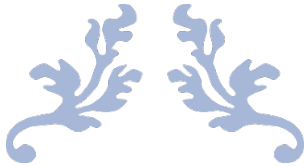


NEW YORK AND TORONTO - AN ANALOGY

COURSERA - CAPSTONE PROJECT REPORT

JULY 8, 2019

SUDHINA S.

Contents

1. Abstract

This project deploys unsupervised machine learning algorithms to compare the cities of New York and Toronto using venue information of the two cities from FourSquare, a Location Based Social Network site (LSBN). It tries to understand the degree of similarity between the two cities while grouping similar neighbourhoods into clusters. This neighbourhood-based comparison of the two cities would be highly helpful to tourists, migrants, businesses and government bodies in recommending locations, urban planning, in the analysis of cities and urban computing.

2. Introduction

2.1 Background

'Stranger you are, Stranger am I...

Nomads in the jungle of civilizations formed by our nomadic forefathers!'

People flock from city to city, country to country in search of better job, better living conditions, better opportunities, sometimes to quench their traveller spirit and so on...To explore new places, new cultures is often life-changing an experience that walk with us to the grave and beyond! So, we see many around us possessed by the gypsy spirit - those who '*travel for travel's sake*!' (R.L. Stevenson, 1878) When we voyage, we often seek novelty. Imagine spending your time, money and effort to end up in a place which is strikingly similarly to the last place you had been to, which has nothing new to offer!

Nuances of 'migration' extend the limits of just a geographical shift to a cultural shift. Being a-few-days visitor is far different from being some-long-years resident. To uproot one's self from the conversancy of homeland to the boons and perils that the novelty offers is easier said than done. Even as enticing and developed the new city/country/neighbourhood may be or perilous and under-developed motherland might be, one's heart always yearns for home. Like people say, a part of your home will always travel with you. Afterall '*homeland is one of the magical fantasy words like unicorn and soul and infinity that have now passed into language*'(Zadie Smith, 2000). How wonderful it is to be able to relocate to a better city which is reminiscent of our home - to be able to fly for our dreams under the aegis of familiarity that our homeland offers!

2.2 Business Problem

This project seeks to cushion the transition for people migrating between the cities of Toronto and New York. Toronto and New York are urban cynosures coveted by many-an-urbanizing youth. They house major industrial, educational and cultural hubs. These cities are growing every passing second! So are the number of people travelling/relocating to/between the two cities!

The exercise clusters the two cities and compares the similarities and dissimilarities between the two based on their major venues. It addresses the question 'How similar are Toronto and New York?' It takes the goal a step further to group similar neighbourhoods in the two cities, for instance, which neighbourhoods in New York are more similar (read ' more comfortable') for a person from East York, Toronto?

2.3 Business Interests

Comparison across cities helps business, people and government alike!

Owing to the growing globalization (and hence the growing numbers of MnC's) and urbanization, workforces transcend geographic boundaries. Business today demands travel and relocation (however long or short a period it may be). Employees are deployed to distant cities frequently and quickly! So, the migration trends are booming. To add to that, people today (in general) relish travel and change. They yearn for heights and are ready to strive for it, shift for it. Any person travelling to a new city looks it up in advance, tries to prepare oneself for the change. While tourists try to check out what is different in the place, people who are relocating try to select a place as close to home. In short, all of us benefit from knowing what is similar and what is not between cities.

City-to-City comparisons helps a firm to zero in on its best location to start/expand to, a tenant to decide on the best rent available in a new place, a professional to identify prime zones/cities to work etc.

Comparing cities also helps in urban planning - to understand what is going on where, to realize best inter-city plans, to analyse what strategy might do well and what might not, to help maintain balanced, decentralised growth and so on...

3. Data Acquisition and Cleaning

3.1 Data sources

To compare and cluster the cities of New York and Toronto, Canada postal code dataset available in the wiki page (coupled with geospatial data) and New York city neighbourhood dataset (source: New York city Spatial Data Repository site) are used. Further, venue information for each city is obtained from FourSquare, a Location Based Social Network site. The venue information is easily accessible through a public REST API.

3.2 Data Preparation and Wrangling

The borough, neighbourhood, postal code data obtained from Canada Postal code Wikipedia page are combined with latitude and longitude data from the geospatial data file, cleaned and formatted to form a pandas dataframe. Only the rows of data that have a borough assigned are processed for further use. All cells with no borough value or value of 'Not Assigned' are ignored. For rows having borough data but no neighbourhood data, the neighbourhood is taken to be the same as the borough. Since the geospatial data file used provides latitude and longitude data per postal code, the borough-neighbourhood data for Toronto is grouped by postal code to reduce redundancy.

