Mavi N. Polatoglu

Project: “Feel right at home”

Neighborhood Similarity According to Foursquare

Contents

[Introduction 2](#_Toc44166022)

[Problem Definition 2](#_Toc44166023)

[Audience / Customer Base 2](#_Toc44166024)

[Data 2](#_Toc44166025)

[Methodology 4](#_Toc44166026)

[Results 6](#_Toc44166027)

[Discussion 7](#_Toc44166028)

[Conclusion 7](#_Toc44166029)

# Introduction

The last course of the Professional Certificate in Data Science by IBM on Coursera is titled “Applied Data Science Capstone”. As its name suggests, this part consists mostly of hands-on project work with goals of practicing the usage of tools learned throughout the previous lectures and exercises. Some examples include, to (i) extract data from tables/links/records online, (ii) import the data as a dataframe into Jupyter Notebooks, (iii) manipulate the data (cleaning, processing, if necessary, modelling, evaluating, and so on), (iv) use various visualization techniques. To demonstrate all these gained abilities and hard skills, students were required to define a problem that could be solved by using Foursquare location data. In the below, you will find the definition of the problem that I chose to work on along with other required report sections.

## Problem Definition

From inspecting data that can be obtained from Foursquare endpoint calls, I saw that comparing different neighborhoods could be a suitable task for the project. As a person who is happy with the location I live in, I wanted to develop a small piece of code to find a similar neighborhood in another city **in case I needed to move to another place.** The two cities that were taken as examples are Toronto(destination) and NYC(origin). However, these two cities could have been any other pair that has accurate Foursquare data. So, the problem can be posed with a simple question as: **“Which neighborhood in Toronto is the most similar to the one I am living in right now (NYC, Fordham)?”.**

## Audience / Customer Base

From personal experience, I can say that **people who are pursuing their graduate studies or young professionals are most likely to move**. So, the target audience would probably be people of all genders ages 20-32.

# Data

To be able to compare neighborhoods, the names were necessary for two cities.

The data are:

* Neighborhood names of Toronto from <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>
* Neighborhood names of NYC from <https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City>
* The locations (latitude and longitude) of each neighborhood was obtained by using the neighborhood names and the city name (e.g., “Parkwoods, Toronto”) as a geocode (as in past assignments and <https://towardsdatascience.com/geocode-with-python-161ec1e62b89>).



Figure 1: Some category and main-category examples

One other piece of information that was not related to city or location were **main category** assignment. Foursquare has a hierarchy of categories belonging to each venue. For example, if a venue category is “Chinese Restaurant”, the primary, or main category is “Food”. Another example is “Bus Line” or “Bus Station” that belongs to “Travel & Transport”. The comparison between neighborhoods were done based on the main categories. Therefore, there is a data extraction step where the hierarchy of categories are acquired using the ***categories* endpoint (**<https://developer.foursquare.com/docs/build-with-foursquare/categories/>) and a processing step in the code that matches each venue to its main category. Fig1 shows other examples.

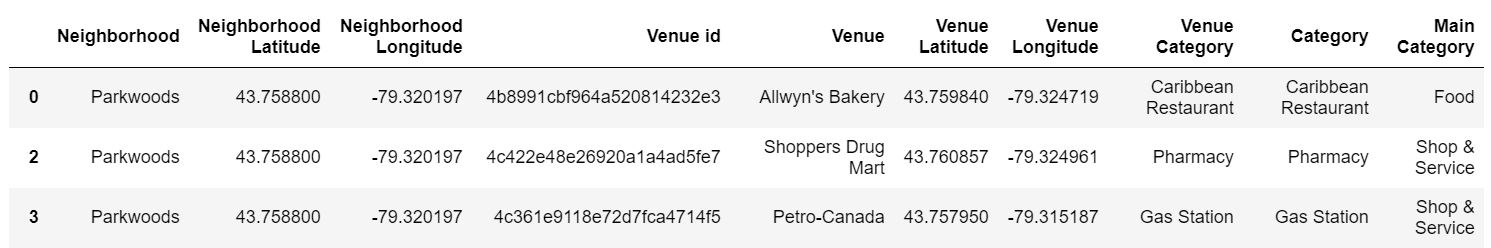


Figure 2: First three rows from a dataframe showing the collected data. This information exists for all neighborhoods of Toronto(city) and NYC, here seen is for 3 venues in Parkwoods, Toronto. Apart from the neighborhood name, each venue’s exact location and main category are listed.

Fig2 above shows in an orderly fashion, how the neighborhood names, location, venues, and categories are seen after acquisition and main category assignment.

# Methodology

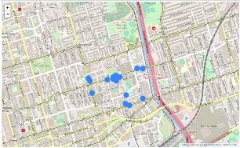
The aim of the project was to be able to compare one neighborhood (the current neighborhood of an individual) to multiple other neighborhoods (potential neighborhoods to move to). So, what needed developing was a type of numerical guide to “how similar” each pair is.



Fordham, NYC



Central Bay Street, Toronto



Regent Park, Toronto



Rogue Hill, Toronto

Other neighborhoods in Toronto

Error{Fordham, Central Bay Street}

Error{Fordham, Regent Park}

Error{Fordham, Others}

Error{Fordham, Rogue Hill}

**Increasing Error**

**Less similarity**

Figure : Schematic explaining pairwise comparison. Each pair has an error value that calculates the dissimilarity between 2 neighborhoods depending on the venues extracted using Foursquare. Blue dots in small icons are the venues Foursquare returned.

