Finding the Right Neighborhood: Toronto and the Twin Cities

Alexander Stetzer

June 3rd, 2021

Introduction

The Twin Cities of Minnesota are a culturally diverse region of the Upper Midwest. Made up of the cities of Minneapolis and St. Paul, the Twin Cities are closely linked and seem almost like the same city, hence the name Twin Cities. Toronto, being a key city in Canada, also has this diverse culture. With many similar characteristics between these cities and the economic growth of these communities, movement between the two is possible. In order to find similar neighborhoods between the two, the goal is to use venue data to find similarities between neighborhoods. If a person needs to move from Toronto to the Twin cities or vice versa they might want to live in a place with similar venues. Using KMeans clustering and cosine similarity, neighborhoods with similar venues can be found.

The interested parties for the project are the people moving in between the cities and looking for new residences. A neighborhood with similar venues may allow for an easier time getting familiar with the surrounding area. Also the project could be used in reverse, instead of finding a similar neighborhood, one could use it to look for neighborhoods to avoid.

Data

The data used for the analysis primarily comes from Wikipedia. Wikipedia was chosen for the source of the data as it is easy to web scrape and gather cleaner data from. The main goal for gathering the data was to use postal codes, or zip codes in the case of the United States, to find the names for the neighborhoods. Then, using the postal codes, location data can be found. A problem started to arise because in the United States zip codes often overlap multiple cities or a single city has around five zip codes to itself. The scarcity of the zip code locations would mean that multiple neighborhoods would not be able to be separated and precision of the data would be too wide. To remedy this, neighborhood lists from wikipedia would be used and then the location data extracted.

The data for Minneapolis neighborhoods, gathered from here, was quite simple to webscrape as it is in easy html code structure. Using BeautifulSoup, the communities were under h3 tags and the neighborhoods were under table tags. The communities were placed into one list and the neighborhoods into another. After the lists are constructed, the lists were combined with the city of Minneapolis placed in a city column.

For St. Paul, the data from the wikipedia page was quite messy (St. Paul Wiki). Districts, as they are called in St. Paul, were listed as the neighborhoods contained inside of them. Some of the districts were listed as the correct name, but in the paragraph explaining the district there was no mention of the neighborhoods contained inside. Due to the odd makeup of the data and

the small size of neighborhoods, it was decided that a brute force method was best. Using the wiki page as a guideline, all seventeen districts were put into a dictionary with districts as keys and neighborhoods as values. In the case that a neighborhood did not appear in the district description, a google maps search was used to find the correct neighborhood names. Once all the necessary data was collected, the lists were combined, St. Paul was placed in a city column, and a dataframe constructed.

Since Canada uses a more accurate postal code system, Toronto data was an easier endeavor. Using the wikipedia page here, the postal codes were stripped from the website using BeautifulSoup. The data was easy to gather as it was all under tables and minimal cleaning was needed to get a complete dataframe with postal codes, communities, and neighborhoods. After the dataframe was constructed and a city column with Toronto placed in every row. After all the neighborhood dataframes were constructed they were concatenated into a single dataframe to complete the neighborhood set.

Once all of the neighborhoods were found, location data for each neighborhood was acquired. Geocoder was initially used to find the data, but was found extremely unreliable and took forever to get a proper response for any neighborhood. To alleviate the location problem Google's Geocoder API was used directly to find the location data for the Twin Cities neighborhoods. Looping through the neighborhood dataframe all of the latitude and longitude data was easily found. To find the location data for the neighborhoods of Toronto, a spreadsheet of postal code geospatial coordinates was used. This was done as it would best line up the way the neighborhood data was collected. The postal code was read into the system and was merged with the Toronto Neighborhood data only keeping the postal codes that were included in the neighborhood data.

With all location data collected, the next step was to find venue data. Using the Foursquare API, a master list of every venue from all of the neighborhoods was created. For ease of machine learning, one hot encoding is used to separate the venue categories into multiple columns. Since the main goal of the project is to find similarities in neighborhoods, the venues were grouped under each neighborhood that they belong to. Along with grouping the data, the mean of each neighborhood venue categories is used. This creates the dataset that is used for the final KMeans Clustering operation.