18.1.2021

Small Czech towns clustering

Coursera Capstone

Jaroslav Tuma

Introduction

This capstone project for coursera course. Assignment is to use Foursquare API to create something interesting. But here is a BIG problem with this assignment:

1) Foursquare API isn't mentioned to return you such data, we were trying to get in previous parts of this course. As result you will get only limited number of venues in all calls.
2) In Foursquare policy is point we are breaking:

   > You shall not use any automated means (for example scraping or robots)

But I have still tried to think out some conditions that would make it possible to get some meaningful results. The biggest problem is that there is maximum results, you can get in one query. So I decided to carry out survey of smaller towns (population 35 000 – 200 000) in Czech Republic so this limitation won't be such issue.

Problem definition

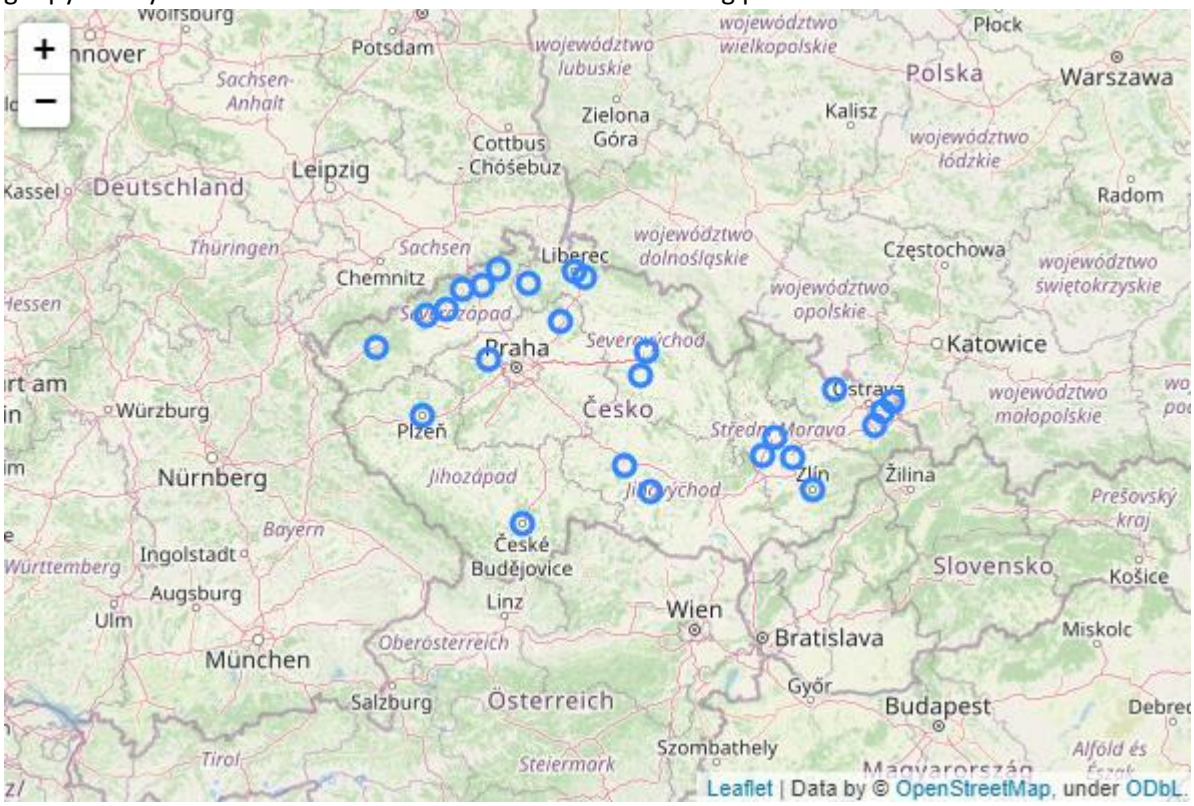
Let's say, that someone has several job offers in different towns. Those offers are almost the same, but each is in different city so he will have to move. He loves his city, so he wants to move to some similar place. The task is to find out what smaller cities in Czech Republic are similar?

Data

We can easily find list of Czech cities and their population on Wikipedia. We will find venues in those cities using Foursquare. With foursquare API we can get several information about each venue, but we will work only with category of venue. We will need count of venues in each category in the city.

