Comparing North America's Biggest Cities

Roberto Cabrer

1 Introduction

Data Science techniques are becoming ubiquitous because of their potential to extract 'hidden' patterns in data. As we've seen in the courses in this specialization, diverse computational abilities are needed in order to extract valuable information from data.

As now is our turn to propose an idea to use the acquired knowledge, for my Capstone project I'll use Foursquare location data from the largest city in each of the top 3 most populous countries in North America, to see if there are similarities among them. The cities to be compared are Mexico City in Mexico, New York in the United States and Toronto in Canada.

With the location (longitude and latitude) of the neighborhoods in this cities, we will extract venue information using Foursquare, that will be used as features in the clustering technique. The kind of information to be extracted is that of the presence of different business included in the Foursquare City Guide (`https://foursquare.com/city-guide`). Once the neighborhoods are clustered through the unsupervised machine learning technique, we will try to find similarities and dissimilarities between cities.

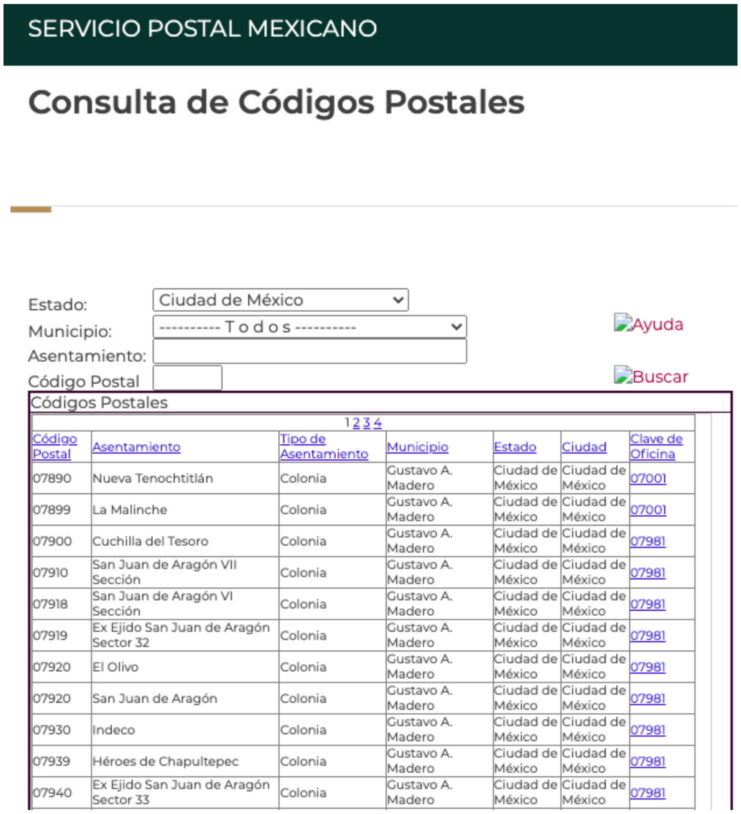
From the clustering of neighborhoods and comparison of these cities, huge amount of information can be extracted, we will focus on the expansion of a business or a brand from one of the cities to the other. Obtained insights can help stakeholders decide wether if it is feasible to expand a known successful business into another city, trying to reduce risks by knowing if the cities are similar or not.

As the data is inherently related to its location, the results from this experiment can also give valuable information about the business offerings by neighborhood that can be used, for example, to decide where to locate a physical store in the case this is needed.

2 Data

The data I'll be needing is the location of neighborhoods in each of the cities, that is latitudes and longitudes, from where we will extract the venue information using Foursquare.

