The Battle of Neighborhoods

IBM Data Science by Coursera

Applied Data Science Capstone
Peer Graded Project

Stefan Schäfer, Dec. 16th, 2019

Introduction / Problem

I want to pick up suggestion 2. of our Coursera tutor Alex Aklson. Let's suppose I am a business consultant and I have a new client who wants to open an Italian restaurant in Toronto. His name is Roberto and he is an Italian chef who worked for several well-known restaurants and hotels everywhere in the world. Not long ago he had an engagement in Toronto, Canada. During this time, he realized that most of the local Italian restaurants there only offer Pizza and Pasta. Born in Rome, Roberto knew that no one in Italy would call a Pizzeria a 'Ristorante'. If you step into a Restaurant in Italy and ask for a Pizza, he told me, you might really earn some unfriendly words. Genuine Italian food is much more than Pizza and has a very broad range of ingredients and flavors, differing from region to region. So, Roberto had the idea to open his own original style Italian restaurant and serve all the fine meals of the regions between Milan and Sicily with monthly highlights of a special meal of one region like Ossobuco Milanese with Saffron Rice etc.

When planning his new engagement in Toronto, Roberto came to a point where he had to decide where to locate his restaurant in Toronto. As we have a long-term friendship, he called me and asked me if I could help him finding the best place. Of course, I assured him of my help and started this notebook to collect, process and analyze specific data and information about Toronto. In particular Foursqare might help me, I thought, because it could provide specific venues data for the boroughs and neighborhoods of Toronto.

To find the best place and make a recommendation we must take a closeer look at the available data and decide what could be the best strategy, e.g.:

- are there already other Italian restaurants
- could this be a chance because the people who go there seem to like Italian food
- or could this be a problem due to the competitive situation
- would it be better to choose a location with e.g. more Asian restaurants to give the people an alternative
- or might a higher no. of Asian restaurants mean that the people there prefer Asian food
- would it be a good idea to look for a location with only few restaurants
- or might there be reasons for the low no. of restaurants there
- we can basically use the rating information of Foursqare to get a better judgement of the situation
- can we suppose a floating population due to some specific venues nearby
- etc. The data will give us a clearer view.

Data

I will scrape the Canadian postcode site of Wikipedia using the Beautiful Soup library and get a dataframe with postcodes, boroughs and neighborhoods like this:

Postcode	Borough	Neighborhood
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront
M6A	North York	Lawrence Heights
M6A	North York	Lawrence Manor

...

Then I will apply latitude and longitude using the file provided by 'cognitive class'. The result will look like this:

M6P	West Toronto High Park	43.6616	-79.4648
M6R	West Toronto Parkdale	43.649	-79.4563
M6S	West Toronto Runnymede	43.6516	-79.4844

...

Now I use Foursquare to get venue data for all neighborhoods. The result will look like this:

Neighborhood, Neighborhood Latitude, Neighborhood Longitude, Venue, Venue Latitude, Venue Longitude, Venue Category
The Beaches, 43.676357, -79.293031, Glen Manor Ravine, 43.676821, -79.293942, Trail
The Beaches, 43.676357, -79.293031, The Big Carrot Natural Food Market, 43.678879, -79.297734, Health Food Store
The Beaches, 43.676357, -79.293031, Grover Pub and Grub, 43.679181, -79.297215, Pub
The Beaches, 43.676357, -79.293031, Glen Stewart Ravine, 43.676300, -79.294784, Other Great Outdoors
The Beaches, 43.676357, -79.293031, Domino's Pizza, 43.679058, -79.297382, Pizza Place
...

This is the starting point for further data analysis. We can sort, filter and aggregate the data in different ways to gain a deeper understanding of the situation. We can also visualize the data with maps for a better overview. Doing all this we can collect information to determine the best place for the new Italian restaurant.

