Capstone Project: Segmenting and Clustering Neighborhoods in Toronto City

Pashkov Daniel

22.10.2020

Segmenting and Clustering Neighborhoods

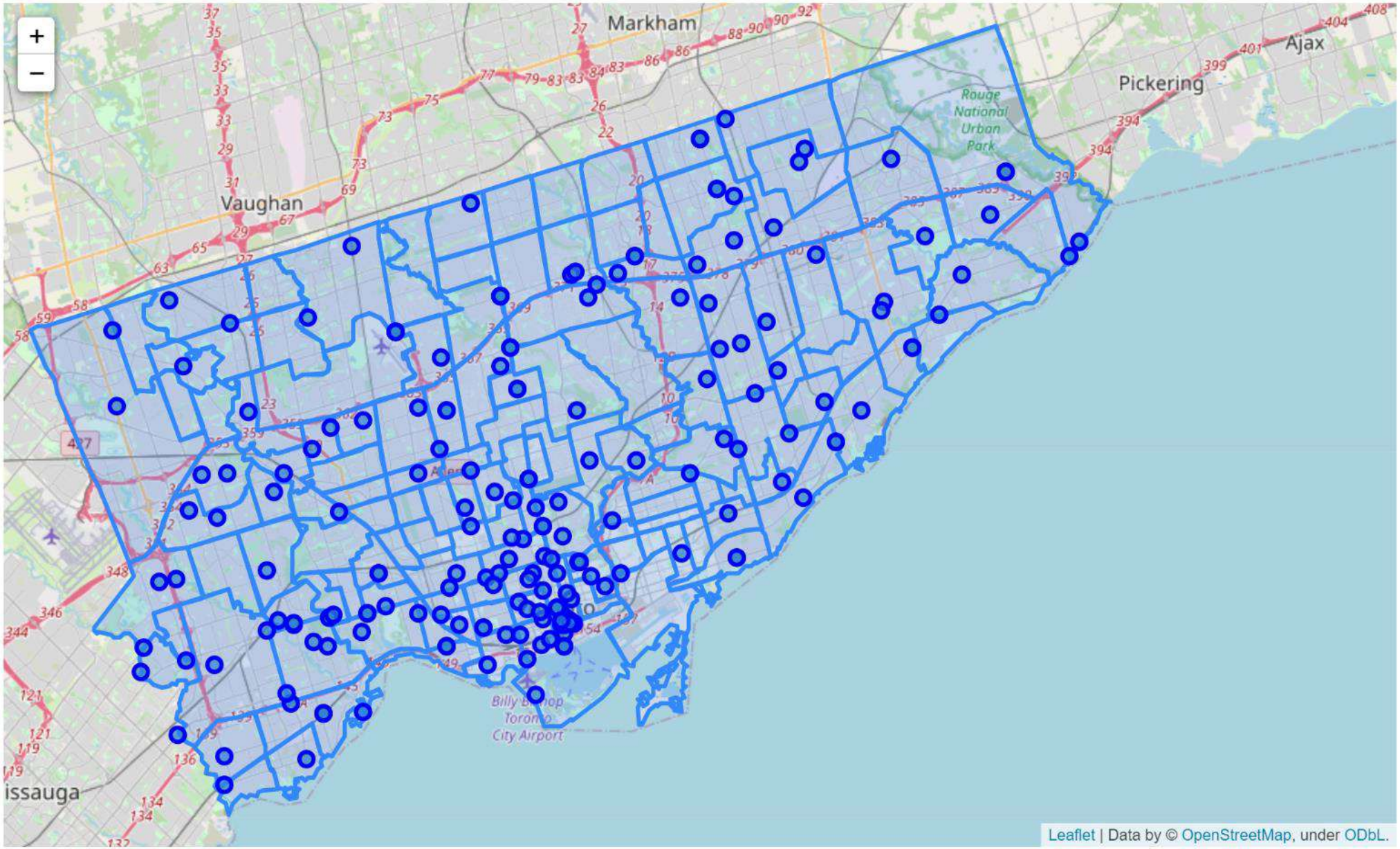
- The main purpose of this project is clusterisation, that dividing city on clusters, which often unequal to districts borders, based on most using places in that area.
- Who might be interested in this?
  - City government, which planning districts specialization or areas statistics.
  - Businessman, who decide open new spa or mall and etc.
  - Citizens, who, for example, want to buy new home.

Data cleaning and feature selection

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.761124	-79.324059
1	M4A	North York	Victoria Village	43.732658	-79.311189
2	M5A	Downtown Toronto	Regent Park	43.660706	-79.360457
3	M5A	Downtown Toronto	Harbourfront	43.640080	-79.380150
4	M6A	North York	Lawrence Manor	43.722079	-79.437507
...	...	...	...	...	...
193	M8Y	Etobicoke	Royal York South East	43.648183	-79.511296
194	M8Z	Etobicoke	Mimico NW	43.616677	-79.496805
195	M8Z	Etobicoke	The Queensway West	43.623618	-79.514764
196	M8Z	Etobicoke	South of Bloor	43.670489	-79.386465
197	M8Z	Etobicoke	Royal York South West	43.648183	-79.511296

198 rows × 5 columns

- For this project neighborhoods data was scraped from Wikipedia page using BeautifulSoup library. This data includes postal code for each borough.
- For exploring neighborhoods in this project was use Foursquare API.
- Further, get the coordinates for every neighborhood using Nominatim search engine and join this data with original dataset.
- After that, clean data includes 198 neighborhoods.



Neighborhoods in Toronto map. Straight line denotes districts borders.

