NEIGHBOURHOOD CULTURAL OPPORTUNITIES INDEX

COURSERA - IBM Applied Data Science Course

Peer-graded Assignment:
Capstone Project - The Battle of Neighborhoods

September 2019

By
Anatolie Poiata

ABSTRACT

This report describe the realization of the Capstone project “The Battle of Neighborhoods” of Coursera course “IBM Applied Data Science Course.

The report consist of two parts, corresponding to 2 weeks.
Part 1:

- A description of the problem and a discussion of the background.
- A description of the data and how it will be used to solve the problem.

Part 2:

- Methodology section which represents the main component of the report where is described exploratory data analysis applied, and used machine learnings methods
- Results section with discussion of the results
- Discussion section with observation during execution and recommendations based on results
- Conclusion section

The final report in Notebook format, with code and data can be found on Github repository at

https://github.com/AnatolPoiata/Capstone-Report

1 INTRODUCTION

Business Understanding

More and more cities trying to identify and evaluate their cultural potential in order to attract more tourists and visitors. Historically, in the same city, different neighborhoods have different number of cultural objects. Some of these attractions are popular, some are not. Thanks to social networks and mobile devices, we know where and when people are active within neighborhoods.
In 2005 by Toronto City Council addopted Toronto Strong Neighbourhoods Strategy (TSNS) 2020. The City uses 140 social planning neighbourhoods for designing programs and services. Each of the City's social planning neighbourhood has at least 7,000 residents and respects important boundaries such as major roads and rivers. More information about these neighbourhoods at http://www.toronto.ca/demographics/neighbourhoods.htm.
The scope of this report is to help to identify neighborhoods with lowest rating of cultural attractions and to recommend a list of these Priority Neighbourhood Areas for Investment (PNIs). The rating of the neighborhoods will be based on number of venues in the neighborhoods per capita, as well as the popularity of venues, according to data on Foursquare. Rated neighborhoods will be clustered and the recommendation for investment will be done for clusters where neighborhoods are close located, as new attractions affect not only neighborhood where is located, but also surrounding neighborhoods.

Analytic Approach

According to the question formulated above we should organize neighborhoods into similar groups based on their ranking scores of cultural potential and popularity. This is a typical K-meaning clustering problem of machine learning. It examines the entire set of interdependent relationships to discover the similarity relationships between the objects in order to identify the clusters.

2 DATA SOURCE

2.1. TORONTO NEIGHBOURHOOD

Neighbourhood refers to the City of Toronto's 140 social planning neighbourhoods. The boundaries of these social planning neighbourhoods are described in GEOJson format and can be uploaded at

https://open.toronto.ca/dataset/neighbourhoods/

Data are located also at
https://github.com/AnatolPoiata/Capstone-Report/blob/master/Neighbourhoods.geojson

2.2. TORONTO NEIGHBOURHOOD PROFILES

In these profiles, "neighbourhood" refers to the City of Toronto's 140 social planning neighbourhoods. The boundaries of these social planning neighbourhoods are consistent over time, allowing for comparison between Census years. We will use the most recent data which refer to 2016, limited to census data . As we will use index calculated as number of objects per capita, only data regarding population will be used.
Data can be uploaded at
https://open.toronto.ca/dataset/neighbourhood-profiles/

Data in .csv format are located also at
https://github.com/AnatolPoiata/Capstone-Report/blob/master/neighbourhood-profiles-2016-csv.csv

Neighbourhoods from mentioned above databases can be linked by common key Neighbourhood Number

2.3. ATTRACTIONS

The data for attractions will be used from Foursquare API. As we are interested in cultural attractions we will use only 3 categories of venues
- Arts & Entertainment
- Event
- Outdoor & Recreation

and 50 subcategories

For each neigbourhood will be identify the list of attractions from Foursquare, located in this neighbourhood.

3 ACQUIRE DATA

3.1 Data Collection

Information about profiles of neighbourhood are stored at https://open.toronto.ca/dataset/neighbourhood-profiles/.

Each data point in this file is presented for the City's 140 neighbourhoods, as well as for the City of Toronto as a whole. The data is sourced from a number of Census tables released by Statistics Canada. The general Census Profile is the main source table for this data, but other Census tables have also been used to provide additional information.

Data are stored in .csv, .json and .xml format. We will use csv format.

