**Relocation Application**

**(Proof of Concept for Finding Similar Neighborhoods in Different Cities)**

**IBM Data Science Professional Certification: Applied Data Science Capstone Exercise**

**March 14, 2020**

**Introduction**

Have you ever been faced with a move to a new and unfamiliar city? How did you choose where you would focus your search for housing? This project conceptually allows a user to find the best match between neighborhoods comparing the current neighborhood/city and its available amenities to multiple neighborhoods in the gaining city by finding the closest matching neighborhood profile.

Use of this application would be targeted to individuals, families, and/or businesses who like their current neighborhood and desire to find a similar neighborhood in a new city when faced with a move. In a more robust version, it could be adopted by and/or packaged for a relocation assistance business.

**Data**

Neighborhood data was harvested from online data sets to identify neighborhoods by name and geographic (latitude/longitude) coordinates. Example:

|  |  |  |
| --- | --- | --- |
| **Neighborhood** | **Latitude** | **Longitude** |
| The Beaches | 43.676357 | -79.293031 |
| Studio District | 43.659526 | -79.340923 |
| Lawrence Park | 43.728020 | -79.388790 |
| Davisville North | 43.712751 | -79.390197 |

Popular venue data was harvested from Foursquare and normalized to determine the “profiles” of each neighborhood. Example:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Neighborhood** | **Art Gallery** | **BBQ Joint** | **Bagel Shop** | **…** | **Thai Restaurant** | **Vegetarian / Vegan Restaurant** |
| Berczy Park | 0.017857 | 0.017857 | 0.017857 | … | 0.017857 | 0.017857 |

Sources:

Foursquare venue data <https://foursquare.com/>

New York City neighborhood data <https://cocl.us/new_york_dataset>

Toronto neighborhood data <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

**Methodology**

For this proof of concept, the current city and gaining city were pre-selected (and loaded) from separate datasets of neighborhoods in Toronto, Canada, and the Manhattan (borough) of New York City, New York. While the current neighborhood in Toronto was coded as the “Studio District”, there is opportunity within the code to easily change this to another Toronto neighborhood for a different comparison.

The geographic coordinates of the current neighborhood were passed to Foursquare to identify popular venues within a 500 meter radius of that neighborhood, and normalized to create a neighborhood venue “profile” (1 row by 32 columns):

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Neighborhood** | **Yoga Studio** | **American Restaurant** | **Bakery** | **…** | **Thrift / Vintage Store** | **Wine Bar** |
| 0 | Studio District | 0.02439 | 0.04878 | 0.04878 | … | 0.02439 | 0.02439 |

The geographic coordinates of all of the gaining city neighborhoods were then similarly passed through Foursquare to also create normalized neighborhood venue profiles for each neighborhood in the target (gaining) city (40 rows by 285 columns):

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Neighborhood** | **Accessories Store** | **Adult Boutique** | **Afghan Restaurant** | **…** | **Women's Store** | **Yoga Studio** |
| 0 | Battery Park City | 0 | 0 | 0 | … | 0.02 | 0 |
| 1 | Carnegie Hill | 0 | 0 | 0 | … | 0 | 0.04 |
| … | … | … | … | … | … | … | … |
| 39 | Yorkville | 0 | 0 | 0 | … | 0 | 0 |

The intent of the project was to match neighborhood venue profiles to the current neighborhood’s profile. The current neighborhood had 32 columns (venue types) while the

the number of venue types (columns) in the gaining city were far greater in number/diversity. For a match, only the same type of venues as in the current neighborhood are relevant so a left join was performed to combine the gaining city neighborhoods with the current neighborhood venue profile on venue type. This data manipulation process utilized transposing dataframes and setting/re-setting indices, along with removal of all “NaN” points representing missing data in the gaining city venues (e.g. “Breweries” was not a venue category in the gaining city, resulting in “Nan” in the joined dataframe).

Once the data had been cleaned and formatted, the original intent was to use K-Nearest-Neighbor (KNN) to discern the closest match to the current neighborhood venue profile. However, it was determined that an easier approach to this problem only required a subset of KNN to determine closest “match” – specifically Euclidean Distance (using the venue profiles of the current and gaining neighborhoods).

|  |  |
| --- | --- |
| **Neighborhood** | **Match\_Distance** |
| Hamilton Heights | 0.129366 |
| Lower East Side | 0.134505 |
| Washington Heights | 0.135156 |
| … | … |
| Roosevelt Island | 0.196466 |
| Soho | 0.214257 |
| Stuyvesant Town | 0.215462 |

The smaller the “Match\_Distance”, the more similar the gaining city neighborhood venue profile is to the current neighborhood profile.

