**Neighborhoods businesses Comparison Using K-Means Clustering, Case of New York vs Toronto**

**1 Introduction**

**1.1 Background**

Cities around the World contain certain number of neighborhoods formed of venues naturally distributed in many different way. The historic need to identify city’s driver factors such as financial incentives, quality of urbanisation, social and food diversities, crime level, housing, education, health, etc, often lead to the exploration of similarity or dissimilarity between cities.

Comparing venues within neighborhoods in cities helps the orientation of decision making in terms of business investment, immigration or relocation, job hunting and much more. The analysis in this work is applied the two most populous and diverse cities of the United State and Canada: New York vs Toronto.

**1.2 Problem Definition**

The Neighborhoods businesses comparison of two or more different cities is a type of unsupervised classification where venues have to be segmented and grouped into a certain number of dissimilar and non-overlapping clusters. These clusters should contain similar venues of common characteristics, without any internal structure or label. K-Means algorithms is one of the most popular tools of segmentation of unsupervised data that will be used here coupled with foursquare API to make essentially venues calls to retrieve needed information.

**1.3 Stakeholders**

This work is particularly targeting business people who need to invest by giving them insight views of market in neighborhoods. This work concern also at large job seekers to explore possibilities that offer one city over the other. The results of this work should also advice on people movement, immigration and relocation, because positive numbers on neighborhoods businesses in a city, will most probably influence the influx of people.

**1.4 Data Source**

Data used in this work were collected from open sources on the internet:

* For New York, a total of 306 neighborhoods with a total of 5 boroughs in neighborhood as well as the latitude and the longitude of each neighborhood fount on the web under the link: <https://geo.nyu.edu/catalog/nyu_2451_34572>
* For Toronto, City neighborhoods were found on Wikipedia website under the following link:

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

This link provided postal codes of venues that could be transform into needed data by using pandas library function. In case the geocoder package couldn't work, the geographical coordinates of postal code can be found under the link:

<http://cocl.us/Geospatial_data>

These data will be cleaned and formatted accordingly to adequately feed the k-Means algorithm for convergent clustering.