Finding best neighborhoods

Manasvi Dobariya

1. Introduction

1.1 Background

Canada is a very big country and Toronto is a business capital of the country. Many people from the country or from abroad move here in search of job or start their own business. It can be very problematic for a new person to find best neighborhood that satisfy all his/her daily needs without any help from the person living in the city.

1.2 Problem

The idea is to find neighborhood in Toronto city of Canada that has all the basic necessity shops within kilometers of the living place. People who are new to the city or shifting from another city to Toronto may require a place to live in. It might be difficult for them to find the neighborhood with all their necessities. The aim of the project is to divide the city neighborhoods in different categories according to shops and facilities available in the neighborhoods. The Foursquare API will be used to find all the nearby venues in neighborhoods and retrieve categories and count of shops in each category for each neighborhood.

1.3 Interest

For this project, basic necessities are considered as required factors to divide neighborhoods in the categories. It would be very helpful for people new in the city as they will be able to identify neighborhood according to their needs.

2. Data acquisition and Cleaning

2.1 Data Sources

Data of boroughs and neighborhoods of the Toronto City would be retrieved from Wikipedia (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Toronto). The data is there in form of tables with postal codes and names of neighborhoods in each of the Borough. The Geospatial data would be used to retrieve Longitude and Latitude of each neighborhood. Then, Foursquare API would be used to retrieved nearby venues of each neighborhood.

2.2 Data Cleaning

The Wikipedia data was retrieved by web scraping the page with BeautifulSoup library of python. The venues data for all neighborhoods retrieved using Foursquare API was converted into count of venues in each category for all the neighborhoods. The mean values of category values were calculated to construct the final data. The formed dataset contained zero mean value rows for selected features as these rows were not adding much information they were removed from the dataset.

2.3 Feature Selection

- Wikipedia Data: Columns Retrieved: Borough, Postal Code, Neighborhoods
- Foursquare Data: Latitude, Longitude, Venues, Category

Example:

Consider North York Borough of Toronto City. It has many neighborhoods in it which it has multiple postal codes for multiple Neighborhoods. For ex, M3A postal codes belongs to Parkwood neighborhood and M4A belongs to Victoria Village. Foursquare will provide longitude and latitude of Parkwood which is -79.329656 and 43.753259 respectively. Foursquare will also provide venues near Parkwood like Cafes, Parks etc.

Here, there are multiple venue categories as columns in the data. The categories selected for this project were basic necessity categories for ex., Gym, Grocery Store, Bank etc. These features can be selected according to user's need of facilities. The selected features would determine how clusters would be made of neighborhoods.

3. Methodology

The first step was to retrieve Neighborhood's data of Toronto city. The data of all the boroughs and their corresponding neighborhoods with postal codes were available on Wikipedia page. This data was retrieved from the page using web scraping by Beautifulsoup library of python. The data has 3 columns, Borough, Postal code and neighborhoods.

The aim of the project was to get best neighborhood according to places near it. For that, the longitude and latitude of each neighborhood were required. The GeoSpatial data contains Postal codes wise Longitude and Latitude data for all the 103 Boroughs of the Toronto city. The two datasets were combined and the new dataset with boroughs, neighborhoods and latitude and longitude was prepared.

The next step was to retrieve nearby places of each neighborhood. The Foursquare API was used for this purpose. The explore request was used to get nearby venues. Limit of 100 was set for each neighborhood nearby venues. The Foursquare API returned a JSON response of the explore query for all the neighborhoods. The information needed from the JSON response was name, longitude, latitude and category of each venue retrieved. The new data frame containing Neighborhood name, longitude, latitude, Venue name, Venue Category, Venue latitude and Venue longitude. As a cleaning step neighborhood with less than 5 nearby venues were removed from the dataset. The reason behind the step was to provide neighborhoods that has possibility of covering all the facilities and so neighborhoods with less than 5 venues were not perfect fit for the solution.

