THE BATTLE OF NEIGHBOURHOOD

A.BHARATH KUMAR

1:INTRODUCTION

Different cities in the world are filled with numerous kinds of venues that in turn define the cultures of the cities. Despite having dissimilarities, it is somewhat possible to group the similar kind of neighborhoods in different cities. It is possible to segment the different venues in a neighborhood. Having grouped similar kind of neighborhoods may serve as a variable to help to decide when people consider moving out of a city to another.

1.1: BACKGROUND

Many people are working in various cities (say New York and Toronto) across the world. Let's say a person wants to move from one city to another city because of his family situation. I think a person would love to shift a location which is exactly or almost similar to his/her last location because he/she loves the great amenities and other types of venues that exist in his/her current neighborhood like school, gym, swimming pool, Amusement park, restaurants, coffee-shops, spencer, etc.. If a person is shifting from one city to another city then my task would be finding the similar neighborhoods.

1.2: PROBLEM

Finding identical neighborhoods in different cities to help provide a perception of similar neighborhoods which may provide with plenty of insights to decide to choose a neighborhood that is far away, yet somewhat feels like home. Here I came with two solution

a)Finding a similar neighborhood by clustering the neighborhood based on the city surrounding

b)by fixing an area in one city and finding the recommended neighborhood in another city based on the similarities between these two cities

2.DATA ACQUISITION AND CLEANING

2.1: DATA SOURCE

This project works with two sets of data.

2.1.1: NEWYORK DATA

The first dataset consists of New York's different neighborhoods and their respective geometric coordinates, which can be found https://cocl.us/new_york_dataset

2.1.2: TORONTO DATA

The second dataset consists of Toronto's different borough and their respective postcodes, which can be found https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and info about longitude and latitude https://cocl.us/Geospatial_data

2.2: DATA CLEANSING

The data which we got is in JSON format which was found under the "feature" category. And it consists of four-column namely Borough, Neighbourhood, Latitude, and Longitude.

The second data source is a Wikipedia page that contains postcodes of the city of Toronto in an imitable. To scrape the data from the URL, Beautiful Soup has been used to extract the table data.

data frame has some values under the column 'Borough' which were not assigned in the first place. So, the rows with no assigned value in the 'Borough' column were dropped.

there were a few rows in the 'Neighbourhood' column that too had no values assigned to it. As a solution, the value from the 'Borough' column of the respective row was copied into the 'Neighbourhood' column.

