Analysis Neighbourhoods of Toronto to find Neighbourhoods in which to open a new coffee shop

IBM Coursera Data Science Specialisation Capstone Project Report

Ash Rosvall

2019

I. Introduction

This report is for the final course of the Data Science Specialisation Professional Certificate, created by IBM and hosted by Coursera. For this certificate the student needs to clearly define a problem or idea of their choice where they will leverage the Foursquare location data to solve or execute. The Foursquare API will be used in conjunction with other data to present a way in which the chosen problem will be solved, with code supplied in a Jupyter Notebook to back up the findings in this report.

The main goal of this report is to explore the neighbourhoods of Toronto in order to identify neighbourhoods in which the opening of a coffee shop / café is most likely to be successful. This idea comes from the huge popularity of these venues worldwide and the often lucrative business coming from these venues, along with the high consumer rate for coffee meaning that there is a large pre-existing customer base which will be looking for a coffee shop in high traffic areas.

The report will identify key features which have already made coffee shops in neighbourhoods extremely popular, then filter those down to find ideal neighbourhoods which have not already been overpopulated with coffee shops. Finally, it will suggest neighbourhoods in which a new coffee shop will have enough foot traffic and interest to be successful, but not so much competition that it will be an oversaturated market.

The target audience for this report is:

- Potential business investors looking to invest in a pre-existing café in an up and coming neighbourhood with desirable venues around it
- Customers interested in neighbourhoods which are not already oversaturated to find a new location likely to provide something that meets their needs
- New businesspeople looking to open a café that is likely to succeed in a market which can often be too competitive in certain neighbourhoods

II. Data Description

Toronto neighbourhoods were chosen for this project due to the abundance of information given in the prior assignments within the Capstone module. The data that will be used for this project and report are from:

- Website scrapings from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M will be used to identify Toronto neighbourhoods
- Geospatial data will be given both by the geocode python package: https://geocoder.readthedocs.io/index.html and the csv file of Toronto postcode coordinates which can be downloaded from here: http://cocl.us/Geospatial_data
- Foursquare API, which provides the venues which surround the postcode coordinates and can be mined for information

To collect and clean the data

- The Toronto neighbourhood data will be scraped from the website and tidied, combining with the postcode coordinates from the geospatial data to create a dataframe of information
- For each neighbourhood in this dataframe, the coordinates will be passed to the Foursquare API using the explore endpoints to find the venues in a defined radius
- The neighbourhoods will then have the occurrence of each venue counted, with the dataframe transformed using one hot encoding to turn each venue type into a column with occurrence as the value

This will result in a dataframe where each row represents a neighbourhood in Toronto and each column is the occurrence of the venue in that neighbourhood. The data can then be processed for the rest of the project, finding the neighbourhoods where coffee shops are most prevalent, where other venues which may be related to coffee shops are also prevalent (e.g. parks, sports venues), and also to find the correlation between coffee shops and other venues through statistical analysis. Machine learning techniques will then be used to find the neighbourhoods which fit the criteria of suited for a coffee shop but not overcrowded already.

III. Methodology

There is an assumption within this data that the common venues in each neighbourhood have a relationship with each other, and that this relationship can be measured in some way to indicate whether or not a coffee shop could be successful in this neighbourhood. Due to the form in which the FourSquare API returns data, along with the data gathered previously and how it is stored, this will not be a regression or classification problem. Rather this problem will use clustering of the unsupervised data to understand the relationship between neighbourhood and venue, and venues to each other, when deciding if a neighbourhood is optimal for a coffee shop to open or not.

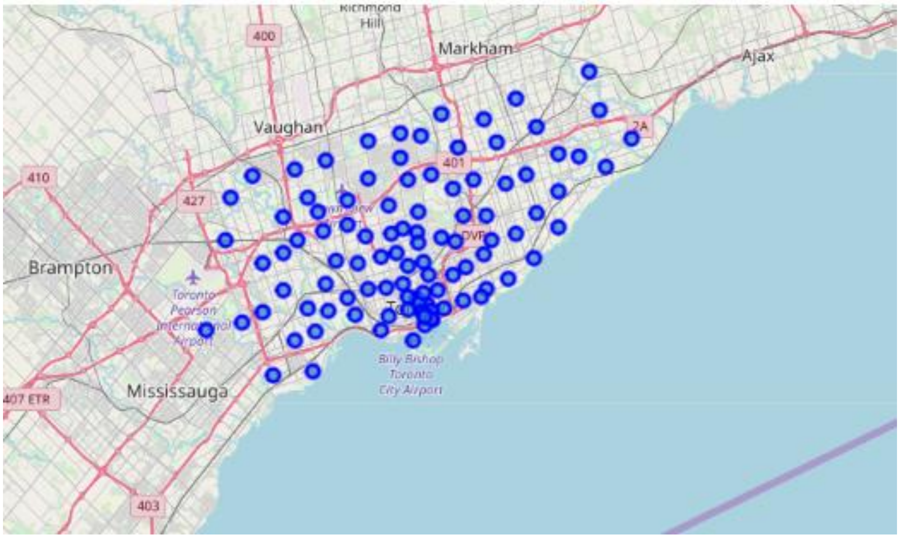
The main way in which the data will be compared is by using a distance matrix between neighbourhoods that already exist within the data to a dummy, ‘perfect’ neighbourhood which will be based on the relationship between the other venues to coffee shop. This distance will then be ranked in order to find the ten closest neighbourhoods to this ‘perfect’ neighbourhood and thus the ten in which a new coffee shop is likely to prove successful.

