Peer-graded Assignment

Yinuo Zeng

December 2019

1 Data

The data for this project are from multiple sources.

The data of neighborhoods and Boroughs with coordinates in New York City is from website https://geo.nyu.edu/catalog/nyu_2451_34572. It includes four features, which are *Borough*, *Neighborhood*, *Latitude*, and *Longtitude*.

The data of neighborhoods and Boroughs in City of Toronto is from website https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. Three variables it contains are *Postcode*, *Borough*, and *Neighborhood*.

The coordinates of neighborhoods in City of Toronto can be obtained using Python library *GeoPy*, which provides attributes *Latitude* and *Longitude*.

Also the other features (venue categories) of each neighborhood are provided by *Foursquare* API.

In order to solve the problem I mentioned in previous section, I am going to perform clustering analysis. Specifically, I will construct a dataframe that combines both neighborhoods in New York City and City of Toronto. Then, K-means and other clustering methods will be used to do neighborhoods segmentation and I will check whether neighborhoods from different cities have more probability to be put in the same group.

In addition, I will examine each group and try to find important features that cause such grouping.

Also, it is possible that there are more features (venue categories) than the number of neighborhoods. Under such condition, I will pay more attention on clustering methods for high-dimensional data.