

Russian Language in Toronto

Sanzhar Seitzkassimov

August, 2020

1. Introduction

Canada is a great place for immigrants and it has the highest immigration rate among developed countries. That 'why a family of 4 people from Almaty, Kazakhstan is considering Toronto for immigration.

This family consists of father (58 year old) and mother (57 year old) with two young children 16 and 10 years old consequently.

The parents are very skilled workers, but not very good at English, so for the first time of adaption in foreign country, they would like to choose a neighbourhood with Russian and Kazakh (if any) speaking inhabitants and café/restaurants with Russian cuisine. Moreover, in the future parents are looking out for a place to open its own café.

Toronto's linguistic diversity is well-established and Russian language is popular in many neighbourhoods. So, we can create language map distribution across neighbourhoods to find the top one with Russian prevailing.

2. Data

To solve this problem, the following sources of data were used for analysis:

- List of Toronto neighborhoods in a geojson format which I found at Github: <https://github.com/adamw523/toronto-geojson/blob/master/simple.geojson>.
- Information about language profile of each Toronto neighborhoods at <https://www.toronto.com/> and <https://open.toronto.ca/>. I found there a csv file with each neighbourhood's profile and filtered out dataset to get the numbers of Russian language speakers.
- And, finally Foursquare API provided me with venues list to get the most common ones, including Russian cafes if any there.

There was a problem with data cleaning. Some neighbourhoods in geojson file and neighbourhoods in Toronto profile had differences, mainly due to changes in official names. Because of this I couldn't get all numbers. It required me to manually check and format to bring it accordingly to the names in geojson file.

3. Methodology

I used Github repository as a database and my main data frame consists of the following columns: *Neighbourhood*, *Number of Russian speaking people*, *Latitude* and *Longitude*.

As we can see from the table below only **Westminster-Branson** with its 5755 people is far cry from other top 5 neighbourhoods.

	Neighborhood	Number	Latitude	Longitude
0	Westminster-Branson	5755.0	43.770392	-79.442111
1	Newtonbrook West	2645.0	43.781663	-79.415981
2	Bathurst Manor	1545.0	43.755480	-79.438390
3	Willowdale East	1220.0	43.766694	-79.388044
4	Waterfront Communities-The Island	1150.0	43.635298	-79.394945

Table 1. Top-5 Russian speaking neighbourhoods

For data visualization purpose, I used **folium** library to create maps and **matplotlib** for plotting graphs. Below, you can see a map of Toronto city with its neighbourhoods (overall there are 140 of them) superimposed on it. To get such map, latitude and longitude values were utilized.