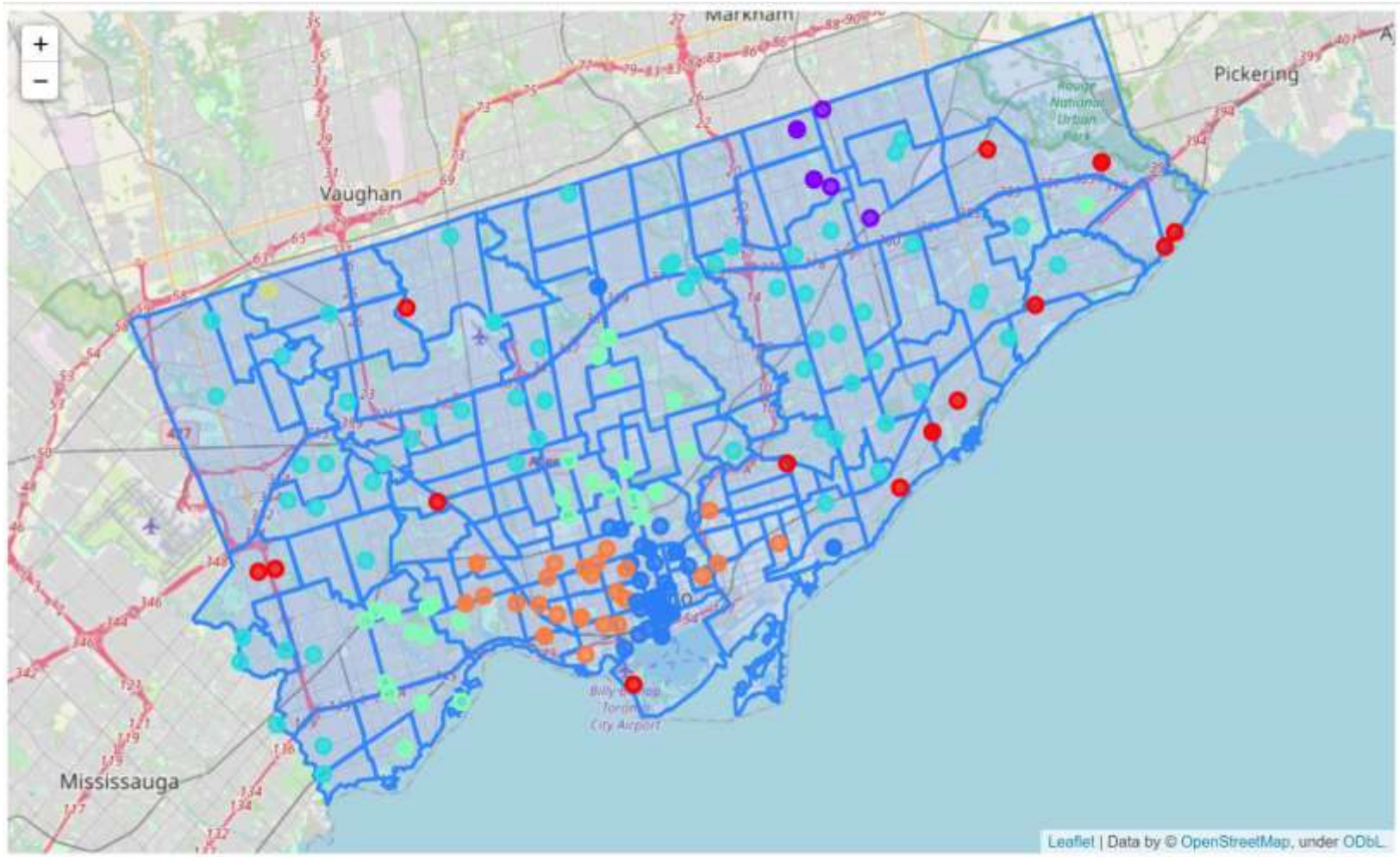
Exploratory Analysis

- To get most common venues for each neighborhoods use Foursquare API explore request for venues with radius, which was calculated in previous chapter (half of 3 kilometers).

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adelaide	Coffee Shop	Restaurant	Café	Hotel	Beer Bar	Japanese Restaurant	Gastropub	Thai Restaurant	Pizza Place	Plaza
1	Agincourt North	Indian Restaurant	Coffee Shop	Pizza Place	Vietnamese Restaurant	Sandwich Place	Chinese Restaurant	Bank	Park	Pharmacy	Bubble Tea Shop
2	Bathurst Quay	Park	Coffee Shop	Gym	Café	Restaurant	Yoga Studio	Italian Restaurant	Bakery	Seafood Restaurant	Dog Run
3	Bloordale Gardens	Coffee Shop	Sandwich Place	Beer Store	Bank	Gym	Pizza Place	Liquor Store	Clothing Store	Pharmacy	Convenience Store
4	Broadview North (Old East York)	Greek Restaurant	Park	Café	Italian Restaurant	Pub	Bakery	Coffee Shop	Pizza Place	Burger Joint	Flower Shop

After grouping, were creating the new dataframe with top 10 venues for each neighborhood.

Clustering using k-means algorithm



- For selecting the k-means parameters use the GridSearchCV library. After training, best models includes 7 clusters. In map each color of markers is different cluster.

Conclusion

- Using **Nominatim search engine**, were receiving coordinates for each neighborhoods and after that, using **Foursquare API**, for each neighborhoods were find **top 10 most common venues** within a radius.
- Further, top venues transfer to ***k*-means algorithm** as a features, and using **GridSearchCV** library, was find **7 clusters** with the best accuracy.
- After analyzing, each clusters was described and labeled.
- In result there are clusters:
  - 'Coasts Areas'
  - 'Chinatown'
  - 'Low Cost Residential Areas'
  - 'Middle Cost Residential Areas'
  - 'City Center'
  - 'Latino-Americans Area'
  - 'Center Residential Areas'