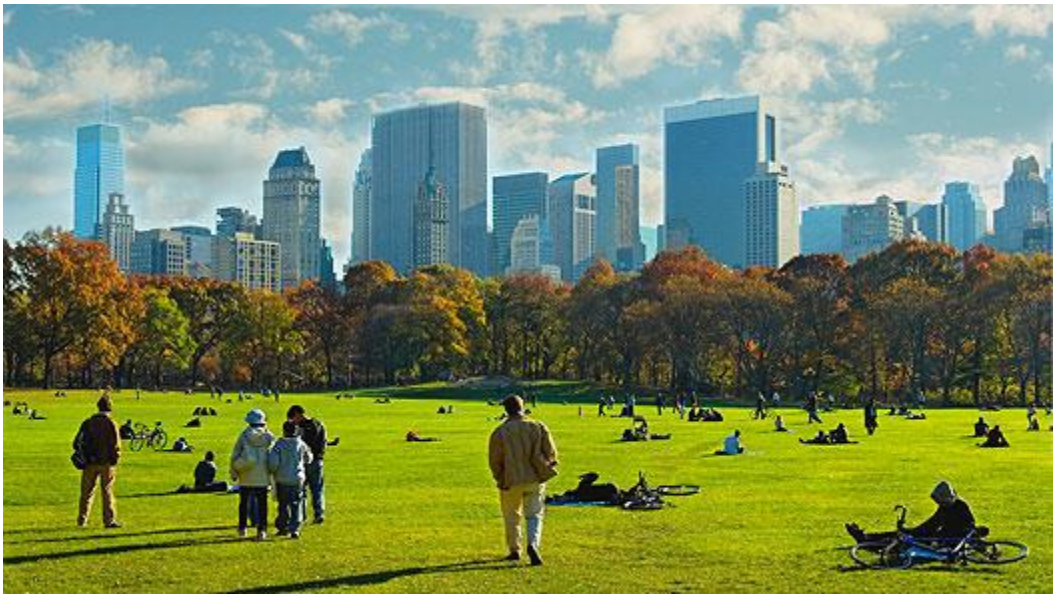
Ecole Nationale de la Statistique et de l'Analyse de l'Information

MSc IN BIG DATA

NoSQL project:
Find the best part of New York City to live in using NYC open data



Author : Amandine Bonizec
Date : 16/12/2016

TABLE OF CONTENT

I – INTRODUCTION

A large set of data about New York City is available for free on the NYC open data website (https://nycopendata.socrata.com).

A woman is moving to New York next month and she ask to find "the best" part of town to move in. These datasets will help to find the ideal borough and even the best street for her to live. She gave the following criteria, by order of importance:

1) A safe environment;
2) Closeness to nature or public parks;
3) Having some sidewalk cafés next to her flat;
4) Living next to a farmer market open on Wednesday since it is her day off.

A tool has been developed to allow the user to import datasets in CSV format and performed queries on it. The datasets and the queries has been defined to make a suitable recommendation to the woman according to her criteria.

The tool is developed in bash scripting language for the data importation and in java for the queries.

II – DATABASE

2.1 Datasets used

Four different datasets have been downloaded in csv format on the NCY open data website:

- "NYPD_Complaint_Data_Historic.csv" includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to 2015. Link: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i
- "2015_Street_Tree_Census_-_Tree_Data.csv" is an inventory of the trees in New York with tree species, diameter and perception of health. Link: https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh
- "Legally_Operating_Businesses.csv" features businesses/individuals holding a DCA license so that they may legally operate in New York City. Link: https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh
- "2012_NYC_Farmers_Market_List.csv" is a Listing of NYC farmer's markets. Link: https://data.cityofnewyork.us/Business/2012-NYC-Farmers-Market-List/b7kx-qikm

2.2 NoSQL database

The NoSQL database MongoDB has been used to import the dataset, process, query and display the data. This choice has been made because:

- MongoDB is suitable to read datasets with a high number of observations;
- It is very easy to import datasets in csv format from the local disk;
- No virtual machine is required to use MongoDB;
- Since MongoDB handles JavaScript, it is very convenient for building a script and load it directly from the terminal.

III – PROGRAM USAGE AND TECHNICAL DETAILS

3.1 Usage information

The script must operate with the original csv dataset files. It can be downloaded from the links given in §2.1.