The borough, neighbourhood, latitude and longitude data for New York city is in already available in a clean JSON format from New York city spatial repository site. This JSON file is loaded and normalized to get the required Pandas dataframe.

Venue information is obtained through Foursquare API using the latitude-longitude data for each neighbourhood. Venue information (upto 25 venues) in each neighbourhood within a radius of 500m are fetched by the Foursquare API in a JSON format. Name, Latitude and Longitude data of venues are filtered from the results (all of these required fields are present under 'Features' key), added to the respective neighbourhood name, latitude, longitude data and then grouped by neighbourhood to form a new dataframe. Since Scikit-learn library doesn't process categorical values, the venue columns (all categorical data columns) are encoded (one-hot encoding) before being passed to the algorithm for data analysis. This dataframe basically contains the mean of each venue category per neighbourhood. However, the usability/readability of the encoded dataframe is minimal. Hence, the unlabelled data is then sorted and grouped to get a new dataframe with 10 most common venues per neighbourhood. After the run of the clustering algorithm, the cluster label is added back to this dataframe for exploring the results.

The encoded dataframe and sorted dataframe for the two cities are concatenated to get the corresponding dataframes for combined dataset. It is also noted that both cities have a neighbourhood with name, 'Rosedale'.

4. Methodology

4.1 Exploratory Data Analysis

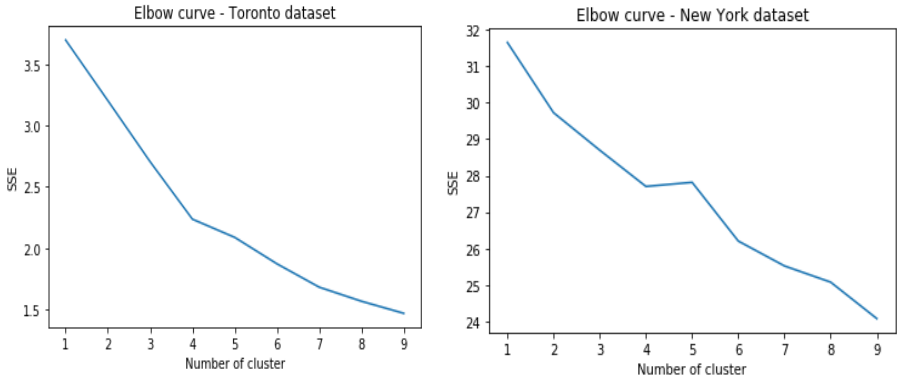
Folium is used to visualize the venues on the map of each city. Being an unsupervised problem and the dataset being normalized, there is little-to-no other data exploration done.

4.2 Clustering the Cities

The cities, Toronto and New York, are sub-divided into groups of their own using clustering algorithm and these clusters are compared to see how similar or dissimilar the trends in the two cities are. A third clustering is done on the combined dataset to group similar neighbourhoods in both cities together.

The first step in data analysis is to choose the appropriate machine learning algorithm. The major clustering algorithms are K-Means Clustering, Hierarchical clustering (Agglomerative and Divisive), Mean-Shift Clustering, Spectral Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM) and Affinity Propagation. Among these, K-Means Clustering and Hierarchical clustering algorithms are the most popular and hence used in the study. Since the dataset is large and cophenetic distance for Hierarchical clustering is an abysmal 0.41, it is concluded that K-Means Clustering is the better alternative.

To run K-Means clustering, the number of clusters(n) has to be set prior to running the algorithm. This number is found out for all 3 datasets – Toronto data, New York data, combined data. The optimal number of clusters is decided using elbow method and computing silhouette coefficient. For the Toronto dataset, the SSE (Sum of squared errors) vs n curve has two dips/elbows at points 4,5 and 7. For the New York data, the curve dips at n values of 2,4,5,6,7 and 8. The curve for the combined dataset has dips at points 2, 6, 7 and 8. For all three datasets, the curve is more linear than desired.