For the exploratory data analysis of this dataframe, a number of techniques were applied.

First, the dataframe was created by using calls to the Four Square API which extracted the occurrence of each venue within a neighbourhood. This was encoded through one hot encoding and a count of each neighbourhood in order to give a picture of how many of each venue already existed in a neighbourhood.

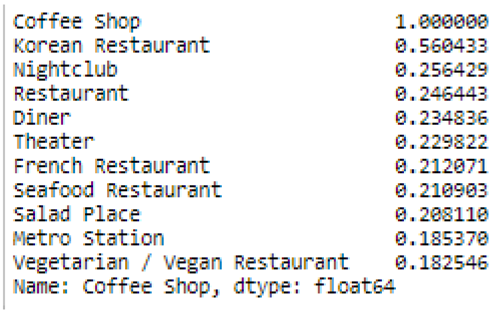
Second Toronto and the location of the neighbourhoods was visualised, with each blue circle indicating the location of the neighbourhood. This gives an overall view of how the city is divided, and also gives a point of comparison towards the end of the project where the results are presented.

Toronto visualisation of neighbourhoods



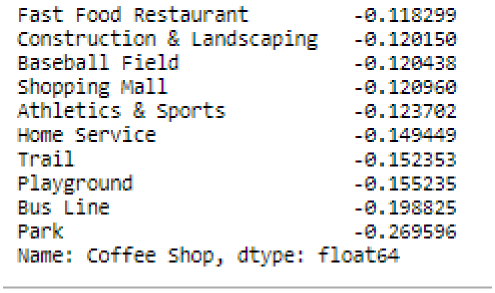
Third, each venue was compared to the coffee shop venue in order to get an overall correlation matrix. This indicated how the presence of other venues will impact upon the neighbourhood and whether or not coffee shops are contained within the neighbourhood already. A strong positive relationship will indicate that the presence of one venue will also indicate the presence of a coffee shop, while a negative relationship shows the opposite. The negative relationship is also a good indicator that there is not an oversaturation of coffee shops in this neighbourhood, as the negative relationship will denote that as the other venue becomes more common, a coffee shop becomes less common and thus less likely to indicate oversaturation in the market. The top ten positive and negative correlations between venues are shown below, with the full results in the accompanying python notebook:

Positive Correlations



```
Coffee Shop                      1.000000
Korean Restaurant                0.560433
Nightclub                        0.256429
Restaurant                       0.246443
Diner                            0.234836
Theater                          0.229822
French Restaurant                0.212071
Seafood Restaurant               0.210903
Salad Place                      0.208110
Metro Station                    0.185370
Vegetarian / Vegan Restaurant    0.182546
Name: Coffee Shop, dtype: float64
```

Negative Correlations



```
Fast Food Restaurant          -0.118299
Construction & Landscaping    -0.120150
Baseball Field                -0.120438
Shopping Mall                 -0.120960
Athletics & Sports            -0.123702
Home Service                  -0.149449
Trail                         -0.152353
Playground                    -0.155235
Bus Line                      -0.198825
Park                          -0.269596
Name: Coffee Shop, dtype: float64
```

After the initial data exploration took place, the KNN technique was needed to be applied. In this case we used KNN due to the simplicity of the algorithm and the fact that the data was unlabelled and can remain that way. The algorithm was applied by creating a dummy neighbourhood from the correlation matrix between coffee shop and other venues with some adjustments made:

1. The coffee shop correlation to itself was reduced down from 1 to 0.2. This decision was made in order to include neighbourhoods which already had coffee shops from their third to tenth most common venue, which were all necessary to be added to ensure that as their markets were not as oversaturated they could be viable locations, and to ensure that there would not be a bias on the relationship a coffee shop has to itself pulling the KNN in the incorrect direction of looking for higher occurrences of coffee shops and thus leaning towards the oversaturated markets.
2. All of the positive correlations were also reduced down if they were over 0.2 as this again gave a bias towards neighbourhoods which already had a high occurrence of coffee shops. They were weighted down to 0 to insure they would not introduce bias towards oversaturation, and to give other venues which had a strong relationship (in the context of this report being approx. >0.2 or <-0.2) and correlation in the negative direction to influence the selection of neighbourhoods.

