Business problem

In my final project I want to explore the Bucharest and Cluj city neighbourhoods. I have formed following task: suppose that I am representing a consulting agency that helps businesses and individuals to find right location for their business and living. I had an order from a person who wants to relocate from Bucharest (the Capital of Romania) to Cluj (also in Romania) and want to find relevant location for buying an apartment. The best location for him will be place that is maximum similar with his Bucharest apartment neighborhoods (Baneasa - district of North-of the city of Bucharest). To solve this task we have analyzed the North regions of Bucharest city and all administrative districts of Cluj city. For Bucharest I decided to take the metro stations as control points of the analysis, since the metro is the basis for the city's general transport infrastructure and provides accessibility to all parts of the city. For Cluj as a control points I have chosen the centers of administrative regions. In this capstone project I have gone with these 5 steps: 1 Explore the neighborhoods of south of Bucharest and all neighborhoods of Cluj (Foursquare API) 2 Join this datasets each other, 3 With k-means clustering method divide explored areas to 5 clusters, 4 Find region in Cluj city that includes in the same cluster with Bucharest apartment of our client.

Data
For our analyze we need data from two cities – Bucharest and Cluj. Bucharest data: As we put emphasis on metro station locations and neighborhoods, we need data about all stations of Bucharest. This type of information we can get from Wikipedia page (https://en.wikipedia.org/wiki/List_of_Bucharest_metro_stations) with “Beautiful Soup”. After gathering information we will clean it and filter stations that included in North regions. Cluj data: Information about Cluj administrative regions we also can gather from Wikipedia page. My next step will be to find correct locations of chosen points with Google geocoding API. Next I will join to each other these two datasets and with Foursquare API find all neighbourhood venues to each point of our dataset. Next I will use K-means clustering method to divide dataset to clusters and find region in Cluj that includes in the same cluster with Bucharest apartment of our client.

Methodology

To compare the similarities of two cities, I decided to explore neighborhoods, segment them, and group them into clusters to find similar neighborhoods in a city like Bucharest and Cluj. To be able to do that, we need to cluster data which is a form of unsupervised machine learning.

We will firstly find correct locations of chosen points with Google geocoding API. Next we will explore neighborhoods in both cities on Foursquare API and get the most common venue categories in each neighborhood.

Next we will use K-means clustering method to divide dataset to clusters and find region in Cluj that includes in the same cluster with Bucharest apartment of our client.