FIND AREAS IN A CITY SIMILAR 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

Abstract

Using Python in a Jupyter Notebooks on Network Skills Lab, K-means clustering, an unsupervised machine learning algorithm, was used to determine what areas of Zurich, CH were similar to specific neighborhoods of interest. In this particular case, the author found the following neighborhoods appealing, and were used as areas of interest in this project: South Kensington - London UK, Greenwich – New York, NY USA and West Loop – Chicago, IL USA. Local venues data from Foursquare was used to cluster the neighborhoods. For better visualization, the areas were mapped and color coded based on their respective cluster. The resulted clusters showed that South Kensington -London was most similar to areas in the intersection of Districts 1, 7 and 8 (also known as Altstadt, Hochschulen, Seefeld, Muhlebach), while Greenwich – New York and West Loop – Chicago, which were clustered together, were most similar to areas of Districts 5 and 10 (also known as Escher Wyss, Industriequartier, Wipkingen).

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 internationally 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 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, 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 a solution.

2. Data

Two Data sets will be used in this project that can be split in two categories: Data that defines areas and data that characterizes areas.

Data regarding the definition of 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 was found, hence the words areas and 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 this 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, for 48 areas in Zurich and 3 areas of interest South Kensington -London UK, Greenwich – New York, NY USA and West Loop – Chicago, IL USA.

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.

3. Methodology

To solve the problem, Python programming was used in a Jupyter notebook on Skills Network Labs.

Preliminary data analysis and method selection

As mentioned in the Data section, 2 data sets were used from 2 different sources. Each data set was imported and stored into a panda dataframe. The two dataframes were joint on the name of the areas and were sorted based on the Foursquare ratings of the venues.

In finding the similarity between areas in Zurich and areas of interest based on characteristics drown from local venues, the similarity can be seen as the dependent value and the characteristics of each area as independent values. At first, the problem sounds like a classification problem, that could be solved with a regression model. However, the data is not right for it, it is actually opposite. There is not enough data to train a model that will then make a prediction regarding classification, as only one area would have to train a model that would be used to predict a lot of areas. In this case, we can take a different route using an unsupervised machine learning algorithm that would cluster all the areas, then search what areas in Zurich are clustered together with the areas of interest. Preliminary data analysis shows that an area can

have up to 250 characteristics, which translates into 250 independent values and thus space dimensions. Keeping this in mind, the K-means method was chosen to be the clustering method.

K-means clustering

K-means clustering was done using a function from the sklearning library available for Python. Before the clustering, two things were done:

- Preparing the data to be used by the function by one hot encoding it
- Finding the ideal K value, the number of clusters

The one hot encoding was done using the integrated python function.

Finding the K value was done using the Elbow method. The elbow method runs K-means clustering for multiple k values and finding the sum of squared distances for each of them. Plotting the results, generally show an elbow in the graph distances vs k values. The graph shows decreasing sum of squared distances as K increases, meaning an improved performance of the algorithm. The elbow in the graph shows the point where increasing the value of doesn't has none to less impact in the distance; making that value of K optimal. The elbow method was run with K values in the 4 to 48 range. These numbers were calculated as the minimum and maximum number of clusters that could be useful in solving the problem. Finding the minimum and maximum value of K is a coloring problem. Minimum useful K was found as each of the 3 areas of interest would be colored differently and all of the areas in Zurich would be colored the same, except for one which be colored as the one of areas of interest. Maximum useful was found as each of the 3 areas of interest would be colored differently, just as each of the areas in Zurich, except 3, which have the same color as the areas of interest.