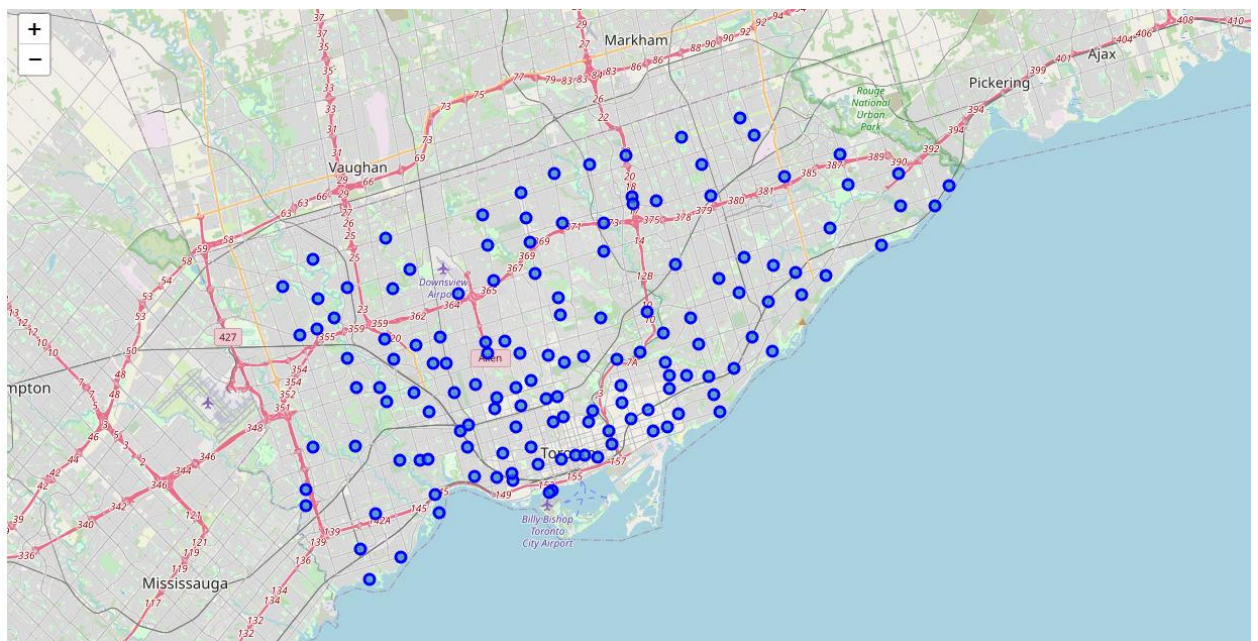


Figure 1. Toronto neighborhoods' map

Foursquare API helped me to get 100 nearby venues within 500 m. radius from each neighbourhood coordinates. The result you can see below. In general, Foursquare returned 2708 venues and there are 298 unique categories.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Westminster-Branson	43.770392	-79.442111	Express Pizza & Grill	43.773310	-79.442337	Pizza Place
1	Westminster-Branson	43.770392	-79.442111	Booster Juice	43.772600	-79.442400	Juice Bar
2	Westminster-Branson	43.770392	-79.442111	Tim Hortons	43.772945	-79.442289	Coffee Shop
3	Westminster-Branson	43.770392	-79.442111	Subway	43.773219	-79.441626	Sandwich Place
4	Westminster-Branson	43.770392	-79.442111	Tuong Phat Supermarket	43.773013	-79.442446	Grocery Store

Table 2. Each neighbourhood venue information

Little Portugal, University and Bay Street Corridor reached limit of 100 venues, but more than half of neighbourhoods have less than 20 venues. This number may depend on latitude and longitude of each neighbourhood. Geojson which I used for analysis has several coordinates for

We can observe from the graph above that the number 3 is optimal for our case.

Also I created a table which shows top 10 venue categories for each neighbourhood. After that, I merged this data frame with cluster labels.

	Neighborhood	Number	Latitude	Longitude	Cluster Labels	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	Westminster-Branson	5755.0	43.770392	-79.442111	0.0	Pizza Place	Coffee Shop	Bakery	Bus Line	Grocery Store	Sandwich Place	Gym	Juice Bar	Farm	Falafel Restaurant
1	Newtonbrook West	2645.0	43.781663	-79.415981	0.0	Korean Restaurant	Coffee Shop	Bubble Tea Shop	Dessert Shop	Playground	Burrito Place	Bank	Food Stand	Smoothie Shop	Salad Place
2	Bathurst Manor	1545.0	43.755480	-79.438390	0.0	Bank	Coffee Shop	Pharmacy	Chinese Restaurant	Middle Eastern Restaurant	Mobile Phone Shop	Shopping Mall	Supermarket	Sushi Restaurant	Sandwich Place
3	Willowdale East	1220.0	43.766694	-79.388044	0.0	Coffee Shop	Café	Spa	Clothing Store	Italian Restaurant	Chinese Restaurant	Shopping Mall	Burger Joint	Liquor Store	Supermarket
4	Waterfront Communities-The Island	1150.0	43.635298	-79.394945	0.0	Coffee Shop	Café	Park	Gym	Tunnel	Japanese Restaurant	Scenic Lookout	Sculpture Garden	Pharmacy	Dance Studio

Table 3. The most common venue categories with cluster labels

4. Results

Examining each cluster shows the following results:

Cluster 0 – popular places: coffee shops, café and Italian cuisine.

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Coffee Shop	23.0	17.0	13.0
Pizza Place	10.0	9.0	6.0
Café	8.0	8.0	4.0
Italian Restaurant	5.0	2.0	2.0
Bar	5.0	1.0	2.0
...
Taco Place	NaN	NaN	1.0
Tennis Court	NaN	NaN	1.0
Theater	NaN	NaN	1.0
Train Station	NaN	NaN	1.0
Vegetarian / Vegan Restaurant	NaN	NaN	1.0

106 rows × 3 columns

Cluster 1 – popular places: primarily park zones.

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Park	12.0	1.0	4.0
Electronics Store	1.0	NaN	NaN
Jewelry Store	1.0	NaN	NaN
Plaza	1.0	NaN	NaN
Stables	1.0	NaN	NaN
Theater	1.0	NaN	NaN
Women's Store	NaN	2.0	2.0
Home Service	NaN	2.0	NaN
Business Service	NaN	1.0	NaN
Convenience Store	NaN	1.0	NaN
Dim Sum Restaurant	NaN	1.0	NaN
Food Service	NaN	1.0	NaN
Grocery Store	NaN	1.0	NaN
Hockey Arena	NaN	1.0	NaN

Cluster 2 – popular places: small retail shops, women’s store and farm.