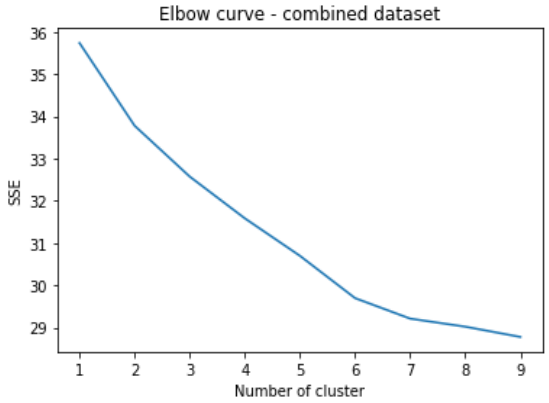


Fig. 1. Elbow Curves for Toronto data, New York data and combined dataset

Silhouette coefficient is computed for all three datasets using Cosine and Euclidean metric. For all 3 datasets, Euclidean metric gives far better values of silhouette coefficient compared to cosine metric. Hence the actual clustering is done based on Euclidean metric. The Silhouette coefficient is highest for n values of 3, 2, 4… (all Euclidean, in descending order of coefficient magnitude) for Toronto data. While the sequence of best n values goes 4, 3, 6, 5… (all Euclidean) for New York data, the same for the combined dataset is 2, 3, 4, 7… (all Euclidean).

Comparing the measures from elbow method and silhouette coefficient method, the n-value is decided as 4 for Toronto data, 4 for New York data and 6 for combined dataset.

K-Means++ is used to initialize the clusters to hasten the convergence. Similarly, random state variable is set to make runs more deterministic. Once clustering is done, the cluster labels are added back to the venue-sorted dataframe and appended with borough, latitude-longitude data for the respective neighbourhood. This helps in understanding the results better as well as viewing the clusters on the map of the cities. The clusters are then examined separately for each run of the algorithm. For the combined dataset, the cluster-labelled sorted dataframe is split into Toronto set and New York set before inserting borough, latitude-longitude data for each neighbourhood. This is to avoid errors rising due to presence of a neighbourhood with same name (Rosedale) in both the cities.

These clusters are then viewed on the map (using Folium).

5. Results and Discussion

It is seen that both the cities are similar to a decent degree. However, New York offers more variety to its people. While New York city has 367 unique categories of venues, Toronto offers 177 unique venue categories. Partly, this can be attributed to the larger size (and hence a greater number of neighbourhoods) of New York over Toronto. While Toronto houses 74 neighbourhoods, the same number for New York is 306.

In the Toronto dataset, top venues of Cluster 1 are Parks, Coffee shops, Pubs/gastropubs, ethnic restaurants (Asian, Greek, Japanese, Indian etc.), Yoga studios, Discount stores, Liquor store/Brewery, etc. Cluster 2 features Cafes/Coffee shops/Diners, Gyms, Performing arts venues, Restaurants (Italian, Eastern European, Falafel, Ethiopian etc.), and so on. Cluster 3 comprises Cafes, Farmers markets, Bars (including Beer Bars), Steakhouses, Restaurants (Asian, Mexican, Seafood, Vegan etc.), Clothing stores and the like. The final cluster (fourth) is a small one consisting of two neighbourhoods which promises Outdoor centres/Parks/Trails/Yoga studios/Dance studios, Eastern European/Ethiopian restaurants, Health food stores among others.

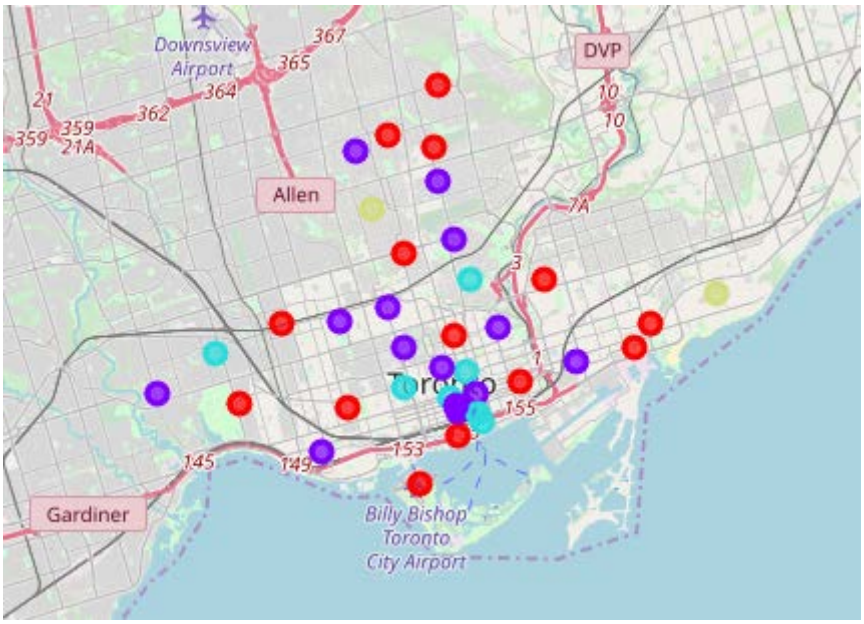


Fig. 2. Neighbourhood Clusters – Toronto (Independent run)

