Clustering and comparing neighborhoods of Newyork and Toronto

Pramudith Karunarathna

1.Introduction

1.1 Background

In this project we will study,analyze,cluster and compare the neighbourhoods of two important cities in the world. Newyork and Toronto are largest as well as financial and tourist capitals of the countries United state and Canada respectively.There are roughly about 8.39million residents in Newyork and 2.93 million residents in Toronto. Newyork and Toronto are both huge,diverse and cosmopolitan cities.
New York City (NYC) is one of the most populous cities in the United States of America. Also, NYC is the most linguistically diverse city in the world: as many as 800 languages are spoken in it. Moreover, NYC plays an essential role in the economics of the USA: if New York City were a sovereign state, it would have the 12th highest GDP in the world. New York City consists of five boroughs: Brooklyn, Queens, Manhattan, The Bronx, and Staten Island.
The second city of interest in this project is Toronto. As with NYC in the USA, Toronto is the most populous city in Canada. It’s recognized as one of the most multicultural and cosmopolitan cities in the world. Toronto also is a very diverse city: over 160 languages are spoken in it. On the economic side, Toronto is an international centre for business and finance and it is considered the financial capital of Canada.

Since both countries are near, people move to two cities for various reasons.they move in order to get a job,to start a business ,to shop,to travel and various other reasons.

1.2.Problem statement

Suppose a person wants to move from Newyork to Toronto for a job.This person does not know anything about Toronto and he would like to move into a place where he lives now.
Is it possible to create a system that can helping our user showing similarities between two cities?

Suppose a businessman wants to move from Newyork to Toronto to start a business.He wants to know what type of what types of businesses are more likely to thrive in both cities, what are the neighborhoods that are suitable for each type of business, and what

types of businesses are not very desirable in each city.If he knows this he can get better and more effective decisions regarding where to open their businesses.
Is it possible to create a system that can helping our businessman showing similarities and differences in businesses between two cities?

1.3.approach

Foursquare is a website where people comment and rank food sites, coffee sites, malls and parks. For instance, let's think that a Foursquare user had to move from New York city, USA to the city of Toronto, Canada. Foursquare location data along with a clustering algorithm can suggest a neighborhood in order to help this user to live in Toronto in a similar place. The neighborhood that will be suggested, will not be a random suggestion, but instead will be a place for his pleasure. Thus, previous data from New York and Toronto will be used to predict a good living neighborhood for him.

1.4.Target audience

People who seek a new job in Toronto/Newyorks
Businessmen start new business in these cities
Residents who move between these two cities.

2.Data

In order to analyse the cities on a meaningful level, they need to be divided into different areas, e.g. neighborhoods, boroughs.A list of neighborhoods in New York and Toronto is downloaded and their respective location in longitude and latitude coordinates is obtained. The sources are the following:

Newyork
https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json

Toronto
https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&direction=next&oldid=942655364

Foursquare API will be used for this project.Moreover, their specific coordinates are merged. Only Manhattan neighborhoods and boroughs that contain the string "Toronto" are taken into account. A Foursquare API GET request is sent in order to adquire the surrounds venues that are within a radius of 500m. The data is formated using one hot encoding with the categories of each venue. Then, the venues are grouped by neighborhoods computing the mean of each feature.
The similarities will be determined based on the frequency of the categories found in the neighborhoods. These similarities found are a strong indicator for a user and can help him to decide whether to move in a particular neighborhood near the center of Toronto or not.

3. Methodology

3.1. Feature Extraction

For feature extraction One Hot Encoding is used in terms of categories. Therefore, each feature is a category that belongs to a venue. Each feature becomes binary, this means that 1 means this category is found in the venue and 0 means the opposite. Then, all the venues are grouped by the neighborhoods, computing at the same time the mean. This will give us a venue for each row and each column will contain the frequency of occurrence of that particular category.

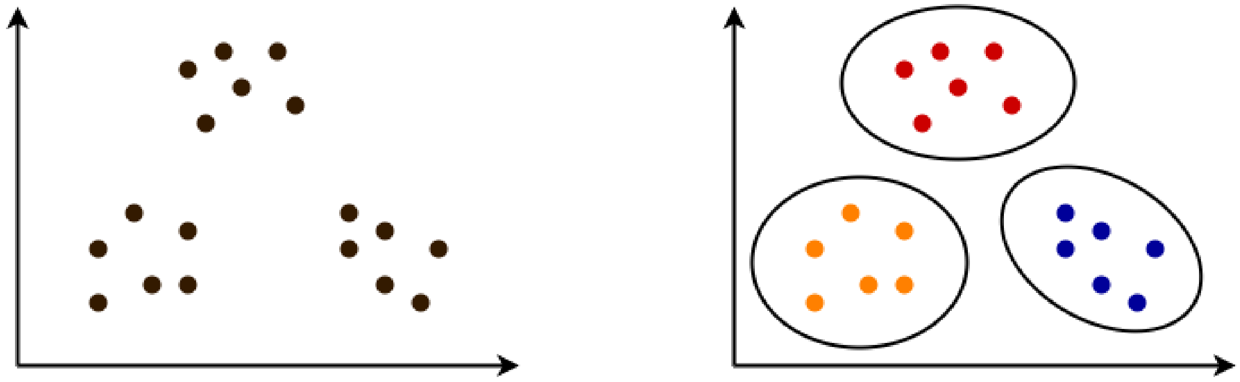
3.2. Unsupervised Learning

For the purpose of doing unsupervised learning to found similarities between neighborhoods, a clustering algorithm is implemented. In this case K-Means is used due to its simplicity and its similiraty approach to found patterns.

- K-Means:

K-Means is a clustering algorithm. This algorithm search clusters within the data and the main objective function is to minimize the data dispersion for each cluster. Thus, each group found represents a set of data with a pattern inside the muldimensional features.

In the following figure there is a graphical example of how a K-Means algorithm works. As it is possible to see, dispersion is minimized by representing all clustered data into one group or cluster.



Before K-Means

After K-Means