Relocation Analysis Tool

# Introduction and problem statement

When preparing a relocation to another city, a usual difficulty is to get some visibility on the differents neighbourhood in destination city.

Services already exists to get some insights on what is available in the different neighborhood.

As an international relocation service company, the objective will however be to offer a personalised analysis of the neighbourhoods at destination.

* Based on the current location the customer is living, a neighborhood profile is mapped.
* According to destination and area of search defined by the customer, different neighborhoods of the destination city will then be analysed:
  + Closest neighborhood, in terms of venues profiles will be identified.
  + On top, destination's neighborhoods are grouped in clusters to provide the customer with other suggestions, analysed by comparison to this reference.
* On top of the overview analysis, tool should allow to refine the search to particular categories of venues.

The relocation service company being working worldwide, the tool needs to be made in a generic way, that can be executed from any city to any city.

# Scenario

Scenario of the analysis will be the following: A customer is moving from Singapore to Toronto and is looking for a location in Toronto. He likes the environmnent he has and first would like to know if there is some neighbourhood with the same overall feeling. His first search will be based on Toronto Central District, where his company is located. The company needs to provide him with some insights on Toronto, in a personalised way, meaning compared to his current neighbourhood. Company has to:

* Identify the most similar neighbourhood in Toronto
* Provide additional suggestions on similar neighbourhoods
* Provide Additional insights, related to Arts & Entertainment that the customer already mentioned, along as being prepared for other extensive analyses on different aspects (eg. Nightlife). Especially, regarding Art & Entertainment, customer is not really satisfied with what the current neighborhood offers and is opened to new suggestions

According to finding, suggest a neighbourhood for the customer

# Data overview

To reach the business objectives, we will use the following data:

Current address of customer (manually decided). Either latitude and longitude, or only an adress and using <https://nominatim.openstreetmap.org/search> to get geocoordinates afterward.

Future area to relocate, defined by it's center and a radius of search.

In order to be as generic as possible, it will not be based on external data, but instead we will be using a manually defined grid.

For resource constraints, this area will then be divided in a regular grid of ~30 "neighborhoods", the size being adapted to the initial radius defined.

On top, foursquare APIs will be used to capture:

Recommended venues around the current address of customer in order to get a reference neighboorhood profile

Recommended venues on the different districts of destination area

# Methodology

The data analysis will be performed as follows:

A - Data Preparation:

* Based on a few manual entries ( Reference/Current negihborhood (address or geocoordinates); Future area to explore (center + radius) )
* The tool will automatically build a grid of neighbourhoods, numbered and located on a map.
* Using foursquare API, the tool will then gather venues around the reference neighbourhood, and future neighborhoods. Neighbourhoods with too low number of venues (for example, it can correspond to areas off shore) will be filtered out.

B - Closest neighbourhood identification

* Tool will then evaluate the difference between future and reference neighborhood, in order to proceed to comparison
* Based on this difference result, the 'closest' future neighborhood will be identified.
  + This neighbourhood is located on the map.
  + A comparison summary is provided to the customer.

C - Neighbourhood clusterisation

* On top, a clusterisation is done on the neighbourhoods
  + The choice on how to proceed to clusterisation is made either using the elbow method when possible, or through a generic manual selection.
  + This clusterisation allows to identify other possible suggestion to the customers
  + The clusterisation is displayed on the map
  + A comparison profile of the different clusters is also provided to the customer

D - Extended analysis: focus on Art & Entertainment categories

* Additionally, steps B and C are performed again, focusing on Art & Entertainment venues.
  + This provides additional insights on the neighborhood feeling.

E - Recommendation

* Based on above studies, a recommendation is provided to the customers, along as possible next steps depending on his feedbacks after this first analysis.

# Data preparation

Geocoordinates:  
Based on manually entered address, the tool determines the latitude and longitude of the reference and future points. This is done by using <https://nominatim.openstreetmap.org/search> api.

Grid construction:

This part if inspired from example made available on coursera: <https://cocl.us/coursera_capstone_notebook>  
  
The code is enhanced to generalize to any place, not only Berlin. This was done using the following documentation on utm zone system:  
<https://www.gislounge.com/how-to-calculate-the-boundaries-of-an-utm-zone/>

The code is also modified to use the radius defined by the user and adapt the grid according to it, keeping 30 Neighborhoods.  
Main advantage of this approach is to be made independent from the city and external data. This preserves from a lot of data cleaning.

