Capstone Project;
The Battle of the Neighbourhoods;

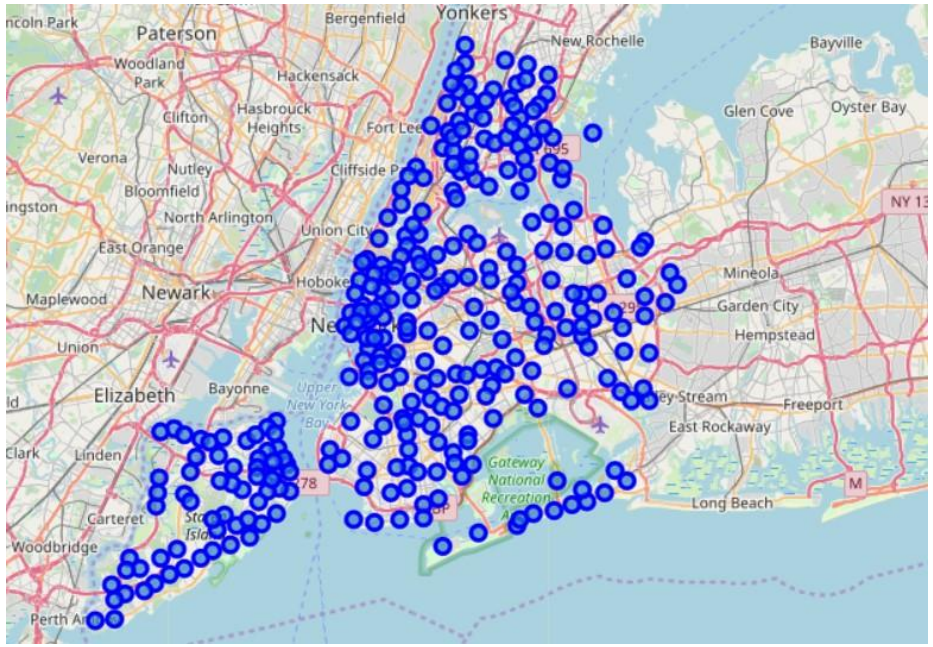
Report prepared by Cristiano Venanzoni

Milan, April 2020

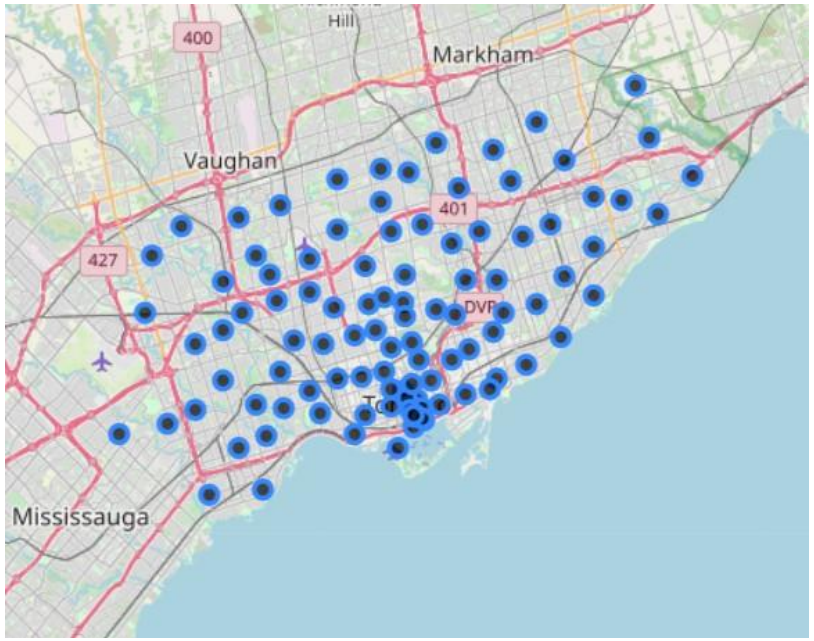
Introduction

- The idea is to build a model that can compare neighbourhoods of 2 cities (A and B) based on the level of similarity of their most popular venues.

