The Battle of Neighborhoods (W1)

Author: ABDULLAH ABDULHADI

1. Introduction

Recently, Machine Learning (ML) algorithms are widely used in the study of data instead of traditional statistics. ML algorithms bring advantages because they offer solutions to problems related to the big quantities of data and set fewer constraints than traditional statistics. In particular, unsupervised learning algorithms are used to find patterns in data in terms of similarity between samples. Depending on the pattern within the data, different algorithms are used. For non-convex data it is used Density-Based Spatial Clustering (DBScan). On the other hand, for convex data it is used a well known algorithm as K-Means.

Foursquare is a website where people comment and rank food sites, coffee sites, malls and parks. For instance, let's think that a Foursquare user had to move from New York city, USA to the city of Toronto, Canada. Foursquare location data along with a clustering algorithm can suggest a neighborhood in order to help this user to live in Toronto in a similar place. The neighborhood that will be suggested, will not be a random suggestion, but instead will be a place for his pleasure. Thus, previous data from New York and Toronto will be used to predict a good living neighborhood for him.

2. Data

For this project the Foursquare API will be used. A list of neighborhoods in New York and Toronto is downloaded and their respective location in longitude and latitude coordinates is obtained. The sources are the following:

- 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. Moreover, their specific coordinates are merged. Only Manhattan neighborhoods

and boroughs that contain the string "Toronto" are taken into account. A Foursquare API GET request is sent in order to adquire the surrounds venues that are within a radius of 500m. The data is formated using one hot encoding with the categories of each venue. Then, the venues are grouped by neighborhoods computing the mean of each feature.

The similarities will be determined based on the frequency of the categories found in the neighborhoods. These similarities found are a strong indicator for a user and can help him to decide whether to move in a particular neighborhood near the center of Toronto or not.

3. Methodology

3.1. Feature Extraction

For feature extraction One Hot Encoding is used in terms of categories. Therefore, each feature is a category that belongs to a venue. Each feature becomes binary, this means that 1 means this category is found in the venue and 0 means the opposite. Then, all the venues are grouped by the neighborhoods, computing at the same time the mean. This will give us a venue for each row and each column will contain the frequency of occurrence of that particular category.

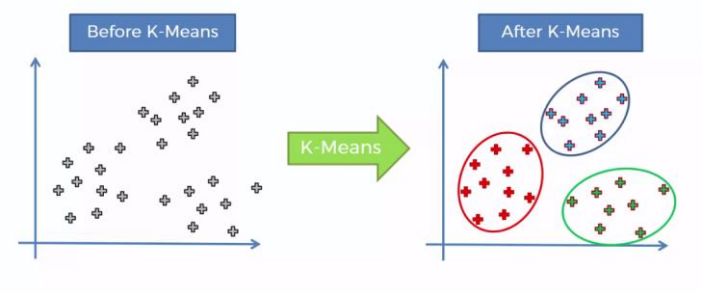
3.2. Unsupervised Learning

For the purpose of doing unsupervised learning to found similarities between neighborhoods, a clustering algorithm is implemented. In this case K-Means is used due to its simplicity and its similiraty approach to found patterns.

- **K-Means:**

K-Means is a clustering algorithm. This algorithm search clusters within the data and the main objective function is to minimize the data dispersion for each cluster. Thus, each group found represents a set of data with a pattern inside the muldimensional features.