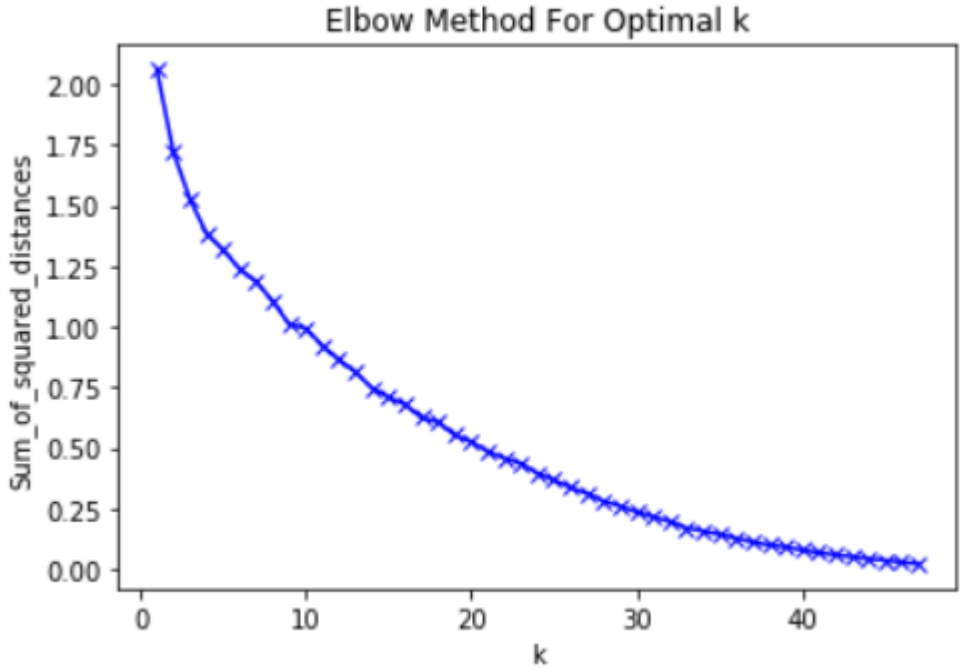

Figure 1: Elbow method graph

Two challenges occurred when making a decision on what K value to use in the K-means clustering. On one hand, as shown in Fig.1 the graph in the elbow method did not shown a clear elbow, making the method less useful. However, a K value of 30 was selected, that corelated to a distance of approximately 0.25. Unfortunately, the K-means clustering with a K value of 30 resulted in no areas in Zurich being clustered with the areas of interest, which meant no solution to the problem; this being the second challenge. Obviously, to have areas of interest being clustered with areas in Zurich, the number of clusters, the K value, needed to be decreased. Iterating with increasingly smaller values of K, areas of interest were clustered with areas in Zurich. Smaller values of K, raised the question of the accuracy of the model. However, a K value of 22, which was found to cluster together areas of Zurich with areas of interest, meant still a distance of under 0.5, which meant that the model was still sufficiently accurate.

Visualisation

Visualizing the solution was done by showcasing the clusters on a folium map of Zurich. Each area was visualized by the location of its center on the map by circle points, and color coded by their clusters. However, 22 clusters meant 22 colors, that were difficult to distinguish. To highlight the solution, the visualized points of the areas of Zurich that were similar to the areas of interest were circled with different shades of gray as the areas of interest respectively. The correlation between the shades of gray on the areas of Zurich and the areas of interest were explained in the legend of the map. The solution, of adding gray circles to the points and keeping the color coding of all areas of Zurich, was adopted because, the information regarding clustering of all areas of Zurich, regardless of the areas of interest was also deemed valuable.

4. Results

Results of the K-means clustering of all areas are shown in Table 1.

Table 1: K-means clustering results

Index	Area	Latitude	Longitude	ClusterLabel
0	ZRHd1_1	47.373055	8.538622	11
1	ZRHd1_2	47.371060	8.546883	11
2	ZRHd2_1	47.362536	8.531325	17
3	ZRHd2_2	47.355346	8.530505	12
4	ZRHd2_3	47.346413	8.528994	16
5	ZRHd2_4	47.338398	8.531827	8
6	ZRHd3_1	47.369357	8.519093	0
7	ZRHd3_2	47.362975	8.517507	17
8	ZRHd3_3	47.367453	8.506836	21
9	ZRHd3_4	47.374435	8.509835	0
10	ZRHd4_1	47.375985	8.527465	0
11	ZRHd4_2	47.380218	8.514601	0
12	ZRHd5_1	47.384446	8.530665	12
13	ZRHd5_2	47.388422	8.520878	12
14	ZRHd5_3	47.391177	8.511289	12
15	ZRHd6_1	47.381567	8.546769	17

