IBM APPLIED DATA SCIENCE CAPSTONE

Gustavo Sousa

INTRODUCTION

The project seeks to return results of a research to a fictional coffee store company that have plans of expanding in New York and Toronto, considering opening near subway stations.

BUSINESS PROBLEM

The moment of choosing and defining the place to begin a new business activity is crutial. An infinity of variables – economic, cultural, social – must be considered and carefully assessed.

I decided to apply this study case considering two extremely popular kinds of venues: coffee shops and subway stations. I exercised the imagination of situation where, beyond the benefit of a huge flow of people nearby a subway station, our fictional entrepreneur is intending to use the subway stations on a two-month advertising campaign with flyers and promotional pamphlets.

The coffee shop franchising intends to open two new venues: one in New York and the other in Toronto. The best neighborhoods in each city to satisfy the business needs and objectives will be analyzed and assessed.

This research is also appliable in similar problems, considering crowded places and venues that can make use of this to gather clients.

DATA

The data sources used on this project are the datasets of neighborhoods from New Your and Toronto, and a sample list of venues distributed on the respective neighborhoods above. The venues' information is obtained via Foursquare, using the cities' datasets as input information.

New York and Toronto Data

The data related to the New York neighborhoods was obtained on the New York University Spatial Data Repository webpage. A 2014 dataset is available for researchers and can be downloaded using the link https://geo.nyu.edu/catalog/nyu_2451_34572 .

The analyzed data was converted to a dataset like the following (just the top ten rows):

	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
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446

The dataset has 306 neighborhoods distributed within 5 boroughs.

In turn, Toronto's neighborhoods data was obtained on the Wikipedia related webpage. The link is https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M .

In this case, it was necessary to extract the data using web scraping techniques and only after that convert it to the dataset. On this specific case, the neighborhoods are listed without coordinates' information, which lead us to a further necessary step to capture these coordinates.

For that, we can use Geopy, a Python client for several geolocation web service. The neighborhood postal code is used as input for obtaining the coordinate for that postal code.

The final analyzed and merged data was converted to a dataset like the following (just the top ten rows):

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern / Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek	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
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	Kennedy Park / Ionview / East Birchmount Park	43.727929	-79.262029
7	M1L	Scarborough	Golden Mile / Clairlea / Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffside / Cliffcrest / Scarborough Village W...	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff / Cliffside West	43.692657	-79.264848

The dataset has 103 neighborhoods distributed within 10 boroughs.

To complete this specific phase, it is necessary to capture the coordinates (latitude and longitude) of New York City and Toronto. For that, we again use Geopy, but now using the name of the city as input. The main coordinates of the city are returned.

These coordinates will be used to visualize the neighborhoods distributed around each city's main coordinates.

FOURSQUARE DATA

Foursquare service provides access to a huge amount of data related to points of interest, named by venues, in almost every city around the globe. A big part of the information provided (and captured) by the service is collaborative, what increases the database day by day.

On this case study we will capture the venues related to each neighborhood of our analyzed cities. After that it will be important to rank the most common venues nearby each neighborhood and proceed with the main purpose of our research.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Allerton	Pizza Place	Deli / Bodega	Cosmetics Shop	Department Store	Supermarket
1	Annadale	Pub	Restaurant	Diner	Pizza Place	Train Station
2	Arden Heights	Pizza Place	Pharmacy	Pool	Playground	Coffee Shop
3	Arlington	Bus Stop	Intersection	American Restaurant	Deli / Bodega	Grocery Store
4	Arrochar	Bus Stop	Italian Restaurant	Deli / Bodega	Cosmetics Shop	Athletics & Sports
5	Arverne	Surf Spot	Sandwich Place	Metro Station	Bus Stop	Playground
6	Astoria	Middle Eastern Restaurant	Bar	Pizza Place	Mediterranean Restaurant	Greek Restaurant
7	Astoria Heights	Chinese Restaurant	Plaza	Laundromat	Bakery	Hostel
8	Auburndale	Furniture / Home Store	Supermarket	Pet Store	Pharmacy	Toy / Game Store
9	Bath Beach	Chinese Restaurant	Pharmacy	Gas Station	Sushi Restaurant	Italian Restaurant

First 10 neighborhoods of New York and their five most common venues

	Neighborhood	1st Most Common	2nd Most Common	3rd Most Common	4th Most Common	5th Most Common
0	Agincourt	Latin American Restaurant	Skating Rink	Lounge	Breakfast Spot	Women's Store
1	Alderwood / Long Branch	Pizza Place	Pharmacy	Skating Rink	Pool	Sandwich Place
2	Bathurst Manor / Wilson Heights / Downsview No...	Coffee Shop	Bank	Gift Shop	Pizza Place	Sandwich Place
3	Bayview Village	Japanese Restaurant	Chinese Restaurant	Bank	Café	Electronics Store
4	Bedford Park / Lawrence Manor East	Sandwich Place	Italian Restaurant	Restaurant	Coffee Shop	Women's Store
5	Berczy Park	Coffee Shop	Cocktail Bar	Bakery	Cheese Shop	Café
6	Birch Cliff / Cliffside West	College Stadium	Skating Rink	General Entertainment	Café	Empanada Restaurant
7	Brockton / Parkdale Village / Exhibition Place	Café	Breakfast Spot	Coffee Shop	Nightclub	Burrito Place
8	Business reply mail Processing CentrE	Gym / Fitness Center	Spa	Auto Workshop	Brewery	Burrito Place
9	CN Tower / King and Spadina / Railway Lands / ...	Airport Service	Airport Lounge	Airport Terminal	Sculpture Garden	Harbor / Marina

First 10 neighborhoods of Toronto and their five most common venues

[updated until week 4]