The Battle of the Neighborhoods

Capstone
Final Project

1.) Introduction

When migrating from one city to another a lot of things have to be considered, e.g. in which neighborhood to live. This needs a lot of time for collecting all the information, comparing the neighborhoods, and evaluating everything that has to be considered.

The objective of this project is to help people who want to move from New York City, USA to Toronto, Canada. Therefore, similar neighborhoods in New York and Canada shall be found for making it possible to make a smart and efficient decision for finding a suitable neighborhood in the new city.

Machine Learning algorithms are a common tool in the study of data and have proven to be an extremely helpful tool when it comes to big quantities of data. Instead of traditional statistics they set fewer constraints. Unsupervised learning algorithms, in particular, can be used for finding patterns in terms of similarity between samples. The algorithms which are used depend on the pattern within the data. Density-Based-Spatial Clustering (DBScan)is used for non-convex data. For convex data a K-Means algorithm is used as a well-known tool.

The social network Foursquare is an established and independent platform for geo-tracking data, to get a good insight into the activities of persons. So, for a user of Foursquare who wants to move from New York to Toronto, the Foursquare location data in combination with a clustering algorithm can suggest a neighborhood in Toronto as a similar place to live in. Thereby the suggested neighborhood will not be a random suggestion, but a very suitable place. For this, previous data from New York and Toronto will be used for the prediction of a good future neighborhood to live in.

2.) Data

The Foursquare API will be used for this project. Lists containing the neighborhoods of New York and Toronto are downloaded and their respective coordinates in latitude and longitude are obtained. The lists are obtained from the following sources:

- Neighborhoods of New York in JSON format
- Neighborhoods of Toronto from Wikipedia

The downloaded data contains the neighborhoods which are located in Toronto and New York. Furthermore, their specific coordinates are then merged with this data. For the analysis are only the neighborhoods of Manhattan in New York and the boroughs which contain the string "Toronto" into account. For acquiring the surrounding venues within a radius of 500m the Foursquare API is used. For this a GET request is sent to the Foursquare

API. Using one hot encoding with the categories of each venue, the data is formatted. After that 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 found similarities will serve as an identifier for the user whether to choose a particular neighborhood near the center of Toronto as new place for living or not.

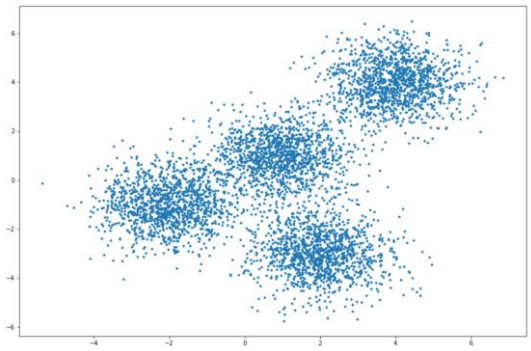
3.) Methodology

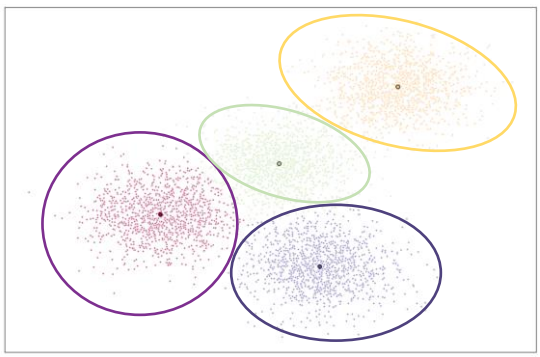
3.1) Extraction of Features

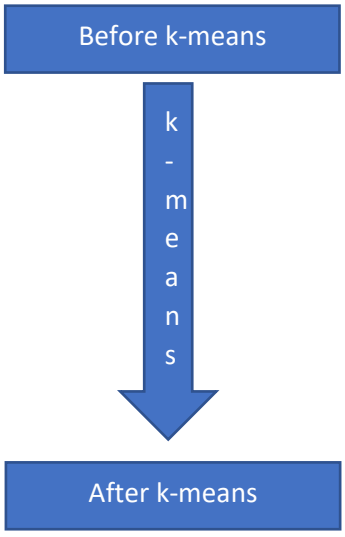
One Hot Coding is used for feature Extraction in terms of categories. Each feature is a category, which belongs to a venue. Each feature is mapped binarily: if 0, the category is not found in the venue and if 1, the category is not found in the venue. In the next step the venues are grouped by neighborhoods, while the mean is computed at the same time. This results in a venue for each row and each column will contain the frequency of occurrence of that particular category.