Gathering data:

Venues data is gathered using foursquare API.  
The search is done using the explore API to get the most recommended places (resource constraint suggesting to focus only on main places) on a specified radius, and focusing on particular venue categories. Categories are retrieved using the corresponding foursquare API, and by mapping all sub-categories with the master category, for example ‘Arts & Entertainment’. This step is done because master categories all have the same categoryId: ‘Food’ cannot be differentiated from ‘Art’directly.  
Neighbourhoods with small number of venues are dropped. They can correspond to areas over sea for example.

Organizing data:

Data are organised in different dataframes:

* Dataframe to store geocoordinates
* Dataframe to store venues detail
* Dataframe to get statistical profile of each neighbourhood. This is done by grouping the venue categories per neighbourhood, and apply mean() function to get the frequency of each category type
* Dataframe used for manual observation of top venues by neighborhoods

To allow comparison between reference and future neighborhood, the structure of the 2 ‘grouped’ dataframes are aligned. Once done, the difference of frequency for each venue category is calculated; and this will be used for further analysis.

# Exploratory analysis

## General analysis

Based on data gathered, cleaned and organized as per previous step, the following analyses are performed:

Determination of closest neighborhood

Based on least square method, the distance of each neighborhood is calculated, closest neighborhood is determined, and the neighborhood is located: Neighborhood 20

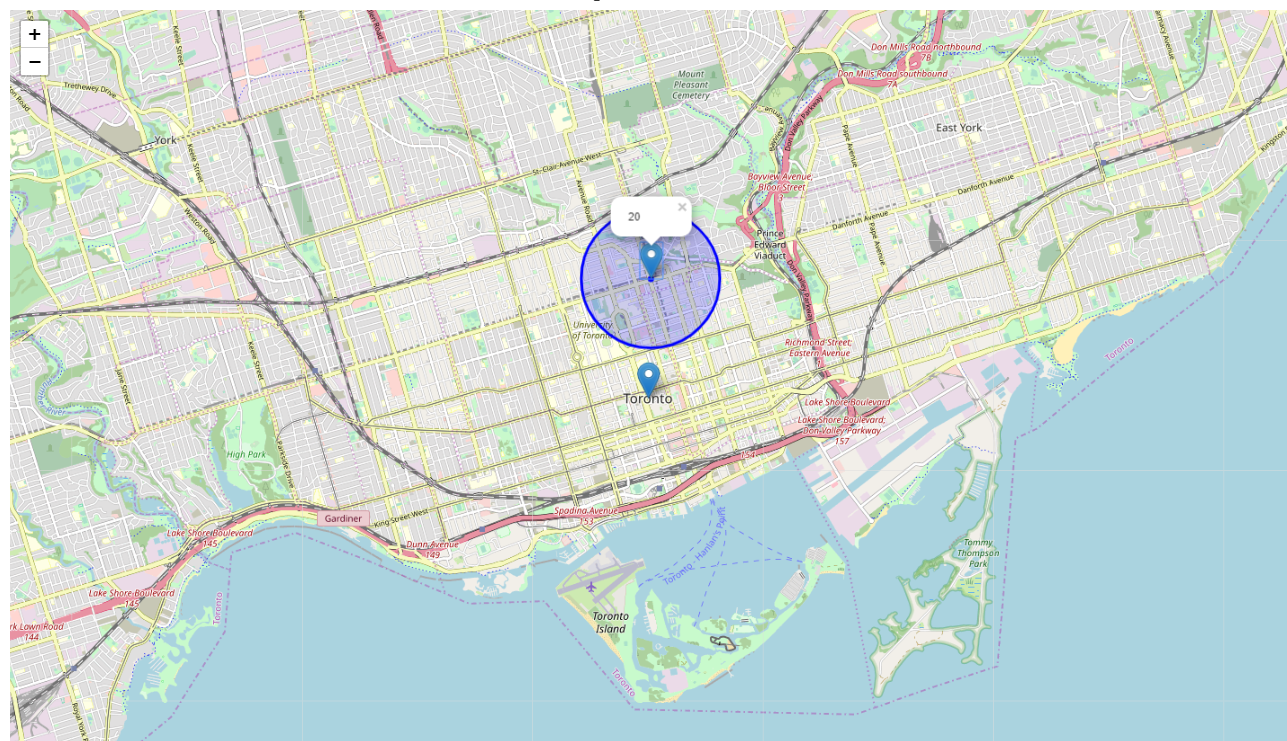


Figure 1. Closest neighbourhood localisation

On top, the different venue categories are listed, sorted; and a report is displaying the main differences between the reference and future ‘closest’ neighborhood:

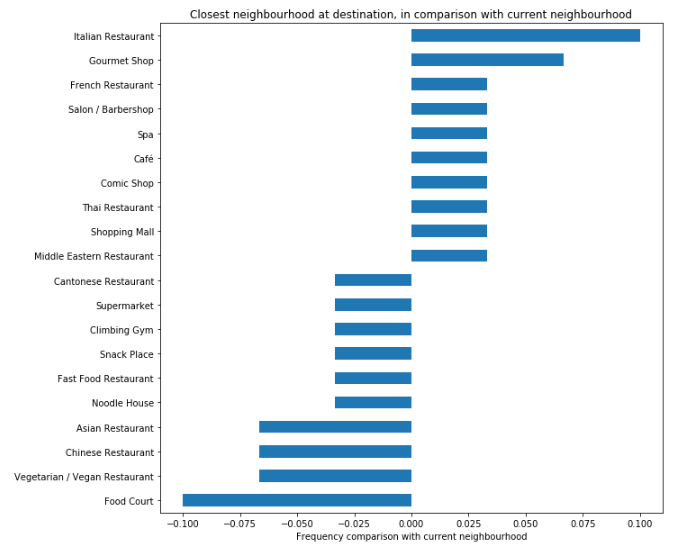


Figure 2. Comparison highlights between closest neighbourhood at destination and reference

Without too much surprise, main differences between reference neighborhood (in Singapore) and destination appears on the food places categories. Food court and asian restaurant are mostly replaced by Italian restaurants and gourmet shops. Other categories show very little difference.

Clusterisation of destination neighbourhoods

To provide additional suggestions to the customer, a clusterisation is performed.

To try and find the optimum k, elbow method is implemented

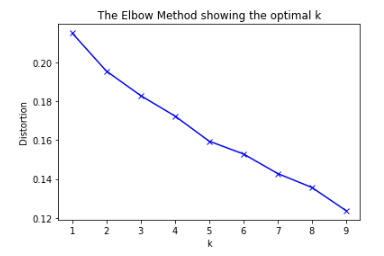


Figure 3. Elbow method – not providing satisfactory results

Unfortunately in our case, there is no obvious elbow. Tool therefore instead a manual selection following a loop going other several clusterisation



Figure 4. Manually going over different possible clusterisation

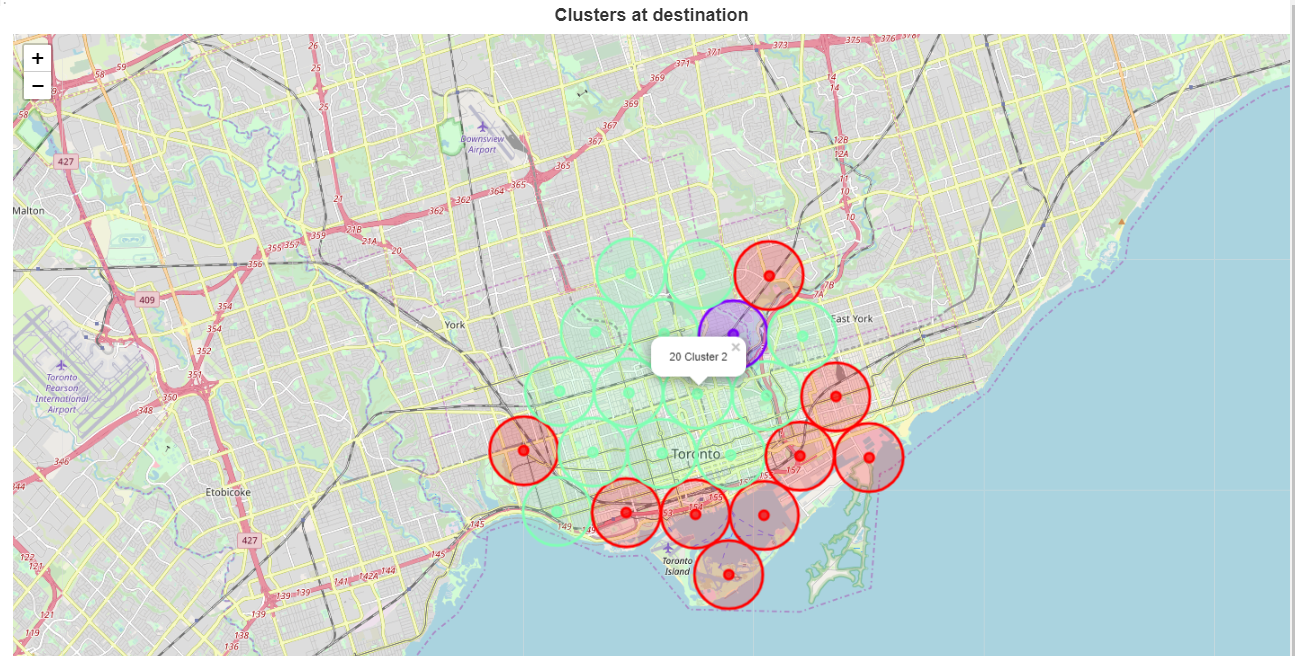
Once selecting a clusterisation that seems to provide interesting results ( at least 2 clusters with reasonable number of neighborhoods), we display the corresponding clusterisation, while identifying in which cluster the ‘closest’ neighborhood defined previously is:   
  