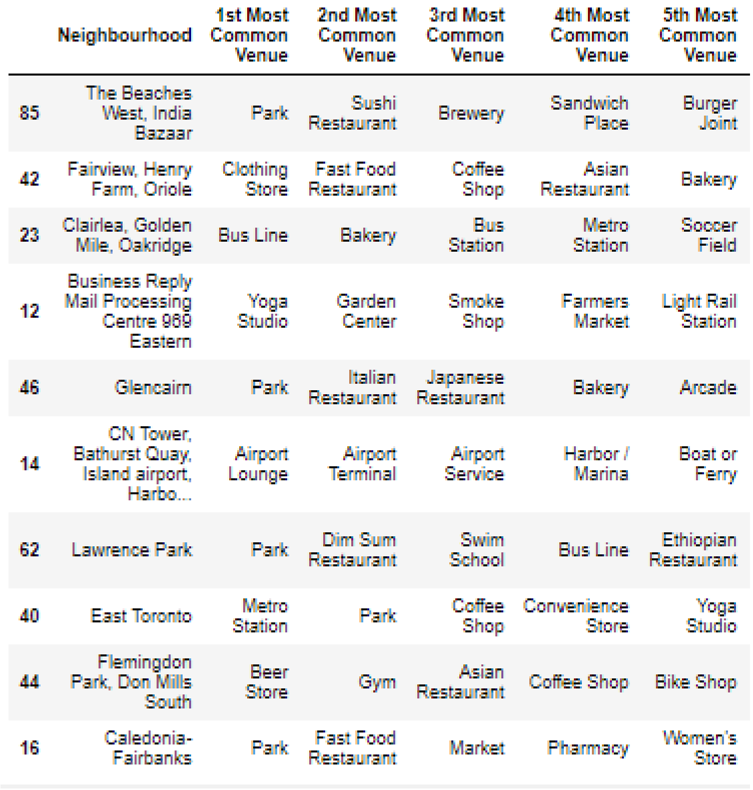
3. Finally, as the dummy neighbourhood was compared to the dataframe which took the mean values of each venue occurring in the neighbourhood, all values were taken as their absolutes to avoid any negative values. They would not have made sense in a distance comparison and would have negated the influence that venues more popular than but related to coffee shops needed to have on the final selection of neighbourhoods.

Once the dummy neighbourhood had been created, all others were compared using Euclidean distance due to the simplicity. The neighbourhood was compared to all actual neighbourhoods within the data set, comparing the dummy to their mean occurrences of venues in order to find a neighbourhood that could meet the criteria laid out in the data section of this report.

IV. Results

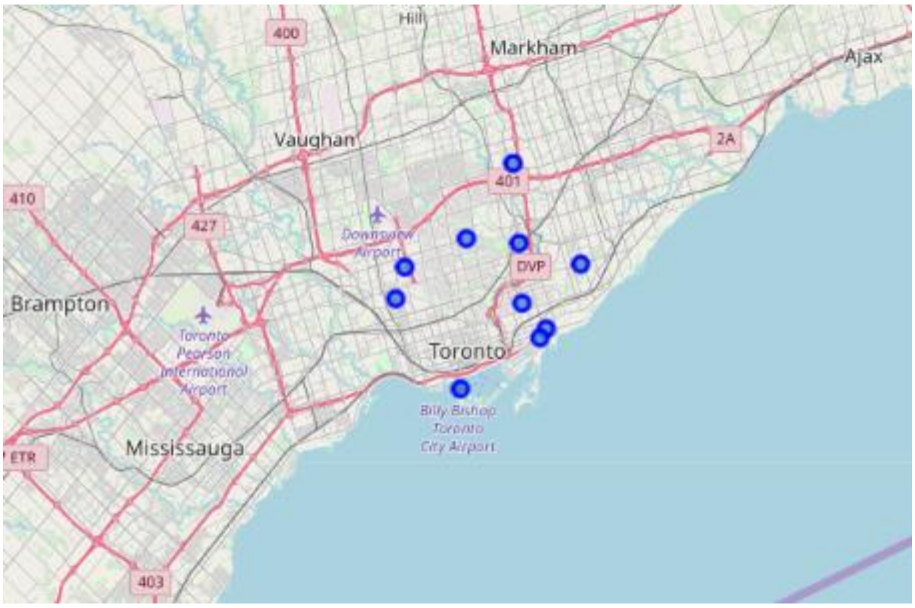
After running the KNN algorithm by getting the distance vectors between the dummy neighbourhood and actual neighbourhoods the neighbourhoods which were added to the oversaturation dataframe earlier in the project were removed as options for creating a coffee shop. From here, the data was then sorted to give the top ten recommended neighbourhoods in which to open a new coffee shop