In the New York dataset, most recurring venues in cluster 1 include Pizza places, Deli/Bodega's, Bus stations, Banks, Donut shops, Pharmacies, Mobile phone shops, Restaurants (Fast Food, Mexican, Spanish, Caribbean etc.), Fried Chicken Joints, Supermarkets, Bakeries, etc. It also boasts of a number of other categories like Bowling alleys, Skating rings, Race tracks, Hobby shops, Martial Arts Dojo's and many more. Cluster 2 mainly comprises Chinese/Italian/Korean/Fast food restaurants, Coffee shops/Cafes/Bakeries, Yoga studios, Farms, Farmer markets and Factories. Deli/Bodega's, Beaches, Parks, Falafel/Italian/Fast Food/Filipino restaurants, Farms, Fields, Farmer markets, Factories etc. appears in Cluster 3 while Cluster 4 is devoted to five neighbourhoods with Harbor/Marina, Fish and chip shops, Falafel/Fast food/Filipino restaurants, Farms, Fields, Farmer markets, factories etc. Clusters 4 is similar to other clusters, especially cluster 3.

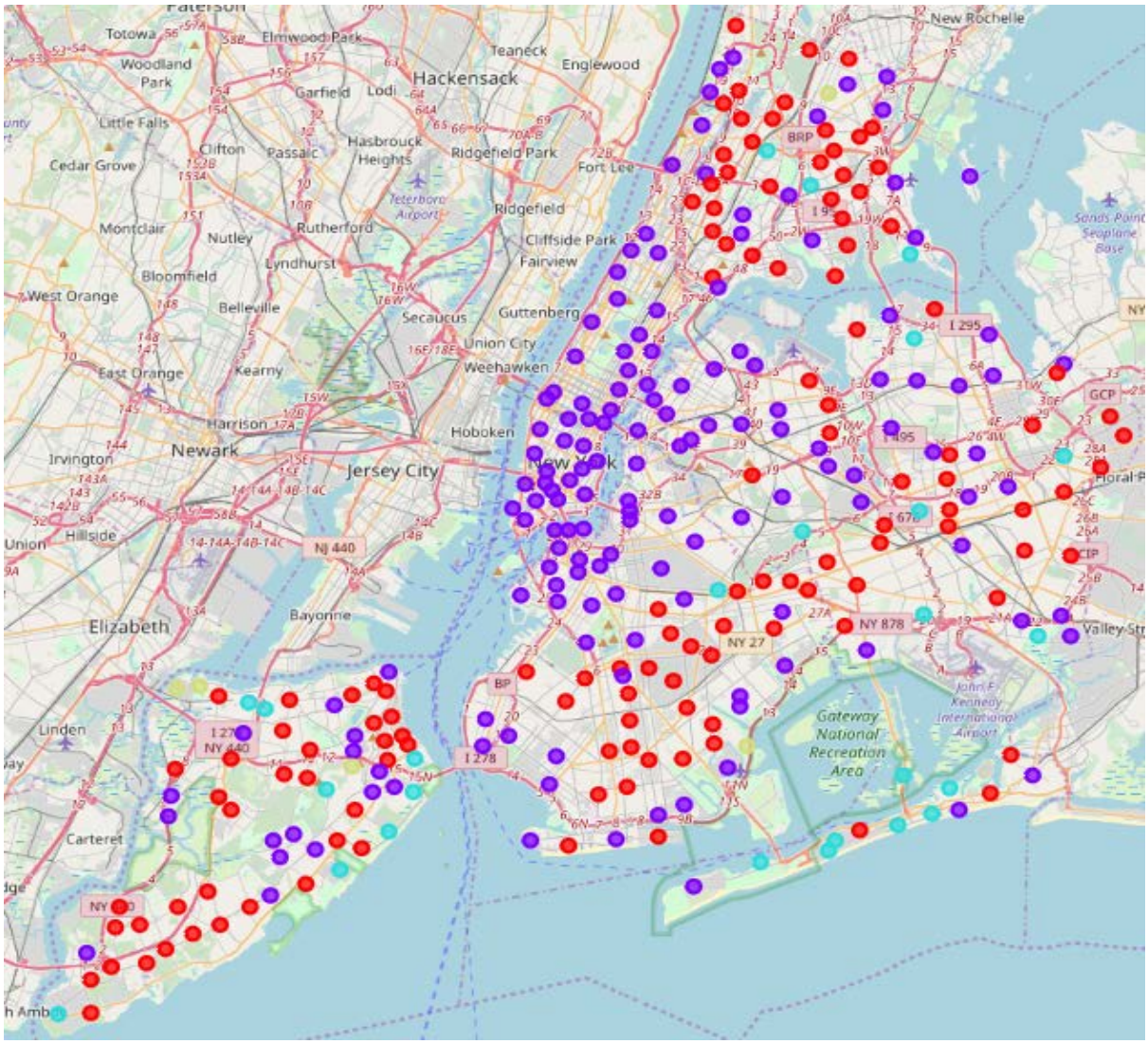


Fig. 3. Neighbourhood Clusters – New York (Independent run)

