Clustering Neighborhoods in New York and Toronto

Phan Nhat Hoang Nguyen

Feb, 2020

I. INTRODUCTION

1. Background

   Cities around the world are filled with many kinds of venues that define the cultures of the cities. A city's inhabitants or services of tourists have put a significant mark on differentiating it from another, not only the mean of global position. Despite of having many different features, it is somehow possible for us to group different venues which has the similar kinds of neighborhoods in different cities. Having grouped together similar kind of neighborhoods may help people make a good decision when they consider moving to another cities.

2. Problem

   Finding identical neighborhoods indifferent cities or countries so that can help provide a perception of similar neighborhoods which provide a large number of insights in order to make a decision of choosing a neighborhood that far away, but somehow feels like home.

II. DATA ACQUISITION AND CLEANING

1. Data sources

   In this capstone project, I use two sets of data. The first dataset consists of Toronto's different boroughs and their respective postcode. The second dataset consists of New York's different neighborhoods and their respective geometric coordinates.

   You can find the tow datasets here:

- Toronto dataset: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- New York dataset: https://cocl.us/new_york_dataset

2. Data cleaning

   The first data source is a Wikipedia page that contains postal code of Toronto city in table. To scrape the data from URL, BeautifulSoup has been used to extract the table data. I would go through some more step, the dataframe obtained consists of PostalCode, Borough, Neighborhood.

   The second data source in the link above is in json format, which consists of many different features. I would format the json file so that the data finally resulted in a dataframe that consists of Borough, Neighborhood, Latitude, Longitude.

   There are some problems with the Toronto dataframe, it has some values under "Borough" column that were not assigned in the first place. So, these rows would be dropped. There are some rows under "Neighborhood" column that had no value assigned to. My solution is copying the value from the "Borough" column of the respective row into the "Neighborhood" column.

3. Features selection

   Now we have the datasets of different neighborhoods and their respective geometric coordinates for the city of Toronto and New York. So, we will come up with different venues that these different venues have to offer.

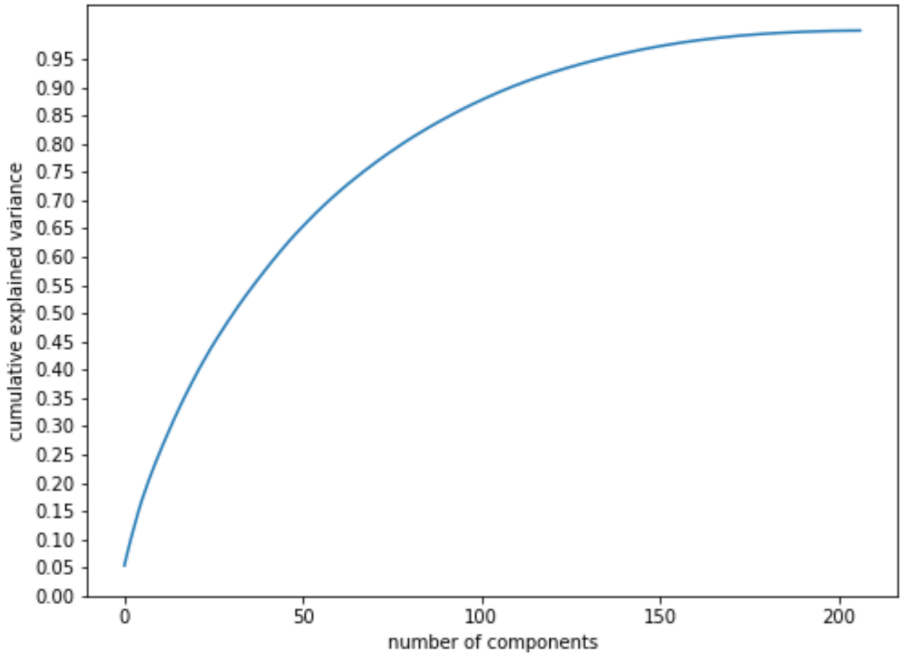
   Foursquare API provides an access to an enormous database consisting of venues from all around the world including rich variety of information such as addresses, tips, photos, etc. Having signed up for a Foursquare Developer account, using the Client ID and Client Secret, it is possible to make an API request in order to retrieve venue's information

   By feeding a function with Neighborhood name and its coordinates, using Foursquare API, different venues were extracted. After performing one-hot-encoding and grouping rows by Neighborhoods, the Toronto and New York datasets seemed to have 250 features. Both the dataframes would be combined into one single dataframe in order to perform clustering operation.

4. Dimensionality reduction

   We use Principal Component Analysis (PCA) as a dimensionality reduction tool to reduce a large set of features to a smaller set that still contain most of information.

Before clustering the neighborhoods, we performed dimensionality reduction using PCA on the dataframe in order to reduce the number of dimensions.



After performing PCA, the number of features was reduced to 150 yet retaining the maximum variance of the dataset.

III. METHODOLOGY

The goal of this capstone project is to group the similar neighborhoods in Toronto and New York.

