The Battle of Neighborhoods

Comparing Toronto, Canada to US Cities

James Walsh

February 17, 2020

Coursera Capstone Project

Introduction

The focus of my project is to use Foursquare data for Toronto, Canada and compare it to that of three US cities: Chicago, Los Angeles and New York, and determine which US city Toronto is most similar to. All four cities are very diverse and have markedly different histories and cultural influences as well as large spaces of land separating them. I will use clustering and classification to compare Toronto to each of the three US cities individually. Determining which US is most like Toronto would be of value to anyone seeking to travel, move or open / expand their business in a new area.

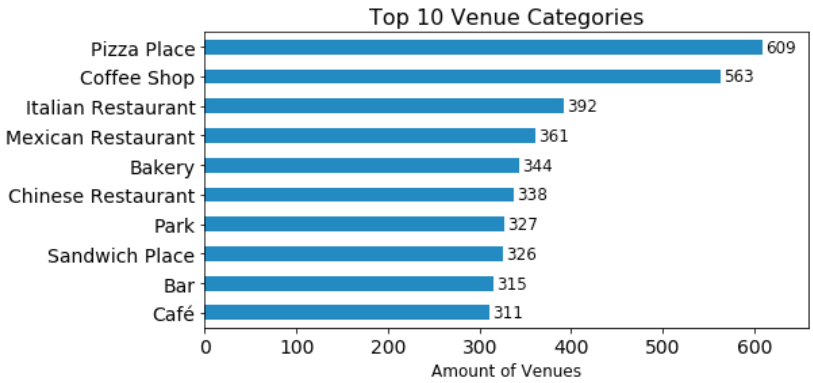
Data

To start, I needed to gather the neighborhoods and their respective coordinate data so I could make use of the Foursquare API. Methods for retrieving this data for Toronto and New York have already been outlined in previous exercises (Wikipedia and cocl.us, respectively) , so I needed to find a way of collecting similar data for Los Angeles and Chicago. Searching the internet led me to data.cityofchicago.org for Chicago city data and usc.data.socrata.com for Los Angeles. Both produced GeoJSON files that had several attributes for the neighborhoods, but I was only after their name and coordinates which I was able to put into a DataFrame. Once I had the neighborhoods and their coordinates for all three US cities and Toronto, I combined all four into one table and pulled the venue data for all neighborhoods from Foursquare which includes up to 100 venues within a 500 meter radius of the geographic coordinates for each neighborhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	City	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Grand Boulevard	41.8168	-87.6067	Chicago	Some Like It Black Creative Arts Bar	41.818432	-87.605251	Juice Bar
1	Grand Boulevard	41.8168	-87.6067	Chicago	Just Turkey Restaurant	41.815248	-87.606212	BBQ Joint
2	Grand Boulevard	41.8168	-87.6067	Chicago	Norman's Bistro	41.816795	-87.601809	Restaurant
3	Grand Boulevard	41.8168	-87.6067	Chicago	Honey 1 BBQ	41.81691	-87.60732	BBQ Joint
4	Grand Boulevard	41.8168	-87.6067	Chicago	Family Dollar	41.8138	-87.606318	Discount Store

This is the primary dataset that I will be using through the course of the project. One issue I uncovered early was the presence of duplicate venues that overlapped multiple neighborhoods, e.g. 2 Bros. Pizza (40.689, -73.981) was listed in both Boerum Hill and Downtown neighborhoods of New York City. Since I am comparing each city as a whole, I only needed one of each venue entry, so any duplicates were removed. I also found that there are some entries of the venue category “Neighborhood” which adds little value and could potentially confuse machine learning later on. Since there are only 13 such entries out of over 16,000 total, I can remove these with little concern. This dataset would be useful for analysis, but not for machine learning. Therefore, I used One-Hot Encoding to convert this table into a numerical vector where each row was a specific venue denoted as either a 1 or 0 in each column that was labeled as each unique venue category. In other words; if a venue was an ice cream shop, there would be a 1 in the “Ice Cream Shop” column and 0’s in all the others. I then aggregated all neighborhoods using the means of all the venues. This would be the final dataset used for machine learning.