The list of neighbourhoods with Boundaries of City of Toronto Neighbourhoods can be found at https://open.toronto.ca/dataset/neighbourhoods/. They are in GeoJSON format, so can be easy converted to Python DataFrame.

The data on venues will be acquire from Foursquare. For each neighbourhood we will use its latitude and longitude, and will search for venues in the radius of 1500 m, as the neighbourhoods are not large than 3 km in diameter. As I am using Personal account, there is a limit of 50 venues per request.

3.2. Data Preparation

The data of neighbourhoods with boundaries from GeoJSON file will be imported to **neighbourhoods dataframe**.

The data of neighbourhoods profile, in csv format will be imported in **n_profile dataframe**.

The neighbourhoods profiles and their boundaries are located in different dataframes, but can be ease joined as they have a common field - Neighbourhood ID. The result will be stored in **neighbourhoods dataframe** with a column of population number for each neighbourhood**.**

For each neighbourhood we will have up to 50 venues from specified categories. Venues are selected from a circle with diameter of 1500 m, but the shapes of neighbourhoods on the map are different. So, after requesting list of venues we have to exclude duplicates in the list of venues. The total number of venues in specified categories - 1792, number of unique venues - 1702.

Result dataframe will be **all_venues dataframe**.

For the next steps we have to associate venues and neighbourhoods. We will use **Mathplot** function **contains_points** to check if venues coordinates are in inside the polygon in boundaries of neighbourhoods. If point is inside the path, the Neighbourhood ID will be indicated.

The result dataframe of venues with associated Neighbourhood ID wil be grouped by Neighbourhood ID. In dataframe neighbourhoods ww will add a column with number of venues in this neighbourhood. It is possible that some neighbourhoods does not contain any venues. This missing data will be replace by 0.
In order to normalize data we will use not number of venues, but number of venues per capita. This index will be calculated as **number of venues / population.**

The result will be stored in in **nb_venues dataframe.**

4 EXPLORATORY DATA ANALYSIS

For the beginning we will plot the boundaries of neighbourhoods and marks corresponding to venues.

