FIND SIMILAR AREAS IN A CITY COMPARED TO AREAS OF INTEREST BASED ON FOURSQUARE LOCAL VENUES DATA USING K-MEANS CLUSTERING

Dan-Stefan Florescu

Author Note

This is the report for the final project in the IBM Applied Data-Science course on Coursera

1. Introduction

This is the report for the final project in the IBM Applied Data-Science course on Coursera, where the brief mentioned leveraging Foursquare data to solve a Data-Science problem, of the student's choosing.

Background

For multiple reasons (that are outside the scope of this project and will not be discussed) the rate of people moving has been rising and continues to rise. People are changing jobs more often. People are traveling more. People are living in more places during their lives. People are studying abroad more. All this means that people are also looking more for information regarding different places. Based on these trends more people might need to move to a place that is completely new to them. They might want to find out what area in that place they would like, or be similar to an area that they know they like. Or even find out what area they wouldn't like or be similar to places they dislike.

Problem

The problem tackled by this project is finding areas in a city that are similar to other areas of interest, with the particularization of the city to Zurich, CH, and the areas of interest to South Kensington -London UK, Greenwich – New York, NY USA and West Loop – Chicago, IL USA.

This type of challenge is often found in the relocation process. Usually, if an employee is relocating to a new city, their employer hires a real estate agency that works closely with the employee to find them a new place in the city they are relocating to. The real estate agency, by

simply asking the employee their favorite neighborhoods, with this solution, the real estate agency could easily find starting search areas for finding a place in a location that the employee will enjoy. So, a real estate agencies in the relocation business have the problem and could find value in this solution.

2. Data

Data regarding the areas

To define an area, the longitude and latitude of the center of the area was used. Data regarding areas of Zurich were difficult to obtain as no list of neighborhoods was found. However, a list of districts were found, hence the words areas, neighborhoods are used interchangeably (https://www.zuerich.com/en/visit/about-zurich/zurichs-districts). Unfortunately, the dimensions of the districts were far larger than the dimensions of the neighborhoods of interest. The assumption being made is that the difference stems from the population, and population density difference of the cities where the areas of interests are, and the city of Zurich. Hence, the districts were split into smaller areas, similar in size to the areas of interest, making the radius that will define the areas together with the center location, 800m. Unfortunately, this data collection and manipulation was done by hand in an excel file that was exported as .csv to be imported in a dataframe. The .csv file had the name of the area and its longitude and latitude.

Data regarding the characteristics of the areas

Data from Foursquare will be used to characterize the areas, becoming the dimensions of the data-set that will be clustered. Using a query to the Foursquare API, data regarding the top 100 venues located in the area (800m, around the center of the area) were collected and stored in a

dataframe. Data regarding the venues was one hot encoded, sorted and then jointed with the areas dataframe for clustering.