3.2) Unsupervised Learning

Unsupervised learning is used for finding similarities among the neighborhoods. Therefore, a clustering algorithm is implemented. Here the k-means algorithm is used. It offers the advantages of simplicity in usage and of dividing the data in non-overlapping subsets (clusters) without any cluster-internal structure. Thereby the objects within a cluster are very similar, while objects across different clusters are very different.
With the following figures an example of the k-means method shall be demonstrated:



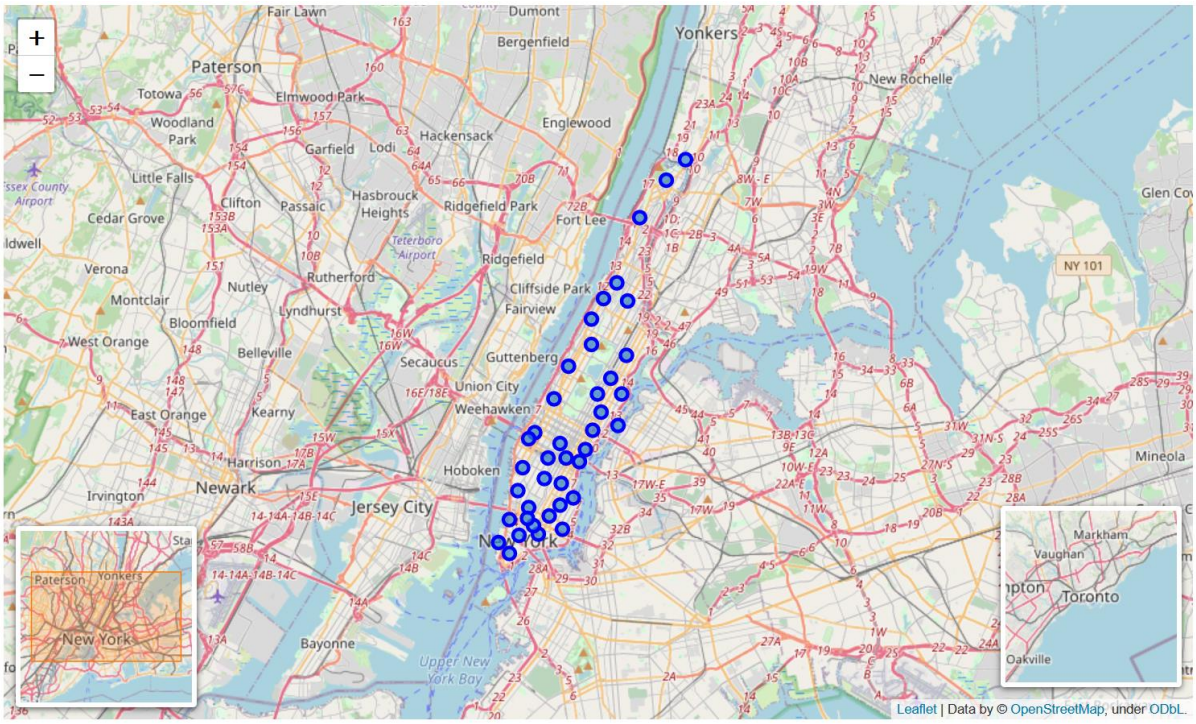


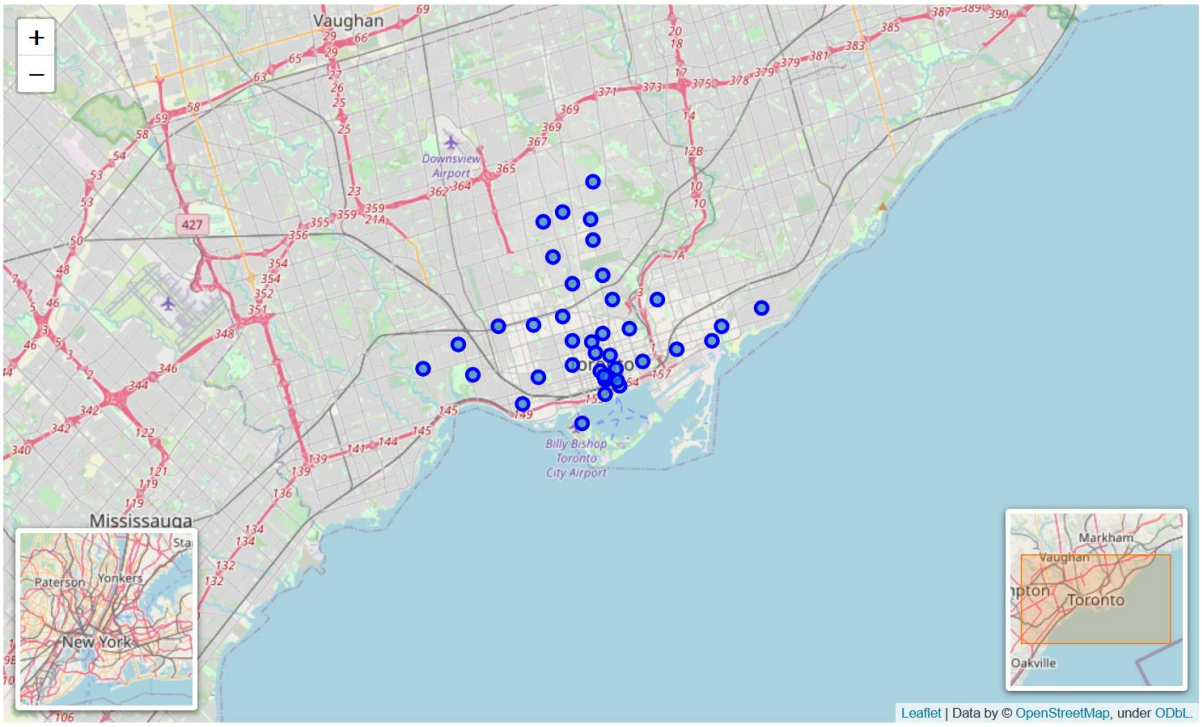