Methodology



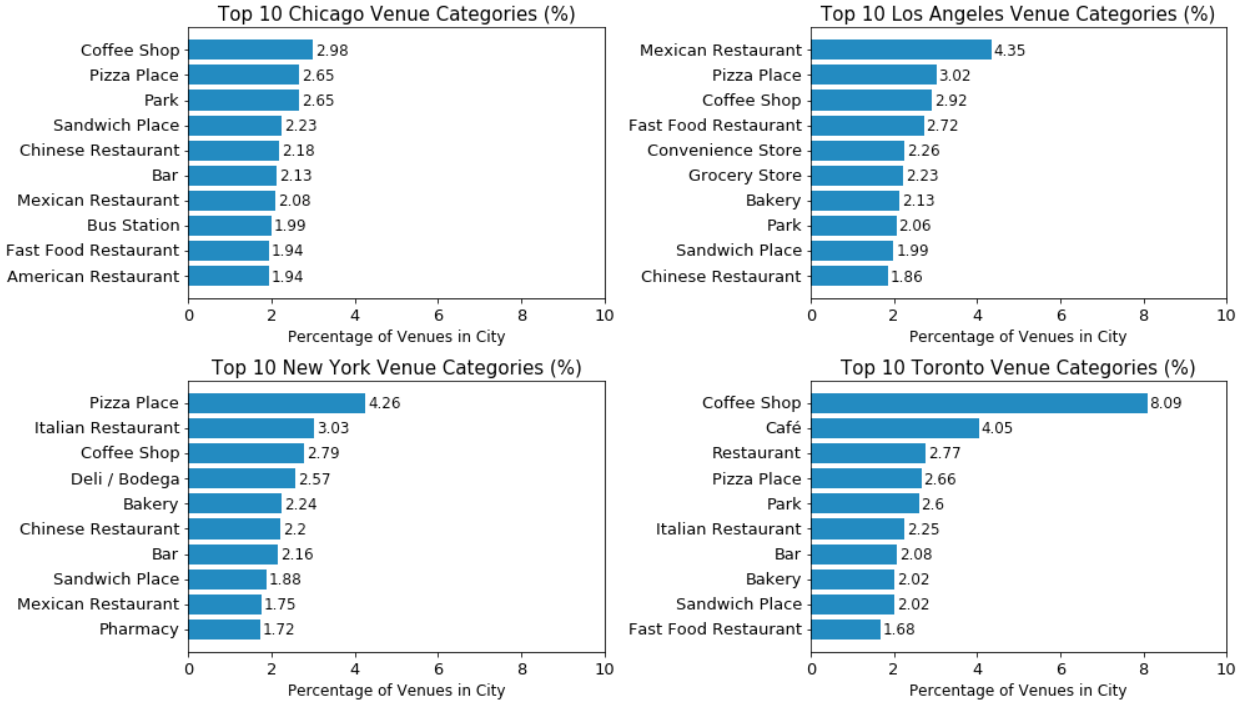
I return to my dataset prior to the One-Hot Encoding for some exploratory analysis. Across all four cities, there were 504 unique venue categories and Pizza Places were the most common, followed closely by Coffee Shops. There is a considerable dropoff after those top two that begins to level off towards the bottom of this top 10 list. Overall, the results are not shocking; Italian, Mexican and Chinese are all common non-continental cuisines and cafes, bars and other such small eateries are common as well. I am intrigued by the presence of Parks in the top 10. I did expect that there would be many, but not the 7th most out of all 504 venues ahead of the likes of bars and cafes. In fact, Parks are the only non-dining venue in the top 10.

These findings led me to wonder if any of these venue categories were strongly correlated to the others. I presumed there would be some significant correlation amongst this group, but I found that there was actually very little. In fact, out of the top 10 venue categories for all four cities, Pizza Places and Parks had the strongest (although negative) correlation at -0.157. It seems that there are not any significant correlations amongst this top 10 group, but surely each venue category is correlated to some other type outside the top 10. These results were a bit more interesting. Sandwich Shops and ATM's shared the strongest correlation at 0.398 followed by Chinese and Dim Sum (a particular style of Chinese cuisine) at 0.318. Cafés and Coffee Shops being correlated to college facilities makes sense as both are very common fixtures on / around college campuses in North America. I have already noted the negative correlation between Parks and Pizza Places, and it is interesting that Mexican Restaurants' greatest correlation is to Strip Clubs.