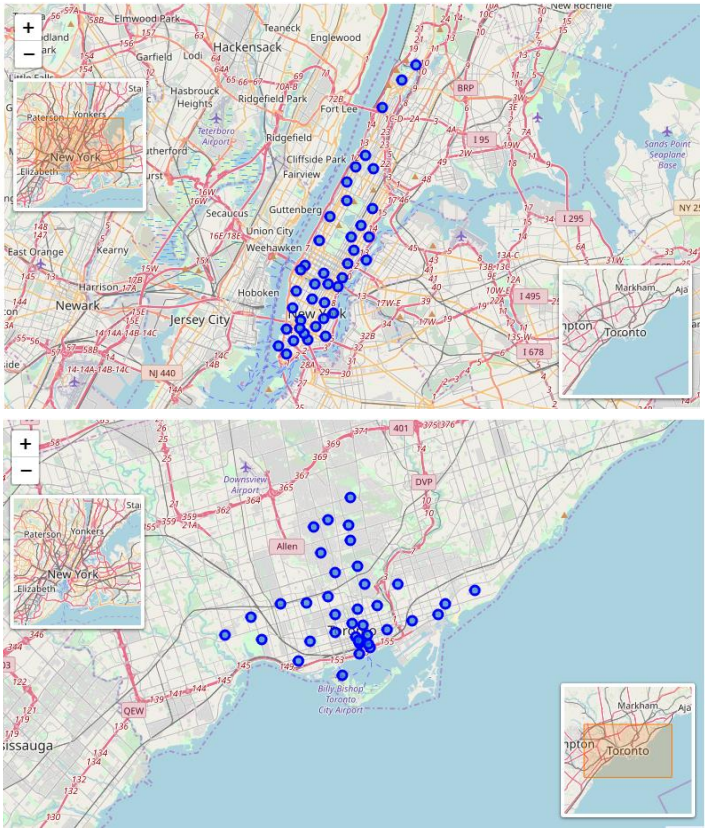
In the following figure there is a graphical example of how a K-Means algorithm works. As it is possible to see, dispersion is minimized by representing all clustered data into one group or cluster.



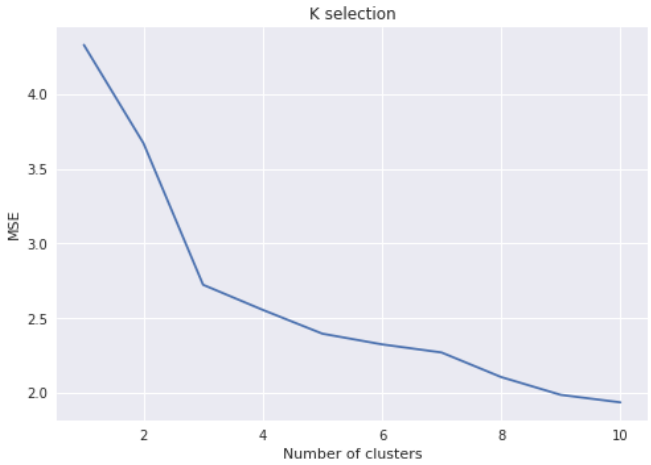
It is necessary for this algorithm to have a prior idea about the number of clusters since it is considered an input of this algorithm. For this reason, the elbow method is implemented. A chart that compares error vs number of cluster is done and the elbow is selected. Then, further analysis of each cluster is done.

4. Results

Firstly, data is plotted in a geographical map to get a notion of the world location. In the two following images are shown the neighborhoods in Manhattan and Toronto.

