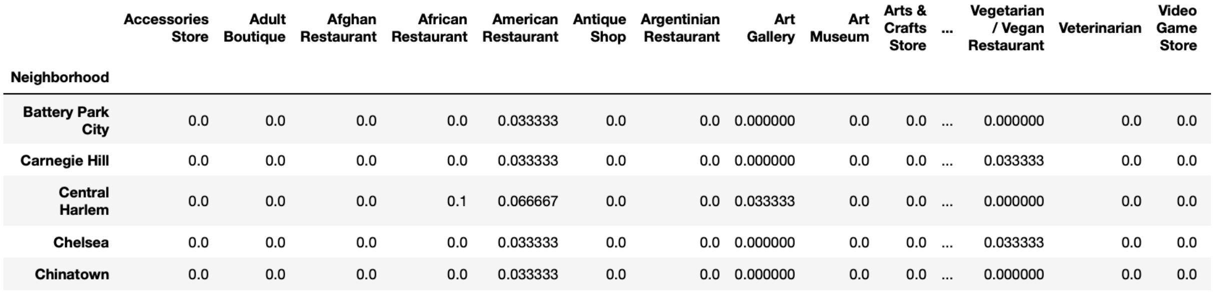
Recommendation system Based on Similarities Between Venues

Introduction/Business Problem

When people want to relocate another place, they usually look for same city life. The new place should has similar facilities and venues. In this project, new neighborhood will be recommended between Toronto and Manhattan. The similarity metric is venue categories and their frequencies in neighborhoods. The venue data is fetched via Foursquare API.

In business point of view, this kind of recommendations can be needed by real estate companies. They can make recommendations between any cities for a customer. Besides, recommendation algorithm does not have too much time complexity. So, they can make instant recommendations by creating a simple API.

Data



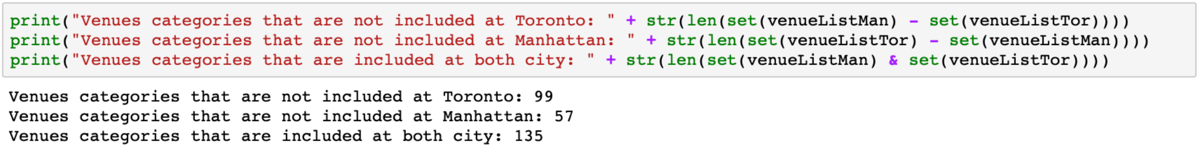
	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	...	Vegetarian / Vegan Restaurant	Veterinarian	Video Game Store
Neighborhood														
Battery Park City	0.0	0.0	0.0	0.0	0.033333	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0
Carnegie Hill	0.0	0.0	0.0	0.0	0.033333	0.0	0.0	0.000000	0.0	0.0	...	0.033333	0.0	0.0
Central Harlem	0.0	0.0	0.0	0.1	0.066667	0.0	0.0	0.033333	0.0	0.0	...	0.000000	0.0	0.0
Chelsea	0.0	0.0	0.0	0.0	0.033333	0.0	0.0	0.000000	0.0	0.0	...	0.033333	0.0	0.0
Chinatown	0.0	0.0	0.0	0.0	0.033333	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0

Above figure is a sample from Manhattan dataset. It consist of neighborhoods and venue categories. Cells represent the frequency of the venue category. For example, “Central Harlem” has 10% African restaurants in its most popular venues. Venue data is fetched from Foursquare API. Fetching parameters are the location of neighborhood, radius of searching area for venues and limit of number of venues. Searching area radius is 500 meters, limit is 30. So, for any neighborhood the number of popular venues can be 30 at most.

Manhattan dataset has 40 rows and 234 columns. Same dataset is fetched for Toronto as well. The Toronto dataset has 38 rows and 192 columns. So, this means Manhattan has more variety of venue categories than Toronto.