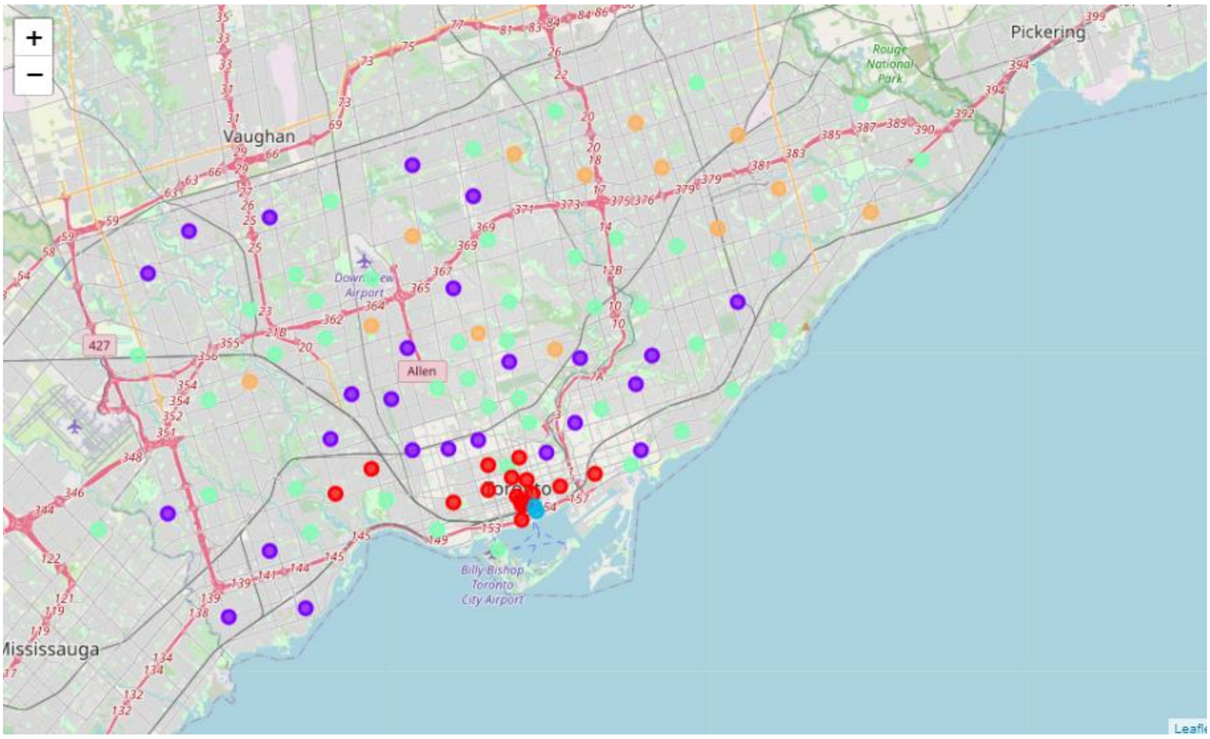
2.3 PRINCIPLE COMPONENT ANALYSIS

Principal Component Analysis is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information from the large set. Principal component analysis (PCA) is a mathematical procedure that transforms several (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components.

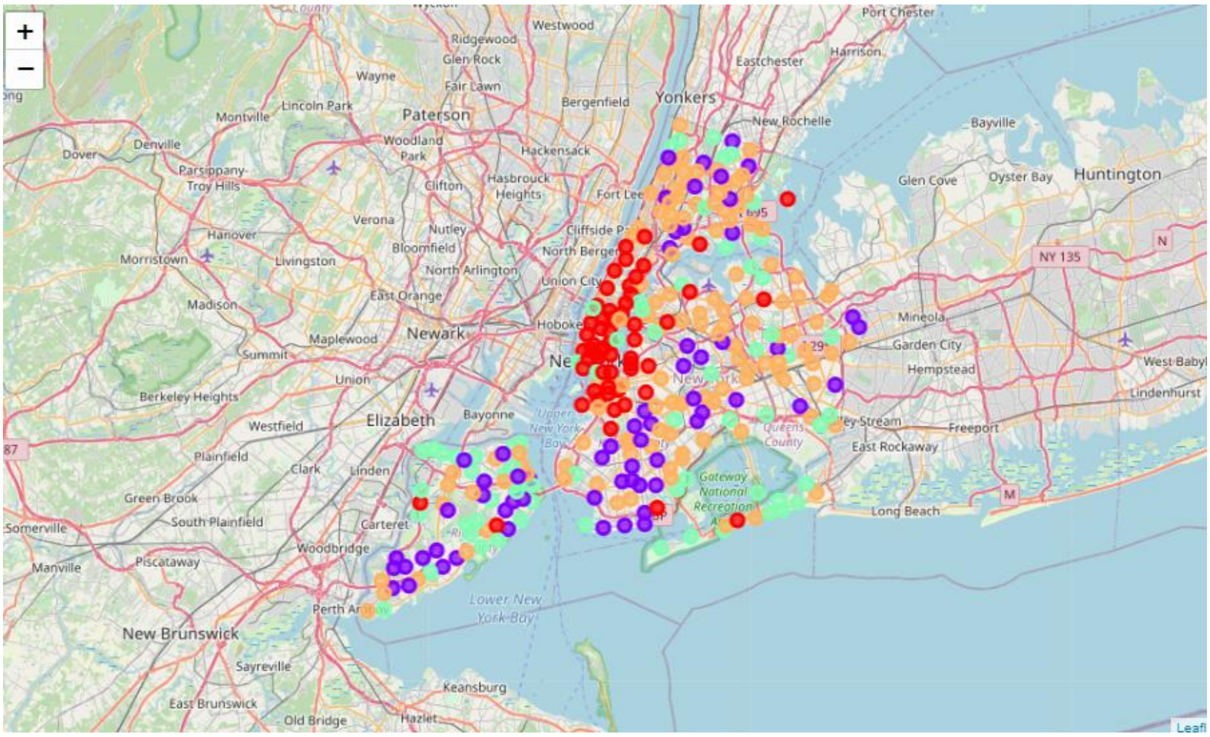
3.METHODOLOGY

After getting the above data, we’ll find the nearby venues to each borough and neighborhood pair in both cities. Foursquare API will be used to get the nearby venue. To use the Foursquare API we’ll need the developer account and it will give the client id and client secrets. Both cities may yield in the different number of venue categories, but we’ll take only common venue categories. Now we can find a borough and neighborhood of a city is how much similar to the boroughs of another city. In other words, we can find top similar boroughs and neighborhoods in another city. We’ll use cosine similarity to find the similarity between two cities and also to cluster the whole city into 5 different groups based on the similarity I used k-means algorithm

