Capstone Project;
The Battle of the Neighbourhoods;

Report prepared by Cristiano Venanzoni

Milan, April 2020

Summary

Disclaimer

The following report has been prepared as part of the IBM Data Science Certification course. The assignment was requested as final assignment of the Applied Data Science Capstone module.

The report contains consideration from the author, which is a beginner in the data science field.

It should be used solely for learning or practice purposes.

Introduction/Business problem

- The objective of the assignment is to leverage on Foursquare location data to compare neighbourhoods.
- 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.

- The model will offer insights among neighbourhood of different cities and can be used in case of relocation but also for tourism purposes: when visiting city A, tourists will be able to pick a neighbourhood similar to one they know of their hometown (city B).
- The scope of the project is to build a model that can be easily understood and customized by users, adding data from any city they would like to test the model with.

Data

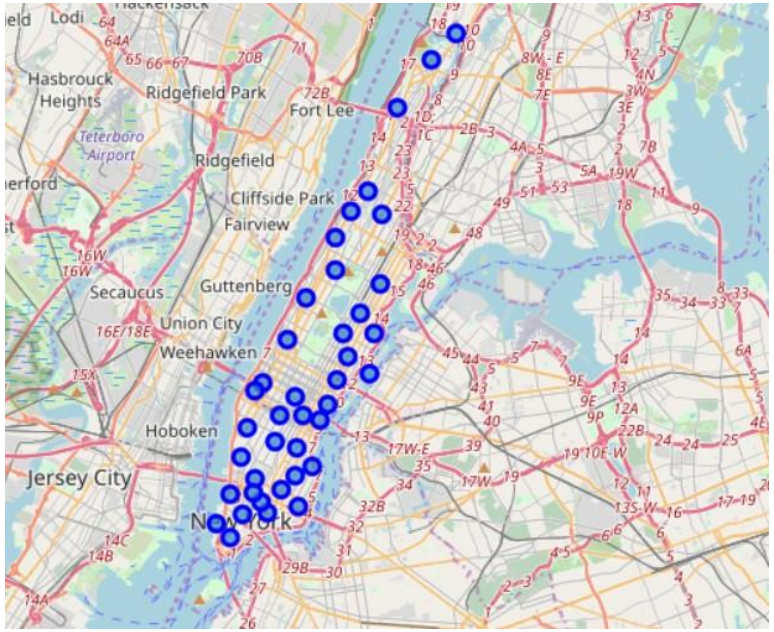
- The model will combine data from cities and datasets of venues with datasets of cities based on neighbourhood geographical codes.
  - Venues:
    - Venues data will be retrieved using Foursquare (www.foursquare.com).
    - In particular the venue/explore? function will be used to retrieve data of popular venues based per neighbourhood.
  - Cities:
    - The model will use 2 given cities data (for simplicity we used Toronto and New York), but as said before, the model can be easily customized and any city data can be inserted by the user.
    - Data of Toronto will be 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 will be retrieved from the following csv file: https://cocl.us/new_york_dataset
    - All datasets are publicly available; user credentials (client_ID, client_secret_ version) are needed to access Foursquare data.

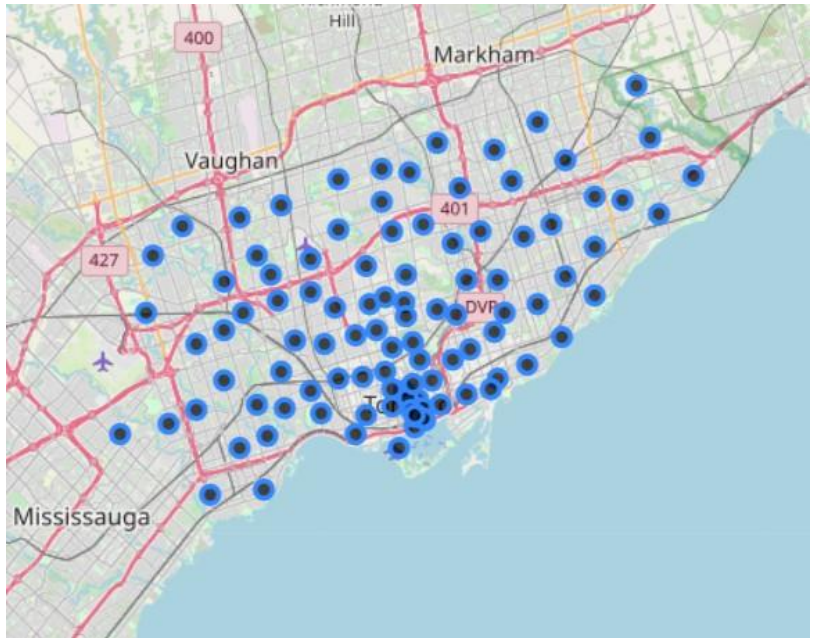
Methodology

- Data analysis:
  - Data exploration:
    - The main problem we faced was determined by the different composition of the 2 input files; in fact while the city A file had geospatial location per neighborhood, the city B file assigned geospatial location per postal code, which encompassed more than one neighborhood.
    - As a consequence the 2 datasets were composed by 306 rows (NYC) and 103 rows (Toronto).
    - We had 2 options:
      - One option could have been to use different criteria (radius and limits) of the Foursquare feature "explore" to adjust the size of the dataset based on the retrieved number of venues, but we thought this would have affected the result making the 2 datasets not comparable.
      - The other option was to consider to slice the datasets based on boroughs and use a comparable amount of boroughs both in term of area and density of venues.
      - The downside of the 2nd option is to exclude some areas from the analysis, but as the objective of the analysis is to segment the 2 cities based on their popular venues typology, we thought a comparable number of venues was the most important attribute for the accuracy of the model.