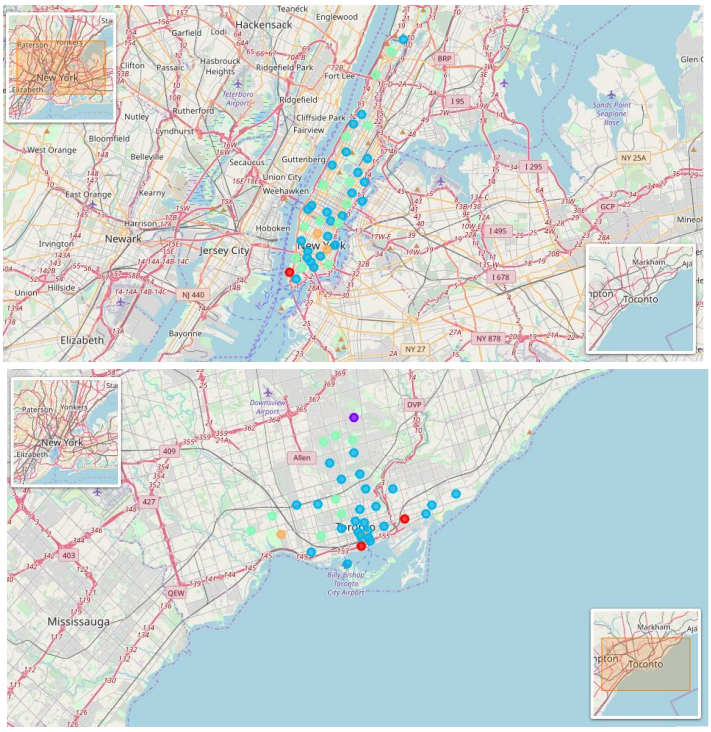


Secondly, the cluster algorithm is implemented. For this purpose, it is necessary to have a prior idea about the number of clusters. Therefore, the mean squared error (MSE) is plotted vs the number of clusters. The number of clusters start with a value of 1 increasing until a value of 10. This chart is shown in the image below.



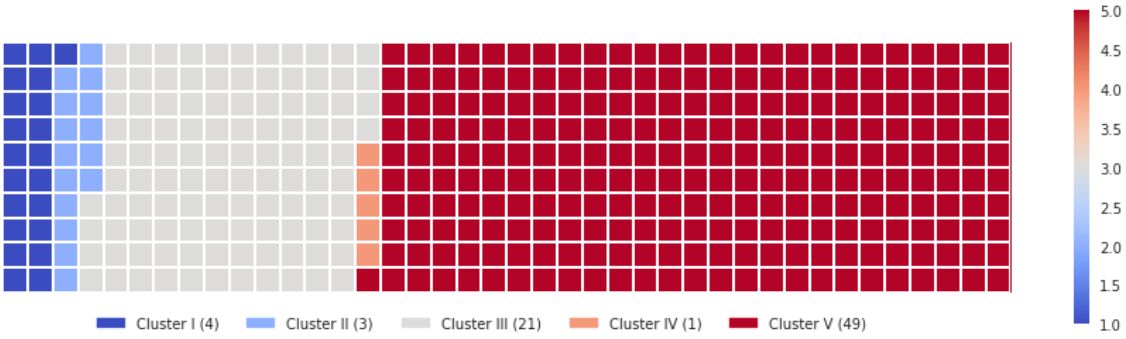
As it is expected, the MSE decreases over the number of clusters. The elbow method here is implemented in order to select the appropriate number of groups. In this case, it is possible to see that the elbow is found more or less around 5. The MSE found below this number shows little changes rather than big ones. Finally, once the number of clusters is fixed, the clustering algorithm is repeated through samples and each neighborhood is labeled according to the clusters found.

For visualization purposes, the geographical data is again plotted but with different colors. Each color represents the cluster for which that neighborhood belongs. This image is shown below.

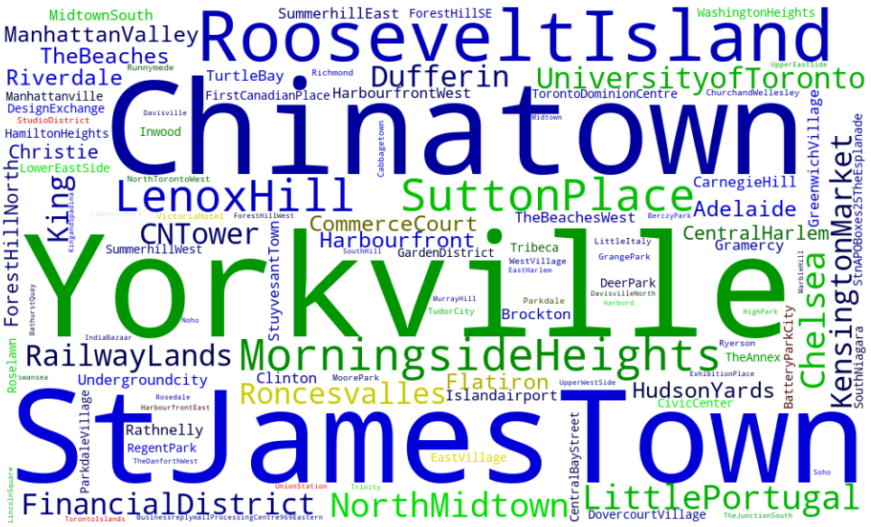


In this image it is evident that cluster algorithm is not segmenting the neighborhoods for location areas. This means that it is not true that geolocation of neighborhoods is correlated with the categories of the venues around each neighborhood. Yet, it is possible to see which neighborhoods within Manhattan, New York are more similar to the neighborhoods within Toronto. Those neighborhoods that are similiar among them belong to the same cluster. Hence, they have the same color in the image above.

