Capstone Project Report

Segmentation and Clustering Budapest Neighborhoods

Péter Szakos

June 6., 2020

1. Introduction

1.1. Background

This report was created to document my Applied Data Science Capstone Project at coursera.com. Requirement for this assignment was to pick and solve a data science problem using location data.

1.2. The problem

The data science problem for this capstone project is to collect Budapest neighborhood data, cluster the neighborhoods based on their similarities, and present the results.

I chose this topic, because as a Hungarian citizen, I wanted to demonstrate what/how we can learn about the Hungarian capital city - using machine learning and location services data.

Getting to know a city is not always straightforward, especially when we talk about country capitals with doz ens of neighborhoods. Many times it can be useful to group similar neighborhoods, so one can have a better understanding what to expect when visiting certain parts of the city. In case of Budapest, there are more than 160 neighborhoods. It is a huge overhead to process and group them manually. On the other hand, it seems obvious to use location services data for the analysis, where people are already indicating what are the points of interests to them.

1.3. The audience

Several businesses can utilize such insights. For example, travel agencies could offer recommendations to their customers based on their preferences; real estate agencies could select top spots for new homes; city management could monitor citizens' preferences and allocate investment fundings accordingly.

2. Data collection and preparation

2.1. Data sources

For a list of Budapest neighborhoods, I was using the below WikiPedia article:
https://en.wikipedia.org/wiki/List_of_districts_in_Budapest

GPS coordinates for the neighborhoods were coming from the Nominatim database:
https://nominatim.openstreetmap.org/

For some neighborhoods, there was no Nominatim GPS data. For these, I used google maps together with terkepem.hu:
https://www.google.hu/maps
https://terkepem.hu/

Venues data was coming from the the FourSquare location service database:
https://foursquare.com/

2.2. Data wrangling

Throughout the capstone project I was using python code in a Jupyter notebook. You can check the code in my GitHub repository:

https://github.com/pszakos/IBM-Data-Science-Capstone-Project

As a start, I extracted the 166 neighborhood names from the WikiPedia article. The article didn't contain neighborhood coordinates, so I collected the missing GPS info from the Nominatim (+ Google and terkepem.hu) database. Once I got the negihborhood details, I used FourSquare location service API to collect the top 100 venues for each neighborhood. With this, I got to know the top 10 venue categories per neighborhoods, as shown in Table 2.1. below.

	Neighborhood	Latitude	Longitude	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Tabán	47.490893	19.042639	2	Park	Pub	Lounge	Trail	Hungarian Restaurant
1	Gellérthegy	47.484565	19.038769	0	Bus Stop	Garden	Playground	Lawyer	Ice Cream Shop
2	Krisztinaváros	47.496865	19.029776	2	Café	Bistro	Bakery	Playground	Park
3	Adyliget	47.547550	18.938984	0	Pharmacy	Bus Stop	Park	Pizza Place	Playground
4	Budakeszierdő	47.510273	18.951182	3	Train Station	Mountain	Yoga Studio	Food Court	Fish Market
...	...	...	...	...	...	...	...	...	...

Table 2.1.: dataframe piece showing top venues per neighborhood

3. Data processing

In order to segment the neighborhoods, I implemented k-means clustering. Originally I planned to have five clusters, but even with four clusters, there was a cluster with only one neighborhood, so I decided to run k-means only for three clusters. In the end, I actually ended up with four clusters, because there were some neighborhoods that didn't have any FourSquare venues. These became the fourth cluster. In order to have some meaningful interpretation I also performed wordcloud analysis on the clusters. You can check the results in the next section.

4. Exploratory data analysis and results

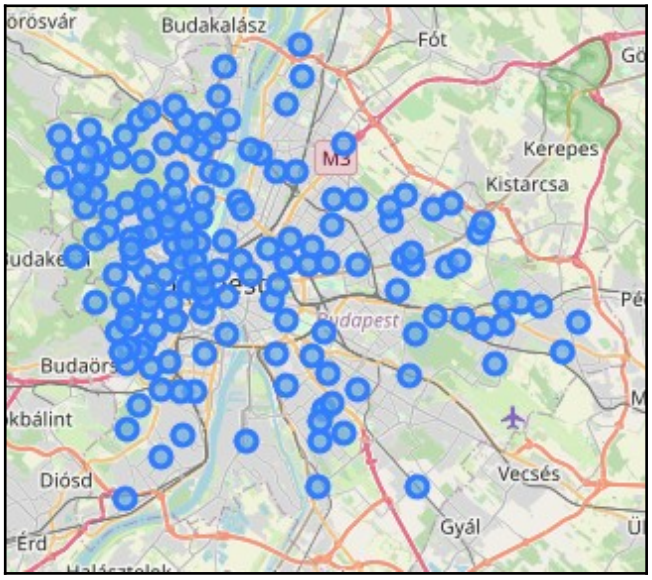

Figure 4.1.: Map of Budapest neighborhoods (sources: WikiPedia, Nominatim)

Figure 4.1. shows the neighborhoods I collected from the WikiPedia article and the Nominatim database. In the next sections, I will share the results of the clustering analysis.

Neighborhoods in general

Figure 4.2.: WordCloud for all clusters and all clusters on a single map

Figure 4.2. shows that FourSquare users find Budapest generally strong in (flea-, fish- and flower) markets, but the city appears to be also the foodies' favourite, regardless of the visited neighborhood.

Cluster 0 - Vibrant, frequented areas

Figure 4.2.: WordCloud and map for Cluster 0

On Figure 4.2. we can see that Cluster 0 is for frequented, vibrant Budapest areas. Markets and restaurants are dominant, but we can also see a colorful selection of venues eg. bars, cafés, drink shops etc.

Cluster 1 - Quiet spots

Figure 4.3. WordCloud and map for Cluster 1

In Cluster 1, there are much less neighborhoods in the city center or near to a major road. Please note the densitiy of neighborhods on the left is higher. This is Buda hill, which offers plenty of natural sceneries. There are a still a lot of venue types to chose from. However, markets rstaurants etc. are less dominant. With more green areas, you can find some quiet spots in the city noise.

Cluster 2 - Commuters' outposts

Figure 4.4. WordCloud and map for Cluster 2

Cluster 2 apparently includes neighborhoods from the outer parts of the city. The venue types suggest that people can collect what's needed for home or do some workout after work hours. Therefore these areas seem to be the target for commuters.

Cluster 3 - Infrequent areas with no FourSquare Data

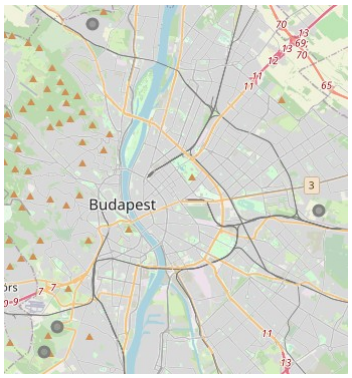

Figure 4.4. Map of neighborhoods without FourSquare data

Neighborhoods on Figure 4.4. are apparently of no interest to FourSquare users, in this meaning these locations were excluded from the analysis.

5. Discussion and conclusion

We can say that the findings are in alignment what I know about / expect from Budapest. The number of clusters can be further increased to see if more insights can be gained. However, the presented clusters demostrated what /how can be extracted from location data with the use of machine learning. In this meaning the project accomplished its target.

Thank you for your interest in this capstone project and reading this report!