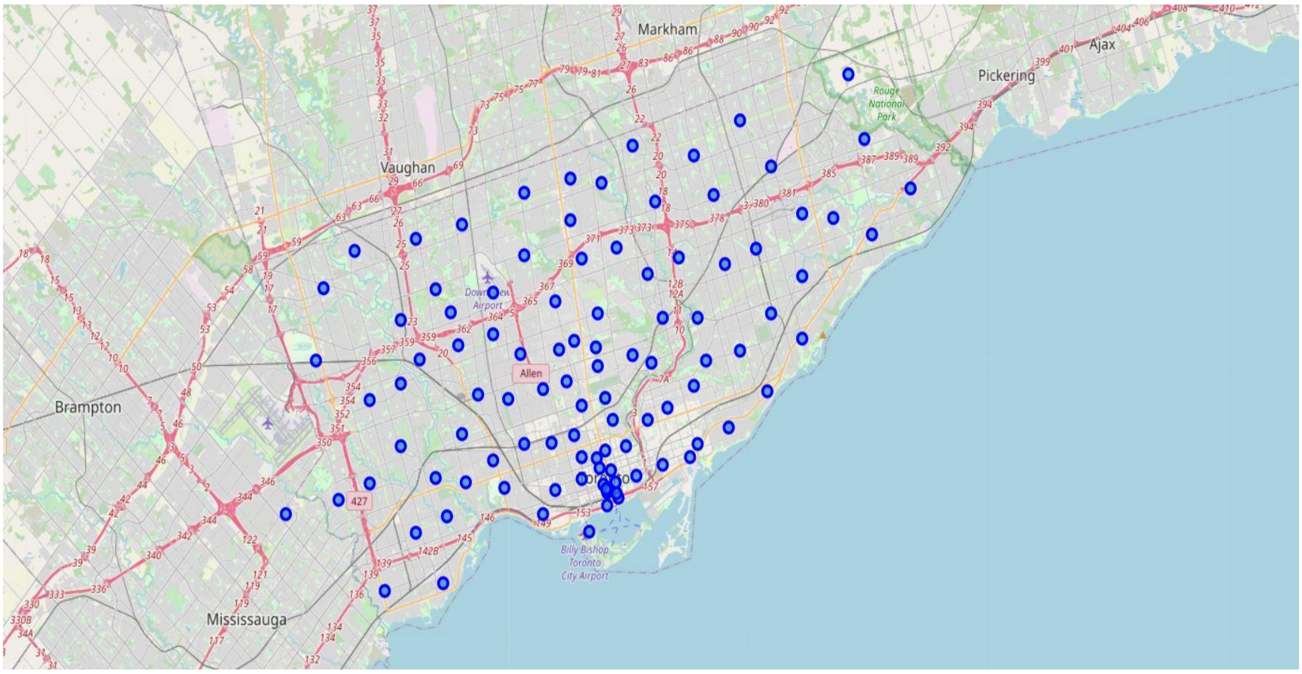
The category data was to be converted to numerical data for modeling the data. Categories data was one-hot encoded using pandas get_dummies function. Now, Data has Neighborhoods and each numerical category data. In the dataset, some neighborhoods were repeated as they had multiple venues and to compare neighborhoods, we have to combine all the same neighborhoods data into one row. For that purpose, mean of each neighborhood for each category was retrieved.

There were 313 different categories of venues. All the categories were not important for the analysis. For a particular person, 6 or 7 categories would be of more importance than other categories. Here, we don’t have particular personal need so I have considered 7 basic necessities categories to cluster neighborhoods. Those 7 categories are Gym/Fitness Centre, Grocery Store, Bank, ATM, Pharmacy, Shopping Mall and Restaurant.

After cleaning null and all zero values, the dataset was ready for clustering. K-means clustering with 5 clusters were used on the dataset. The features of clustering were those 7 categories retrieved on previous step. The frequency of occurrence of each category determined clusters of neighborhoods. The cluster which has high frequency of occurrence of these categories are better. These clusters will help in recognizing neighborhoods with needed category shops.

4. Results

This map is of all the neighborhoods before clustering. The neighborhoods are dense in some areas and scattered in some areas.



The following map is of Toronto city after performing clustering. There are 5 different clusters of neighborhoods. Red and Purple clusters have more neighborhoods compared to other clusters. There are basically 5 different types.

