The Battle of Neighborhoods

Introduction

New York City, also known as the City of New York or simply New York (NY), is the most populous city in the United States. New York City comprises 5 boroughs sitting where the Hudson River meets the Atlantic Ocean. At its core is Manhattan, a densely populated borough that's among the world's major commercial, financial and cultural centers.

Toronto, the capital of the province of Ontario, is a major Canadian city along Lake Ontario's northwestern shore. It's a dynamic metropolis with a core of soaring skyscrapers, all dwarfed by the iconic, free-standing CN Tower.

The New York City and the city of Toronto are very diverse and are the financial capitals of their respective countries. In this project we are exploring the neighbourhoods of both cities with the help of Foursquare Places apis. In exploring the Neighbourhoods, we will figure out how similar or disimilar their neighbourhoods are. In doing so we will have answer to several questions such as Which locations would be best for me if i move from toronto to New York ? or which location would be the best for opening the another branch of my company ?

Data Description

New York city Neighbourhood Data

Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segement the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the the latitude and logitude coordinates of each neighborhood.

This dataset exists for free on the web. Link to the dataset is:
https://geo.nyu.edu/catalog/nyu_2451_34572

Toronto City Data

For city of toronto we are going to extract information regarding it's neighbourhoods from wikipedia page : https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Geographical coordinates data of both the cities will be utilized as input for the Foursquare API, that will be leveraged to provision venues information for each neighborhood. We will use the Foursquare API to explore neighborhoods in both cities

Methodology

Data cleaning and processing

Data mentioned in above section is used to extract the neighboudhood information of each city.

The New york city data is available in Json format. The data is processed in to the pandas dataframe to get all the neighbourhoods of New york city :

```
In [26]: neighborhoods.head()
```

Out[26]:

	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

```
In [27]: print('The Dataframe has {} boroughs and {} neighborhoods.'.format(
             len(neighborhoods['Borough'].unique()),
             neighborhoods.shape[0]
         ))
```

```
The Dataframe has 5 boroughs and 306 neighborhoods.
```

The neighbouhood data is than processed and foursquare places apis is used to get the top 100 venues located in each locality.One hot encoding is applied and data is than grouped on neighbourhood to get the dataframe ready for comparison.

Similar method were applied on Toronto city data to get the dataframe ready for comparison. Neighbourhoods of toronto city :

```
toronto.shape
```

Out[91]: (287, 3)

```
In [92]: toronto.head()
```

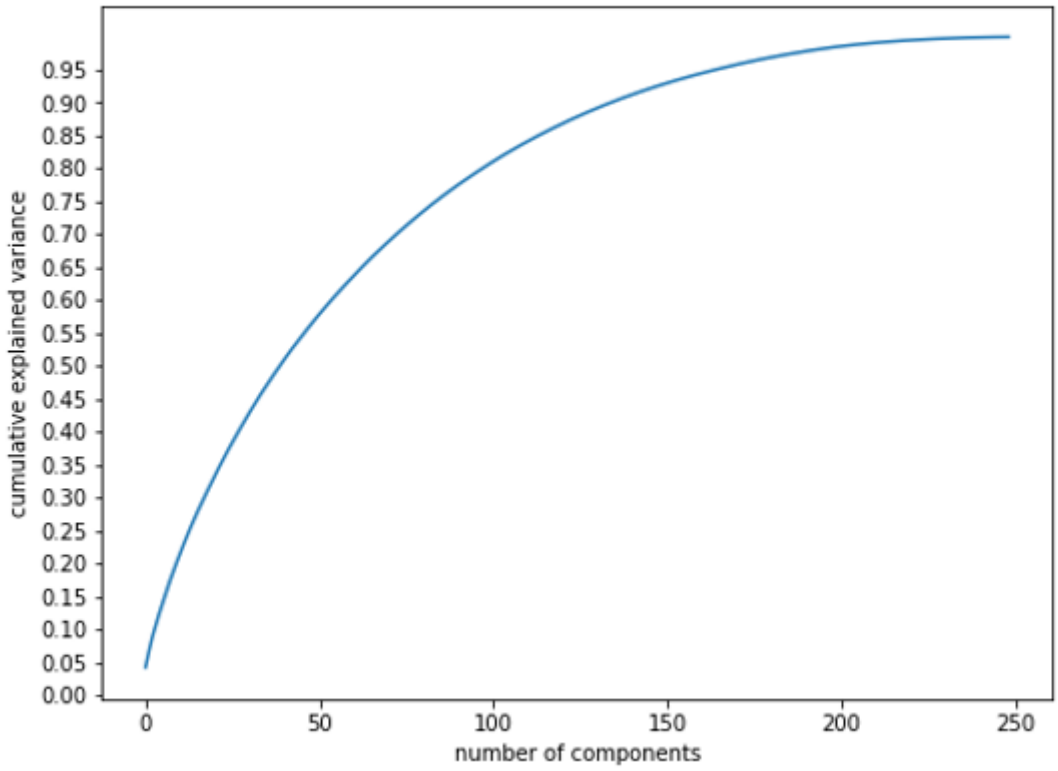
Out[92]:

	Postcode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront

Dimensionality Reduction

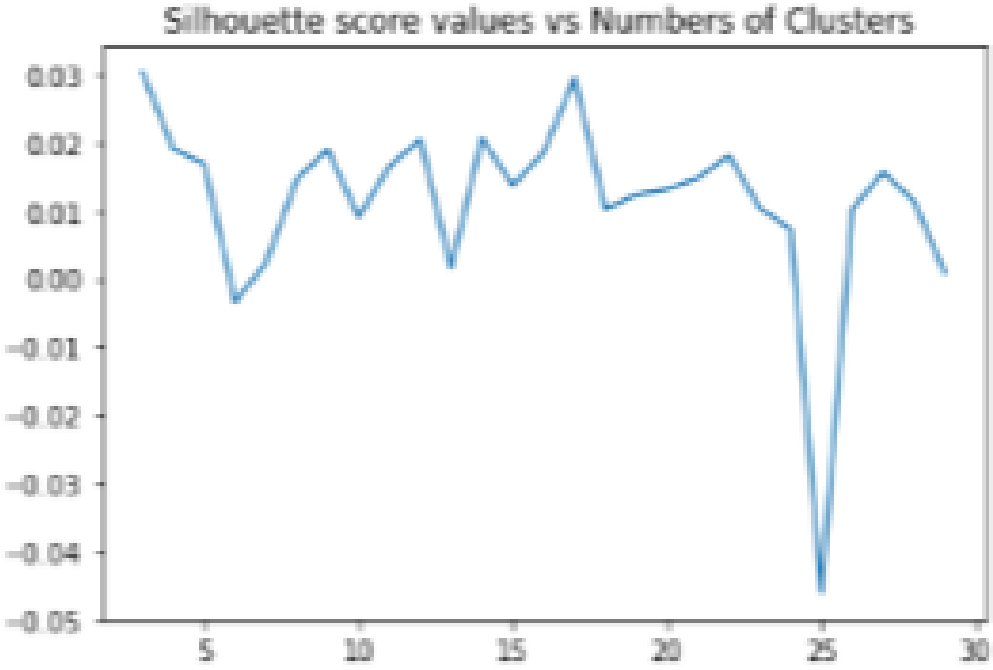
Principal Component Analysis (PCA) is an unsupervised linear transformation technique that is widely used across different fields, most prominently for feature extraction and dimensionality reduction. Dimensionality reduction is performed on merged dataframe of neighbourhoods of both the cities using Principal Component Analysis on the dataframe in order to reduce the number of dimensions. we are able to downscale the number of features yet retaining the same variance.

```
]: Text(0, 0.5, 'cumulative explained variance')
```



Silhouette score method

Silhouette score method is used in order to find out the best k or best optimal number to perform K-means clustering Algorithm. Here is the result




```
Optimal number of components is:
3
```

Using above method, we reached at the conclusion that the best value of K is 3 for performing kmeans algorithms on merged dataframe .Hence , we have performed Kmeans algorithms to classify all the neighbourhoods of both the cities in to three clusters .

Results

After performing the KMeans function on the processed dataframe , neighbourhoods of both cities are assigned to their respective cluster groups. Here are the results calibrated on map :

New york Neighbourhoods :

Toronto Neighbourhoods :

Discussion

Through this analysis , neighbourhoods of both cities were segmented in to three groups using the limited data from foursquare places api. Hence it's scope and applications is limited to some extent .But still ,this model can be further refined by using different open source data sets to gain more information regarding the neighbourhoods of both cities to further crystalise it's findings.

Conclusion

The study has given us the similarity of neighbourhoods and the solution of various problems we were facing at beginning. If a company headquatered in Manhattern wants to open second branch in toronto and wants similar ecosystem , by looking at results , the company can have shortlist of neighbourhoods to begin with, which can further be refined according to it's preferences . A person seeking relocation from Toronto to New york can make better choices.