**Results**

From this approach, it was determined that Hamilton Heights, in Manhattan, was the nearest match to the Studio District in Toronto based on Foursquare venue profiles. Note: While Hamilton Heights was determined to be the “closest” match, the data shows that it is not a “perfect match” as there are venues in the Studio District that are not found in Hamilton Heights.

|  |  |  |
| --- | --- | --- |
| **Neighborhood** | **Studio District** | **Hamilton Heights** |
| Yoga Studio | 0.02381 | 0.04 |
| American Restaurant | 0.047619 | 0 |
| Bakery | 0.047619 | 0.04 |
| Bank | 0.02381 | 0 |
| Bar | 0.02381 | 0.02 |
| Bookstore | 0.02381 | 0 |
| Brewery | 0.047619 | 0 |
| Café | 0.095238 | 0.08 |
| Cheese Shop | 0.02381 | 0 |
| Clothing Store | 0.02381 | 0 |
| Coffee Shop | 0.071429 | 0.08 |
| Comfort Food Restaurant | 0.02381 | 0 |
| Convenience Store | 0.02381 | 0 |
| Coworking Space | 0.02381 | 0 |
| Department Store | 0.02381 | 0 |
| Diner | 0.02381 | 0 |
| Fish Market | 0.02381 | 0 |
| Gastropub | 0.02381 | 0.02 |
| Gay Bar | 0.02381 | 0 |
| Gym / Fitness Center | 0.02381 | 0 |
| Ice Cream Shop | 0.02381 | 0 |
| Italian Restaurant | 0.047619 | 0.02 |
| Latin American Restaurant | 0.02381 | 0.02 |
| Middle Eastern Restaurant | 0.02381 | 0 |
| Park | 0.02381 | 0.04 |
| Pet Store | 0.02381 | 0 |
| Sandwich Place | 0.02381 | 0.02 |
| Seafood Restaurant | 0.02381 | 0.02 |
| Stationery Store | 0.02381 | 0 |
| Thai Restaurant | 0.02381 | 0 |
| Thrift / Vintage Store | 0.02381 | 0 |
| Wine Bar | 0.02381 | 0.02 |
| Match\_Distance | 0 | 0.128363 |

**Discussion (observations/recommendations)**

This was an interesting project using the Foursquare API for a novel purpose. As a proof of concept, the approach and scope were definitely limited, if not rudimentary.

The key assumption to this proof of concept was that users would want to closely mirror the venues of their current neighborhood with that of a gaining locale. While this may be true for some, it would not be appropriate as a universal law. Many enjoy the adventure of a move, and that includes exploring a new city and finding new places. Others may, indeed, like the amenities of their current neighborhood and desire similarity, albeit at a lessor priority to other aspects of a neighborhood (for example, friendly neighbors, clean streets, and newly built homes).

From an approach perspective, there are likely many other/better ways to find a nearest matching neighborhood. Originally, my intent was focused on using KNN though I determined that using the KNN component of Euclidean distance would be more suitable/sufficient.

Basic functionality of this proof of concept could be enhanced by acquiring and/or maintaining a library of worldwide neighborhood datasets and allowing for users to select any current neighborhood in any city as well as any gaining city as inputs. There are multiple sites for aggregating this information available today, though preliminary investigation reveals a dearth of comprehensive data sets (at least with free access).

A far more robust tool (perhaps one that could even be used by relocation services) would include many more factors when considering which new neighborhood to relocate to. A sample of those other considerations includes but is not limited to:

* Cost of Living Indices
* Housing (Type, Availability, and Affordability) – perhaps linking to a local real estate URL
* Transportation (Statistics, Types/Public Options, Availability)
* Demographic Information
* Other Economic Information (Taxes, Median Income)
* Etc.

From a process perspective, certain aspects of the coding took seemingly inordinate amounts of time to run using the free version of Watson Studio – specifically, the calls to install different libraries (I suspect the greatest culprit to this was folium) required approximately 25-30 minutes each time I had to re-run it.

Lastly, the coding involved in this project certainly has opportunity for efficiency/improvements which will come with additional exposure and experience with Python as a tool. I found the process educational and feel at least somewhat familiar with the basics of Python and data science as a result.

**Conclusion**

If you are moving from the Studio District in Toronto, Canada, to Manhattan in New York City, New York, your best bet to feel “at home” is to live in Hamilton Heights. While this conclusion is consequent to one of many possible ways to choose a new place to live, nothing beats actually visiting several target locations, though this proof of concept can give people a great way to create a short list of places to consider.