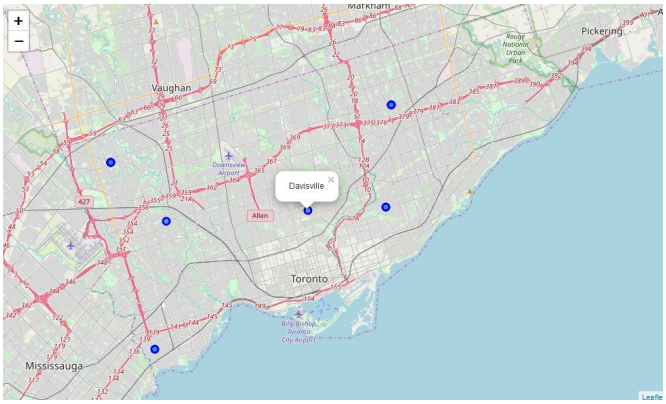
My results and observations will then lead to a recommendation for my client Roberto.

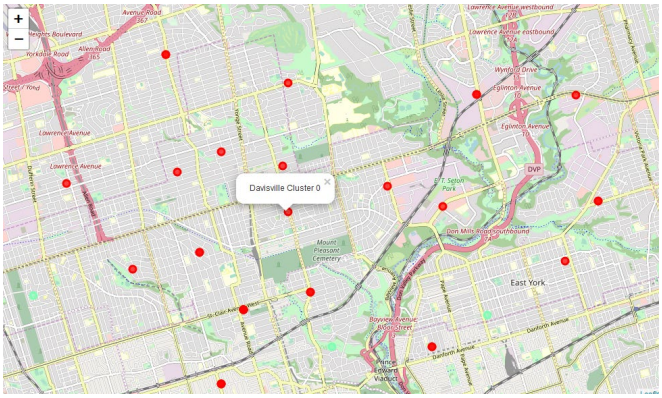
Methodology

- First, I will collect the data necessary to set up the model.
- I'll use Wikipedia to get the borough and neighborhood information
- I'll clear up the data and show some information
- Then I'll load the corresponding geo coordinates from the COCL CSV file
- Now I'll show a first map with the neighborhoods
- The next step is to load corresponding data (venues) from Foursquare
- Now in combination with geo data and venues I'll look for Italian food in Toronto
- When found some good fitting locations I'll cross check them by clustering the neighborhoods using their venues
- Now I'm able to select a good fitting location for the new Italian Restaurant

Results

The result is that I would like to recommend **Davisville** as location for the new Italian Restaurant. Davisville has an „urban touch“, is no pure residential area, there are some other restaurants (pizza as no. 1 venue) and it is relatively close to Toronto downtown, the highway and train stations.





- Toronto has 210 Neighborhoods in 11 Boroughs
- There are 272 unique venue categories based on Foursquare data
- 50 neighborhoods have a Pizza Place in the top 10 venues
- 27 neighborhoods have an Italian Restaurant in the top 10 venues
- At 17 neighborhoods a Pizza Place is the no. 1 venue
- 3 of these 17 neighborhoods have also an Italian Restaurant in the top 10 venues
- No neighborhood has an Italian Restaurant as venue no. 1

Observations and Recommendations

I observed that there are several good fitting locations and without deeper analytics it will be hard to make good recommendation. I used Google and Google maps to check the locations I found by data analytics and concluded that, other than first supposed, I will not recommend one of the locations with top ranked venues like banks or shopping malls. These places seemed to be a little bit solitary in between residential areas or along streets. This shows that we can't only rely on figures or ranks but must look outside the box.

Conclusion

Our opportunity is to chose a neighborhood where Italian Food is already established (Pizza and Restaurant) and where we have some other restaurants, which is a signal for people going out there for lunch or for dinner. As there is no neighborhood with an Italian Restaurant as venue no.1, our goal will be to conquer this position.

Data Analytics gives a good foundation for decisions. The more data we get the better we can decide. Doing evaluation and statistics is quite simple with Python, all the libraries and Jupyter Notebook. But when it comes to Machine Learning (or statistical learning) we see that the problem is to choose the best method or model. I used KMEANS and that gave me a result where most of the neighborhoods were in 1 cluster due to the similarity of the venues. Nevertheless, KMEANS showed that there are Neighborhoods where the top ranked venues are parks, baseball fields or print shops. So even if clustering did not really help to choose the best location it helped to discard some other locations.