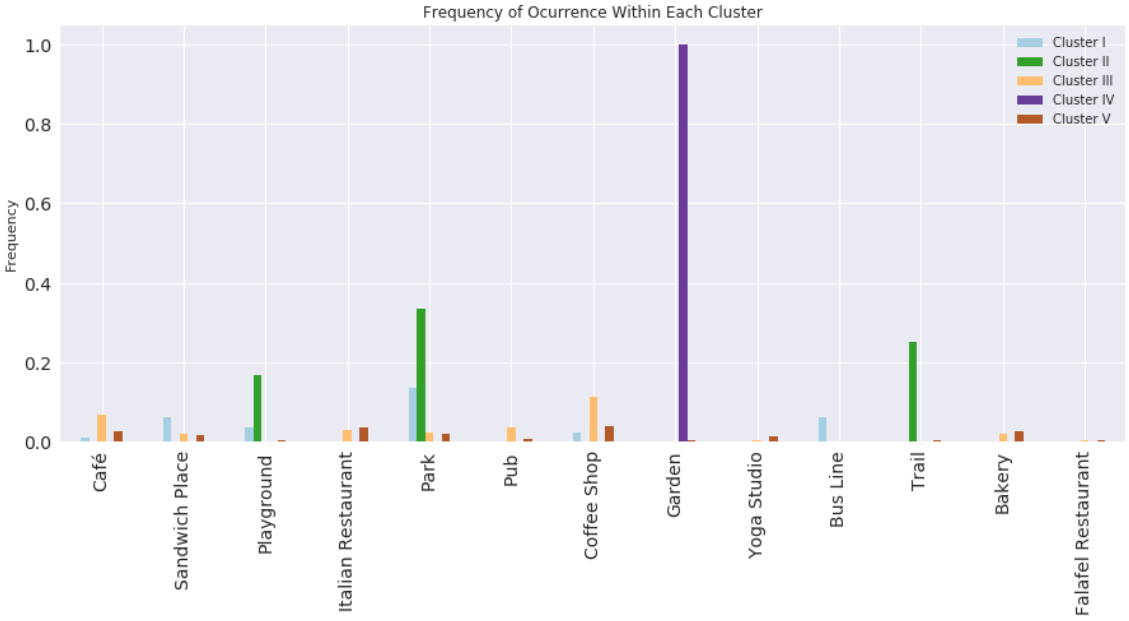
In the image below it is found the proportion of the neighborhoods assigned to each cluster. For this reason a waffle chart is implemented. There are two major clusters and two minor clusters. Moreover, there is a cluster that has only one neighborhood. This neighborhood is Lawrence Park from Toronto. This neighborhood has no similarity with New York city.