4:RESULT

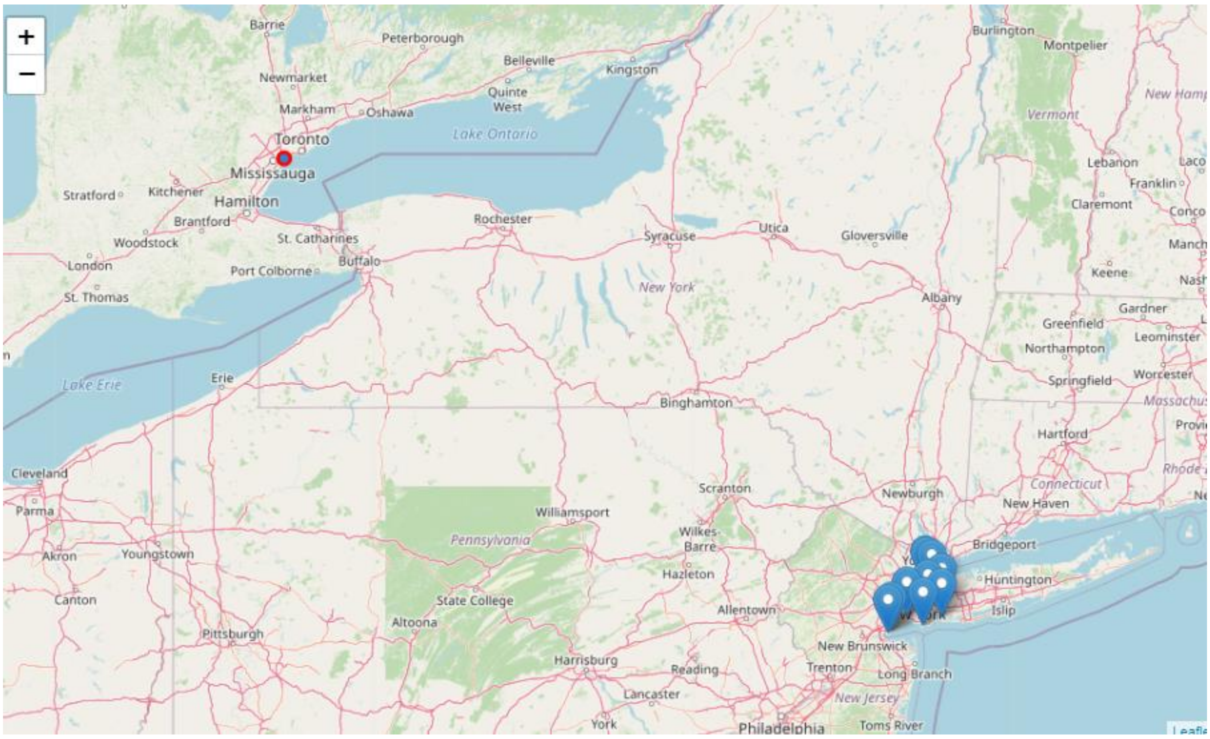


a)Venues pinned on Toranto after Clustering



b)Venues pinned on NY after Clustering

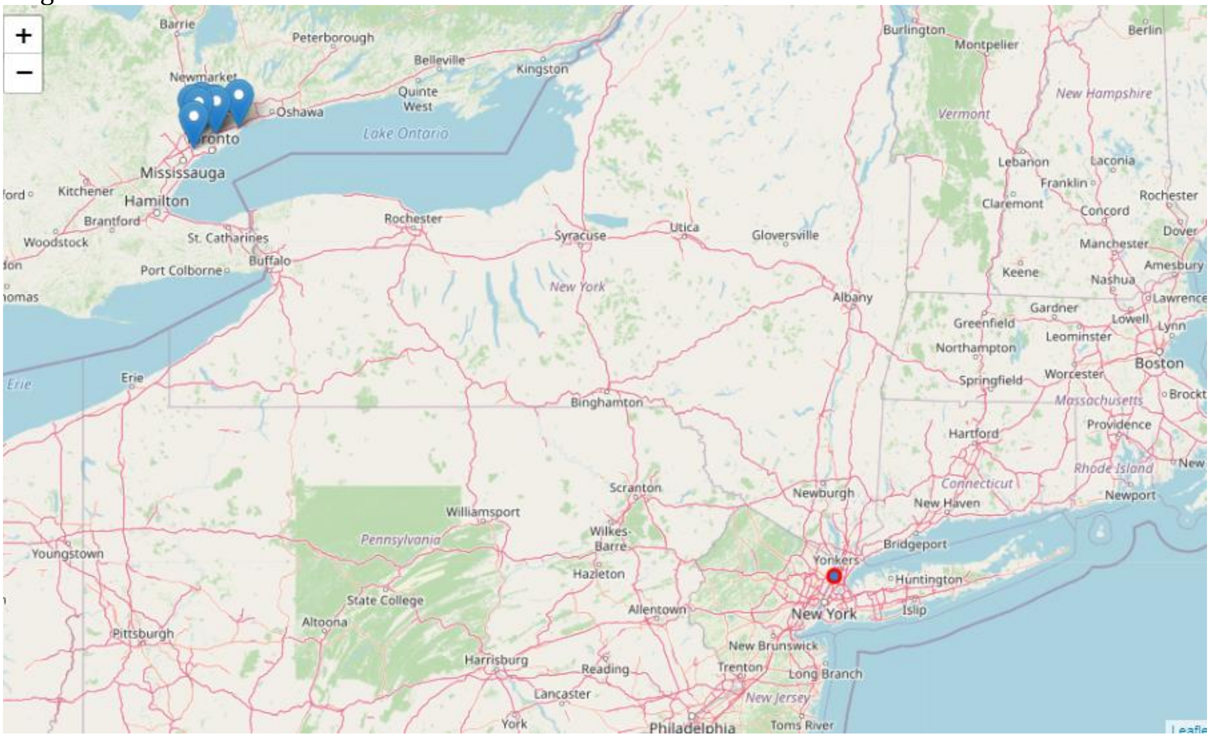
5 distinctive cluster has been created and also it represented by 5 different colors to differentiate it and to visualize it easily so now by comparing both city one can easily find a similar neighbor in another city



From New York to Toronto

Current Location: Bronx, Riverdale, New York

Here one decides to move from Bronx, Riverdale, New York so in the map it pinned the neighborhood in Toronto which almost has similar features



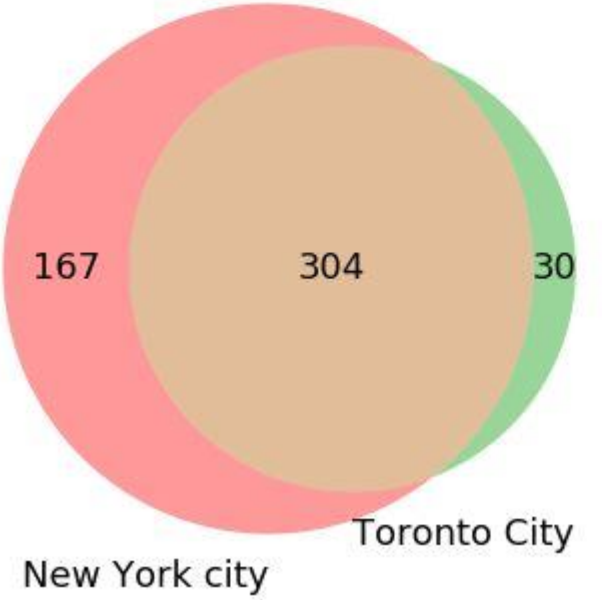
From New York to Toronto

Current Location: Etobicoke, 'Alderwood, Long Branch', Toronto

Here one decides to move from Etobicoke, 'Alderwood, Long Branch', Toronto so in the map, it pinned the neighborhood in Newyork which almost has similar features

5:DISCUSSION

There are some common venue categories in both data. The figure below will show that there are 304 common venues in both cities



.

The project was only done on the zip codes of New York and Toronto, each having 150 features, even after performing dimensionality reduction. Having more samples may result in a better clustering The study here is being ended by visualizing the data and clustering information on the map of the City of New York and Toronto

6:CONCLUSION

People are frequently moving into new cities. And in this ever-growing world filled with technology, having a neighborhood recommendation based on location data is something to be considered basic nowadays. And the application of neighborhood segmentation lies beyond this application too. This can serve to be an impressive tool to better organize city resources. Furthermore, it can be used as a tool for security measurement if combined with crime data.

Cosine similarity is used here to find how similarity between two boroughs. This model can be implemented within a city also. This will help to find a suitable place for people. This will give the most similar borough and neighborhood.