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Convenience Store	2.0	NaN	NaN
Women's Store	NaN	2.0	NaN
Farm	NaN	NaN	2.0

Additionally, I searched for venues with Russian restaurants or cafes as well (with 'Russian' in its name), but unsuccessfully. Perhaps there are some Russian venues but it will require another approach which is far away from my case.

As a result of K-means algorithm you can see a clustered map of Toronto below.

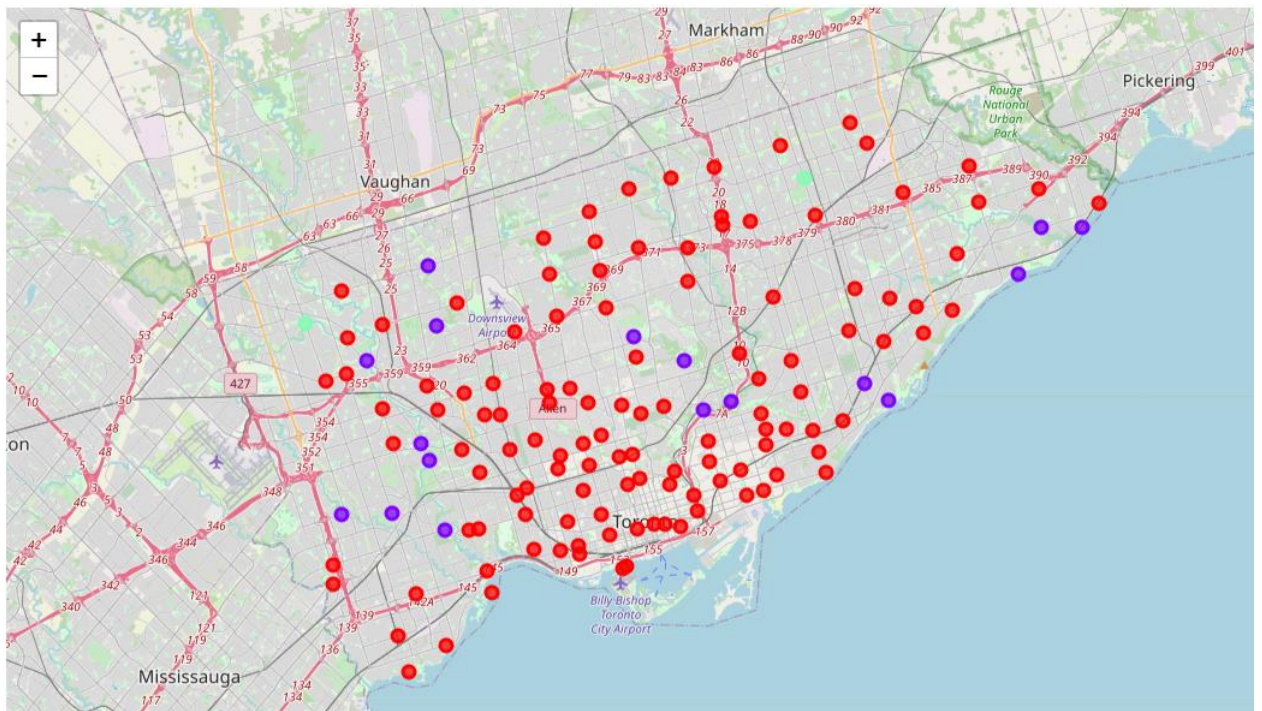


Figure 4. Toronto clusters

The main task was to find an area of language homogeneity for our family. I used choropleth map to see how many people who speak Russian are distributed on the city map. The result you can see below.