1. Determining Optimal Cluster Number

K-Means is a simple unsupervised machine learning algorithm that group data into a specified number (K) of clusters. When using K-Means clustering, we need to determine the optimal number of clusters in order to maximize the accuracy of the results.

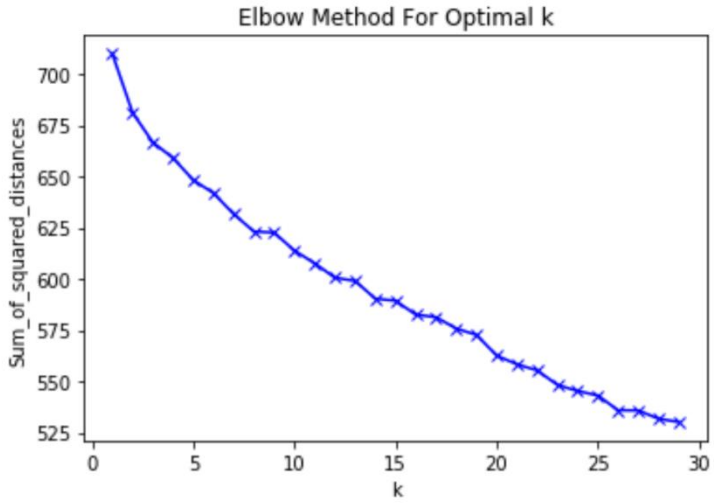
The first method I use is elbow method. The idea of this method is to run K-Means clustering for a range of values of K (from 1 to 30), for each value I calculate the sum of squared errors (SSE). The ‘elbow’ is the value of K to be used. The goal is to choose a small value of K that still have a low SSE, and the elbow usually represents where we start to have diminished return by increasing K.

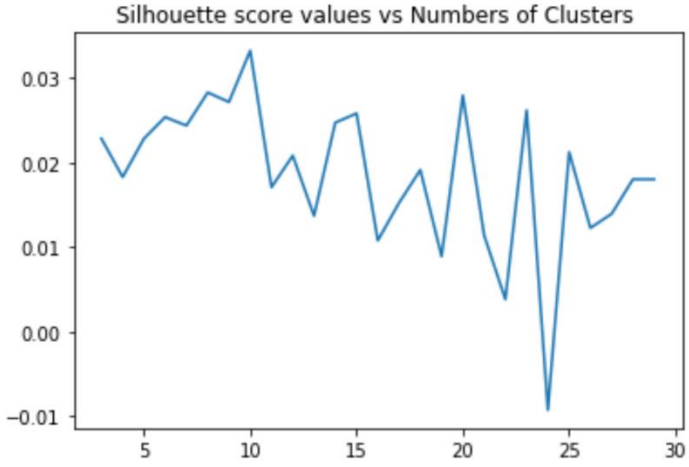
The second method is silhouette analysis. This analysis is a way to measure how close each point in a cluster is to the points in its neighboring clusters. It is a neat way to find out the optimum value for K during K-Means clustering.

2. Random Initialization
   Since K-Means incorporates a heuristic approach, it does not ensure converging at global optima at each iteration. Depending upon how the initial positions of the clusters were set, it may converge into different local optima. In order to overcome this issue, numerous iterations on different random initialization were performed in order to find the best set of converge.

IV. RESULTS

1. Optimal number of clusters



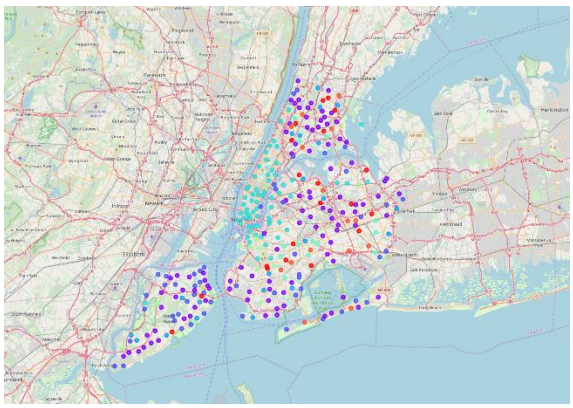



```
Optimal number of components is:
10
```

With the two figures above, we can come to a conclusion that 10 cluster would be a reasonable choice for the algorithm.

2. Visualize the cluster on the map

We create folium maps to obtain a visual perception of how the clusters look on the map when plotted on the map of the two cities, New York and Toronto.



New York

Toronto

The neighborhoods with the same color would belong to the same cluster, that means these neighborhoods share the similarities.

V. DISCUSSION

Since this was an unsupervised clustering work, we have many different approaches can be adopted in order to achieve better results. This project was only done a small dataset, having more samples may results in a better clustering.

Having dealt with location data on a deeper level, such as at neighborhood level may results in better grouping of similar data points which eventually may results is better clustering.

Through this project, we can cluster the similar neighborhoods in many different cities around the world, which means helping people make a good decision when they consider moving to another cities.

The study here is being ended by visualizing the data and cluster information on the map of city of New York and Toronto.