For this project I will use datasets from Mexico City, New York and Toronto. Datasets from Toronto and New York have been used previously in other labs,


SERVICIO POSTAL MEXICANO

Consulta de Códigos Postales

Estado: Ciudad de México

Municipio: ---------- T o d o s ----------

Asentamiento:

Código Postal

Códigos Postales

1 2 3 4

Código Postal	Asentamiento	Tipo de Asentamiento	Municipio	Estado	Ciudad	Clave de Oficina
07890	Nueva Tenochtitlán	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07001
07899	La Malinche	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07001
07900	Cuchilla del Tesoro	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07981
07910	San Juan de Aragón VII Sección	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07981
07918	San Juan de Aragón VI Sección	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07981
07919	Ex Ejido San Juan de Aragón Sector 32	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07981
07920	El Olivo	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07981
07920	San Juan de Aragón	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07981
07930	Indeco	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07981
07939	Héroes de Chapultepec	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07981
07940	Ex Ejido San Juan de Aragón Sector 33	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07981

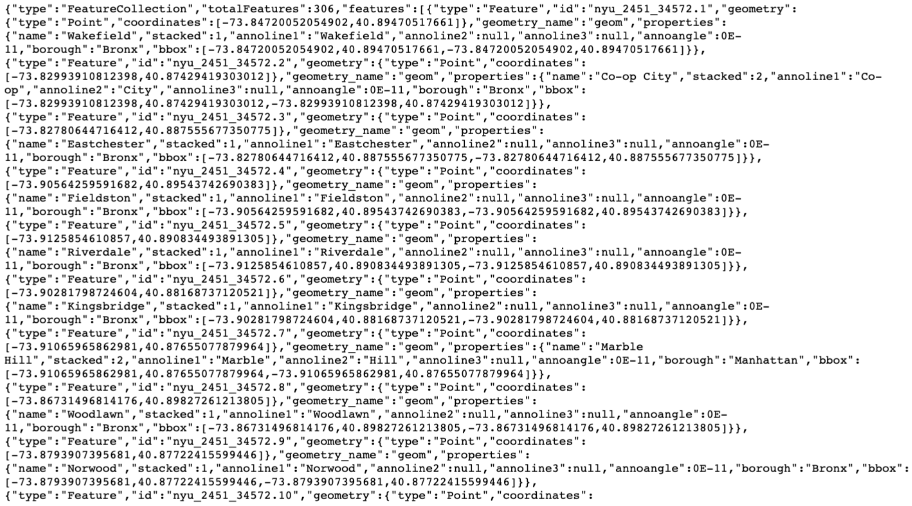
Figure 1: Mexico City's data and the query used to extract the Postal Code data for Mexico City. The website belongs to the Mexican Postal Service.

besides these two datasets I will include data from Mexico City obtained from an HTML source. As these datasets come from different sources they will need slightly different cleansing procedures.

2.1 Mexico City data

The dataset for Mexico City is extracted from the Mexican Postal Service website (found at the url `https://www.correosdemexico.gob.mx/SSLServicios/ConsultaCP/Descarga.aspx`). This dataset contains Postal Codes that will help us retrieve location data for each neighborhood in Mexico City using the geocoder package.

