IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

APPLIED DATA SCIENCE CAPSTONE PROJECT

THE BATTLE OF NEIGHBORHOODS

RISHI SAXENA

INTRODUCTION

- In the year 2019, approximately 36 million travelers crossed the U.S.-Canadian border.
- It is expected that these numbers would increase exponentially in the coming years following the 2020 bottleneck. This would include travel for business, employment, immigration etc.
- A key challenge of this travel is identifying neighborhoods that suit one's preferences. This program aims to provide a high-level solution to this problem by employing the use of Data Science.

- Key features of this program:
  - Compares neighborhoods of the city of Toronto with that of New York City and identifies similarity.
  - Uses Cosine Similarity to compute similarity index.
  - Is very flexible and can be used to support various different purposes with a few changes in the source code.
  - Is FREE!

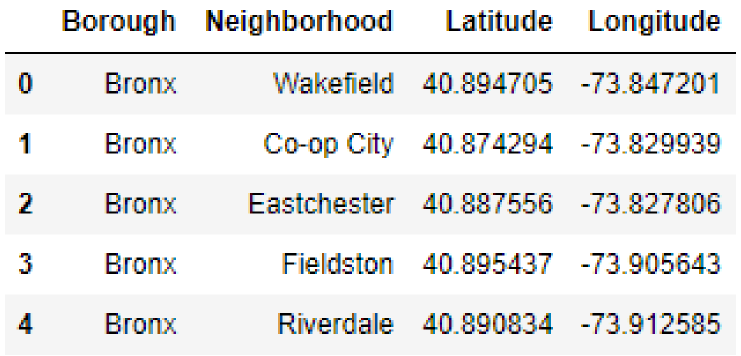
DATA

- Adopts the use of the industry-leading location tracking platform service, the FourSquare API to fetch relevant data for a given neighborhood.
- FourSquare API requires geospatial data – the coordinates for each location. This location data for the city of Toronto and New York are obtained from the following sources –
  - New York University – Spatial Data Repository (https://geo.nyu.edu/)
  - Wikipedia (https://Wikipedia.org)
- This raw data is processed by the program to generate a data frame that can be used with FourSquare API. However, this process is not very straightforward.
- The detailed process is explained in the next slide.

DATA – PROCESSING FOR FOURSQUARE API

- Data obtained from the two sources listed above needs to be processed before FourSquare API can be called.
- New York city raw data is directly obtained from the following link –

  https://cocl.us/new_york_dataset
- Unfortunately, Toronto data is not so easily obtained. The relevant data is obtained from Wikipedia using the following links –
  - Postal Code – Neighborhood Data (https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&direction=prev&oldid=926287641)
  - Postal Code – Geospatial Coordinate Data (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
  - This data is scrapped from Wikipedia using Beautiful Soup and processed using Pandas library.
- Once fully processed, the data resembles –



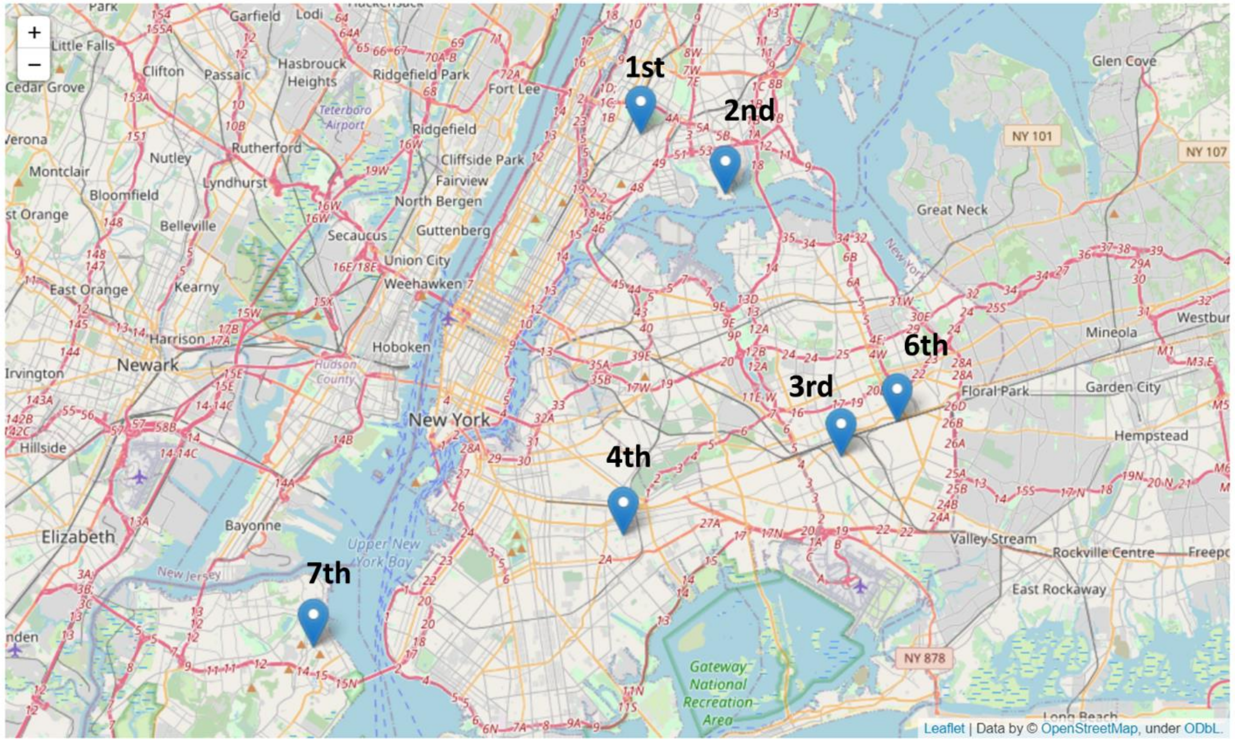
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

RESULTS

- This program can be used to identify similar neighborhoods for any pair in New York and Toronto. Thus, this program has to demonstrate full functionality in both directions of application.
- In order to demonstrate results, a random neighborhood is first selected in the city of Toronto. The first analysis uses this neighborhood as a source and identifies the most similar neighborhoods in New York city.
- This similarity is calculated using the cosine similarity index. Additionally, the similarity is based on the number and quality of the venues available locally at a given neighborhoods and compares similarity to the source neighborhood in the city of Toronto.
- These results are presented using a high fidelity data visualization python library called Folium.

