process is complete, the data is suitable to be merged with the supporting data scrapped from the internet – generating the final data frame that is used to for the city of Toronto in this program.

Once the normalization is complete, the processed data frames look like this –

1. For New York City:

	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

2. For Toronto:

	PostalCode	Borough	Neighborhood	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
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

It should be noted that for the purposes of this program, the PostalCode column is not used. Furthermore, in order to fetch the desired results, FourSquare API does not explicitly need the Borough and Neighborhood names. It just needs the geospatial coordinates for a given location. However, for the sake of the results and for optimum visibility, this data is retained. The results from this program are intended for humans and humans are the not the best at identifying neighborhoods by their latitudes and longitudes.

Following this clean-up process, the program finds the nearby venues for each borough and neighborhood pair in the two cities. As mentioned earlier, FourSquare API is used to provide this information. However, in order to use FourSquare API, a developer account is needed along with the account credentials. These credentials consist of a *client id* and *client secret*, and must be supplied when calling the API.

As every city is different, they are categorized individually by the FourSquare API. This means that each city may yield different number of venue categories. For the sake of clarity and accuracy, only similar venue categories are considered. This outlook aligns with the intention of the service this program aims to provide. Once FourSquare API provides the relevant information regarding the venues available for a geolocation, this data can be analyzed to identify the most similar pair of neighborhoods between the two cities of New York and Toronto.

Using FourSquare API, the venues within a one-kilometer radius from the provided geolocation are identified. The following is the resulting dataset for each city –

1. For New York City, the shape of the resulting data set is 306 rows and 488 columns; with 484 different venue categories. A snapshot of the dataset is attached below –

	Borough	Neighborhood	Latitude	Longitude	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Amphitheater
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	0.014925	0.0
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	0.022222	0.0
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	0.0
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	0.0
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	0.017857	0.0

2. For Toronto City, the shape of the resulting data set is 103 rows and 334 columns; with 330 different venue categories. A snapshot of the data set is attached below –

	PostalCode	Borough	Neighborhood	Latitude	Longitude	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	American Restaurant	Amphithea
0	M7Y	East Toronto Business	Enclave of M4L	43.662744	-79.321558	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.021277	
1	M9W	Etobicoke Northwest	Clairville, Humberwood, Woodbine Downs, West H...	43.706748	-79.594054	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
2	M5W	Downtown Toronto Stn A	Enclave of M5E	43.646435	-79.374846	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.020000	
3	M7R	Mississauga	Enclave of L4W	43.636966	-79.615819	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	

As this program only uses the common venue categories between New York City and Toronto, the extra categories are removed.

The number of unique venue categories in New York City: 183
The number of unique venue categories in Toronto: 28
Thus, the number of common venue categories in both New York City and Toronto: 301

The following Venn Diagram shows the shared venue categories between New York City and Toronto –

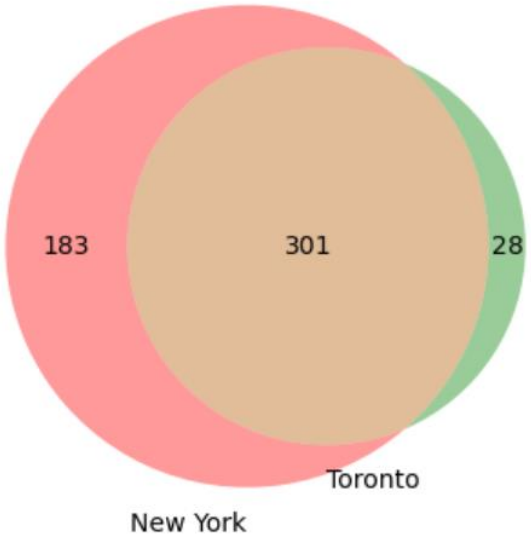

Figure 1: Venn-Diagram for Venue Categories in Both Cities

Once the data has been prepared, the analysis can be carried out. The similarities between neighborhoods in the two different cities is determined using the **Cosine Similarity**. A function determines the top *'n'* most similar neighborhoods in the opposing city, provided the source neighborhood in the initial city.