Methodology

- Based on previous considerations, the perimeter we decided to consider has been the following:
  - Toronto: entire city (103 “superneighborhoods”, ~ 2170 venues)
  - New York City: (Manhattan borough, 40 neighborouhoods, ~ 3240 venues)





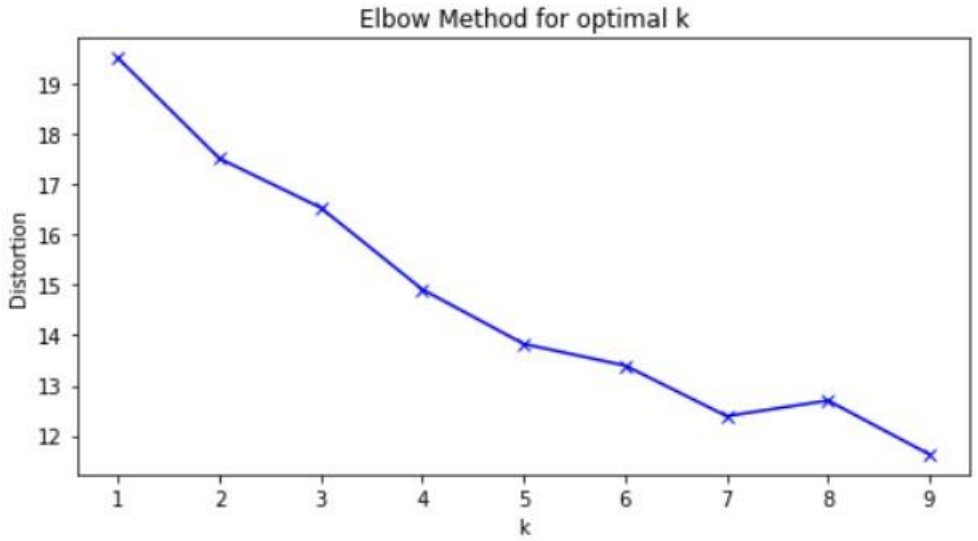
Methodology

- Data wrangling:
  - A moderate amount of data wrangling was requested to prepare the data set, especially because Toronto's data were retrieved from 2 data set while New York City had a unique one.
  - The 2 Toronto datasets were joined based on the "Postal Code" column.
  - The 2 cities datasets have then been sliced in the same shape to facilitate the merging.

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>

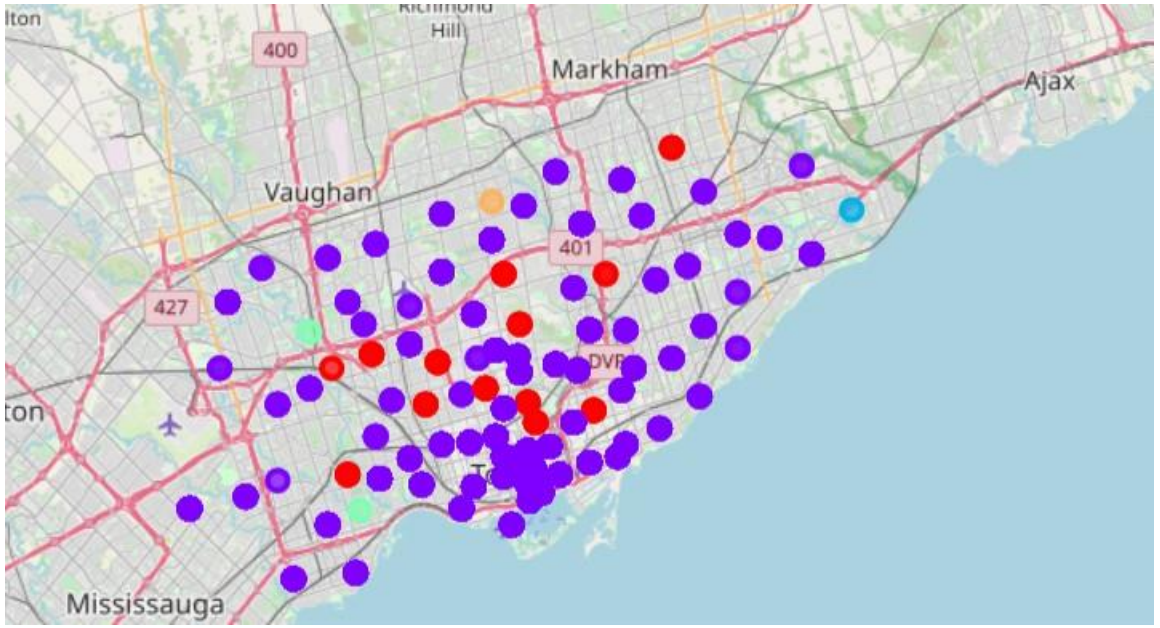
- Even though no clear indications were provided by the method we observed that for K=5 the model retrieved the most accurate output in terms of dataset segmentation.

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

Results

- The model showed a predominance of a single cluster which included more than 95% of the sample.
- 2 clusters were composed by 1 or 2 data points, clearly representing outliers.
- The analysis has been confirmed by the DBSCAN method that retrieved 1 cluster for the data set.
- Surprisingly enough, all the Manhattan data points appear to belong to the same cluster, while Toronto returned a more diverse composition of the neighbourhoods (see next slide).

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).

Discussion

- The most relevant point of attention of the present project is represented by the selection of city perimeter phase.
- Perimeters have to be comparable both in terms of borough/neighborhoods amount but also from a venues density perspective.
- A not appropriate selection of the above mentioned features will affect the solidity of the conclusions we can derive from the model.

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.