For this algorithm it is necessary to have a prior idea about the number of clusters since it is a necessary parameter for the k-means algorithm. To find the best number of clusters the elbow method is applied: An evaluation, which compares the mean squared error (MSE) to the number of clusters is performed and the number of the elbow is selected for the number of clusters. After this each cluster can be further analyzed.

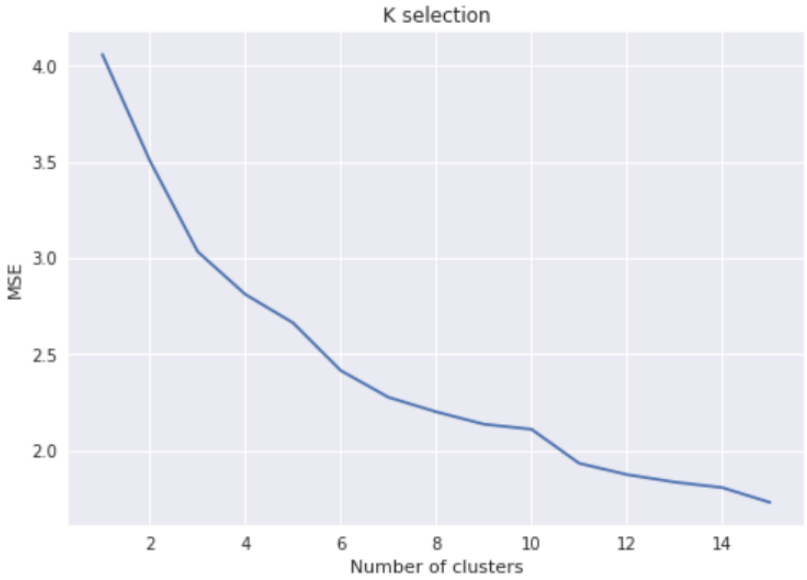
4.) Results

For getting an overview of the locations, the neighborhoods of New York, Manhattan and Toronto are plotted in geographical maps:



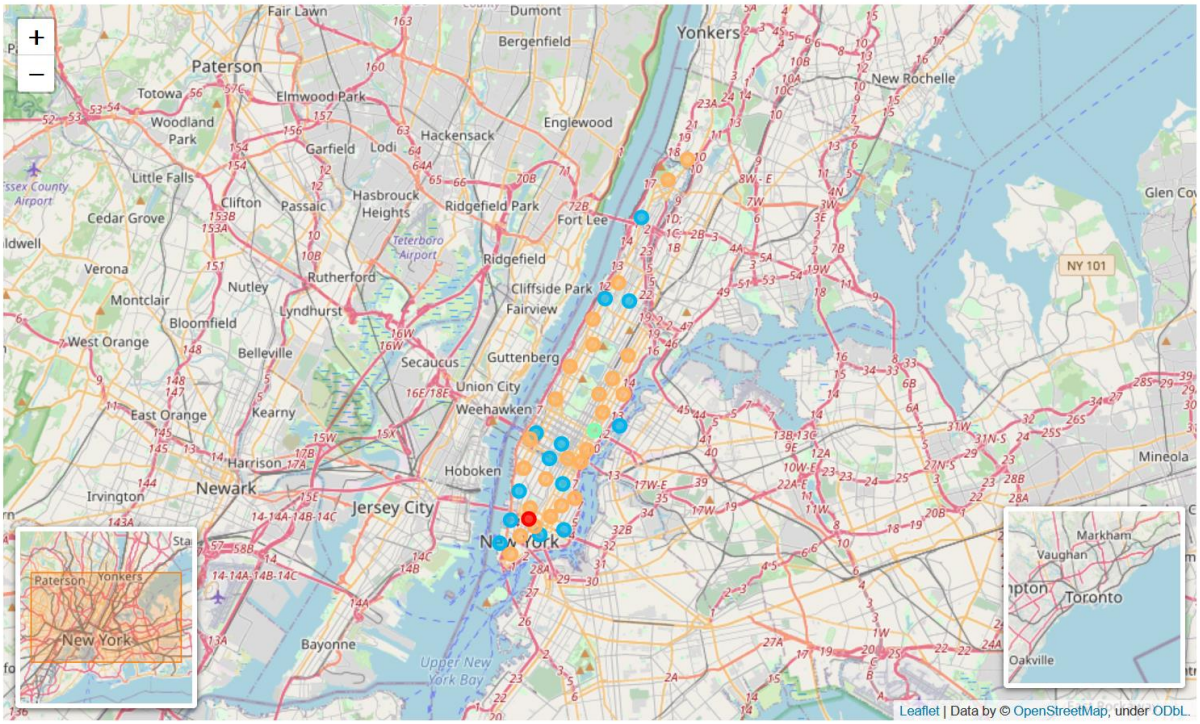


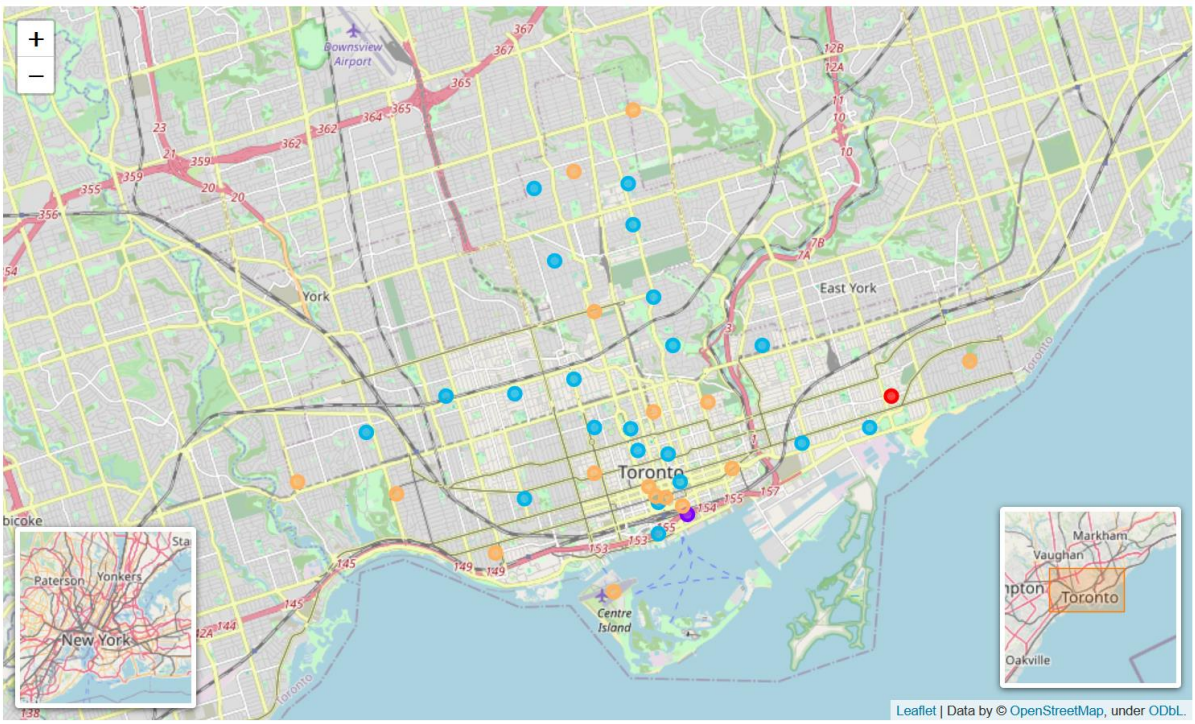
In the next step the procedure for the k-means algorithm is performed, beginning with the determination of the appropriate number of clusters. Therefore, the MSE is plotted to the numbers of clusters in the range from 1 to 15:



The MSE decreases with the number of clusters. With the elbow method, for the selection of the