Capstone Project – The battle of neighborhoods

1. Introduction

The intention of this capstone project is to explore the neighborhoods of the cities Toronto and New York. Therefore Foursquare location data will be leveraged to identify characteristics of the neighborhoods of the cities. In a first step the data will be mined and wrangled before the characteristics will be visually explored. Afterwards the usage of clustering methods will allow to find similarities in the data and answer the question of the stakeholder i.e. solve the business problem.

The concrete question that will be answered is:
What are the characteristics of someones neighborhood in NY (e.g. in Riverdale, NYC)? If he/she would like to move to Toronto, which neighborhoods are comparable and due to which characteristics?

2. Data

For this task the following data is used:

- New York City

  URL: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

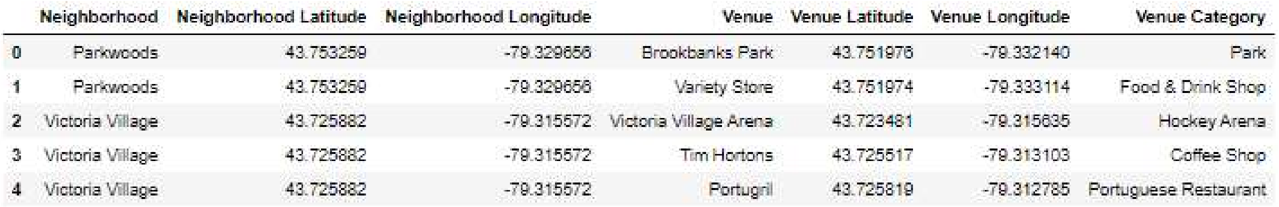
- Toronto (merged of two datasets)

  URL: 1. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

  2. http://cocl.us/Geospatial_data

The New York City dataset consists of 306 examples i.e. neighborhoods, while the Toronto dataset contains 103 neighborhoods. The Foursquare API is used to get nearby venues for each neighborhood. Parameter for the search are a limit of 30 venues and the radius of 500. In total 7739 venues are added and considered for the further analysis.

The first 5 (out of 7739) examples of the gathered data:



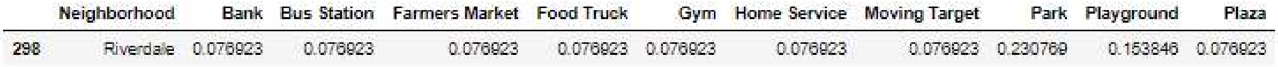
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

There are 424 different venue categories which is also the potential number of features which can be used later for the data clustering.

3. Methodology

Before clustering one hot encoding was performed to make the data fit the k-means algorithm. Further for every neighborhood the mean of the amount of each venue category was calculated.

This gives for example the following values for the Riverdale neighborhood (only the venue categories unequal to zero are presented):

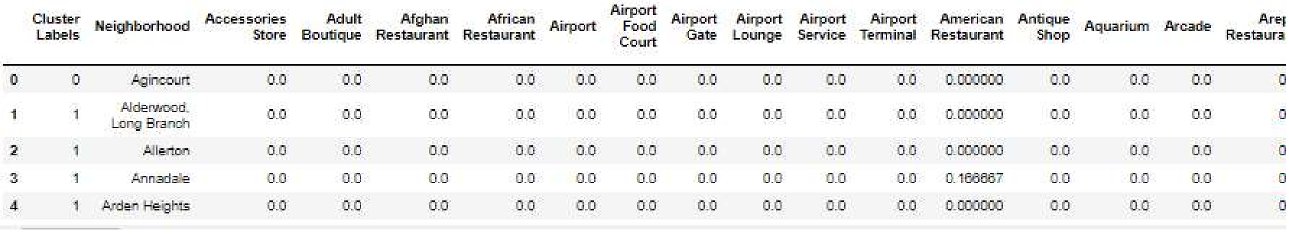


	Neighborhood	Bank	Bus Station	Farmers Market	Food Truck	Gym	Home Service	Moving Target	Park	Playground	Plaza
298	Riverdale	0.076923	0.076923	0.076923	0.076923	0.076923	0.076923	0.076923	0.230769	0.153846	0.076923

The parameters of the k-mean-clustering were chosen as in the following:

- Number of n-clusters: 5
- random state: 0 (start point are chosen randomly)

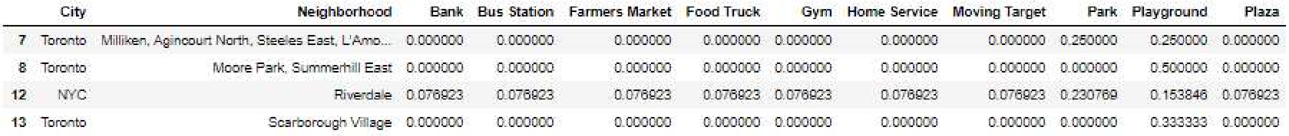
Cluster labels are added to the dataset:



	Cluster Labels	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Arcade	Arep Restaura
0	0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0
1	1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0
2	1	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0
3	1	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.166667	0.0	0.0	0.0	0
4	1	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0

4. Results

After applying the algorithm the cities were added to the results, giving the following table of neighborhoods, which are in the same cluster as the regarded neighborhood Riverdale and located in Toronto.



	City	Neighborhood	Bank	Bus Station	Farmers Market	Food Truck	Gym	Home Service	Moving Target	Park	Playground	Plaza
7	Toronto	Milliken, Agincourt North, Steeles East, L'Amo...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.250000	0.250000	0.000000
8	Toronto	Moore Park, Summerhill East	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.500000	0.000000
12	NYC	Riverdale	0.076923	0.076923	0.076923	0.076923	0.076923	0.076923	0.076923	0.230769	0.153846	0.076923
13	Toronto	Scarborough Village	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.333333	0.000000

The Riverdale neighborhood is characterized by those venue ccategories. One can see that Park and Playground have a higher rate than the others, which means those are more represented in Riverdale. According to the clustering there are several comparable neighborhoods to Riverdale in Toronto. Namely those are Moore Park, Summerhill East, Scarborough Village, Milliken, Agincourt North, Steeles East and more.

5. Discussion

The analysis gave several neighborhoods in Toronto which are comparable to Riverdale according to the venue categories which can be found there. These neighborhoods can be used as a first approach to look for livable neighborhood when moving from Riverdale, NYC to Toronto.

When taking a deeper look it appears that the similarity of the neighborhoods ground mainly in the similarity of two feature i.e. venue categories: Park and Playground. So one should make sure if these categories are the right reasons for living in this neighborhood. Otherwise there should be done a refinement regarding the features. Reasonable could be a weighting of venue categories by considering personal preferences.

6. Conclusion

In this report it was shown how geospatial data of Toronto and New York City can be used to find similarities between neighborhoods. Several similar neighborhoods for Riverdale were found in Toronto and could be considered if someone wants to move there from New York City. The same approach could also be used for other cities and in general could be further refined regarding the weighting of the venue categories.