Pairwise-error is based on a distance measure that is calculated with the below formula. It is essentially a root-mean-squared error in multidimensions where each dimension is the venues’ main categories.

Where,

N is the total number of main venue categories in Fordham

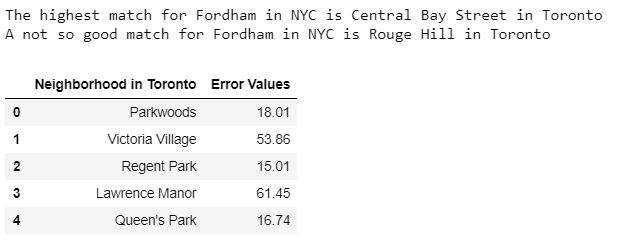
is the number of venues in a that are in the category n

In English, the formula translates into the following:

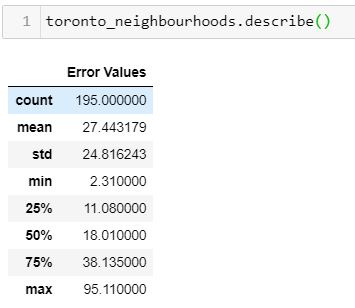
* For the categories that exist in both neighborhoods, the error contribution is the Euclidean distance between two points in the coordinate system where each axis is the number of venues belonging to separate main categories.
* For categories that do exist only in Fordham, the power of 3 applies a harsher penalty on the neighborhood to make it dissimilar from Fordham.
* For categories that exist only in neighborhood X, there is no contribution.

Finally, the individual contributions are summed and divided by the number of unique categories in Fordham which is N. **The higher the Error{Fordham, Neighborhood X}, the less similar Fordham and neighborhood are.** Therefore, the neighborhood which scored least is proposed as the neighborhood to move to.

# Results



(a)



(b)

Figure : (a) Printed output of the best and worst candidate neighborhood as well as head of the erro dataframe. (b) The errors described using the “describe” method.

There are as many error values as many candidate neighborhoods in Toronto. In Fig4a there are 5 neighborhoods and their respective error values listed. The printed output is for the user to see which neighborhoods are the best and worst match. In Fig4b, the minimum and maximum error values can be seen as well as other metrics which explains the distribution.

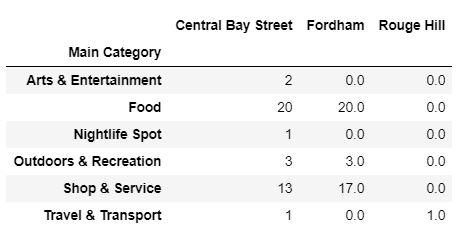


Figure : Central Bay Street, Fordham, and Rouge Hill compared according to venues.

The best candidate for someone who lives in Fordham, NYC is Central Bay Street and the worst is Rouge Hill. As seen in Fig5, the error metric successfully reflects the differences in the venue types.

# Discussion

For the mathematical expression of the error and the results, it can be seen that the algorithm not only tries to match category types, but also the number of venues. In Fig5 it can be seen that Fordham has 20 Food locations and 13 Shop&Service. If a neighborhood A , say, had 40 Food and 25 Shop&Service and another neighborhood B had 30 Food and 15 Shop&Service, although both A and B fulfilled the requirements of the main categories, neighborhood B would be the better choice since the “density of venues” would be more similar. However, altogether the results seem plausible.

Given that the algorithm seems to be giving reasonable results, it is worth mentioning that the explore endpoint was given radius of 500 meters and limit of 50. There are two assumptions that need to be explicitly stated in accordance.

* First pertains to the radius of 0.5km: the venues are assumed to be distributed mostly in the circular area with that radius where the latitude and longitude is the center given by geocode. The neighborhoods on the other hand do not always have circular borders. However, considering people do not really care about crossing (virtual) neighborhood borders as long as the venue is close, this did not pose a problem.
* Second is the limit of 50 venues which means that if the chosen cities/neighborhoods are very very dense, this limit would be achieved for many neighborhoods and the number of venues (in addition to the distribution of categories) will not be a determining factor anymore.

Some simple suggestions for the future are increasing the limit to the number of venues and making a neighborhood dependent radius. A bit more complicated suggestions would be to incorporate criminal data from police departments to assess neighborhoods also for safety and to find average house rent prices for each neighborhood in order to find a much more realistic match. These dimensions of the problem is unfortunately missing.

One pother idea is that the algorithm can be customized depending on the person. One example would be that if the user, say, needs to be close to a dialysis center, distance to the hospital can be penalized heavily in the mathematical error expression.

# Conclusion

In the light of the above, this project can give initial insights into which neighborhoods may look/feel alike since types and density of venues already give a lot of information into what type of a neighborhood is. For instance, if there are more Arts&Entertainment venues in a neighborhood it is more likely to be in the heart of the city center. Or if a neighborhood does not have many venues, it could be an indication of a residential area. **These effects are successfully observed in the algorithm developed.** This project can be used also to narrow down candidate neighborhoods. The error values are not categorical variables, so, **it would be plausible for someone to also be interested in the neighborhoods that have the second or third lowest error.**

All in all, the code written during this project was a very good practice into what was learned during the certificate program.