The program must be started from a Linux terminal. First, the user must connect on MongoDB with the following command line:

```
sudo service mongod start
```

The program *NYCopendata.sh* must be launched at its root with:

```
bash NYCopendata.sh [options]
```

The following options are implemented:

```
---------------------------------------------------------------------------------------------------
-f  <dataset_folder> : Path to the folder containing the datasets
-i : Importing the datasets into a MongoDB database called NYC
-q : Running the queries and printing it on the terminal
---------------------------------------------------------------------------------------------------
```

It is mandatory to specify the dataset folder (option f) for the importation of the datasets (option i). The queries can be run (option r) without specifying the folder as soon as the datasets are loaded.

Examples:

- *bash NYCopendata.sh -f /home/.../data -i* : import the datasets
- *bash NYCopendata.sh -f /home/.../data -i -r* : import the datasets and run the queries
- *bash NYCopendata.sh -r* : run the queries

The script *MongoDB_queries.js* is loaded from the main script *NYCopendata.sh* for the queries under MongoDB. Both scripts must be in the same folder.

The database on MongoDB called NYC. The queries in output are printed on the terminal.

3.2 Technical details

Datasets importation

Command lines in bash have been used to import each dataset:

mongoimport -d NYC -c <collection_name> --type csv --file $folder/<dataset_name>.csv --headerline --drop

The *headerline* option allows to take the first line as column names. For some datasets, the first line has been modified before importation with more synthetic column names.

The drop option is set to delete a collection if it exists already before importation.

Queries

The JavaScript code for the queries is written in a separate script (*MongoDB_queries.js*). Several MongoDB methods have been used to query the data such as the aggregate, update, find and pretty methods.

Error handling

If an error is identified in the command line or on the database, the program won't be executed and an error message will appear in red.

The following errors are handled:

- The dataset folder does not exist;
- No argument/path is given for the option -f;
- The importation is introduced without the definition of the dataset folder;
- The user wants to run the queries but the dataset NYC doesn’t exist, the datasets haven’t been loaded yet.

IV – RESULTS

Six queries were processed to find the best place to live in New York. Find bellow the results by query. The size of the population by borough in 2014 has been used to calculate the crime rates and the trees rates. The figures can be found at this address: https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Boroughs/9mhd-na2n

****** Number of crimes per borough ******
{ "_id" : "QUEENS", "nb_crimes" : 1011002, "pop" : 2230722, "crime_rate" : 0.4532173888095424 }
{ "_id" : "STATEN ISLAND", "nb_crimes" : 243790, "pop" : 468730, "crime_rate" : 0.5201075245877157 }
{ "_id" : "BROOKLYN", "nb_crimes" : 1526213, "pop" : 2504700, "crime_rate" : 0.6093396414740289 }
{ "_id" : "MANHATTAN", "nb_crimes" : 1216249, "pop" : 1585873, "crime_rate" : 0.7669271120701342 }
{ "_id" : "BRONX", "nb_crimes" : 1103514, "pop" : 1385108, "crime_rate" : 0.7966988855742657 }

The number of crimes per inhabitant is the lowest in Queens. It is then the safest borough to live in.

****** Number of trees per borough ******
{ "_id" : "Staten Island", "nb_tree" : 62971, "pop" : 468730, "tree_rate" : 0.13434386533825443 }
{ "_id" : "Queens", "nb_tree" : 184137, "pop" : 2230722, "tree_rate" : 0.0825459201101706 }
{ "_id" : "Brooklyn", "nb_tree" : 115783, "pop" : 2504700, "tree_rate" : 0.04622629456621551 }
{ "_id" : "Bronx", "nb_tree" : 46075, "pop" : 1385108, "tree_rate" : 0.033264554099752514 }
{ "_id" : "Manhattan", "nb_tree" : 37697, "pop" : 1585873, "tree_rate" : 0.02377050369102696 }

The number of trees per inhabitant is the highest in Staten Island, but Queens has a quite high rate too. It means that some parks or green areas can be found there.