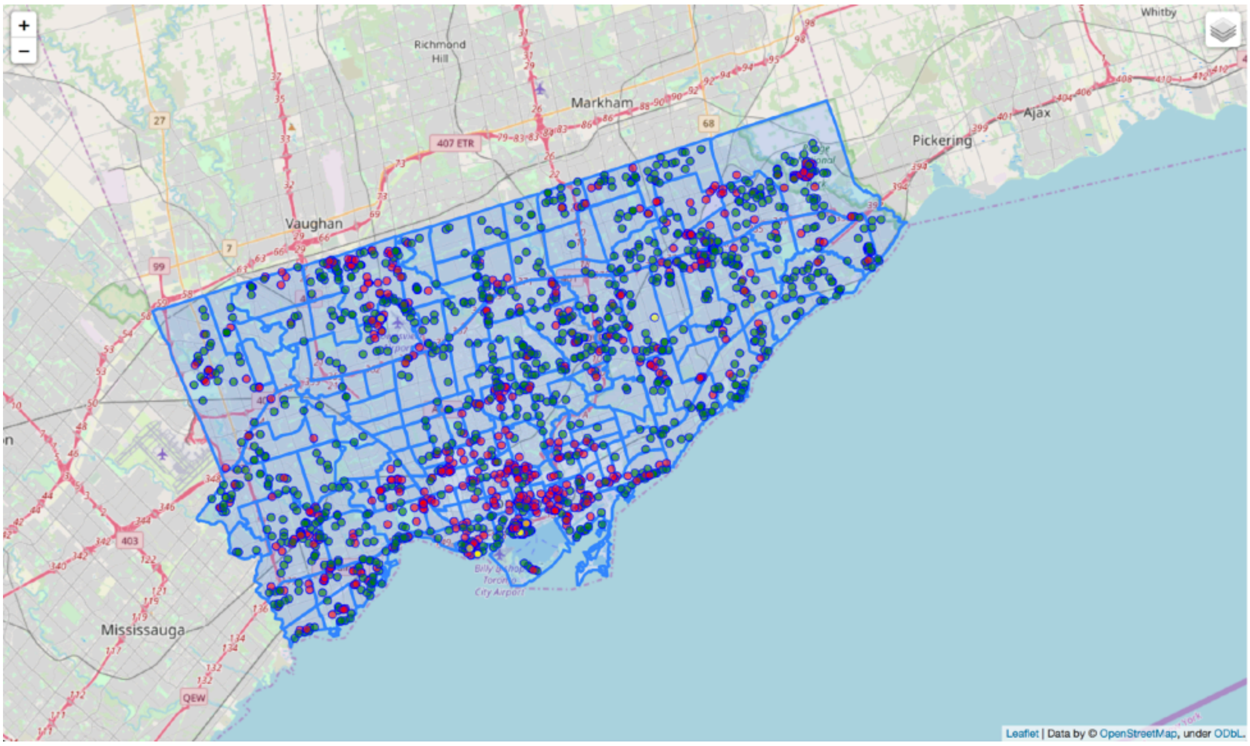

+
−
Whitby
Richmond Hill
Markham
Pickering
Ajax
Vaughan
407 ETR
Mississauga
QEW
Billy Bishop Toronto City Airport
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 1. Boundaries of City of Toronto Neighbourhoods and venues

As we can see the density of venues differ from neighbour to neighbour. Some of them have no any venues of specified categories.

Plot a box plot with statistic representation of the the distribution of the data through five main dimensions - Minimum, First quartile (middle number between the minimum and the median, Second quartile (Median), Third quartile (Middle number between median and maximum, Maximum. *(See Figure 2)*.
According to the result there are some neighbourhoods outliers and majority are distribute between minimum and median.
At the same time it’s confirmed that the minimum is 0, that’s mean there are several neighbourhoods without any venues.

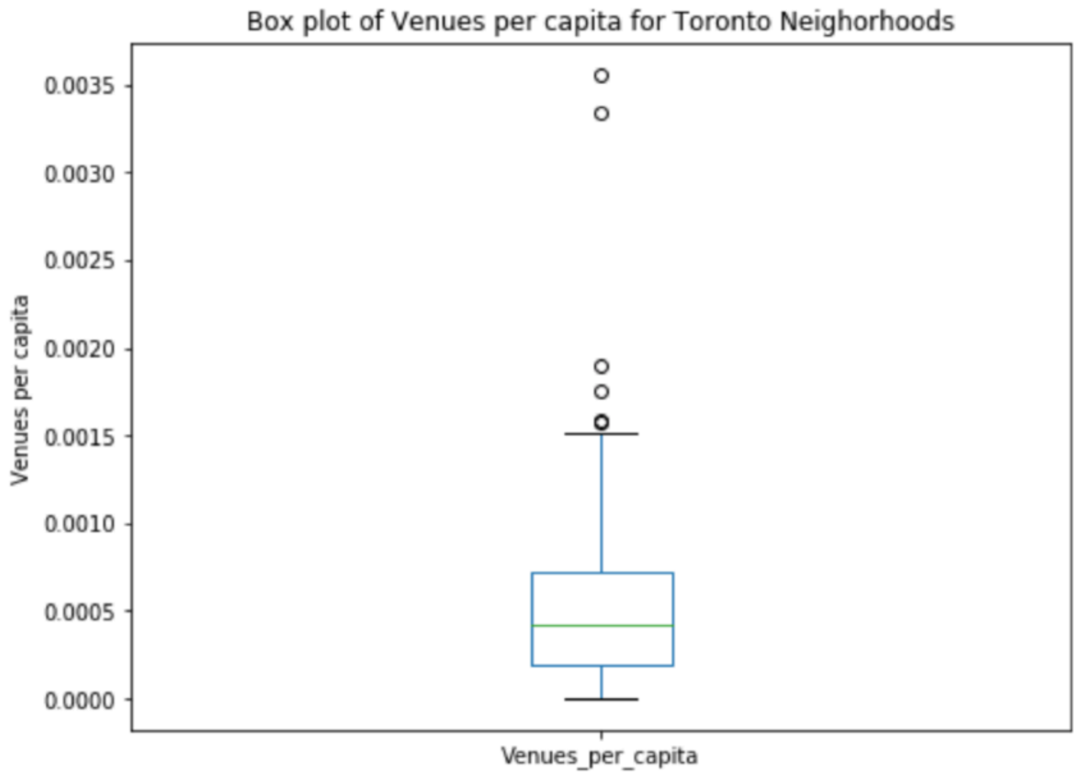


Figure 2. Box plot with statistic representation of the distribution of the data

Check the histograme for distribuition of number of venues per capita for neighbourhoods.

