The Battle of Neighborhoods: New York vs Toronto

1.Introduction

1.1 Background

Different cities in the world are filled with numerous kinds of venues that in turn define the cultures of the cities. A city not only differs from another by means of global positioning, what it to showcase to its inhabitants or tourists has put a significant mark on differentiating it from the rest. Despite of having dissimilarities, it is somewhat possible to group together the similar kind of neighborhoods in different cities. It is possible to segment the different venues in a neighborhood according to venue category, and then to group neighborhoods together that incorporate similar kind of neighborhoods. Having grouped together similar kind of neighborhoods may serve as a variable to help make a decision when people consider moving out of a city to another.

1.2 Problem

Finding identical neighborhoods in different cities in order to help provide a perception of similar neighborhoods which may provide with a great deal of insights in order to make a decision of choosing a neighborhood that is far away, yet somewhat feels like home.

2. Data Acquisition and Cleaning

2.1 Data Sources

This project works with two sets of data. The first dataset consists of New York's different neighborhoods and their respective geometric coordinates, which can be found here. The second dataset consists of Toronto's different borough and their respective postcodes, which can be found here.

2.2 Data Cleaning

The first data source (for New York) in the described link is in json format. It initially consisted of many different classes of data. Upon examining them, the data that we are interested in was found under 'features' category. Further formatting of the json data finally resulted in a dataframe that consists of 4 columns, namely: **Borough, Neighborhood, Latitude and Longitude**.

The second data source (for Toronto) is a Wikipedia page that contains Postcode of the city of Toronto in a wikitable. To scrape the data from the URL, BeautifulSoup has been used to extract the table data. After going through a few more steps, the dataframe was obtained which consists of: **PostalCode, Borough and Neighborhood**.

But the problem with this dataframe was, it has some values under the column 'Borough' which were not assigned in the first place. So, the rows with no assigned value in the 'Borough' column were dropped. Another problem was there were a few rows in the 'Neighborhood' column that too had no values assigned to it. As a solution, the value from the 'Borough' column of the respective row was copied into the 'Neighborhood' column.

2.3 Feature Selection

Now that we have obtained the different neighborhoods and their respective geometric coordinates for the city of New York and Toronto, it is time to come up with different venues that the different venues have to offer.

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

By feeding a function with Neighborhood name and its geometric coordinates, using Foursquare API different venues (Restaurants, Coffee shops, etc) were extracted. After performing One-Hot-Encoding and grouping together the rows by neighborhoods, the NY dataset and Toronto dataset seemed to share 250 features. Both the dataframes were combined into a single dataframe in order to perform clustering operation.

2.4 Dimensionality Reduction

Principal Component Analysis is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information from the large set. Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components.

Before diving into clustering operation, first we performed dimensionality reduction using Principal Component Analysis on the dataframe in order to reduce the number of dimensions.

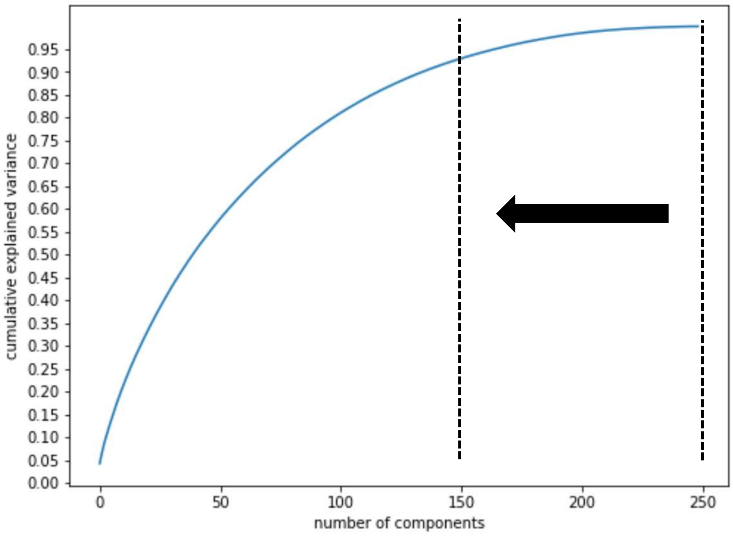

Figure 1 - Selection of number of Principal Component Analysis

Having performed PCA, the number of features was reduced to 150 from 250 yet retaining the maximum variance of the dataset.