The Battle of the neighborhoods:

Segmenting and Clustering Neighborhoods in Chicago based on Venues Data and Selected Socioeconomic Indicators of Chicago Community Areas

Introduction

Chicago is the third-most-populous city in the United States and an international hub for finance, culture, commerce, industry, education, technology, telecommunications, and transportation. The city has 77 distinct community areas, which can further be subdivided into over 200 informally defined neighborhoods. The city's waterfront location and nightlife has attracted residents and tourists alike. Chicago's 58 million domestic and international visitors in 2018 made it the second most visited city in the nation, as compared with New York City's 65 million visitors in 2018. The city has many upscale dining establishments as well as many ethnic restaurant districts. [https://en.wikipedia.org/wiki/Chicago]

In this data science project, a successful ethnic restaurant (which is named as R1 here) in one community area (which is named as CA1) would like to open its 2nd restaurant (named as R2) to expand the business, and we will use machine learning to help it to find the best location candidates for its new restaurant (R2) in terms of community areas. We will use the unsupervised clustering method such K-Means for segmenting and clustering community areas in Chicago. Our assumption is that the best locations for the 2nd restaurant R2 should be in the community area that is in the same cluster containing CA1. Or in other words, CA1 and the new community area (denoted by CA2) for the best location candidates should be within the same cluster. After clustering, we will further analysis how similar the two community areas and why the new one CA2 is a great choice.

Data Section

For this project, we plan to leverage on two sets of data. The first dataset is the venues data obtained from Foursquare. Given the geographical coordinates of each community areas in Chicago, we will request the venues data within a certain proximity to the center of each community area from Foursquare via API calls and extract the top venues for each community area. The venues data is essential for clustering. In addition to the Foursquare data, the 2nd dataset is called Selected Socioeconomic Indicators in Chicago, 2008 – 2012, which is available online. This dataset contains a selection of six socioeconomic indicators of public health significance and a “hardship index,” by Chicago community area, for the years 2008 – 2012. The 2nd dataset will be a great complementary to the Foursquare data. We will use the socioeconomic dataset to improve the clustering results and provide more meaningful explanations to the results.

Note: The link of the 2nd dataset is https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2 Census_Data_-Selectedsocioeconomic_indicators_in_Chicago__2008___2012.csv