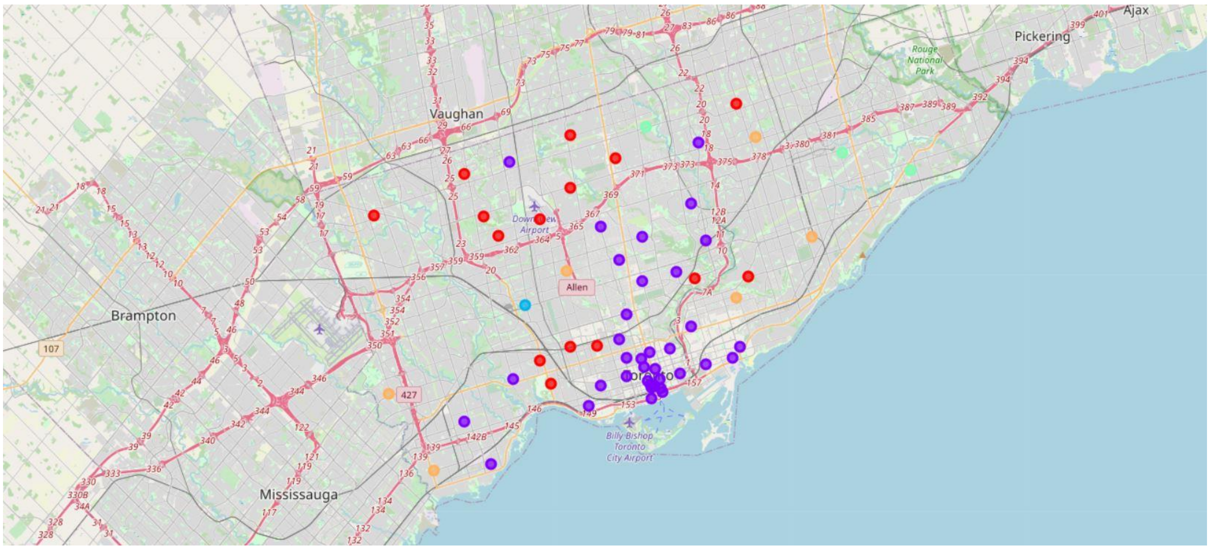
The red clusters are mostly on the airport side of the City which seems less populated.

Purple neighborhoods are near University of Toronto and beach side. This side is more dense than other sides.

The yellow cluster is of neighborhoods which are very far from main city area.

The sea blue cluster has only one neighborhood in it which is inside city region but it is only one neighborhood in the area.

The Cyan clusters are nearly on the border of the city.



5. Discussion

The results include 5 clusters and are of different properties and characteristics. The sea blue cluster has only one neighborhood and it is very deserted area. This area does not all the necessary facilities which makes it very weak candidate for the selection of this neighborhood. The Cyan cluster is at very end of the city which makes it very obvious for having less amenities so it is also not good for selection. The yellow cluster has very similar properties as Cyan so it is also a very bad candidate. There are two clusters remaining for the selection Red and Purple. The red cluster has no ATMs. The purple has few ATMs but is scarce in terms of Gyms and Shopping Malls. The red cluster is very scattered and purple is very dense in the area. The decision of choosing neighborhood now depends on distance, area of choice and which facilities are more important than others. For example, if Gyms and Shopping malls are more important and more frequently visited than ATMs and the person like to live in scattered area with some free space then neighborhoods from Red clusters will be more good choice over purple clusters. Then, to choose a neighborhood from the selected cluster would consist of consideration of proximity of work place. The one thing that was not considered in the discussion was number of restaurants. The reason was that there were many categories of restaurants in the City so it would clearly depend on the person to choose type of restaurant with his/her favorite food types. Here, I have considered generic restaurant category for clustering.

6. Conclusion

The project overall helps person select best neighborhood to live in. The other aspect of the project may help shop owners and businessmen to determine what kind of shops would be required in the area. If a person could identify basic needs of people living in the neighborhood than one place with all those facilities can be built and would give guaranteed business. The one limitation I can identify of this approach is that some small shops in small cities may not be registered on Foursquare and it would become difficult to take them into consideration while finding best fit of neighborhood. Overall, this project would help all the stakeholders to solve the problem and get the best solution.