The Battle of Neighbourhoods

1. Introduction/Business Problem

There are many people who are working in various cities (say New York and Toronto) across the world. Let's say a person got a job offer from a great company with great career prospects in other city or other borough of his/her current city. If person accept the job offer then he/she has to shift to new location. I think a person would love to shift a location which is exactly or almost similar to his/her last location because he/she loves the great amenities and other types of venues that exist in his/her current neighbourhood like school, gym, swimming pool, Amusement park, restaurants, coffee-shops, spencer etc. So I'll find out what are borough-neighbourhoods are very similar to give current location. A person can shift within the city or from one city to another city. If person is shifting within the city then he/she must be changing the borough and my task would be finding neighbourhoods in that borough which are very similar to his/her last location. If a person is shifting from one city to another city then my task would be finding the similar neighbourhoods.

2. Data

2.1 New York Data

I'll be using the borough and neighbourhood data of two cities. My first city would be New York and data of this can be downloaded from this link (https://cocl.us/new_york_dataset). This data would be json format and it can be very easily transformed into pandas dataframe. I would take only those are important for me like borough, neighbourhood, latitude and longitude. For more information of this data you can visit this site (https://geo.nyu.edu/catalog/nyu_2451_34572).

	Borough	Neighbourhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Figure - 1: New York Data

2.2 Toronto Data

Second data of Toronto city cannot be directly downloaded. Postal Code, borough and neighbourhood can be scrapped from wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). There are some rows where "Not assigned" is written. Drop the rows where borough is "Not assigned" (only). If Neighbouthood is 'Not assigned' but Borough is assigned then make corresponding Borough as Neighbourhood. Merge the rows if Postal Code and Borough of two or more rows same and merged neighbourhood will be separated by comma ",". Latitude and longitude information can be downloaded from here

(https://cocl.us/Geospatial_data). Finally, this data and Toronto data can be merged together. Once both data are ready we'll use Foursquare API to get the venues near each neighbourhood. To get the nearby venues data we must have Foursquare account and app which will give the credentials.

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Figure - 2: New York Data

3. Methodology

After getting above data, we'll find the nearby venues to each borough and neighbourhood pair in both cities. Foursquare API will be used to get the nearby venue. To use the Foursquare API we'll need the developer account and an app which will give the client id and client secrets. Both cities may yield in different number of venue categories, but we'll take only common venue categories. Now we can find a borough and neighbourhood of a city is how much like the boroughs of another city. In other words, we can find top similar boroughs and neighbourhoods in another city. We will use cosine similarity to find the similarity between the two cities.