Data preparation

I have used pandas to read the Wikipedia table and dropped all unnecessary columns. Thanks to geopy library I was able to find location of all cities. Following picture shows cities I have used.



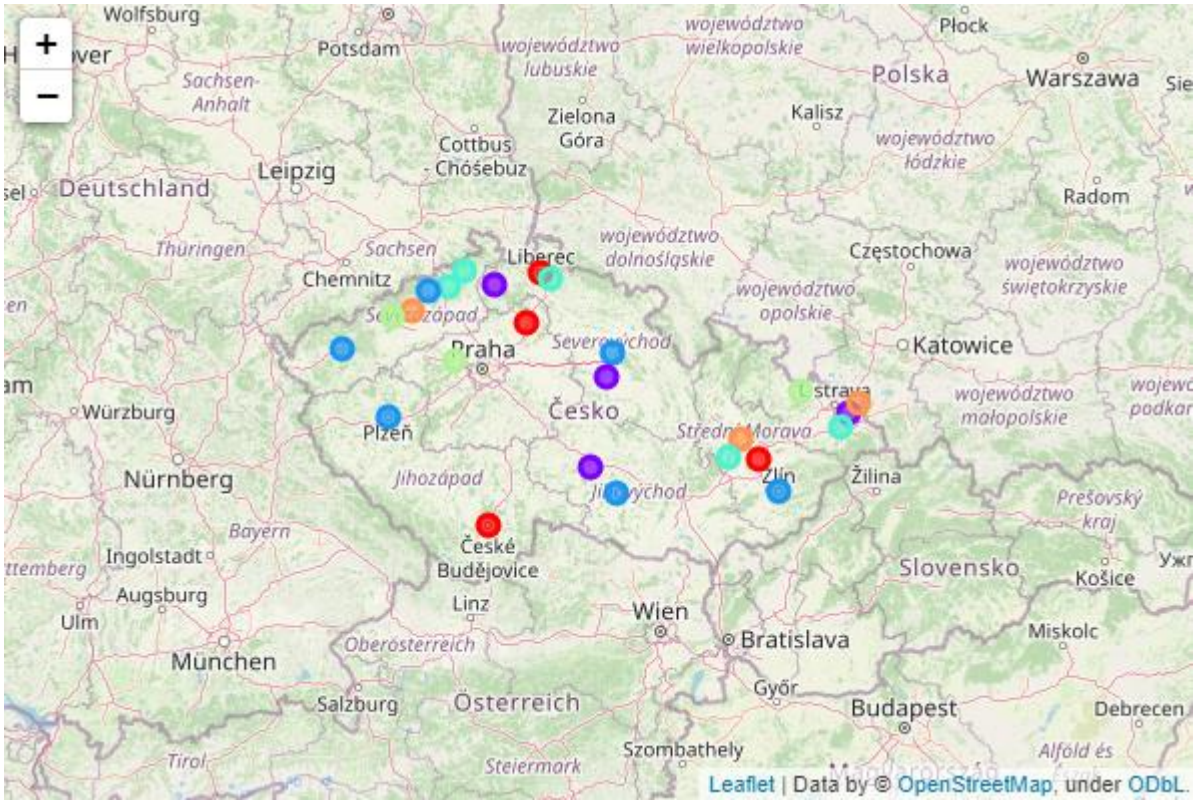
Then I have use foursquare API to get as many venues in all cities as possible. I have extracted information about category of each venue and created pivot table containing count of venues in each category for each city.

Methodology

I have clustered cities by number of venues of certain category in each city. To categorize cities, I have use kmeans algorithm from Scikit learn library. Kmeans algorithm have sorted cities into 6 clusters.

Results

To show results of clustering, I have created a map in Folium with different colour for each cluster.



Discussion

Unfortunately, data about venues you can get from Foursquare are not good enough. As I have said in the beginning this is because Foursquare API is not intended for such job. In this case there is a lot of venues missing, so results are very misleading. As solution I would suggest use different source of those data.

Conclusion

For me this part of course, which I otherwise liked a lot, was disappointing. I could think a lot of useful machine learning applications, but not if I must use Foursquare. This API is intended to do absolutely different job, so data you can get from it will always be incomplete for such analysis. All you can do is somehow alternate the example from earlier lessons and I am not interested in that at all. At least not this way.