The dataset comes as a HTML file with Borough, Postal Code and Neighborhood information as shown in Figure 1. Other information is included in the file, nonetheless it does not add value so it won't be used.



```
{"type":"FeatureCollection","totalFeatures":306,"features":[{"type":"Feature","id":"nyu_2451_34572.1","geometry":
{"type":"Point","coordinates":[-73.84720052054902,40.89470517661]},"geometry_name":"geom","properties":
{"name":"Wakefield","stacked":1,"annoline1":"Wakefield","annoline2":null,"annoline3":null,"annoangle":0E-
11,"borough":"Bronx","bbox":[-73.84720052054902,40.89470517661,-73.84720052054902,40.89470517661]}},
{"type":"Feature","id":"nyu_2451_34572.2","geometry":{"type":"Point","coordinates":
[-73.82993910812398,40.87429419303012]},"geometry_name":"geom","properties":{"name":"Co-op City","stacked":2,"annoline1":"Co-
op","annoline2":"City","annoline3":null,"annoangle":0E-11,"borough":"Bronx","bbox":
[-73.82993910812398,40.87429419303012,-73.82993910812398,40.87429419303012]}},
{"type":"Feature","id":"nyu_2451_34572.3","geometry":{"type":"Point","coordinates":
[-73.82780644716412,40.887555677350775]},"geometry_name":"geom","properties":
{"name":"Eastchester","stacked":1,"annoline1":"Eastchester","annoline2":null,"annoline3":null,"annoangle":0E-
11,"borough":"Bronx","bbox":[-73.82780644716412,40.887555677350775,-73.82780644716412,40.887555677350775]}},
{"type":"Feature","id":"nyu_2451_34572.4","geometry":{"type":"Point","coordinates":
[-73.90564259591682,40.89543742690383]},"geometry_name":"geom","properties":
{"name":"Fieldston","stacked":1,"annoline1":"Fieldston","annoline2":null,"annoline3":null,"annoangle":0E-
11,"borough":"Bronx","bbox":[-73.90564259591682,40.89543742690383,-73.90564259591682,40.89543742690383]}},
{"type":"Feature","id":"nyu_2451_34572.5","geometry":{"type":"Point","coordinates":
[-73.9125854610857,40.890834493891305]},"geometry_name":"geom","properties":
{"name":"Riverdale","stacked":1,"annoline1":"Riverdale","annoline2":null,"annoline3":null,"annoangle":0E-
11,"borough":"Bronx","bbox":[-73.9125854610857,40.890834493891305,-73.9125854610857,40.890834493891305]}},
{"type":"Feature","id":"nyu_2451_34572.6","geometry":{"type":"Point","coordinates":
[-73.90281798724604,40.88168737120521]},"geometry_name":"geom","properties":
{"name":"Kingsbridge","stacked":1,"annoline1":"Kingsbridge","annoline2":null,"annoline3":null,"annoangle":0E-
11,"borough":"Bronx","bbox":[-73.90281798724604,40.88168737120521,-73.90281798724604,40.88168737120521]}},
{"type":"Feature","id":"nyu_2451_34572.7","geometry":{"type":"Point","coordinates":
[-73.91065965862981,40.87655077879964]},"geometry_name":"geom","properties":{"name":"Marble
Hill","stacked":2,"annoline1":"Marble","annoline2":"Hill","annoline3":null,"annoangle":0E-11,"borough":"Manhattan","bbox":
[-73.91065965862981,40.87655077879964,-73.91065965862981,40.87655077879964]}},
{"type":"Feature","id":"nyu_2451_34572.8","geometry":{"type":"Point","coordinates":
[-73.86731496814176,40.89827261213805]},"geometry_name":"geom","properties":
{"name":"Woodlawn","stacked":1,"annoline1":"Woodlawn","annoline2":null,"annoline3":null,"annoangle":0E-
11,"borough":"Bronx","bbox":[-73.86731496814176,40.89827261213805,-73.86731496814176,40.89827261213805]}},
{"type":"Feature","id":"nyu_2451_34572.9","geometry":{"type":"Point","coordinates":
[-73.8793907395681,40.87722415599446]},"geometry_name":"geom","properties":
{"name":"Norwood","stacked":1,"annoline1":"Norwood","annoline2":null,"annoline3":null,"annoangle":0E-11,"borough":"Bronx","bbox":
[-73.8793907395681,40.87722415599446,-73.8793907395681,40.87722415599446]}},
{"type":"Feature","id":"nyu_2451_34572.10","geometry":{"type":"Point","coordinates":
```

Figure 2: For New York City if the URL is accessed, the json file will pop-up presenting this structure.

2.2 New York City data

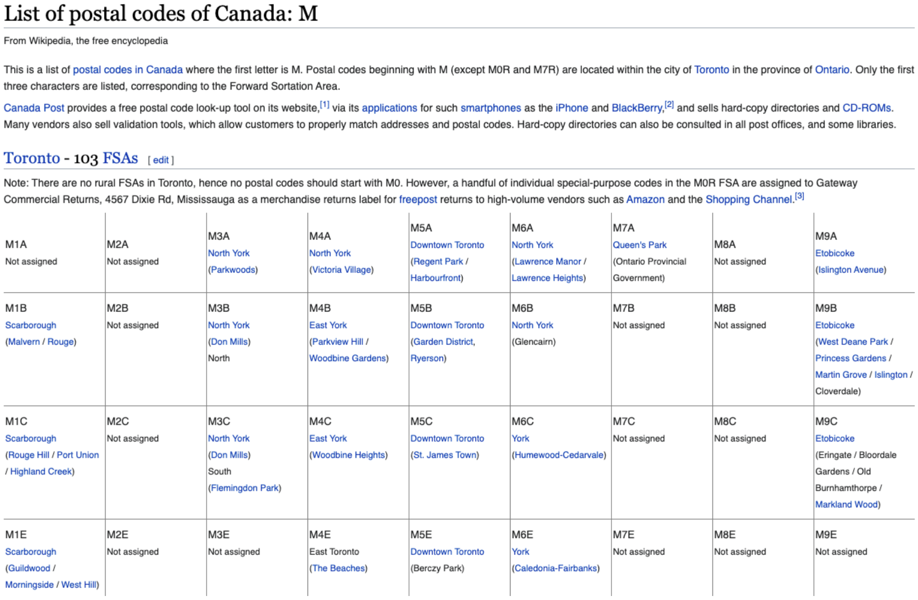
The data of New York City is stored in the IBM cloud and is requested from the URL `https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json`. The data comes in the form of a json file (Figure 2). In this file the neighborhood information with latitudes and longitudes can be found to create the desired dataframes.

