Data description

The data needed for this project would be a list of all Berlin metro stations. To perform a cluster analysis on this data additional information about the characteristics of each station is necessary. These characteristics might be the types of surrounding venues in the neighborhood of the metro station which are representing the possible destinations for travelers arriving at this station. For example, a metro station which has a lot of office buildings and cafes in its area could be interpreted as business category where a lot of commuting traffic is to be expected.

For this, a specific radius around each station must be defined in the analysis, in which the venues are lying. To get a list of venues around a specific location, Foursquare API can be used. Foursquare is a company which provides location-based data to allow users to explore their vicinity. It provides an API, whose ‘explore’-function can be used for this purpose. This function needs to get passed the API credentials of the user, the area where to search (given by coordinates and a radius) and a maximum limit of venues to return. The response of the API call is a .json file with the information about the venues like name, category and exact location.

In conclusion, the dataset needed to perform a cluster analysis would be a list of all metro stations in Berlin with their respective geographical coordinates and the venues lying in the defined area around the station. The list of metro stations can be retrieved from the corresponding Wikipedia page. The coordinates are found on this page, too, but they needed to be cleaned and modified to be passed to the API function. The venues for each station are responded by the Foursquare API call.