Top ten neighbourhoods by their common venues



	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
85	The Beaches West, India Bazaar	Park	Sushi Restaurant	Brewery	Sandwich Place	Burger Joint
42	Fairview, Henry Farm, Oriole	Clothing Store	Fast Food Restaurant	Coffee Shop	Asian Restaurant	Bakery
23	Clairlea, Golden Mile, Oakridge	Bus Line	Bakery	Bus Station	Metro Station	Soccer Field
12	Business Reply Mail Processing Centre 969 Eastern	Yoga Studio	Garden Center	Smoke Shop	Farmers Market	Light Rail Station
46	Glencairn	Park	Italian Restaurant	Japanese Restaurant	Bakery	Arcade
14	CN Tower, Bathurst Quay, Island airport, Harbo...	Airport Lounge	Airport Terminal	Airport Service	Harbor / Marina	Boat or Ferry
62	Lawrence Park	Park	Dim Sum Restaurant	Swim School	Bus Line	Ethiopian Restaurant
40	East Toronto	Metro Station	Park	Coffee Shop	Convenience Store	Yoga Studio
44	Flemingdon Park, Don Mills South	Beer Store	Gym	Asian Restaurant	Coffee Shop	Bike Shop
16	Caledonia-Fairbanks	Park	Fast Food Restaurant	Market	Pharmacy	Women's Store

Visualisation of the suggested neighbourhoods in Toronto in which to open a coffee shop



By looking into these neighbourhoods through their top ten common venues, we can see that a number have coffee shop as their 3rd-10th most common venue. This meets the criteria laid out in the project, where there is enough of a presence to suggest that the new coffee shop will be successful but not so many that they would be entering an oversaturated market. In contrast three of the neighbourhoods have park as their most common venue, with coffee shop not appearing in the top ten. This could suggest two things

1. That the algorithm and dummy neighbourhood could be adjusted with more data (more than the potential 68 neighbourhoods found) or
2. From human intuition we know that a venue such as a park is likely to be frequented by people relaxing and enjoying time with friends or at work. In the real world a coffee shop is likely to be nearby in order to provide drinks to these park customers, and thus though it is not in the top ten common venues a coffee shop is likely to be successful due to the high foot traffic and habits of many people to purchase coffee or drinks and take them into a park to enjoy. This would need more statistical proof than human intuition, but it suggests that the results from the comparisons to the dummy neighbourhoods should not be dismissed out of hand

V. Discussion

There are a number of changes which could be made to strengthen the result of this project. More data is the most common one across the board, with a few ideas including

- Getting data on the statistics of coffee shops, what makes them profitable, the percentage of new businesses that open and fail or fold over years
- Getting data on the statistics of other neighbourhoods in the wider Toronto area
- And gaining data on similar cities to Toronto to see the relation between neighbourhoods, coffee shops and other venues to see their correlation.

The dummy neighbourhood used to calculate the distance and thus likeliness of a coffee shop succeeding in other neighbourhoods could also be further developed. In this project it was based on the correlation of the venues but correlation does not always imply causation. With further time and data spent into the project the results could be refined and also provide better indicators for the venue correlation which could allow the methodology to applied accurately to other neighbourhoods.

Based on any data which may be gained about the function of coffee shops, it might even be an option to classify the neighbourhoods as likely/unlikely to succeed and through that gain a model which could be applied to other cities or adjusted to other venues. This however would take a great deal more data and would also require potentially sensitive and ethically ambiguous data that relates to more specifics of neighbourhoods, such as the general income and other census based date of the populations.

VI. Conclusion

Analysis of the neighbourhoods of Toronto and their associated venues managed to provide a number of options for where a coffee shop could be opened with a relative chance of success, avoiding the pitfalls of the oversaturated market and identifying some venues which had a strong correlation with coffee shops. While the project did not provide an easily applicable model which takes the advantage of a predefined algorithm from a source such as scikit learn the steps which were taken to make a KNN-based distance measure would be applicable to other neighbourhoods if the data were appropriately shaped and analysed.

This project has shown that analysis of neighbourhoods with FourSquare API is both doable and can provide information which can be used and applied in a real world sense. There are many ways in which this project could be extended, most through the use of extra data or more forms of comparison, but the final neighbourhoods chosen by the KNN based method are a good starting point for stockholders interested in opening a new coffee shop.

Thank you for reading this report, and looking through the associated code which can be found through a link here:

https://github.com/arosvall/Coursera_Capstone/blob/master/Complete%20Capston%20Project%20Code.ipynb