- New York City



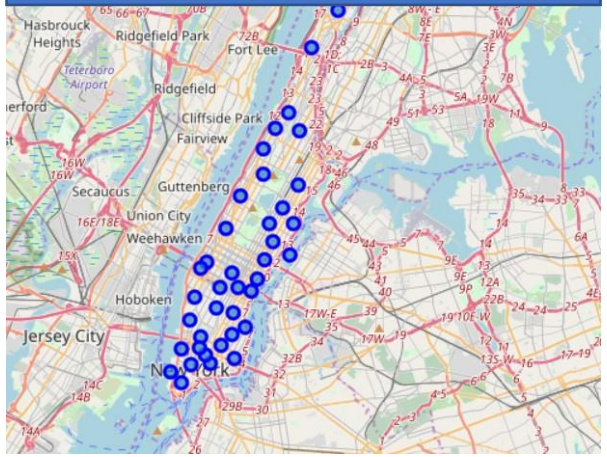
Toronto



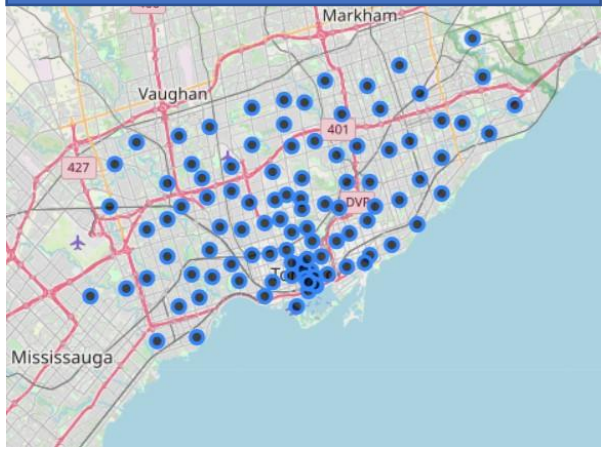
Data

Comparable perimeters identification:

Manhattan area



Toronto area

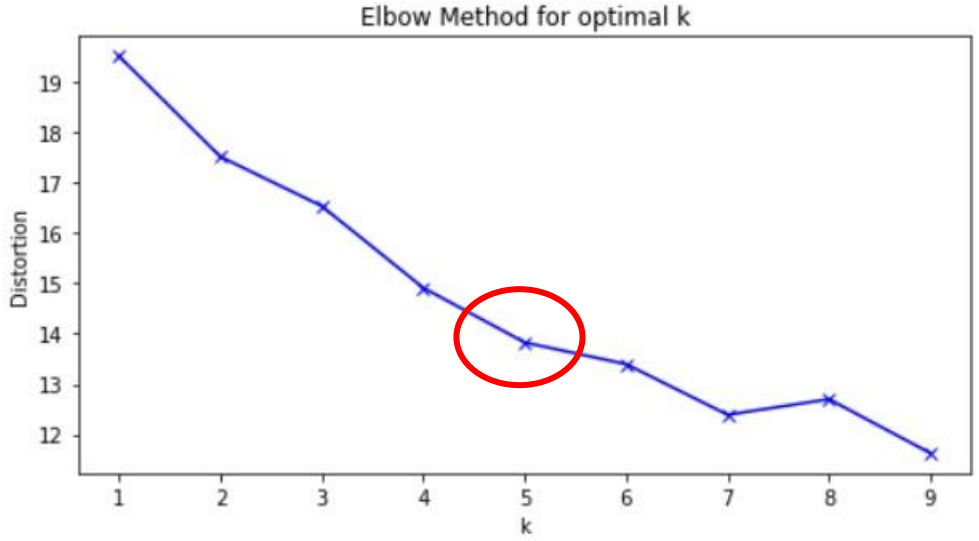


- Venues:
  - Foursquare (www.foursquare.com).
- Cities:
  - Data of Toronto have been retrieved from: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M , while the geo data will be obtained by the following csv file: http://cocl.us/Geospatial_data
  - Data of NYC have been retrieved from the following csv file: https://cocl.us/new_york_dataset

Methodology

- Modeling:
  - In order to identify clusters, the model we used a K-means algorithm. K-means is a partition based clusterization technique that aims at identifying clusters by minimizing intra cluster distance while maximizing inter cluster distance.
  - One hot encoding technique has been used to prepare the dataset and weight the occurencies of the venues categories in the different rows.
  - Centroids were defined randomly.
  - The model has been trained using with different value of K and the best K was found through the “Elbow method” (see next slide).