In the merged dataset, six clusters are there, of which most of the Toronto neighbourhoods fall in the fourth cluster (Fig. 4). Many New York neighbourhoods fall into the same cluster as well (Fig. 5). So, it can be inferred that the two cities are indeed similar. The other clusters are more or less similar to the New York city clusters. The prominent venues in the fourth cluster (or in other words, the features that make the two cities similar) are Cafes/coffee shops/bakeries, Restaurants (Italian, Sushi, American, Fast food, Mexican etc.), Parks, Yoga studios, Pizza places, and all that screams 'Millennials and Generation Z'. Notably, Pizza places are much more prevalent in New York than Toronto. It has to be also noted that the much larger scale of New York city wields a profound impact on the comparison. In every cluster, the number of New York neighbourhoods outweigh the number of Toronto neighbourhoods. Therefore, New York neighbourhoods have the upper hand in deciding the prominent venues in each cluster. In short, the two cities are similar in their iconizing of American culture and different in their degrees of doing so – with New York being a bigger and better torch-bearer.

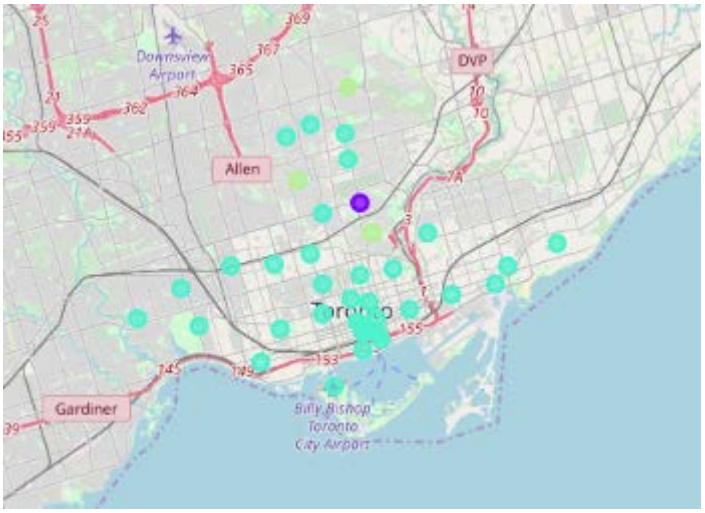


Fig. 4. Neighbourhood Clusters – Toronto (Combined Dataset Clustering)