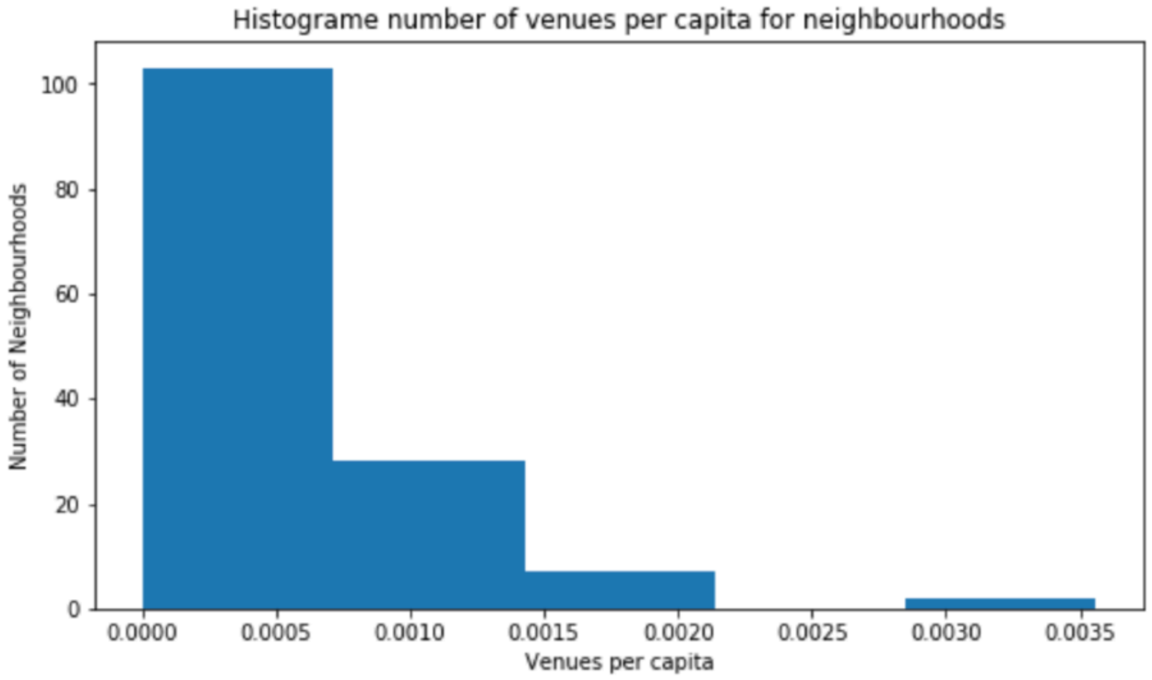


Figure 3. Histograme for number of venues per capita for neighbourhoods

According to the histogram, the majority (100 from 140 neighbourhoods) have a low index of venues per capita, in limits 0.0000 - 0.0007. At the same time there are some neighbourhoods with high index (0.0030 - 0.0035)

Use scatter plot to compare venues per capita ie each selected category. For a better presentation the inxed of venues per capita is shown x1000