Methodology

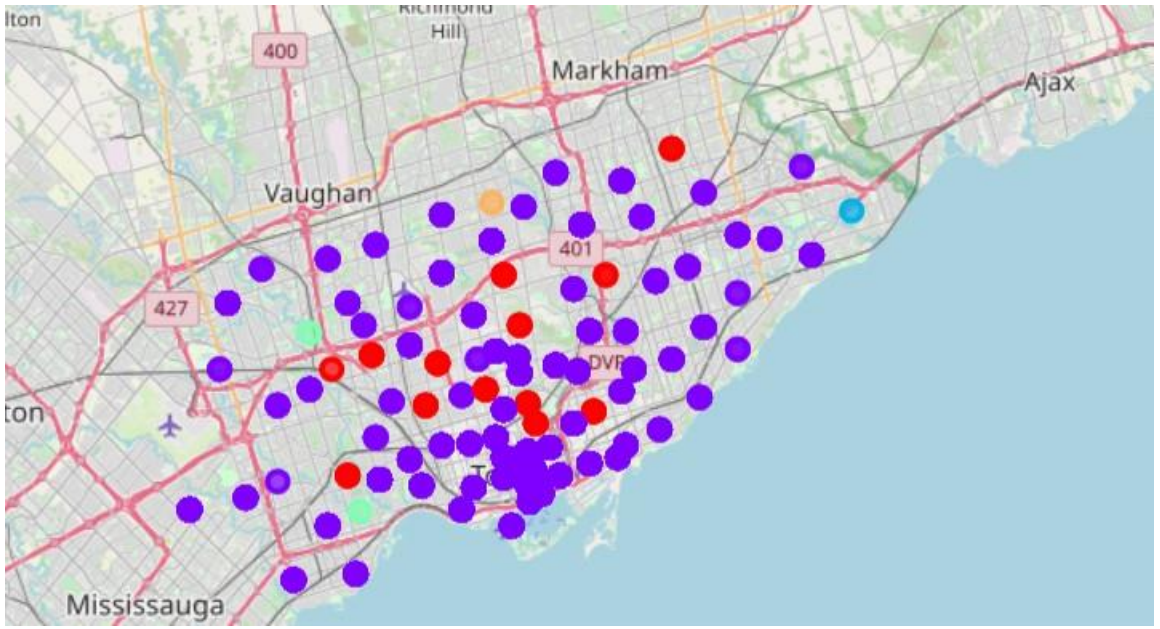


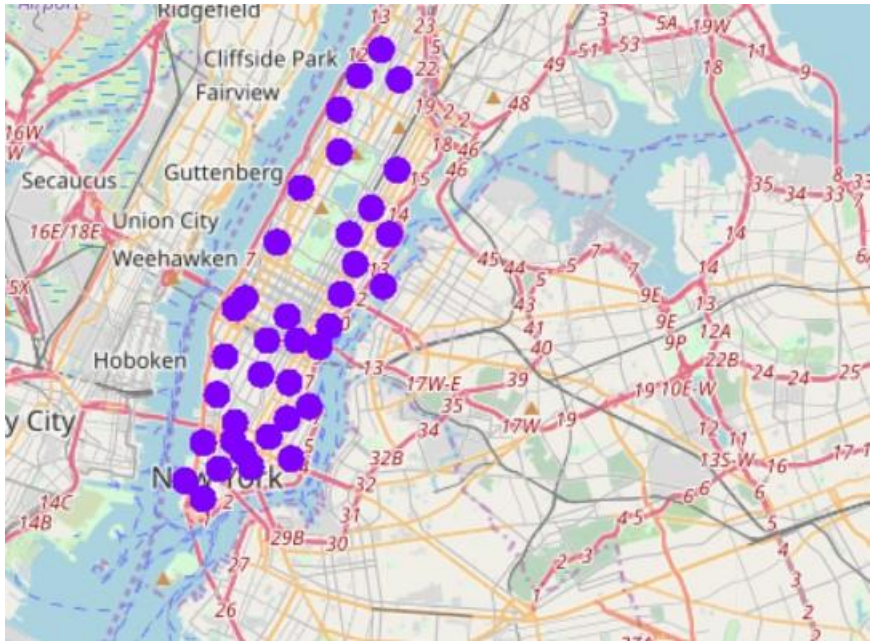
<Figure size 432x288 with 0 Axes>

- Even though no clear indications were provided by the method we observed that for K=5 the model retrieved the most accurate output in terms of dataset segmentation.

- The dataset has also been trained with a DBSCAN algorithm.

Results





- The main difference among clusters appears to be represented by the presence of green areas/parks among the top 3 venues; this can be seen as a direct consequence of the choice to compare different areas in term of extension (the entire city for Toronto, the Manhattan borough for NYC).

Conclusion

- The current report was built as part of the Capstone project assignment.
- Its purpose is to compare neighbourhoods from different cities based on the similarity of their most popular venues as retrieved from the Foursquare API.
- It can be used for tourism purpose in case travellers wants to pick neighbourhoods of the visiting cities based on their similarity with the ones they know in their hometowns.