The battle of neighborhood in Bloomington

December 19, 2019

0.1 Introduction

0.1.1 Background

There are 58 neighborhoods in Bloomington, IL based on the information from localWiki. There are State Farm headquarter, Pepsi Ice Center, Grossinger Motors Arena in the city. Different neighborhoods have different majority venues. Visitors coming to the city usually have limited time. How to recommend the visitors to the right neighborhood is important to the development of the city's tourism.

0.1.2 Description of the problem

If a visitor comes to Bloominton and he likes outdoor activities and Chinese food, which neighborhood he should go. So he can play and eat well. The solution of the problem can be utilized in Bloominton's city website as a guidance to the city.

0.2 Data

0.2.1 Sources of Data

There are three main sources of data.

The first is neighborhood data scraped from NextDoor website, it has 90 rows, each row represents one neighborhood.

The second is coordinate data. Neighborhood name will be used to get latitude and longitude using geopy package in Python. There are 23 listed neighborhoods whose coordinate information is not available. So the joined data set has 67 rows(neighborhoods), 3 variables(neighborhood name, latitude, longitude). After checking the summary statistics of latitude and longitude, there are a few points which are potentially incorrect. After verifying the points with visualization on the map. The final joined data set has 59 rows, 3 variables.

The third is venue data. By using neighborhood coordinates information, we can get venues data from FourSquare API. After data transformation, the neighborhood_venues data set has 345 rows(venues) and 7 variables(neighborhood name, neighborhood latitude, neighborhood longitude, venue name, venue latitude, venue longitude, venue category). ### Utilization of Data By combining venue data and neighborhood data, we will group neighborhood with similar venue category distribution into different clusters and visualize them on map. We will also use the combined venue data and neighborhood data to find the special neighborhood which can meet the visitor's requirement.

0.3 Methodology

0.3.1 K means cluster

After getting the transformed data which has 310 rows(per neighborhood and venue), 111 columns(first column is neighborhood name, other columns is indicator of whether the venue exists). By grouping the neighborhood, we get the frequency of each venue category in each neighborhood. The number of cluster is set to 3. K means cluster method will inialize three center points and group points which are closer to the center points into different cluster. And recalculate the new center point inside each cluster until the result is converged or reach specied iteration time.

0.3.2 Summary statistics

Rank the venue category for each neighborhood based on the frequency. For each neighborhood, there will be five columns which represent the top five venue category within each neighborhood. Based on the top five venue, find which neighborhood has the most Chinese restaurant and outdoor activities.

Group by the neighborhood, count the number of venues for each neighborhood.

When searching for neighborhoods which have the most Chinese restaurant and outdoor activities, also consider the number of venues in the neighborhood, preventing neighborhood has high frequency while only has few venues.

0.4 Results

From the clustering result, the first cluster is a group of neighborhood which contains mixed outdoor venues, restaurants, and stores.

The second cluster is a group of neighborhood which contains outdoor venues.

The third cluster is a group of neighborhood which contains restaurants.

By searching for Chinese token inside the most common venue categories, it first appears in the 3rd Most Common Venue in neighborhood Clobertin Ct. Looks like Bloomington doesn't have lots of Chinese restaurant. Clobertin Ct is the fourth neighborhood which has the most venues. Clobertin Ct's two most common venue categories don't have outdoor activities. So we need to find another neighborhood.

By looking through the ranked neighborhood based on venue count, E Grove ST is the fifth neighborhood and the 1st common venue is Park. So E Grove ST will be another recommended neighborhood.

0.5 Discussion

The two selected neighborhoods both belong to the same cluster which is mixed of outdoor venues, restaurants and stores. The features we chose is not good enough to distinguish different clusters. I may try other clustering method like DBSCAN clustering or hierarchical clustering. Or collect other features about the neighborhood except the venue category.

0.6 Conclusion

By doing analysis on the location data of Bloomington, IL, we successfully cluster the neighborhood into three parts and each part has different characteristics. And based on summary statistics, we succesfully find Clobertin Ct has more Chinese restaurant compared to other neighborhoods and E Grove ST has more outdoor activities. We can recommend the two neighborhood to the visitors.