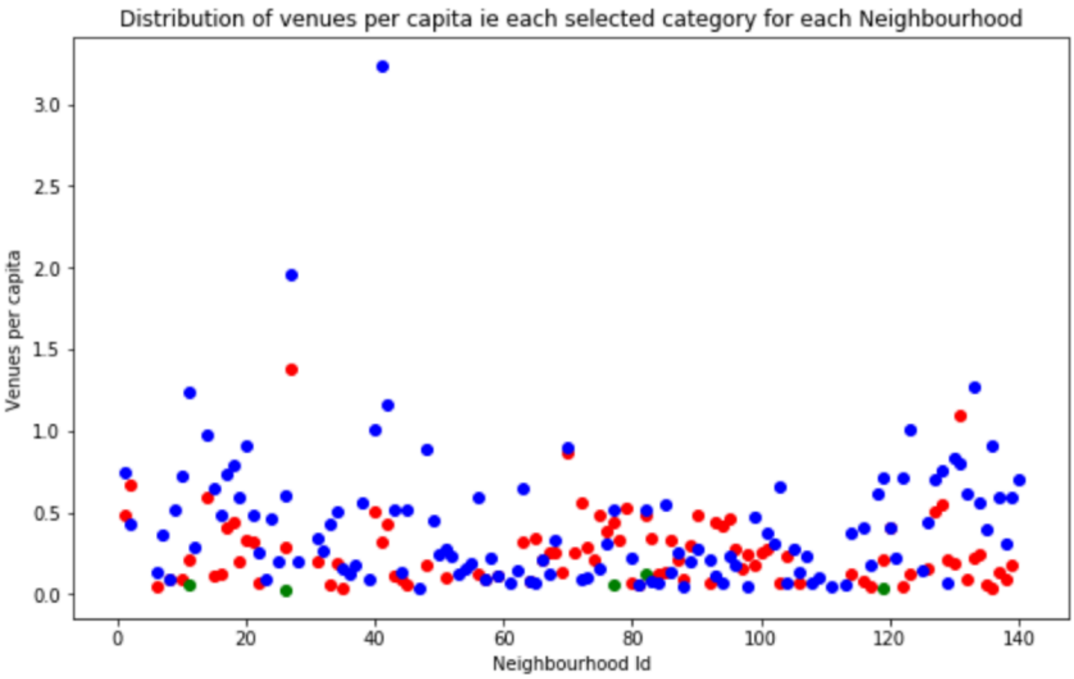
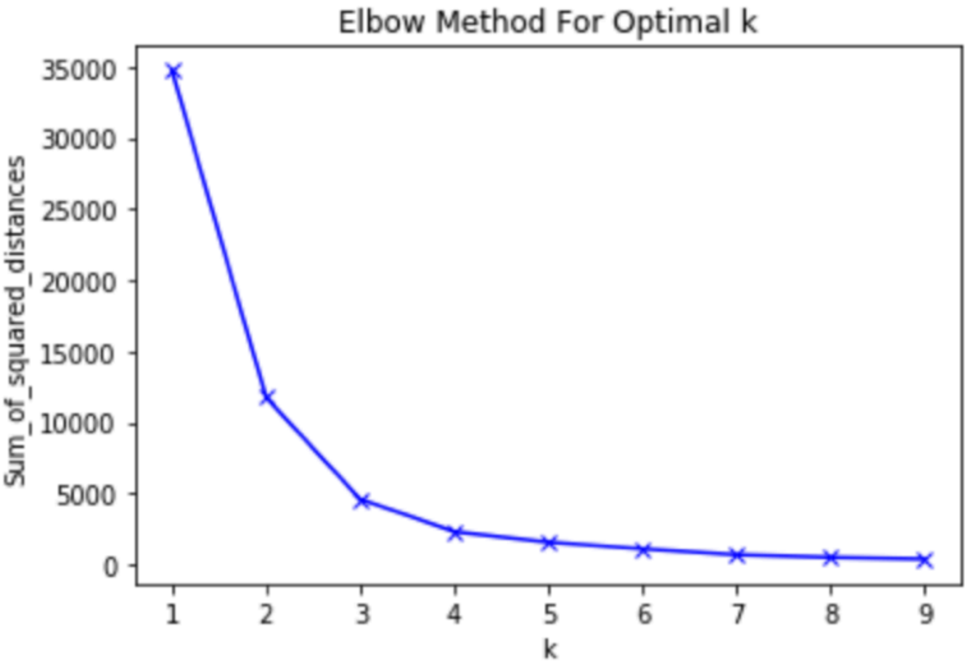


