II. Data

In order to solve the business problem described previously, I am going to collect and analyze the following data:

- Names of districts in Prague. The data is collected by parsing the *Wikipedia* page of Prague using the *Beautiful Soup* Python package.

- Geographic location of districts. The data is collected using the *Nominatim* Python library to find the geographic coordinates by name and address (geocoding).

- Name, category and geographic location of the existing food venues in Prague. The data can be retrieved with *Foursquare* - a location data provider with information about venues and events within an area of interest that can be obtained through the API. The Foursquare uses a defined search radius to sweep and discover the restaurants within its range, however the response is limited to 100 venues per call, therefore the search location need to be spitted in smaller portions.

I am going to start with collection of names of the districts in Prague, then I can find out the geographical location of their centers using *Nominatim.* Having those, I can us *Foursquare* to explore the food venues at given coordinates within 1.5km search radius.

Based on the collected data, I have sufficient information to solve the business case. I will cluster the locations of existing restaurants and food venues using DBSCAN data clustering algorithm to find out where are the popular spots for eating out. Afterwards, I am going to find out in which clusters the pizza restaurants are underrepresented. That will allow me to choose a popular spots that is not occupied by the competition. Using this results, our stakeholder can take the necessary decision about opening the pizza restaurant.