	Dependant	Correlation
Sandwich Place	ATM	0.398361
Chinese Restaurant	Dim Sum Restaurant	0.317655
Bar	Drugstore	0.31589
Café	College Stadium	0.266386
Coffee Shop	College Auditorium	0.253476
Italian Restaurant	Spa	0.225283
Pizza Place	Empanada Restaurant	0.200907
Mexican Restaurant	Strip Club	0.17039
Bakery	Dance Studio	0.135972
Park	Pizza Place	-0.156821

While these figures apply to the dataset of Toronto, Chicago, Los Angeles and New York combined, I wanted to get a better idea of how each city's venues compared to each other, since that is the focus of the project. Since the venue totals of each city varies considerably (New York: 428, Los Angeles: 320, Chicago: 302, Toronto: 265) I looked at the percentage of the respective total for each city rather than the

raw totals. These results were actually quite in line with my assumptions for each city. The highest percentages in Chicago are Coffee Shops, Parks and Pizza Places. Coffee Shops and Parks are common in most places, and Chicago does claim fame for their pizza. It is interesting to see Bus Stations so high.



Los Angeles' most common venue by a sizable margin is Mexican Restaurants, which makes sense given the proximity to the US / Mexico border and the heavy cultural influence. This is also the only city where I saw Grocery Store and Convenience Store in the top 10. New York's most common venue is Pizza Places, which is not surprising as New York, like Chicago, takes great pride in its pizza. (Ask a local of either city whose pizza is best. I promise it will be entertaining.) Italian Restaurants and Deli / Bodegas are also very characteristic of New York. For Toronto, it was surprising not that Coffee Shop was the most common venue, but that it was so by nearly double that of the next category. Aside from that, there is some Italian, but not the same prevalence of ethnic cuisine that can be found in the US cities. From this data alone, it is hard to discern which US city Toronto is most like, so I'll turn to machine learning.

My first exercise will be with clustering, where I will employ k-Means. I will create three data sets that pair Toronto with each of the three US cities individually. I will use k=2 since there are two cities in each set and I want to determine how similar each city is to Toronto. If the cities are clustered exclusively (i.e. Toronto in one cluster and the US city in the other) than the cities are not very alike. The level of mixing in the clusters will give an indication of the similarities of the cities. To account for class imbalance, I prefer undersampling and will set the sample amount equal to that of the city with the lowest neighborhood count. After one attempt, the results were not quite what I was hoping for. The clusters were not clearly demarcated by city and one cluster is roughly 10x the size of the other. At a quick glance, Los

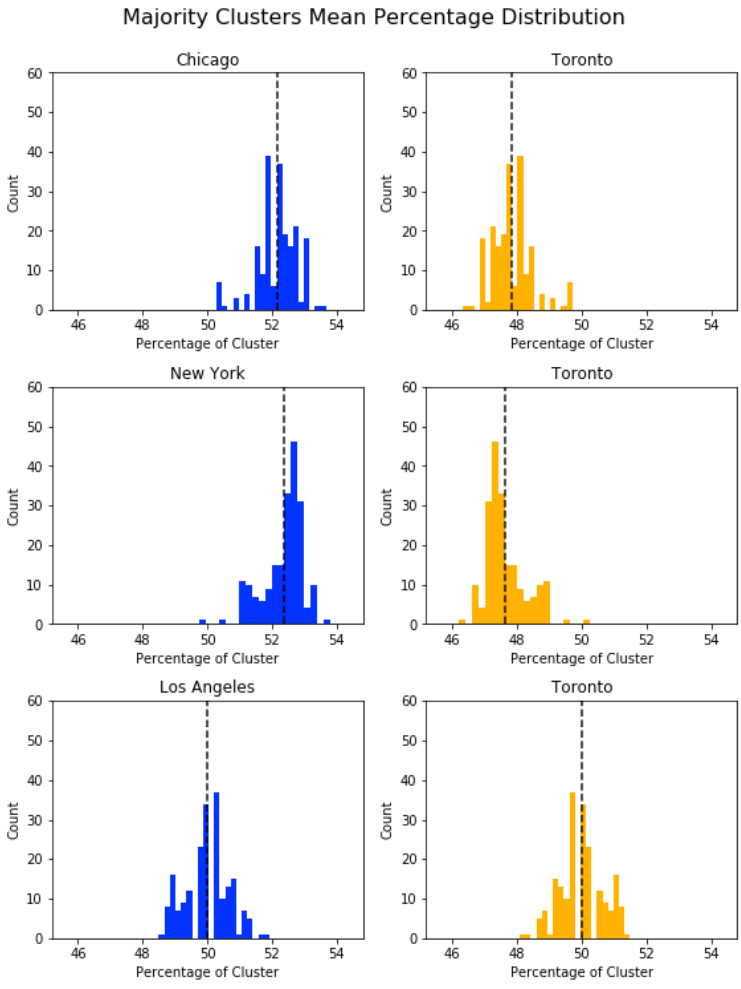
```
TORONTO COMPARED TO LOS ANGELES
Cluster 1
Los Angeles     94
Toronto         93
Name: City, dtype: int64
Total: 187

Cluster 0
Toronto          7
Los Angeles      6
Name: City, dtype: int64
Total: 13
```