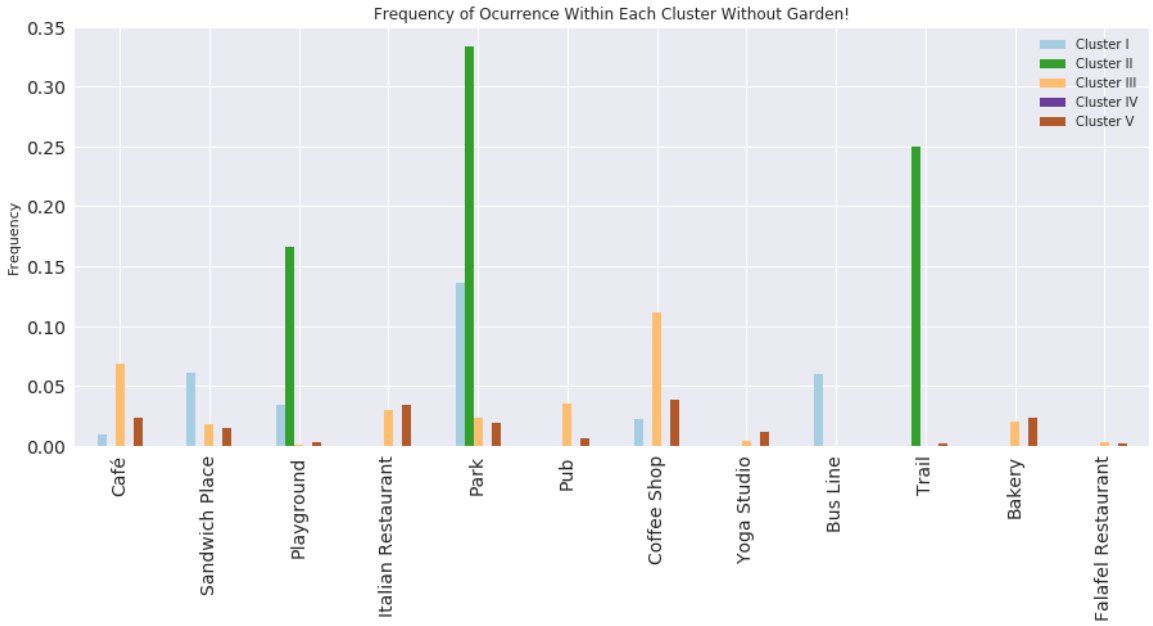
For practical purposes, a word cloud is shown in the image below. In this way, a person trying to locate a similar neighborhood in Toronto can locate it looking for the neighborhoods with the same color. Here we can see that Chinatown, Roosevelt Island, St James Town and so on belong are similar among them. On the other hand, York ville, Sutton Place, North Midtown and so on are different from those we mentioned earlier but again they are similar among them.



For research purposes, bar charts are employed to found insights within the clusters. The bar chart that is shown below shows the features with higher frequency in the centroids found by the algorithm. In this way we can learn what the algorithm is finding.



The image above shows a particular category "garden" with a high frequency of 1. This category is related to the IV cluster, which is the one that has just one neighborhood. This cluster does not add further information. Hence, we can remove it and analyze the other ones. The figure below shows the new bar chart without the IV cluster.



It is possible to see that I cluster focuses on neighborhoods that have around parks, bus lines and sandwich places. On the other hand II cluster focuses on neighborhoods that have around parks, playgrounds and trails. Third (III) cluster focuses on neighborhoods that have around coffee shops, pubs and italian restaurants. The last (V) cluster focuses on neighborhoods that have around coffee shops, parks and bakeries.

5. Discussion

It is worth to note that this work is useful only for those who live in Manhattan, New York or in the neighborhoods near the center of Toronto. The reason is because there is a limited amount of data we can request using de Foursquare API. Consequently, it will has a greater cost than the Lite version.

Moreover, 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.

6. Conclusion

Section where you conclude the report.

In this work a segmentation between two different countries is done. This segementation involves the neighborhoods in Manhattan, New York and the neighborhoods near to the center of Toronto. The data is downloaded and the venues around the neighborhoods is adquired using the Foursquare API. One Hot Encoding is used for converting the categories of the venues into a feature matrix. Then, all venues are grouped by neighborhoods and at the same time the mean is calculated. Hence, the resulting features used are the frequency of occurrence from each category in a neighborhood.

The K-Means clustering algorithm is used for finding similiraties 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 adition, there is one group that contains only one neighborhood that is isolated from others. The description of the clusters is the following:

Cluster

- I: Neighborhoods that have around parks, bus lines and sandwich places.
- II: Neighborhoods that have around parks, playgrounds and trails.
- III: Neighborhoods that have around coffee shops, pubs and italian restaurants.
- IV: Neighborhood that have around gardens.
- V: Neighborhoods that have around coffee shops, parks and bakeries.

Finally, any user who wants to move from manhattan to toronto and viceversa can use this system to get a notion or idea about what is the best suitable place for him.