Introduction and Data Description

September 9, 2019

0.1 Part I Introduction/Business Problem

The question that I would like to leverage the Foursquare location data to solve is about categorizing and identifying neighborhoods by amenities, events and activities within a neighborhood.

Daniel Silver, Terry Nichols Clark and Clemente Navarro (2007) raised a concept “Scene”, which represents a place with the meanings expressed by the people and practices there, to help analyze social characteristics of places. In their theory, a "Scene", is constructed by physical structure, persons and activities together, can be analyzed under 15 dimensions, such as "Legitimacy", "Traditionalistic","Self-Expressive","Glamorous" etc. A “scene” with high scores in specific cultural characteristics can attract the in-flow of the population that feel afliated to corresponding cultural characteristics and cultivate a cultural environment which encourage corresponding values such as creativity/innovation/cautiousness/etc thus influencing the social and economical performance in an area. For example, when artists locate in more self-expressive, glamorous, and charismatic scene,general economic growth are stronger (Silver and Clark 2016).

The framework behind the "Scene" theory, which categorizes different places by their cultural dimensions and examine how places of different categories affect social and economical performance, is a interesting and useful method. However, the 15 cultural dimensions which are used to categorize places seems too heavily based on hypothesis and theory and have some weakness in emprical evidence. Therefore, inspired by the week 3 segmenting and clustering project, I would like to try machine learning categorization algorithemes to provide a new perspective on neighborhoods categorizing and help understand their social characteristics, based on activities and amenities data in neighborhoods.

0.2 Part II Data Description

For amenities data, I will leverage the Foursquare API to access venues information in neighborhoods.

For activities data, I will employ the data from Meetup.com website. Meetup is an online platform with more than 40 millionusers,320,000 active groups, and an average 12,000 events each day that allows participants to connect with people in their geographic area through public events (Meetup.com).It is a so-called “Event Based Social Network” (EBSN), which allows users to find and join groups unified by a common interest and participate the events held by these groups. Meetup API allows any registered user to request for and store the information (including precise date, time, geography coordinate, host group, category, description, the number ofattendees,rate etc.) of Meetup events that happen within a certain distance from cities. For this project, in order to simplify the data collecting process, I plan to use the Meetup dataset on Kaggle (https://www.kaggle.com/ruosiwang/meetup), which was originally collected through Meetup API.

By Ying Cai