appropriate number of groups, a value of 5 is chosen as number of clusters.
With the appropriate number of clusters found, the k-means clustering algorithm is performed for the samples and each neighborhood is labeled, according to the clusters found.

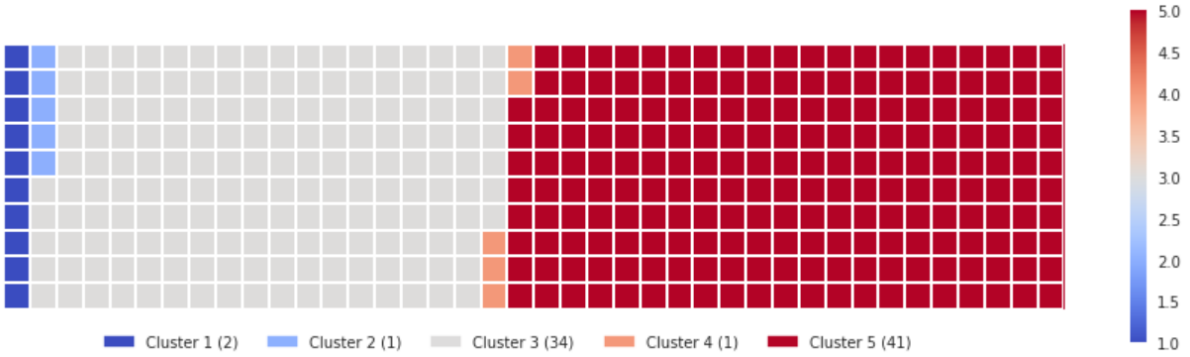
The following maps display the clustered neighborhoods. Hereby each cluster is represented by a different color:



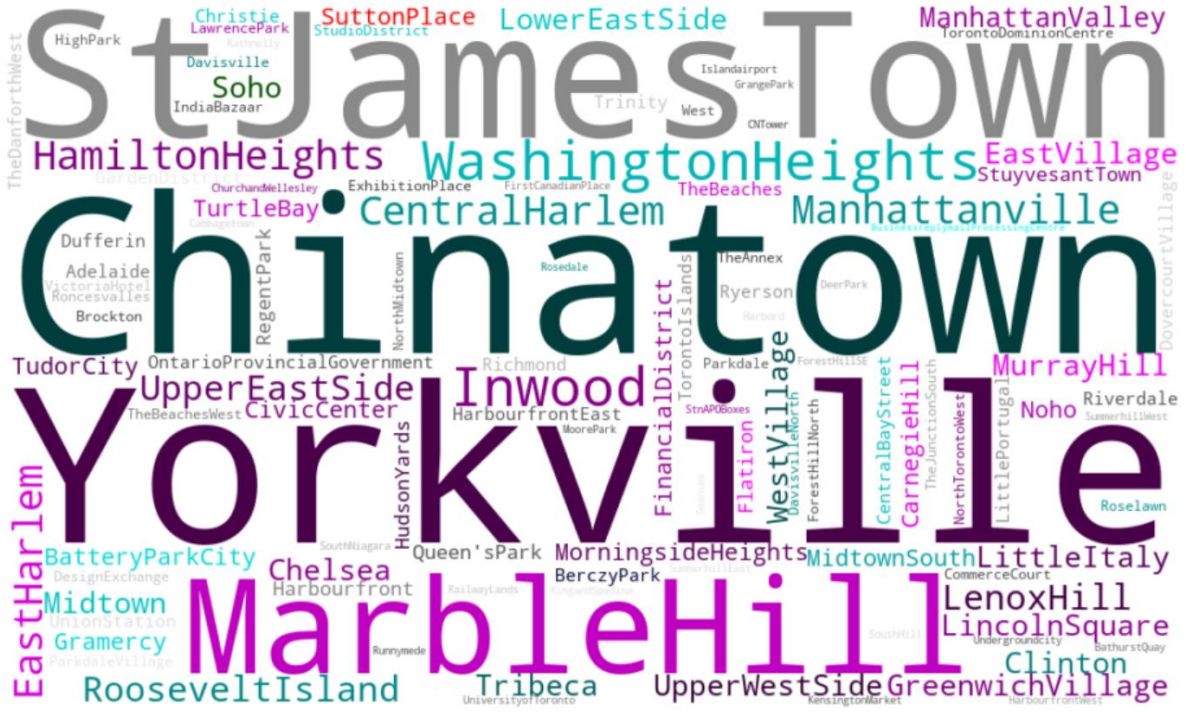


From these figures it can be determined that the k-means cluster algorithm does not segment the neighborhoods for their location. This shows that the geolocation of the neighborhoods does not correlate with the venues around each neighborhood. Within these maps it is possible to identify which neighborhoods in New York, Manhattan are similar to neighborhoods in Toronto. The similar neighborhoods belong to the same cluster and are therefore displayed by the same color.

