

A description of the data and how it will be used to solve the problem.

Data sources

To address this problem, we first need data from different neighborhoods of the city of Toronto. We need the latest data on the restaurants and their geo-locations. We also need the fire-stations and their geo-locations as well. To get the data on restaurants first we obtain the data on the neighborhoods and their geo-locations for the city of Toronto. We used the Wikipedia page for the city of Toronto.

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

The geo-coordinates (latitude, longitude) data is been taken from source made available by the course instructors.

http://cocl.us/Geospatial_data

Next, by using the neighborhood data we obtain the latest data on restaurants and their geo-coordinates from Foursquare API. This task requires one to create an account as foursquare developer via the link:

<https://foursquare.com/developers/apps>

Getting the latest data is important for restaurants as they open and close daily.

Finally, the fire-stations and their geo-coordinates data is made available by the city of Toronto at their Open Data portal that can be accessed via weblink:

<https://portal0.cf.opendata.inter.sandbox-toronto.ca/dataset/fire-station-locations/>

Data cleaning

I started with the neighborhood data so to get the different neighborhoods in the city of Toronto we looked at the postal codes and grouped neighborhoods in terms of postal codes. This is done by scrapping through the webpage on Wikipedia and saving the data-table where information is provided for the neighborhoods, the postal code and the boroughs they belong to.

For details, please refer to following link to the jupyter notebook:

https://nbviewer.jupyter.org/github/ViditAg/Restaurant_fire_stations_clustering/blob/master/Capstone_Project_Get_neighborhoods_Toronto.ipynb

To this data-table we also add the information about the latitude and longitude geo-coordinates for each postal code.

Next step is to use these geo-coordinates of different neighborhoods to access the foursquare API and perform a search for 'query = restaurants' in the geo-location of each postal code. This search returns a json file containing details about the restaurants such as address, name, geo-coordinates etc. We extract name and geo-coordinates

only as per the requirement of current project. One must be careful while getting data from Foursquare as there is a daily limit on the number of calls that can be made.

For details, please refer to following link to the jupyter notebook:

https://nbviewer.jupyter.org/github/ViditAg/Restaurant_fire_stations_clustering/blob/master/Capstone_Project_Get_Restaurants_Toronto.ipynb

Finally, we get the data for fire-stations as their geo-locations from the Open data Portal for the city of Toronto. The geo-locations are provided in the UTM 6-degree coordinates (x,y) format. We convert these to latitude and longitude so that the clustering and comparison with restaurant data can be performed.

Please refer to following link to the jupyter notebook:

https://nbviewer.jupyter.org/github/ViditAg/Restaurant_fire_stations_clustering/blob/master/Capstone_Project_Get_Firestations_Toronto.ipynb

All the 3 datasets are saved in csv format and then finally loaded into data-frame to perform final analysis.