A Tale Of Two Cities

Differences Between New York and Toronto Neighborhoods By Clustering Analysis of Neighborhood Venues Data

Introduction

New York City and Toronto are both the most populous metropolitan areas of the respecting countries. New York, the larger one of the two in terms of populations (over 23 millions vs. roughly 6 millions) boasts to be the cultural, financial, and media capital of the world and the center of commerce, entertainment, research, technology, education, tourism, arts, fashion and sports. Toronto is also the center of business, finance, arts and culture and recognized to be one of the most multicultural and cosmopolitan cities in the world. Despite the difference in total populations, the two cities embrace very similar characteristics of their neighborhoods as a whole, but are they really that similar if we are to compare the categories of venues in all the neighborhoods in both cities? If we are to perform cluster analysis for each city alone and for the combined neighborhood data of the two cities, what will the clusters distributions look like in individual city analysis and in the combined data analysis? What are the best numbers of clusters for each city and for the combined data? Will the majorities the two cities fall into the same clusters? A main goal of the project is to see the distributions of the clusters in each city in terms of venue categories of the neighborhoods when the data of both cities are combined. For example, does Manhattan Borough of New York City share the same characteristics with the 4 inner Boroughs in Toronto (East, Centra, West and North Toronto)? Or are the two cities fall into very different clusters? If the two cities turned out to be dominated by different clusters, how are they different? This project will attempt to determine the best numbers of clusters for each city alone and for the combined data using Kmeans Elbow Method.

Data

The New York geographic data are the newyork_data.json provided in the Applied Data Science Capstone Class of the Coursera IBM Data Science Professional Certificate Specialization. The Toronto geographic borough and geographic data are from Wikipedia and Geospatial_Coordinates.csv file provided in class. The geographic data of both cities will be used to construct the dataset of neighborhoods in each borough and their coordinates and convert to two pandas dataframes for New York and Toronto. Each row of the data frame is one neighborhood with its coordinates. A function looping through each row of each dataframe sends requests to Foursquare API to request a maximum of 100 venues in a radius of 500 meters for each neighborhood coordinates. The returned venue data for every neighborhood in each city are converted to a data frame for each city. These data will be cleaned and pre-processed with on-hot encoding to get ready for Kmeans clustering analysis.