Using Folium, the neighborhoods that are most similar to the selected neighborhood are highlighted on a map of the two cities. This visualization is selected because it provides a clear picture of the location of the selected neighborhood and also the similar neighborhood in the other city.

RESULTS

To demonstrate the results of this program, two scenarios are considered. The first moving from Toronto to New York City, and the second moving the other way around, from New York City to Toronto. In both cases, a sample neighborhood is randomly selected to act as the source that this program will use to find similar neighborhoods in the following city.

TORONTO TO NEW YORK CITY

The source neighborhood in Toronto is chosen to be Alderwood/Long Branch, Etobicoke, Toronto. The program is used to identify the top seven, most similar neighborhoods in New York City. The results can be seen in the following image –

Figure 2: Most Similar Neighborhoods in New York City

Figure 3: Most Similar Neighborhood

Upon hovering over any of the pinheads, the name of the neighborhood(s) is revealed. Looking to the image on the left, the neighborhood in New York that is the most similar to Alderwood, Toronto is Claremont Village.

(It should be noted that the 5th most common neighborhood in figure 2 is outside of the screenshot area. This view can be zoomed out in the program.)

FROM NEW YORK CITY TO TORONTO

The source neighborhood in New York City is chosen to be Riverdale, Bronx, New York City. The program is used to identify the top seven, most similar neighborhoods in the city of Toronto. The results can be seen in the following image –

Figure 4: Most Similar Neighborhoods in Toronto

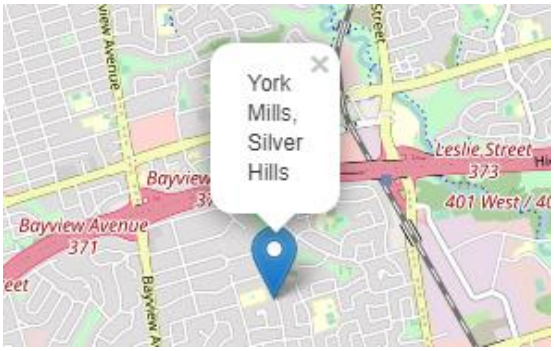

Figure 5: Most Similar Neighborhood

Upon hovering over any of the pinheads, the name of the neighborhood(s) is revealed. Looking to the pinhead marked "1st", the neighborhood in the city of Toronto that is the most similar to Riverdale, Bronx is York Mills, Silver Hills.

DISCUSSION

As illustrated in the results presented in the previous section, it is clear that this program provides an excellent first-look at the similarities between neighborhoods between New York City and the city of Toronto. The input parameters are simply the name of the current city, the current borough, the current neighborhood and the number of most similar neighborhoods to find. These are all information that should be easily available for any user that intends to use this platform.

As this program uses FourSquare API, it can be modified to support various different scales of operation. This program is already capable of determining the most similar neighborhoods within the same city, whether New York or Toronto – provided minor edits in the source code. However, this program can also be tailored to explore different pairs of cities across the world, as long as FourSquare has the venue data for a given location and the geospatial data of the location is publicly available. Additionally, this program can also be expanded to compare entire cities, counties or provinces (states). With a few edits, this program and also be tailored to support businesses with their expansion plans.

Unfortunately, one of the key strengths of this program is also it's greatest weakness. As this program uses publicly available data, the quality of this data would directly affect the results given by this program. Furthermore, the geospatial data may not be available for most locations. This renders this program helpless. However, this is something that can be built up in the coming years.

Finally, it should once again be noted that this program only provides a first-look in comparing neighborhoods. It uses venue data provided by FourSquare API and creates results that are the most similar on paper. However, in reality there are a lot more factors that affect a neighborhood that are impossible to account for. In addition to amenities, the social culture of a neighborhood, it's history and its inhabitants might be far more important for most relocators. Thus, this program should only be used as a guide and its results should be supplemented with further research.

CONCLUSION

This program provides a way for travelers between the cities of New York and Toronto to assess the most similar neighborhoods in their destination and assist them in their relocation efforts. This program uses cosine similarity to determine the most common neighborhoods and presents this information in a graphically intuitive way using Folium library.

This program has the potential to be able to support a wide variety of purposes, from individuals to large scale businesses. There is a lot of work to be done, but the future is full of possibilities.