Data Preprocessing

Toronto and Manhattan data frames do not have same types of venues. For computation of similarity, different venue types are unnecessary.



```
print("Venues categories that are not included at Toronto: " + str(len(set(venueListMan) - set(venueListTor))))
print("Venues categories that are not included at Manhattan: " + str(len(set(venueListTor) - set(venueListMan))))
print("Venues categories that are included at both city: " + str(len(set(venueListMan) & set(venueListTor))))
```

```
Venues categories that are not included at Toronto: 99
Venues categories that are not included at Manhattan: 57
Venues categories that are included at both city: 135
```

New data frames are created by including only mutual 135 venue categories. After this step, method can be applied.

Method

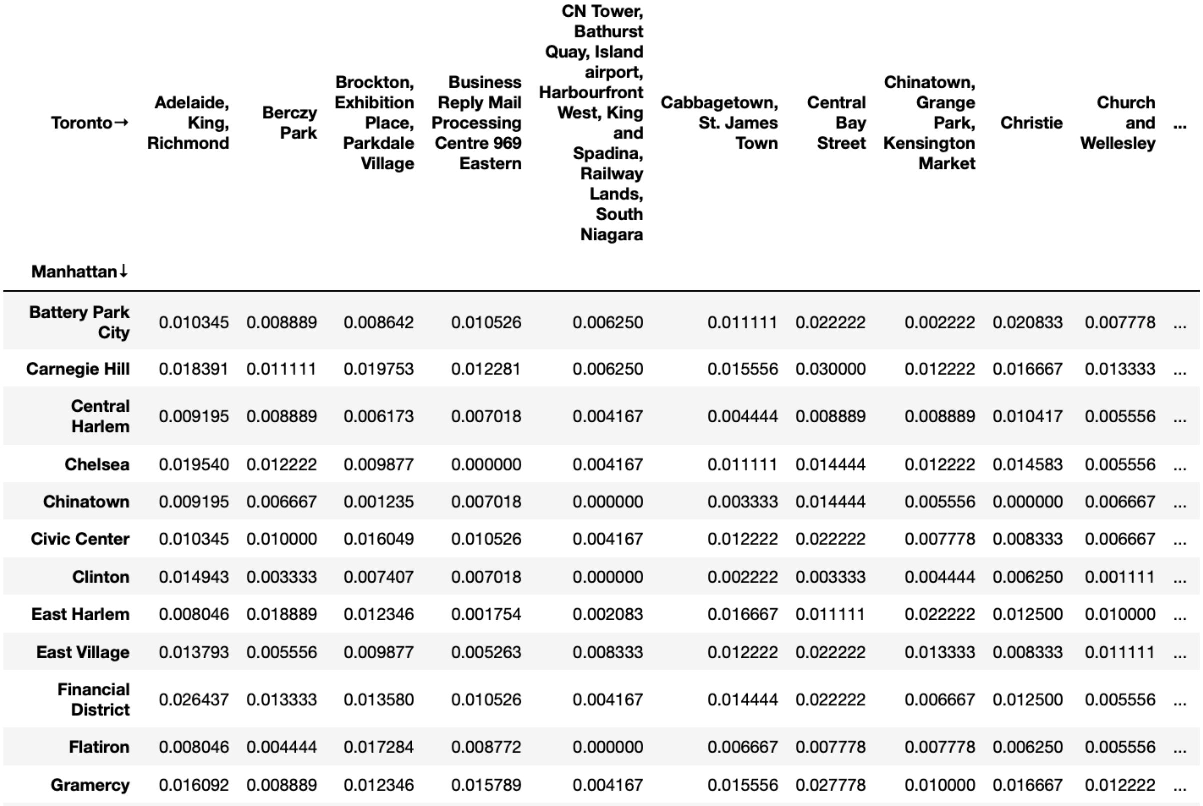
For finding similarity between two neighborhoods, same venue frequencies should be multiplied one by one. Sum of the products will give similarity score. All of similarity scores can be calculated by a matrix multiplication. Transpose of second matrix is multiplied by first matrix. Result matrix consists of similarity scores.

$$\underbrace{\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}}_{Toronto} \cdot \underbrace{\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}^{T}}_{Manhattan} = \underbrace{\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}}_{Similarity\ Matrix}$$