Figure 5. Clusterisation result

First, let’s notice that tool successfully removed neighborhood oversea.  
Then, Visualisation on the map seem to indicate a difference between city center and areas close to the sea and harbor. Similar to previous analysis, we can display main highlights of the different clusters, compared to reference neighborhood:

|  |  |
| --- | --- |
|  |  |

Figure 6. 2 main clusters differences with reference neighbourhood

As expected based on map, Cluster 0 has more harbor venues than reference. (The neighbourhood chosen in singapore is not close to the sea).

Cluster 0 will lack food and asian food compared to current neighbourhood. However, it will present much more coffee shops, café, bars, park and harbor venues + Beach.

Cluster 2 does no miracle regarding asian food, except for thai and indian food. Café and Italian restaurants are expected instead. This cluster seems to provide quite extensive choice of food places.

## Arts & Entertainment venues analysis

As explained before, technically, this part requires to retrieve all categories mapped to Arts & entertainment, as the major category ‘Arts & Entertainment’itself has the same categoryId.

The rest of the analysis is made almost automatic by the tool, user only has to select the category hto focus on, identified by a number.

To complete previous study, a similar analysis is done, focusing this time on Arts & Entertainment venues. Especially, this is a point where customer highlighted that he might be interesting in getting other suggestions.

Our analysis focused on Arts & Entertainment.

