Help mover to find similar area Report

Problem Overview

Many people like to stay in places with certain features and when they have to move they struggle with finding similar place

This project will use Machine learning clustering algorithm techniques along with real data from Foursquare API to quantify and provide guidance to movers from New York city, USA to the city of Toronto, Canada

Data Overview

For this project the Foursquare API will be used along with A list of neighborhoods in New York and Toronto is downloaded with location in longitude and latitude coordinates

New York
neighborhoods: https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json

Toronto neighborhoods: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The data downloaded are the neighborhoods located in New York and Toronto and will be determined based on the frequency of the categories found in the neighborhoods.

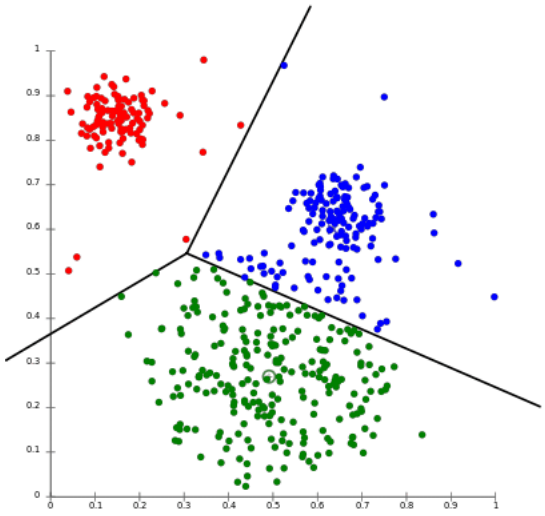
Methodology

feature extraction

Each feature becomes binary, this means that 1 means this category is found in the venue and. Then, all the venues are grouped by the neighborhoods, This will give us a venue for each row and each column will contain the frequency of occurrence of that particular category. computing at the same time the mean.

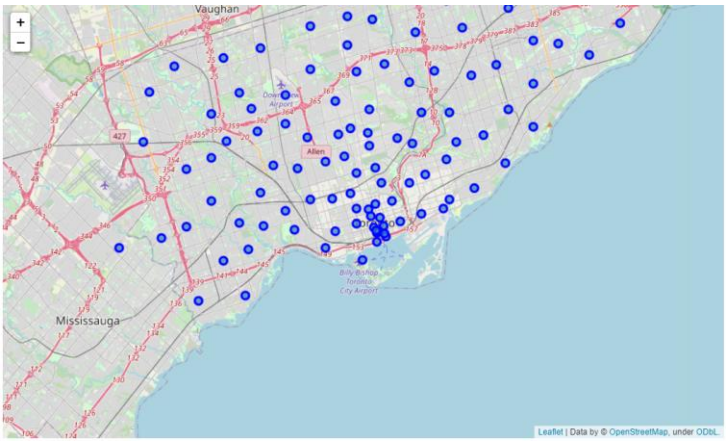
method used is K-means clustering

***k*-means clustering** is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in data mining. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.



Results

data is plotted in a geographical map to get a sense of the world location.



Then the number of clustter was implemented using the elbow method

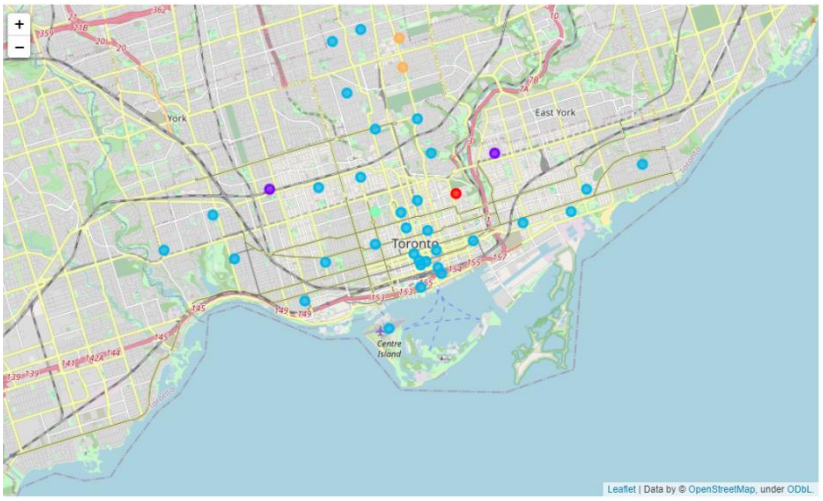
there is a cluster with one neighborhood. In the results we found out that this cluster has a frequency of 1 in garden places. This means the cluster is not segmenting correctly data and the centroid is located in the exact position of that neighborhood. This neighborhood has a high

frequency of garden places around. Hence, we can say the algorithm is doing great since there is no other cluster with similar venues around.

Conclusion

The K-Means clustering algorithm is used for finding similarities between all the neighborhoods listed in the feature matrix. The elbow method is used for selecting the appropriate number of clusters. Hence, the K selected is 5. Results show that there are 2 major groups and 2 minor groups. In addition, there is one group that contains only one neighborhood that is isolated from others.

Clusters



1. Neighborhoods that have around parks, bus lines and sandwich places.
2. Neighborhoods that have around parks, playgrounds and trails.
3. Neighborhoods that have around coffee shops, pubs and Italian restaurants.
4. Neighborhood that have around gardens.
5. Neighborhoods that have around coffee shops, parks and bakeries.