This method can easily be applied by numpy library.

```
resultMatrix = pd.DataFrame(np.dot(mutualToronto, mutualManhattan.T))
```

Results



Toronto→	Adelaide, King, Richmond	Berczy Park	Brockton, Exhibition Place, Parkdale Village	Business Reply Mail Processing Centre 969 Eastern	CN Tower, Bathurst Quay, Island airport, Harbourfront West, King and Spadina, Railway Lands, South Niagara	Cabbagetown, St. James Town	Central Bay Street	Chinatown, Grange Park, Kensington Market	Christie	Church and Wellesley	...
Manhattan↓											
Battery Park City	0.010345	0.008889	0.008642	0.010526	0.006250	0.011111	0.022222	0.002222	0.020833	0.007778	...
Carnegie Hill	0.018391	0.011111	0.019753	0.012281	0.006250	0.015556	0.030000	0.012222	0.016667	0.013333	...
Central Harlem	0.009195	0.008889	0.006173	0.007018	0.004167	0.004444	0.008889	0.008889	0.010417	0.005556	...
Chelsea	0.019540	0.012222	0.009877	0.000000	0.004167	0.011111	0.014444	0.012222	0.014583	0.005556	...
Chinatown	0.009195	0.006667	0.001235	0.007018	0.000000	0.003333	0.014444	0.005556	0.000000	0.006667	...
Civic Center	0.010345	0.010000	0.016049	0.010526	0.004167	0.012222	0.022222	0.007778	0.008333	0.006667	...
Clinton	0.014943	0.003333	0.007407	0.007018	0.000000	0.002222	0.003333	0.004444	0.006250	0.001111	...
East Harlem	0.008046	0.018889	0.012346	0.001754	0.002083	0.016667	0.011111	0.022222	0.012500	0.010000	...
East Village	0.013793	0.005556	0.009877	0.005263	0.008333	0.012222	0.022222	0.013333	0.008333	0.011111	...
Financial District	0.026437	0.013333	0.013580	0.010526	0.004167	0.014444	0.022222	0.006667	0.012500	0.005556	...
Flatiron	0.008046	0.004444	0.017284	0.008772	0.000000	0.006667	0.007778	0.007778	0.006250	0.005556	...
Gramercy	0.016092	0.008889	0.012346	0.015789	0.004167	0.015556	0.027778	0.010000	0.016667	0.012222	...

This is a sample of similarity matrix. Columns are Toronto neighborhoods, rows are Manhattan's. First three recommendation pair which have maximum similarity scores are:

1. 'Stuyvesant Town' - 'Rosedale',
2. 'Battery Park City' - 'Rosedale' and
3. 'Stuyvesant Town' - 'Moore Park, Summerhill East'

These recommendations are reasonable since all of neighborhoods are known by their parks. But different kind of cities must be considered for being sure about recommendations. When we look at Central Bay Street which has many coffee shops, first recommendations are Hamilton Heights, Carnegie Hill, Morningside Heights and Murray Hill. All recommendations have high number of coffee shops as well.

Discussion

Although recommendations are rational, venues are not only parameter to choose a new neighborhood. Real estate companies can utilize this resource but they must also consider other parameters to make better recommendations. As another disadvantage, some of neighborhoods does not have variety of venue categories. As a result, all of similarity scores are zero for Roselawn. Lastly, rather than similarity score, a recommendation score can be created which considers dissimilar venues as well. Let's assume A neighborhood has same similarity scores for B and C neighborhoods. However, if B has more additional venues than C, B should get more score. That is why a new metric can be useful.

Conclusion

In this project, a recommendation system is built between neighborhoods of Toronto and Manhattan. Data including venues for neighborhoods are fetched via Foursquare API. Then, mutual venue categories considered to get similarities between neighborhoods. After this step, similarity matrix is created which consist of similarity scores between neighborhoods. Results are analyzed whether they make sense or not. It is seen that results are successful for recommendation. Finally, disadvantages and possible remediations are discussed.