Similarly as before, closest neighborhood is identified and located

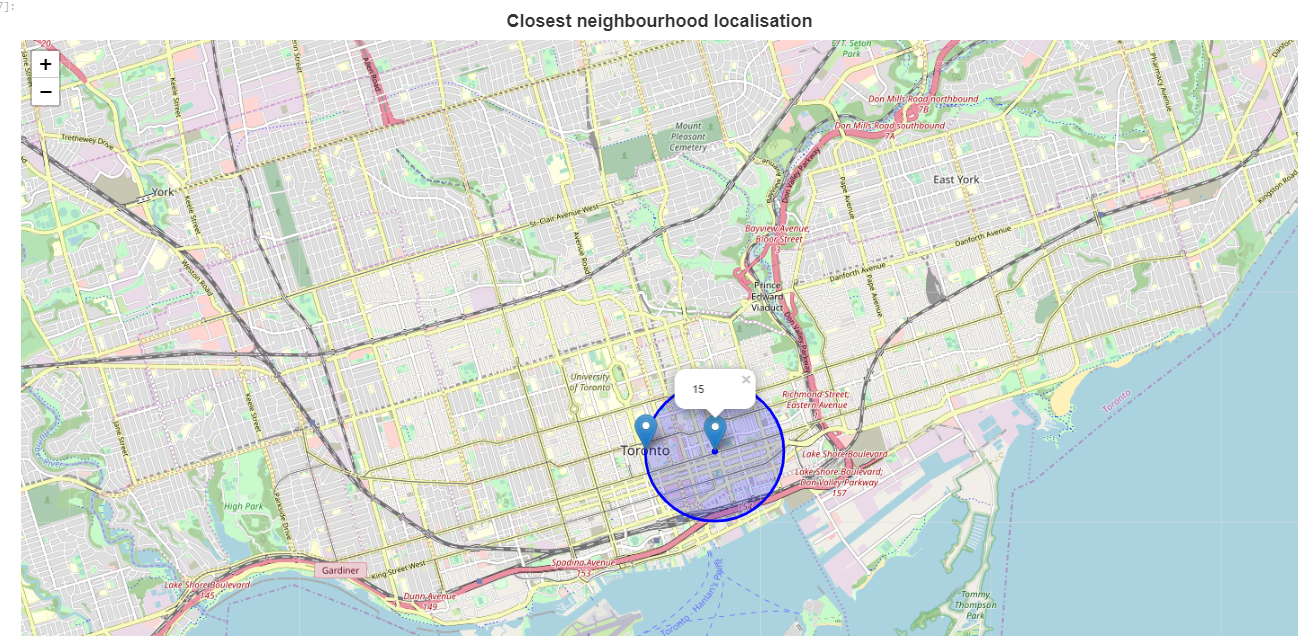


Figure 6. Closest neighborhood in terms of Art&Entertainment venues

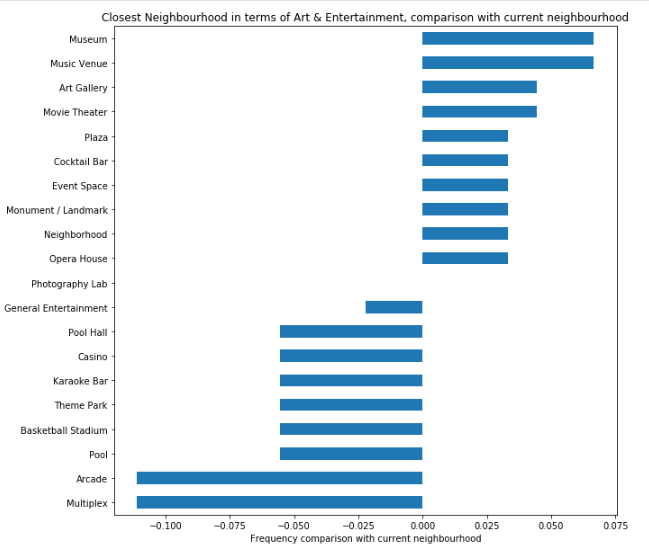


Figure 7. Closest neighbourhood profile comparison

However, the customer will me more interested in getting a clusterisation to evaluate the different profiles of neighbourhoods, as he didn’t explicitely indicated he wanted to get the same as currently regarding arts & entertainment:

Elbow method also didn’t prove to be efficient on this example, we therefore selected manually clusters:

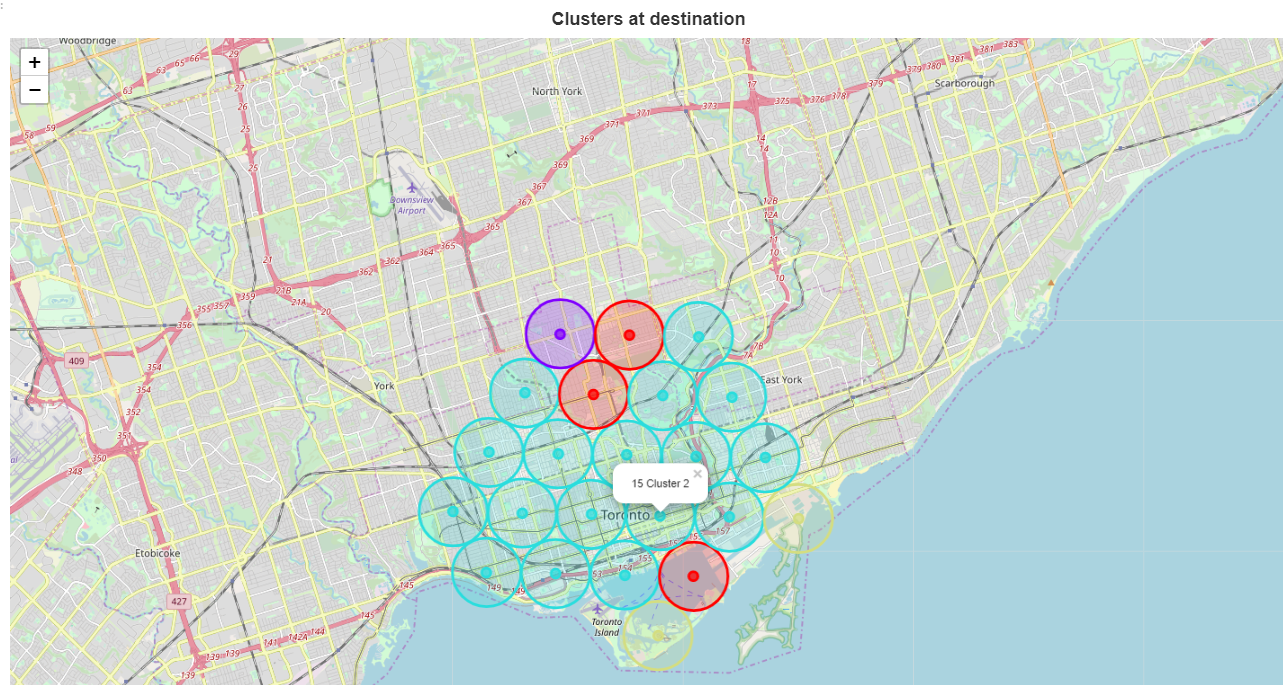


Figure 8. Arts & Entertainment clusters

With corresponding profiles:

|  |  |
| --- | --- |
|  |  |

Figure 8.2 Arts & Entertainment clusters profile comparisons

Main observation at this stage is the following:

The 2 main clusters idetnfied here are very distinguishable: Cluster 2, which included the 'closest neighbourhood' in terms of art/entertainment shows more art and museums but less entertainment areas than current neighbourhood.  
On the contrary Cluster 0 is more oriented on the entertainment venues.

# Results and discussion

We provided an overall overview of what neighborhood in Toronto fits the best to current profile: this was identified in **"Neighborhood 20"; around bloor younge station**. (please refer to corresponding map above)

Clusterisation did not highlight anything specific on this aspect: there is no miracle regarding asian food, when comparing singapore and toronto.

As this overview eventually provided mostly insights on food related aspects; an extended analysis was done focusing on the art/Entertainment aspect.

As a result, it appeared here again that the toronto neighborhoods in the searched area were quite homogeneous, mostly centered around Art galleries and performances. However, even in this area, some neighborhood were really distringuishable, focusing on Entertainment.

Neighbouhood 20, which was the closest overall to current neighbourhood also to be rich in terms of art gallery and performances places, but also at walking distance to Neighbourhood 24, part of the ‘Entertainment cluster’.

At this stage, the recommendation to the customer will therefore be Neighbourhood 20.  
However, decision will be on his side, and further analyses may be requested. We can already anticipate:   
- A larger area of search, the current one is very central in Toronto, leading to some unexpected homogeneity overall.

- Possible search regarding other specific venues category.

The tool is flexible to ensure both.

# Conclusion

The objective was to provide a tool and corresponding analysis, to help customers, moving from one city to another, to identify the destination's cities neighbourhood profile, in comparison with the neighbourhood they are familiar with.

- First, I created a tool, made in a generic way, so that any area can be compared to any city.

- The notebook is also organised so that extended analyses can be easily generated, according to customer feedbacks on what aspects and what venues matter the most for him.

My analysis obviously did not react on any customer feedbacks, I therefore took the assumption of someone moving from a eastern part of Singapore to Toronto, searching in the central part of Toronto.

Toronto was chosen for me to evaluate the results of this modified approach (neighbourhood made of a grid instead of defined neighborhood) compared to previous lab, and also because it presented an additional difficulty, being close to the sea - which is taken into account in data treatment. However, the method and tool is made to be generic, and proved to work, for example on a Paris to Barcelona relocation.

Based on this analysis, there already are good insights to share with the customer. There are definitiely ways to get an even more detailed study, by looking at larger area of toronto, or filtering further the sub-categories expected by the customer. This study however demonstrate the methodology followed, and the corresponding notebook can be used by anyone to do the corresponding search expected.

One might argue that instead of clusterising the neighbourhood we can instead create a map based on ‘distance’ with reference neighbourhood. We actually have all data needed at this stage, as the distance was calculated to determine the closest neighbourhood. I however found more interesting, as part of the coursera course, but also in terms of methodology to dress clusters; that highlight different profiles of neighbourhood that can be interpreted, while the ‘distance’ notion is quite vague.

Note: foursquare API is showing some instability with request includes list of categories, you may have to try several times the corresponding entry in the notebook, but you are more than welcome to give this notebook a try, at least for the overview part.