**Capstone Project - The Battle of Neighborhoods (Week 2)**

**I Introduction**

**I***nvestment is in general a complex and challenging task.One of the most difficulty is to collect right data.Today there are certainly lots of data sources thanks to internet data centers developpement ,data processing technologies which make it possible to collect, process and store data in real time.Although these technologies are present in our day life and more often availables as open source tools,they still need to be carefully used in other to obtain right data.*

# 1.1 Problem

**T**he problem as set consists collecting ,porcessing and analyzing the cities **NEW YORK CITY, Toronto And Paris**. More precisely ,the problem is about Collecting informations about all the venues of different neighborhood of all borough in each of these cities in order to find the similarities between them base on data collected using data anlysis tools ,here is python and external API such as foursquar.The result obtained results will be use to suggest a suitable location for investment .

## 1.2 interest

**Evidently,interested are either poeple interested in investing in one of these tree cities or somme one willing to know a litle bit about kaind of venues in these cities.**

## 1.3 Data socuces

### 1.3.1 problem 1 Cities anlysis

**1.3.1.1-NYC(new york city) all data where provided link\_to\_data\_set:** <https://cocl.us/new_york_dataset>

From this link data are downloaded using python api pandas ,process and transform into usable data format. The foursquare API is then used to generate the final data set including geografical coordinate of each neighbouhood.Thinal data set in the generated (Data frame including geographical coordinates of each neighborhood

**1.3.1.2 -Toronto links To dataset:**  <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> -<http://cocl.us/Geospatial_data>\*\*

From this link data are downloaded usind python api pandas ,process and transform into usable data format. The foursquare API is then used to generate the final data set including geografical coordinate of each neighbouhood.Thinal data set in the generated (Data frame including geographical coordinates of each neighborhood.

**1.3.1.3 Paris links to data sets:** <https://fr.wikipedia.org/wiki/Liste_des_quartiers_administratifs_de_Paris> [<https://opendata.paris.fr/explore/dataset/quartier_paris/download/?format=csv&timezone=Europe/Berlin&use_labels_for_header=true&csv_separator=%3B>

From this link data are download ,loded usind python api pandas ,process and transform into usable data format. The foursquare API is then used to generate the final data set including geografical coordinate of each neighbouhood.Thinal data set in the generated (Data frame including geographical coordinates of each neighborhood

### 1.3.2 Problem 2 (venue sugestion for ristaurant investment)

data set used here is the results of the data collected and transform (using python and foursquere api) into strutured and usable data format in the previous section

# II-Data collection , processing and anlysis for each city

## II-1 New York City.

**AS** previously said,data are collected from links provided,the programming language used is python on IMD cloud,more precisely IBM service.Thanks to these tools we have colletted and tranformed data from unusable data forrmat to structured data frame.Below is the scrreshut of the final data structure where in the columns we have from left to right exactly :**Borough's name ,Neighborhood's name ,Latitude and Longitud**e of the overall newyork city.

Newyork city data set with geogaphical coordinates

*Une image contenant capture d’écran, intérieur, mur

Description générée automatiquement*

The total number of Neighborhood in newyork city is : **103**

Using External fousquare API and further processing data are transformed in the usable the final data structure wifhi can be seen in the following.

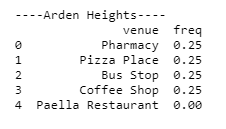
Table including venue categorie

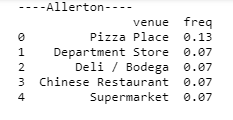
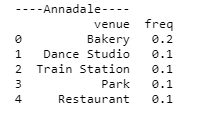
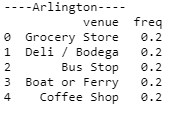
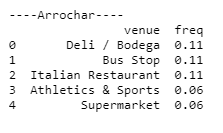
Une image contenant capture d’écran

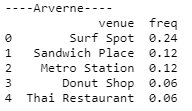
Description générée automatiquementThe total number of venues in newyork city is : 1.22 per 100 inhabitants.

In terms of venues diversity,there is 429 venues categories,all is resumed in the following table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Population | Venues number | Venues per 10000 inhabitants  inhabitants | Different venue  category | Restorants  number | Restorant per  10000 inhabitant | Diferent restaurant categories |
| 8398748 | **10255** | **12,2** | **429** | **2580** | **3** | **92** |

Let look a litle bit deeper into the data set .Here below are the top fives venues for 6 neighbohoods in Newyork city.





Below is the screenshut classification of the most comon venue for each neiborhood.



The next spet consit in clustering the data set in clusters of neighborhoohs sharing similarities between them base on the most common venue in each neighborhood.

Clustering using k-mean with 5 clusters