Mavi N. Polatoglu

Project: “Feel right at home”

Neighborhood Similarity According to Foursquare

Contents

[Introduction 2](#_Toc44099148)

[Problem Definition 2](#_Toc44099149)

[Audience / Customer Base 2](#_Toc44099150)

[Data 2](#_Toc44099151)

[Methodology 3](#_Toc44099152)

[Results 3](#_Toc44099153)

[Conclusion 3](#_Toc44099154)

# 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>).

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 : Some category and main-category examples

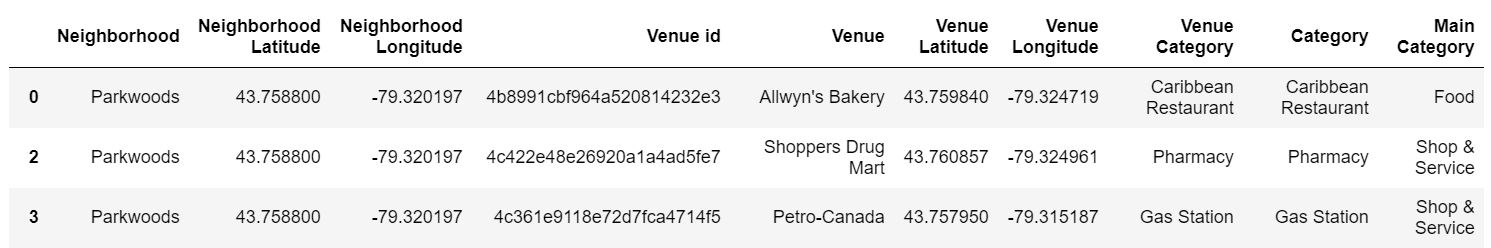


Figure : First three rows from a dataframe showing the collected data. This information exists for all neighborhoods of Toronto(city) and NYC, here seen is for 3 venues in Parkwoods, Toronto. Apart from the neighborhood name, each venue’s exact location and main category are listed.

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

# Methodology

# Results

# Conclusion