2.3 City of Toronto data

Toronto's data is extracted from an article in wikipedia. The table contains Postal Codes, Borough name, and the Neighborhoods present in each Borough (Figure 3). This data can be found at `https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M`. and is extracted using Beautiful Soup, a package useful to parse data contained in HTML format to Python dictionaries that later can be used to create Pandas DataFrames.

As with the Mexico City data, the information extracted does not contain latitudes and longitudes for the neighborhoods, these have to be extracted through the postal codes using the geocoder package.

The final desired dataframes will contain columns 'Borough', 'Neighborhood', 'Latitude' and 'Longitude' as this is the information needed to do requests to the Foursquare API.



List of postal codes of Canada: M

From Wikipedia, the free encyclopedia

This is a list of postal codes in Canada where the first letter is M. Postal codes beginning with M (except M0R and M7R) are located within the city of Toronto in the province of Ontario. Only the first three characters are listed, corresponding to the Forward Sortation Area.

Canada Post provides a free postal code look-up tool on its website,[1] via its applications for such smartphones as the iPhone and BlackBerry,[2] and sells hard-copy directories and CD-ROMs. Many vendors also sell validation tools, which allow customers to properly match addresses and postal codes. Hard-copy directories can also be consulted in all post offices, and some libraries.

Toronto - 103 FSAs [edit]

Note: There are no rural FSAs in Toronto, hence no postal codes should start with M0. However, a handful of individual special-purpose codes in the M0R FSA are assigned to Gateway Commercial Returns, 4567 Dixie Rd, Mississauga as a merchandise returns label for freepost returns to high-volume vendors such as Amazon and the Shopping Channel.[3]

M1A Not assigned	M2A Not assigned	M3A North York (Parkwoods)	M4A North York (Victoria Village)	M5A Downtown Toronto (Regent Park / Harbourfront)	M6A North York (Lawrence Manor / Lawrence Heights)	M7A Queen's Park (Ontario Provincial Government)	M8A Not assigned	M9A Etobicoke (Islington Avenue)
M1B Scarborough (Malvern / Rouge)	M2B Not assigned	M3B North York (Don Mills) North	M4B East York (Parkview Hill / Woodbine Gardens)	M5B Downtown Toronto (Garden District, Ryerson)	M6B North York (Glencairn)	M7B Not assigned	M8B Not assigned	M9B Etobicoke (West Deane Park / Princess Gardens / Martin Grove / Islington / Cloverdale)
M1C Scarborough (Rouge Hill / Port Union / Highland Creek)	M2C Not assigned	M3C North York (Don Mills) South (Flemingdon Park)	M4C East York (Woodbine Heights)	M5C Downtown Toronto (St. James Town)	M6C York (Humewood-Cedarvale)	M7C Not assigned	M8C Not assigned	M9C Etobicoke (Eringate / Bloordale Gardens / Old Burnhamthorpe / Markland Wood)
M1E Scarborough (Guildwood / Morningside / West Hill)	M2E Not assigned	M3E Not assigned	M4E East Toronto (The Beaches)	M5E Downtown Toronto (Berczy Park)	M6E York (Caledonia-Fairbanks)	M7E Not assigned	M8E Not assigned	M9E Not assigned

Figure 3: Table present at the mentioned wikipedia URL. From here postal code information is extracted.