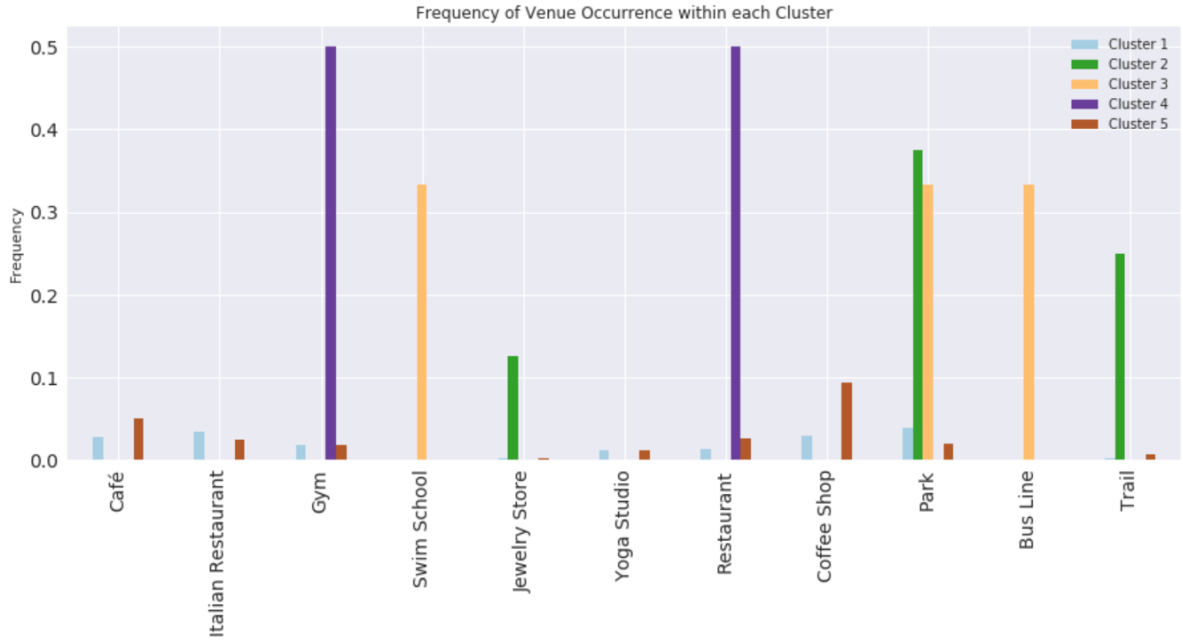
The following waffle charts shows the proportion of the neighborhoods assigned to each cluster. Two clusters contain the major part of all neighborhoods, one cluster with only two neighborhoods and two clusters with only one neighborhood. The cluster with two neighborhoods contains one neighborhood of each New York, Manhattan and Toronto (red color circle in map). The single neighborhood clusters refer to neighborhoods without similarities in the other city, one for New York, Manhattan (green circle) and one for Toronto (purple circle).



With the word cloud shown in the figure below, similar neighborhoods can be detected by the same color. So, it can for example be seen, that Yorkville, LenoxHill and others in the same color belong to the same cluster and are similar, while SuttonPlace is not similar to other neighborhoods.



With a bar chart a more detailed view of the clusters and its venues can be obtained. It shows the features and their frequency in each cluster.



5.) Discussion

It should be mentioned that the results of this work are mostly useful for people who live in New York Manhattan or near the center of Toronto. This is because there is only a limited amount of data, which can be requested using the Foursquare API. Consequently, increasing the amount of data would have a greater cost but would also give the possibility to extend this analysis.

Furthermore, there are two clusters, 1 and 3, with just one neighborhood, one for each city. As can be seen in the bar chart, their venue features are very unsimilar to the other clusters, as all venue feature are very unsimilar. Hence, we can conclude, that the algorithm is performing great, since all clusters do have very unsimilar venue features.

6.) Conclusion

This work presents a segmentation between tow different cities in two different countries. In this segmentation are the neighborhoods of New York, Manhattan, and the neighborhoods near the center of Toronto involved. The data is acquired by downloading and importing from a JSON file and collecting data from a Wikipedia page. The data of the venues near the neighborhoods is acquired via the Foursquare API. With One Hot encoding the categories of the venues are converted into a feature matrix. After that 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.

With the k-means algorithm the similarities between all the neighborhoods is listed in the feature matrix. With the elbow method the appropriate number of clusters is determined and gives here a number of 5 clusters as a good a good value. The results show that there are 2 major groups, 1 group with two neighborhoods, which are similar and one can be found in each city and two groups with each one neighborhood whereby there is one of these groups for each city. These single neighborhood groups therefore represent an unsimilarity to all other neighborhoods.
This gives the following Clusters:

Cluster 1: Neighborhoods that have Cafés, Italian Restaurants, Coffee Shops and Parks around

Cluster 2: Neighborhoods that have Jewelry Stores, Parks and Trails around

Cluster 3: Neighborhoods that have Swim Schools, Parks and Bus lines around

Cluster 4: Neighborhoods that have Gyms and Restaurants around

Cluster 5: Neighborhoods that have Cafés and Coffee Shops around