16	ZRHd6_2	47.386617	8.540370	17
17	ZRHd6_3	47.386942	8.550189	17
18	ZRHd6_4	47.394008	8.539962	12
19	ZRHd6_5	47.400673	8.535893	21
20	ZRHd7_1	47.358830	8.589952	13
21	ZRHd7_2	47.365023	8.564416	10
22	ZRHd7_3	47.369526	8.555543	11
23	ZRHd7_4	47.377577	8.557842	15
24	ZRHd8_1	47.361644	8.551567	11
25	ZRHd8_2	47.356043	8.555821	17
26	ZRHd8_3	47.350862	8.563894	17
27	ZRHd8_4	47.352356	8.574684	7
28	ZRHd9_1	47.373830	8.494644	8
29	ZRHd9_2	47.383398	8.497749	17
30	ZRHd9_3	47.381534	8.485042	1
31	ZRHd9_4	47.389787	8.481952	1
32	ZRHd9_5	47.394815	8.490125	1
33	ZRHd10_1	47.393160	8.529010	12
34	ZRHd10_2	47.397340	8.513729	20
35	ZRHd10_3	47.402125	8.498570	19
36	ZRHd10_4	47.407262	8.488645	3
37	ZRHd10_5	47.408824	8.507229	6
38	ZRHd11_1	47.406606	8.548580	1
39	ZRHd11_2	47.414197	8.551506	1
40	ZRHd11_3	47.423918	8.546135	1
41	ZRHd11_4	47.415105	8.535854	1
42	ZRHd11_5	47.420329	8.506675	5
43	ZRHd12_1	47.408017	8.568017	2
44	Zollikon	47.339297	8.574556	14
45	Kusnacht	47.319856	8.584222	18
46	Kilchberg	47.325200	8.543901	4
47	Ruschlikon	47.310434	8.550874	9
48	South Ken	51.494780	-0.179104	11
49	Greenwich	40.733301	-74.003011	12
50	West Loop	41.886675	-87.656497	12

As mentioned in the methodology section, the results were visualized in a folium map. In Figure 2, the areas of Zurich can be seen as dots on the map, color coded based on their respective cluster. As described in the legend areas represented by circled dots with different shades of gray

highlight the areas of Zurich that were clustered together with areas of interest, showcasing the similarity between them.