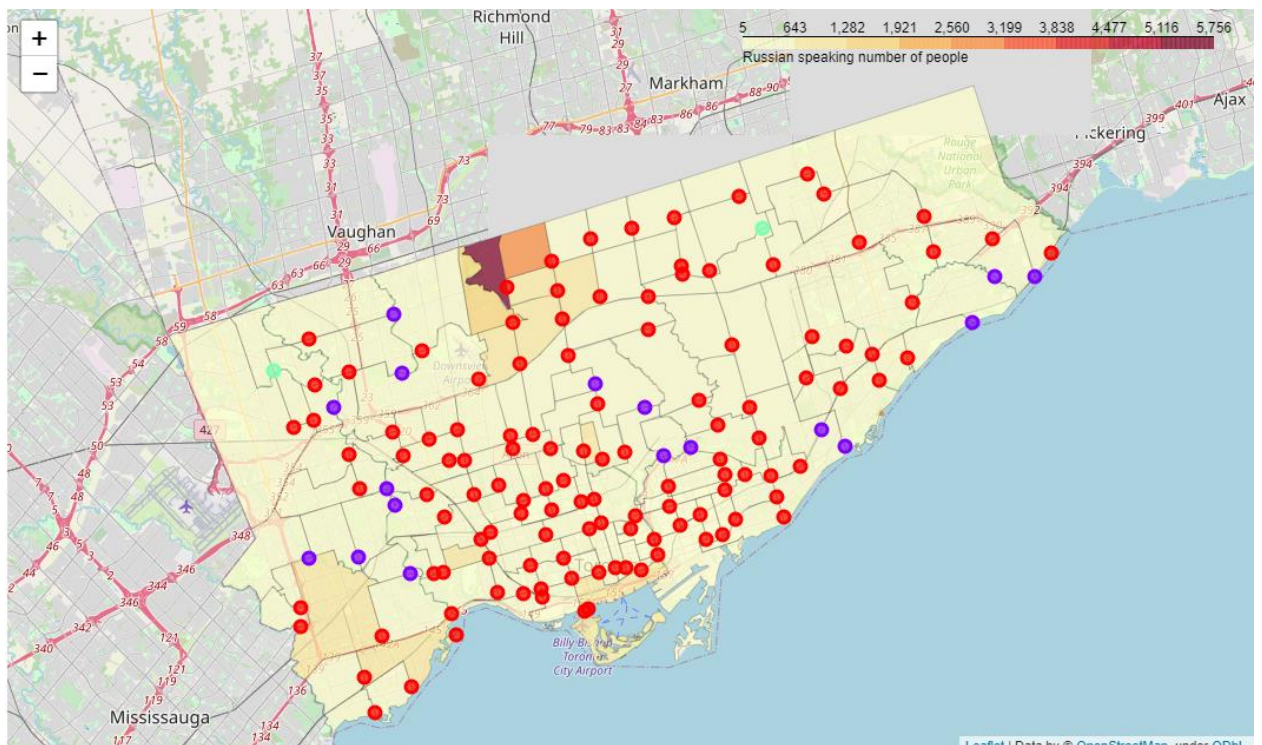


Figure 5. Toronto neighbourhoods by Russian language presence

5. Discussion

While English is the predominant language, Toronto is home to more than 140 languages. Such multi-linguistic diversity makes Toronto an attractive place for people all over the world. Moreover, every new coming citizen of this big city can find a suitable language area at first time of adaptation. For example, the Russian is one the top 15 non-official languages in the city and more than 35,000 people who list it as their Mother Tongue (2011 census).

In our case, given the age of parents it will be hard for them to learn new language. Taking this into account, clustering method was the best one to distinguish the right place to choose accommodation. According to the **Elbow method**, k equal to 3 was the optimal one.

For better analysis all neighbourhoods of Toronto were used. Besides districts coordinates I used census information including language profile. At the end I got 3 different clusters with coloured areas, where the darker ones imply the most populated by Russian speakers.

First and the most numerous one embrace all top Russian speaking neighbourhoods. As we can observe it is the northern part of the city (especially North York borough) where you likely hear Russian speech. Some area in the west and south of the city (close to the harbour) second place after North York. Cluster's another distinguishing feature is also multiple social venues like coffee shops, Italian café.

The second and third clusters are less populated by Russian speakers. However they seem to be family oriented, good for walking and entertaining. For example, common venues of the cluster 1 are parks, besides that other the most common venues you can find there stores necessary for home (electronics, home, food services) or have a good time with friends (hockey arena, theatre or stables).

6. Conclusion

Typically, families with children tend to move to the area which is away from noisy city centre but at the same time has all comforts nearby. In this regard cluster 1 will be the best one to recommend. However, given language barrier for elder parents and for the sake of their convenience cluster 0 (especially North York neighbourhoods) will be optimal one.

Such big decision as moving to another country should also cover other factors than only language preferences. This research takes those factors all else equal.

Finally, the language distribution map would be useful for those who target certain groups as their customer audience. For example, I didn't manage to find Russian restaurant in North York (the top Russian speaking area) and this will be a good food for thought to open it there.