RESULTS – FROM TORONTO TO NEW YORK

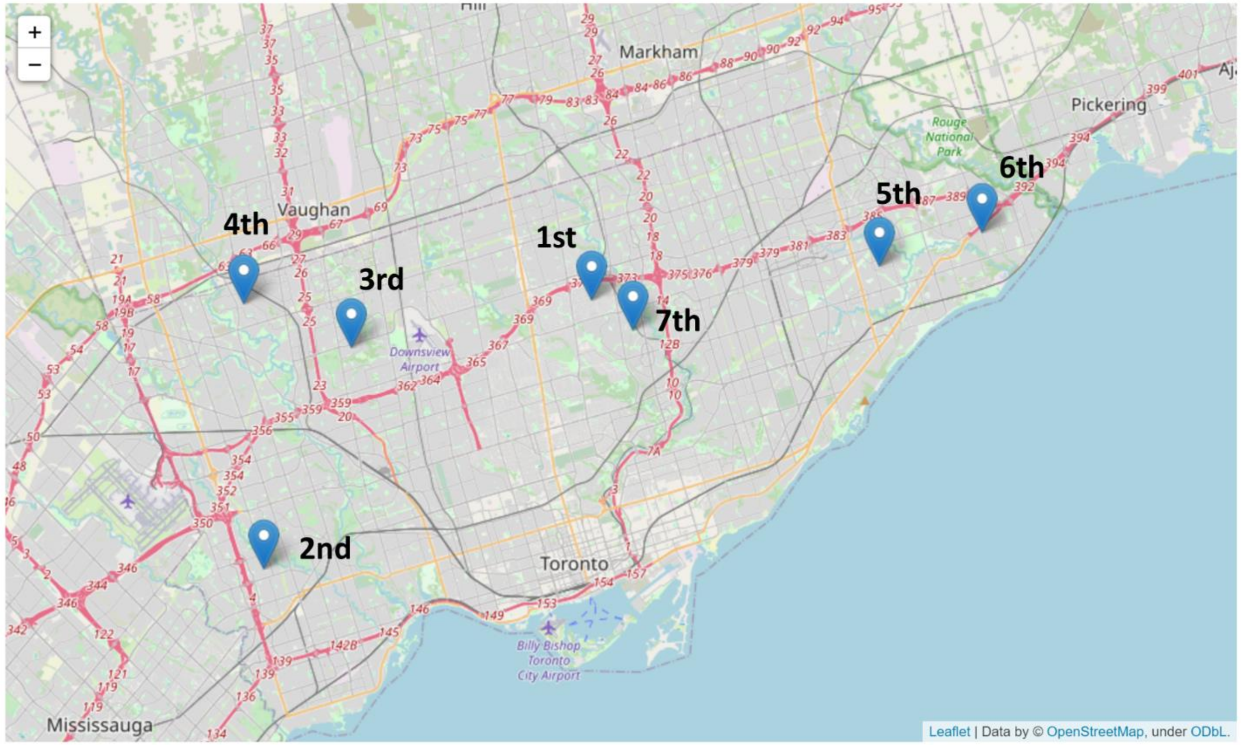
- The source neighborhood in Toronto is chosen as Alderwood/Long Branch, Etobicoke. The program gives the following results as the top seven most similar neighborhoods in New York city –



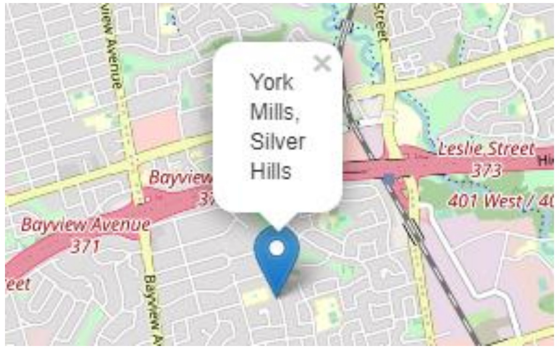
- Upon clicking on the pinhead marked "1st", the most similar neighborhood in New York city is determined to be Claremont Village.

RESULTS – FROM NEW YORK TO TORONTO

- The source neighborhood in New York city is chosen as Riverdale, Bronx. The program gives the following results as the top seven most similar neighborhoods in Toronto –



- Upon clicking on the pinhead marked "1st", the most similar neighborhood in the city of Toronto is determined to be York Mills, Silver Hills.



DISCUSSION

- This program provides an excellent first-look at the similarities between neighborhoods between New York City and the city of Toronto.
- This program is flexible and can be used to support the following with some edits in the source code –
  - Determining the most similar neighborhoods within the same city, whether New York or Toronto.
  - Explore different pairs of cities around the world.
  - Can be tailored to compare cities and counties in the future.
- As this program uses publicly available data, the quality of this data would directly affect the results given by this program. Also, geospatial data may not be available for most locations.
- This program only provides a first-look in comparing neighborhoods and should only be used as a guide and its results should be supplemented with further research.