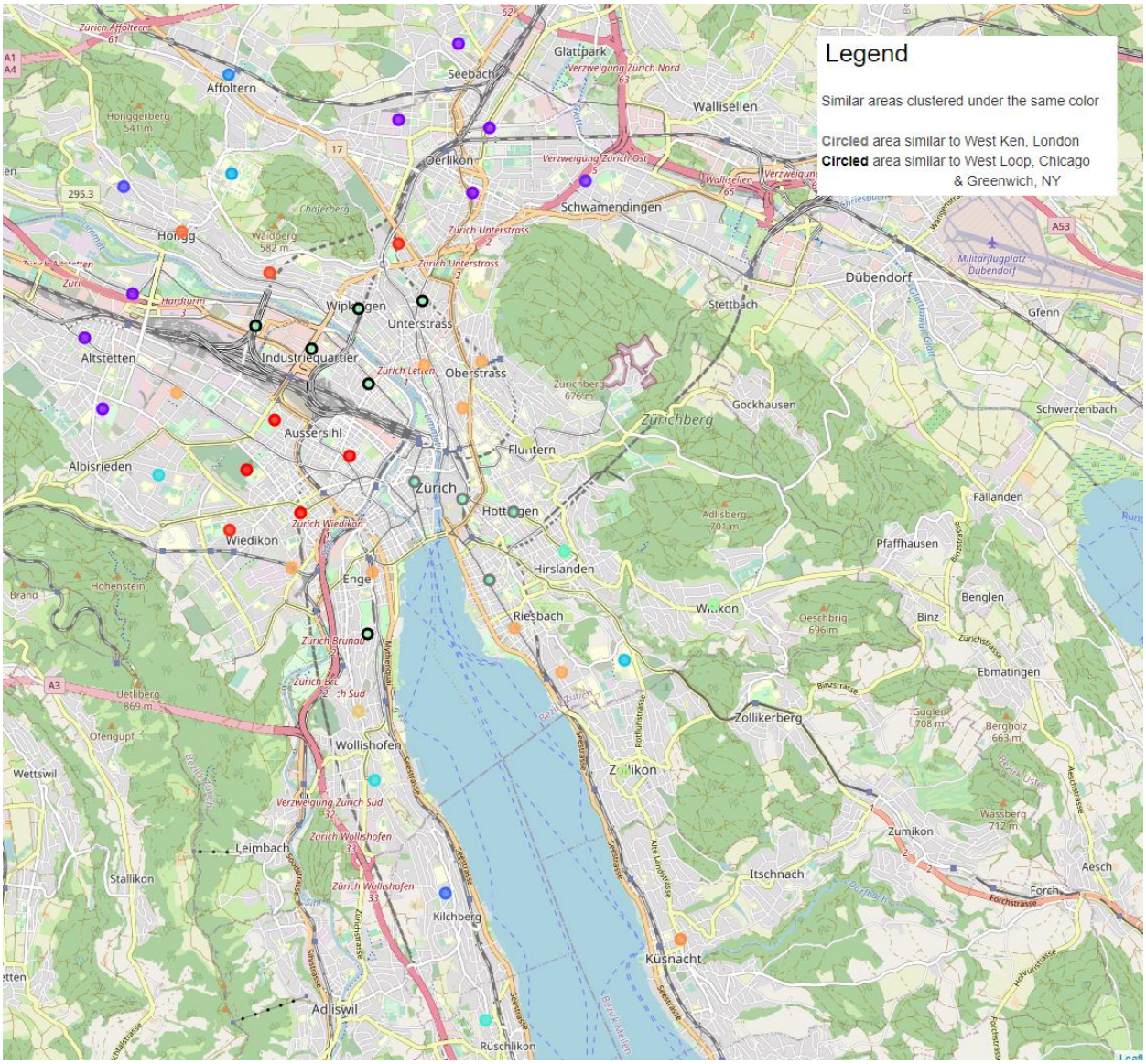

Figure 2: Folium map showcasing the K-means clustering of areas in Zurich by coded by color

From Table 1 and Figure 2, the following result can be drawn for the problem:

In Zurich, areas: ZRHd1_1, ZRHd1_2, ZRHd7_3, ZRHd8_1 are similar to West Kensington, London,

and areas: ZRHd2_2, ZRHd5_1, ZRHd5_2, ZRHd5_3, ZRHd6_4, ZRHd10_1 are similar to West Loop, Chicago and Greenwich, NY which were also clustered together.

5. Discussion

Data

Unfortunately, the data gathering regarding the location of the areas in Zurich was done manually, which is not ideal. The amount of data was relatively small, so it was not a problem, however, for a project that would require larger amounts of data, this could become an issue.

Based on the fast update rates that Foursquare is doing on their data, each quire most likely will return different data, which consequently will result in different results. One challenge regarding this issue was the different timezones of the areas that were analyzed. Chicago, CT (Central Time) and New York, ET (Eastern Time) are 7 and 6 hours away respectively, from CET (Central European Time), the timezone of Zurich. So, if the query was done in the CET morning hours, it would correspond to night time data regarding the North American areas, skewing the data. Optimally, two separate queries should have been done, to match the time of day. However, iterating showed that picking a query time in the CET late afternoon, corresponding to late morning CT and ET, would provide good results.

K – value

As mentioned in the methodology section, setting the value of K for the K-means clustering function presented challenges. The most important discussion point regarding the K value was the fact that the elbow method graph, did not show a clear elbow. No completely accurate explanation is known. However, the assumption being made is that the small and scarce amount

of data regarding some areas of Zurich had this effect; Especially as some areas of Zurich were drastically different from the areas of interest.

Visualization

Some challenges occurred while visualizing the results in a folium map. On one hand, the large number of clusters meant a large number of colors needed to color code the areas, making the colors almost indistinguishable. The colors also raised another unanswered question. What similar colors represent? If anything? Do clusters color coded with similar color show a similarity between the areas in the two clusters? Unfortunately, an answer to this question was not found, as more research is required. However, the assumption is that similar colored clusters show similarities between the areas in those clusters.

The folium map resulted in this project was shared with peers and some difficulties were found in interpreting the map. One reason found, was the multitude of colors and the increased similarity between some of the colors. Circling the areas that were clustered with the areas of interest, increased the readability of the map. However, changing the shape of the markers is assumed to have an even greater positive impact on readability. Unfortunately, after multiple tries, this was unsuccessful.

6. Conclusion

In this project, unsupervised machine learning, in the form of K-means clustering was used to solve the problem of finding similar areas in a city to specific 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. Visualizing what areas in Zurich were clustered together with the areas of interest showed that the solution to the problem

is: South Kensington -London was most similar to areas in the intersection of Districts 1, 7 and 8 (also known as Altstadt, Hochschulen, Seefeld, Muhlebach), while Greenwich – New York and West Loop – Chicago, which were clustered together, were most similar to areas of Districts 5 and 10 (also known as Escher Wyss, Industriequartier, Wipkingen).