Data Science Capstone

Comparison of Neighborhoods in Toronto and New York

Table of Contents

1 Introduction

1.1 Background

New York and Toronto are very diverse cities with a great variety of restaurants, multicultural food supply and recreational venues. Each of these cities has an interest in attracting both businesses and highly qualified workforce.

1.2 Problem and Interest

Suppose you want to develop your career and are thinking of moving from New York to Toronto (or vice versa).

When looking for a place to live, you might want to move to a similar neighborhood that you already know and like from the city you are moving from.

How similar are the neighbourhoods in each city, in terms of restaurants and recreational opportunities and what exactly makes them similar?

2 Data acquisition and cleaning

2.1 Data sources

For this analysis, I use freely available geographical data on the location of neighbourhoods in New York and Toronto.I also use the freely available Yelp Fusion API. Yelp provides a commercial rating system mainly for restaurants and some recreational venues.

IMPORTANT NOTE

Due to a well known looping issue of the Foursquare sign up procedure (see Coursera Forum), I wasn't able to create a developer account to use the Foursquare API in this exercise. To proceed with the course, I obtained a developer account on Yelp and used the Yelp Yelp Fusion API instead.

The following data sources are used:

- New York boroughs and neighborhoods including latitude and longitude of each neighborhood (provided as JSON-file from the Coursera IBM Data Science Course Material)
  https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json
- Toronto boroughs and neighborhoods including postal codes (retrieved from Wikipedia on 27 Dec 2020)
  https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Geocoder using the ERSI ArcGIS API
  https://developers.arcgis.com/rest/geocode/api-reference/overview-world-geocoding-service.htm
- Yelp Fusion API to retrieve information about businesses by location
  https://www.yelp.com/developers/documentation/v3

2.2 Data cleaning

There was only some minor data cleaning necessary:

- From the initial Toronto boroughs and neighborhoods data, only entries with a borough assigned to the postalcode were used.
- PO Boxes in the Toronto dataset where dropped from the dataset.
- Neighborhoods which didn't return any venues from Yelp where removed from the analysis.
- The dataset was also reduced to parts of central Toronto and Manhattan.

After retrieval of venue information based on location of the neighborhood from Yelp, the dataset was built from the following columns:

Column	Type	Description
City	String	New York or Toronto
Borough	String	
Neighborhood	String	
Neighborhood Latitude	Float	location of the Neighborhood
Neighborhood Longitude	Float	location of the Neighborhood
Venue	String	Name of the venue
Venue Latitude	Float	location of the venue
Venue Longitude	Float	location of the venue
Venue Category	String	category of the venue (like type of cuisine etc)
Neighborhood_long	String	Concatenated string of city + borough + neighborhood

2.3 Feature selection

The business data I retrieved from Yelp for each neighborhood came with a business category for each venue. In the combined dataset of Toronto and New York where 287 unique categories.

The categories where one hot encoded, resulting in 287 separate columns.

After that I grouped rows by neighborhood, calculated the mean of the frequency of occurrence of each category and finally sorted the venues in descending order of category occurrence. The resulting dataframe is the basis for a clustering of venues based on categories with K-Means.

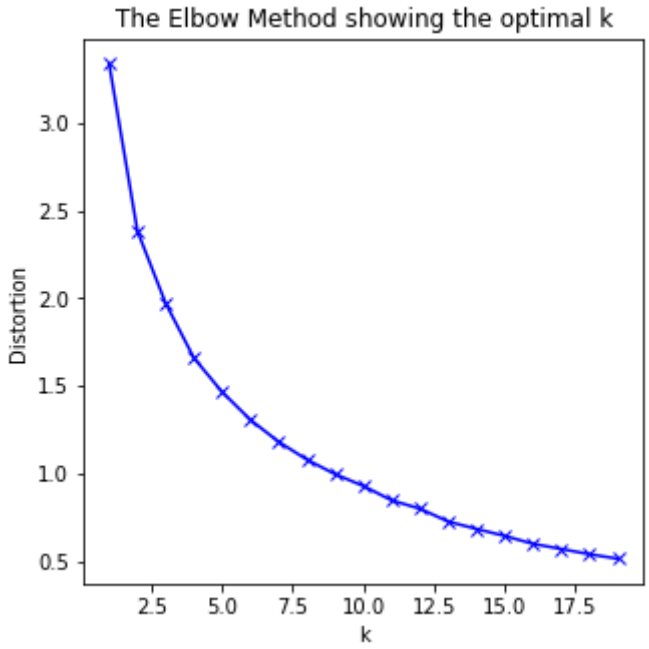
To check single neighborhoods for most common venue categories, I further created dataframe with columns for the top 10 venues for each neighborhood:

	Neighborhood_long	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	New York-Manhattan-Battery Park City	coffee	parks	landmarks	pizza	bars	french	mexican	newamerican	museums	italian
1	New York-Manhattan-Carnegie Hill	coffee	pizza	italian	mexican	delis	japanese	bars	bagels	newamerican	vietnamese
2	New York-Manhattan-Central Harlem	chinese	caribbean	delis	bakeries	hotdogs	chicken_wings	breakfast_brunch	bars	coffee	french
3	New York-Manhattan-Chelsea	coffee	italian	bakeries	pizza	delis	mexican	newamerican	bars	cafes	lounges
4	New York-Manhattan-Chinatown	chinese	bakeries	coffee	italian	desserts	bubbletea	foodtrucks	pizza	bars	newamerican

3 Exploratory Data Analysis

For a cluster analysis with K-Means, one has to figure out the number of clusters. Based on the mean of the frequency of occurrence of each category by neighborhood I used the elbow method as a heuristic for determining the number of clusters in the dataset.

This method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.



From the figure above there is no clear 'elbow' visible, which may be a first hint, that clustering with K-Means based on this features are not the best solution. I chose a reasonable k = 5 to proceed.

4 Results

As expected K-Means with k=5 does not separate the neighborhoods by venue categories, as the distribution of neighborhoods among the clusters shows:

Cluster Labels	Neighborhood_long
0	1
1	70
2	1
3	1
4	1

5 Discussion

This quick analysis shows, that the chosen inner city neighborhoods of Toronto and New York cannot be segmented by venue categories . K-Means is probably suffering from the curse of dimensionality as there are only 74 neighborhoods but 287 categories.

The visual analysis of the distribution of the most common venue categories shows, that there is a great variety of venues, especially food venues, in multicultural cities like Toronto and New York. Nevertheless some neighborhoods have more or less obvious clusters of categories like coffee bars in well known business districts like lower Manhattan.

6 Conclusions

The data and analysis shows that – in the case of Toronto vs New York - distinguishing neighborhoods only by venue categories do not lead to a clear segmentation and therefore will not help finding a suitable new place to live when moving to another city.

There is mor work to be done regarding feature engineering, like aggregating the categories.

On the other hand a lot more freely available information could be taken into account, like population density, type of housing, ….

Reduction of the retrieved categories to food-related businesses, because Yelp mostly delivers information about food venues.