Figure 4. Distribution of venues per capita in each selected category

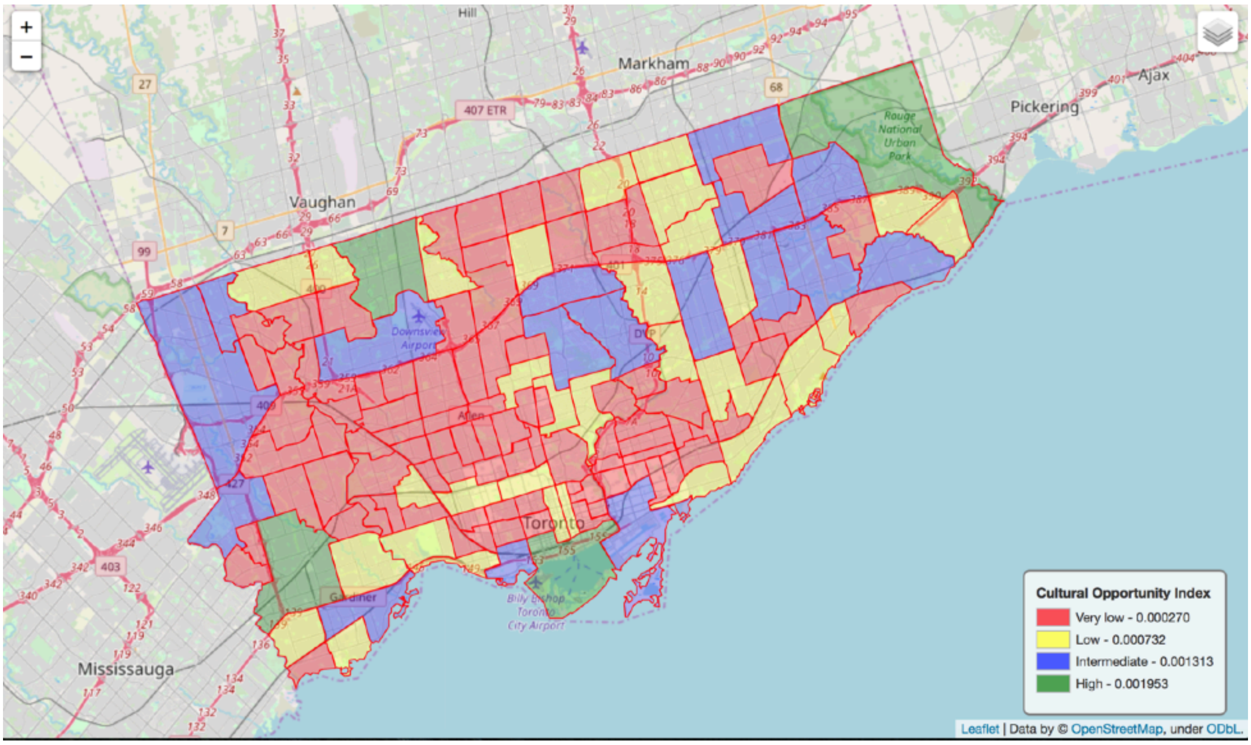
The above plot show that the Event category (green color) are practically missed from majority of neighbourhoods. The Outdoor&Recreation category (blue) are dominated and in some neighbourhoods the index of venues per capita is quite high.

It's clear that are several groups of neighbourhoods. We will use k-meaning clustering to group them. The optimal number of clusters will be determine by Elbow Method For Optimal k.



The optimal value of k is 4

Each neighbourhoodwill be associated to a cluster and a result map of neighbourhoods, colored in the color of respective cluster will be plot.


Markham
Vaughan
Pickering
Ajax
Rouge National Urban Park
Downsview Airport
Toronto
Mississauga
Billy Bishop Toronto City Airport
Cultural Opportunity Index
🟥 Very low - 0.000270
🟨 Low - 0.000732
🟦 Intermediate - 0.001313
🟩 High - 0.001953
Leaflet | Data by © OpenStreetMap, under ODbL.

With red are marked neighbourhoods with low index of cultural opportunity index per capita.

5 CONCLUSION

The “Battle” of Neighborhoods of Toronto, based on Neighbourhood Cultural Opportunity Index, show the following results:

1. Not all neighbourhoods are equal developed and cannot offer cultural activities at the same level.
2. There 4 from 140 neighbourhoods with quite high index of cultural opportunity. As a tourist I will select this neighbourhoods to explore more venues in a shorter time.
3. Neighbourhoods with low index are positioned relatively compacted and construction of new amenities in these regions can positively influenced the entire area, not only neighbourhood.
4. The number of Events are quite low, so It’s recommended to develop and promote Events, especial in neighbourhoods with low index.

6 FUTURE DIRECTIONS

The selected Neighbourhood Cultural Opportunity Index can be used as a subindex for a more complex research, for example in Cultural and Creative Cities Monitor or Quality of Life Index. The set of categories and relationshisp can be extended. Another direction of development can be including in the model the ratings of venues, based on Foursquare database, as well as number of visitors of the venues. This will offer the possibility to estimate not only opportunities, but also the popularity of venues and to calculate more exact the impact of venues on the life of inhabitants of the neighbourhood.