***** Number of farmer's markets open on Wednesday per borough *****
{ "_id" : "Manhattan", "nb_market" : 10 }
{ "_id" : "Bronx", "nb_market" : 10 }
{ "_id" : "Brooklyn", "nb_market" : 8 }
{ "_id" : "Queens", "nb_market" : 1 }

There is one farmer's market open on Wednesday in Queens but none in Staten Island.

***** Number of sidewalk cafe per borough *****

{ "_id" : "Manhattan", "nb_cafe" : 1010 }
{ "_id" : "Brooklyn", "nb_cafe" : 193 }
{ "_id" : "Queens", "nb_cafe" : 124 }
{ "_id" : "Bronx", "nb_cafe" : 27 }

There is no sidewalk cafe in Staten Island, however there are many ones in Queens.
According to the woman criteria, Queens is a suitable place.

***** List of farmer's markets in the Queens open on Wednesday *****

{
"BORO" : "Queens",
"MARKET" : "Astoria Greenmarket",
"ADD" : "31st Ave at 14th St",
"DAY" : "Wednesday",
"HOUR" : "8am-3pm"
}

The only farmer's market open on Wednesday in Queen is in Astoria, 31rs avenue.

***** List of sidewalk cafés in the Queen, Astoria neighborhoods, 31st avenue *****

{ "LICAT" : "Sidewalk Cafe", "BNAME" : "3321 ASTORIA INC.", "STREET" : "31ST AVE", "CITY" : "ASTORIA", "BORO" : "Queens" }
{ "LICAT" : "Sidewalk Cafe", "BNAME" : "SO-TAUN ENTERPRISES, LLC", "STREET" : "31ST ST", "CITY" : "ASTORIA", "BORO" : "Queens" }
{ "LICAT" : "Sidewalk Cafe", "BNAME" : "IFETA CORP.", "STREET" : "31ST AVE", "CITY" : "ASTORIA", "BORO" : "Queens" }
{ "LICAT" : "Sidewalk Cafe", "BNAME" : "MAHAPOCHANAHPHAN INC", "STREET" : "31ST AVE", "CITY" : "ASTORIA", "BORO" : "Queens" }
{ "LICAT" : "Sidewalk Cafe", "BNAME" : "B5 LLC", "STREET" : "31ST ST", "CITY" : "ASTORIA", "BORO" : "Queens" }

In the 31st avenue of Astoria, there are 5 different sidewalk cafés.

V – CONCLUSION

The queries computed from MongoDB database show that Queens is a convenient borough to live in, in terms of safety, natural environment, sidewalk cafés and farmer's markets.

With a closer look on the data, the 31st avenue of Astoria has been defined as the best place for living. Indeed, there is a farmer's market open on Wednesday and several sidewalk cafés.

Figure 1: The city Astoria in Queens borough

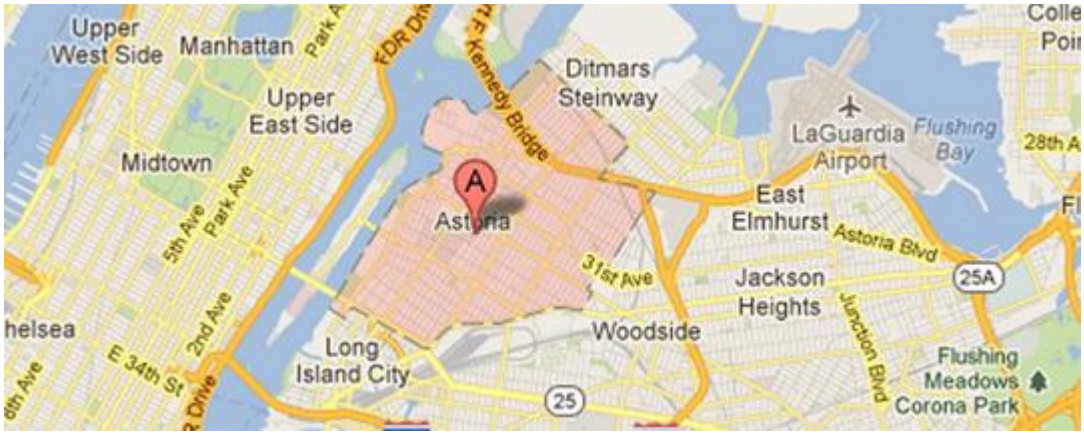


Figure 2: Astoria neighborhoods, 31st avenue