Angeles neighborhoods are the most evenly clustered with Toronto's. I wanted to better understand how the algorithm was splitting the data so I calculated the most common venue categories for the clusters of each Toronto / US city pair. For each pair, the larger cluster had Coffee Shop and Pizza Place as the top two most common venues where the smaller cluster had Parks as the most common venue category followed by the likes of Convenience Stores, Basketball Courts, Lakes and Playgrounds. It seems the major distinction between the two clusters for all three pairs is that the larger is dining options while the smaller is non-dining venues. With this understanding, I will take both clusters into account when comparing how even the clusters compositions are and retain my assumption that clusters closer to 50/50 are indicative of more similar cities.

I also complimented the clustering with some classification. I used the same dataset as in clustering, but used the city names as the labels to predict. I would check which Toronto neighborhoods were labeled incorrectly and what city label was applied. The most common incorrect label will be the US city Toronto is most like. Logistic regression is an appropriate option since this is binary classification and the results are fairly easy to explain. This time, I would use oversampling of the dataset as well as Grid Search for model tuning.

Results

Because I am using undersampling for clustering, I want to get multiple sets of samples and results out of k-Means. I ran the sampling and model fitting 200 times and recorded the mean percentages of the clusters from each iteration. The means of the larger cluster for each Toronto/US city pairing are very close to 50/50, but Toronto/Los Angeles are exactly even at 50% for each city. The New York and Chicago pairings are close to 50/50, but both are just slightly off. The difference is small, yet consistent. Histograms of the distribution of these mean percentages show the the ranges of the Toronto/US city pairs and that Toronto/Los Angeles are perhaps the most “normal” distribution with their means right at the 50% mark. The differences in the smaller clusters are far more pronounced. The percentages of the Toronto/Los Angeles pair have means of



53.14% Los Angeles and 47.33% Toronto, much closer to an even split than either of the two other US cities (Toronto/Chicago: 77.33% Toronto, 23.13% Chicago; Toronto/New York: 82.76% Toronto, 18.78% New York). This lends further credence to the presumption that Toronto is more like Los Angeles than either of the other US cities.

With classification, the tuned Logistic Regression model produced an accuracy score of 96.7%, which is really quite good. The predicted labels show 2.58% of Toronto neighborhood samples were mislabeled as Los Angeles with none mislabeled as either New York or Chicago. The percentage is small, to be sure, but it does provide another indicator that Toronto is more like Los Angeles than either New York or Chicago.

```
Tuned Logistic Regression Parameters: {'C': 100}
Accuracy score is 0.967
Best score is 0.973
```

```
Toronto        97.42
Los Angeles     2.58
Name: Predict, dtype: float64
```

Conclusion

The indicators may not be especially profound, since all four cities are in North America, very large, economically developed and culturally diverse. It would be fun to include even more cities in this exercise and see which are most similar. With that said, the data does illustrate certain patterns and both classification and clustering point to Toronto having more in common with Los Angeles than either Chicago or New York.