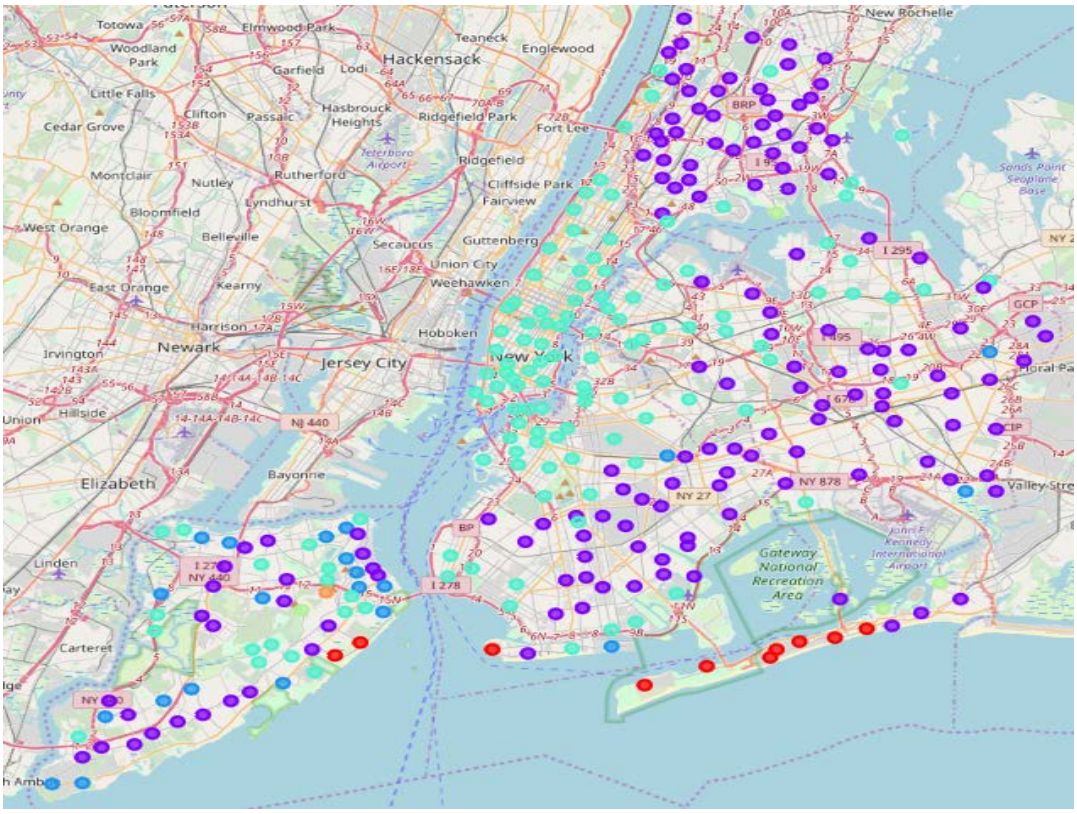


Fig. 5. Neighbourhood Clusters – New York (Combined Dataset Clustering)

In this project, we have used only venue information to compare the two cities. A more holistic comparison can be obtained by incorporating activity data (check-ins/internet data patterns/crime data etc.), census data of the cities and so on. Similarly, only two clustering algorithms has been used/reviewed in the project. Reviewing the efficacy of the other algorithms as well before deciding on one would be a better approach.

6. Conclusion

In this project, neighbourhoods in the cities of New York and Toronto are compared based on their venues. K-Means Clustering algorithm is used in the study. The study concludes that the cities are alike to a fair degree but in different stages of urban growth. Both Toronto and New York are beacons of American culture – open to all, designed to live life to its fullest! There is a multitude of options available to every sect of society – the nature-lovers, the adventurists, the shoppers, the foodies, the sporty, the home-makers and so on, each of these options replete with choices! Toronto is, however, small compared to New York with lesser number of neighbourhoods, venues and venue categories. It is inferred that with the inclusion of more parameters (like activity data, census data etc.) and cross-validation of more clustering algorithms, the efficiency of the study can be improved.

From the user perspective, New York is certainly more enticing to a tourist with its plethora of venues but much more costly. Toronto offers similar venues at lesser costs. It is upto the visitor to choose based on their travel-taste, budget and other preferences. From a business perspective, New York is full of life, offering more customers and competition at the same time. From a governing perspective, New York is reaching the apex of its growth while Toronto is still in the growing phase. Strategic measures have to be taken to avoid growth stagnation for New York without overloading the city while Toronto needs to provide the optimal environment for its growing needs. Venue concentrations throws light on regional imbalances in the growth and can be judiciously used to rectify them. It can also be used to organize successful events. For instance, to organize a farming initiative in New York, it is best to focus on clusters 2, 3 and 4 of New York clusters.

Addressing the relocation problem defined at the start, based on the outcome of this study, migrating between New York and Toronto might be easy to most people, save alone the dimensional difference between them. New York, as mentioned earlier, is bigger, richer (in options provided) and costlier than Toronto. Tourists, therefore, might need to re-consider travelling to both the cities in quick succession as they are fairly alike.