NoSQL Project

Jérémy LE GALL

Février 2017

A la recherche du meilleur endroit pour vivre à New York

1 Définition du meilleur endroit pour vivre et description des données

Tout d'abord, la qualité de l'air est primordiale de mon point de vue. En effet, New York étant une ville sur-peuplée, il y a forcément des zones plus polluées que d'autres. Ainsi, vivre dans un endroit sain et avec un air pur serait vraiment indispensable pour profiter de ma vie à New York.

Ensuite, la sécurité devra être également nécessaire. Un endroit avec peu d'incidents criminels aurait une meilleure ambiance. En effet, les américains étant relativement paranoïaques, si je peux trouver un endroit limité en criminalité cela ajouterai un plus de gaieté et de sécurité !

Comme je vais devoir trouver un logement, on pourrait s'attarder pour chercher les meilleurs prix ou les logements bien situés. Ici ne disposant pas forcément de ces données, on va plutôt chercher les logements où il y a très peu de plaintes des habitants. En effet, moins il y a de plaintes à un certain endroit, plus on peut supposer que la qualité du logement trouvée sera importante. Cela va limiter les mauvaises surprises lorsque j'emmenagerais là-bas.

Enfin, pour déterminer l'endroit exact on va chercher à obtenir un endroit qui bouge. En effet, un endroit où l'on peut trouver un grand nombre d'activités sera idéal. Plus l'endroit sera vivant, plus l'ambiance sera intéressante.

D'autres notions contribuent à obtenir le meilleur endroit pour vivre. Par exemple, une excellente connexion internet au niveau de la latence et du débit est primordiale. Mais aucune donnée ne peut illustrer ce point. De plus, on peut supposer que les connexions internet à New York sont à peu près equivalentes, et on n'observe surement pas de différences entre les différents quartiers voire rues.

Ensuite on aurait également pu s'attarder sur les transports, les infrastructures de santé ou bien l'éducation. Ces points pourraient améliorer le choix du meilleur endroit pour vivre dans une future étude.

2 Choix des technologies

Pour le choix de la base de données, nous avions le choix entre 4 bases :

- Neo4J
- MongoDB
- Redis
- Cassandra

Tout d'abord, nos données ne dépassant jamais quelques dizaines de méga octets, on n'a pas beaucoup d'intérêts d'utiliser les bases Redis et Cassandra. En effet, ces bases sont optimisées pour obtenir des temps de requêtes optimaux sur d'énormes bases de données. Ces bases fortement scalables ne seront pas ainsi utilisées dans ce projet.

Ensuite, on va plutôt préférer MongoDB au final. Neo4J est une base relativement équivalente à MongoDB, mais son point fort est la vue graphe de Neo4J. Or là encore au vu des données choisies, la vue graphe est inutile et les relations apportées par Neo4J n'apportent rien.

De plus, on a un plugin PyMongo fortement utilisé et bien implémenté sous Python. Ceci conforte le choix de MongoDB puisqu'on pourra effectuer le traitement de nos données puis les requêtes dans un seul environnement qui sera Python 3.5.

Ainsi le code sera effectué sous Python 3.5 à l'aide du plugin PyMongo 3.4.0 et la librairie Pandas. On utilise l'IDE Atom.io vraiment adapté à tous les langages de programmation.

3 Méthodologie pour déterminer le meilleur endroit

Au vu des données choisies, les trois premières bases sont adaptées pour effectuer une analyse par grand quartier de New York tandis que la dernière base sera très utile pour choisir une location précise où se trouvera notre futur logement.

En effet, la méthodologie utilisée sera de choisir le meilleur quartier par rapport à la qualité de l'air, la sécurité criminelle et la qualité des logements au vu des plaintes posées par les habitants. Enfin les activités permettront de choisir la rue la plus vivante pour installer notre petit logement sur New York.

Qualité de l'air

Ces données comportent par quartier et sous-quartiers beaucoup de données sur la qualité de l'air. On a une vingtaine d'indicateurs au total qui décrivent l'air des locations de New York.

Dans notre script, on va tout d'abord traiter les données. On s'occupe des valeurs manquantes, on choisit les années les plus intéressantes puis on change les types de certaines variables.

Ensuite on groupe les données des indicateurs de pollution par sous-quartiers. On va scale les données de chaque indicateur de 0 à 100.

C'est à dire que si une rue est très polluée en benzène (c'est un exemple d'indicateur parmi les 20), son score sera proche de 0, tandis qu'une rue ayant une bonne qualité de l'air (faible indicateur en benzène par exemple) obtiendra des scores d'indicateurs proche de 100. Pour conclure, on moyenne les scores de chaque indicateurs par sous-quartiers ce qui nous fait obtenir le classement suivant :

	_id	quartier_id	average
10	South Beach - Tottenville	504	95.406906
21	Willowbrook	503	89.335727
1	Southeast Queens	409	89.252567
2	Rockaways	410	88.580114
0	Port Richmond	501	88.559477
24	Stapleton - St. George	502	87.779433
26	Jamaica	408	87.744921
3	Southwest Queens	407	87.273595
16	Canarsie - Flatlands	208	86.491061
6	Flushing - Clearview	403	86.362271
37	Coney Island - Sheepshead Bay	210	84.331573
19	Bayside - Little Neck	404	83.632652
30	Borough Park	206	82.550616
5	Ridgewood - Forest Hills	405	82.156861
32	Bensonhurst - Bay Ridge	209	81.949632
4	Fresh Meadows	406	81.810290

On retrouve un grand nombre de quartiers de Staten Island, avec quelques quartiers du Queens qui se démarquent du lot avec une excellent qualité de l'air dans le top 8.

Crimes dans les parcs

Sur tous les parcs de New York, on a les crimes commis depuis quelques années. Les données ne comportent pas une forte description de la location à part les grands quartiers de New York qui sont enregistrés dans une colonne.

Pour ces données, on va effectuer un ratio à l'aide de MongoDB. On va chercher à trouver le quartier ayant le plus fort ratio de surface de parcs en acres par rapport aux crimes commis au total par quartier dans les parcs.

On obtient ainsi le résultat suivant :

Quartier	Surface Totale	Nombre de parcs	Meurtres	Ratio
Staten Island	7321	124	3	2440.3
Queens	7502	284	29	258.7
Bronx	7052	217	46	153.3
Brooklyn	3858	327	52	74.2
Manhattan	1975	197	59	33.5

On peut voir que Staten Island se démarque très fortement là encore sur ce nouveau dataset. En effet, Staten Island comporte moins de parcs que les autres quartiers, mais qui sont beaucoup plus grands en surface. De plus, il y a très peu de crimes commis dans ce quartier (10 à 20 fois moins).

Tandis que Manhattan paraît être vraiment le pire quartier pour le moment. En effet Manhattan est le quartier le plus pollué et avec le plus de crimes commis pour une faible surface d'espaces verts.

Qualité des logements

Pour les complaintes sur les logements, on effectue la requête mongoDB suivante. On va compter les nombres de complaintes ayant été jugées ou pas. Puis on effectue les ratios des cas jugés par rapport aux nombre de cas totaux.

```
pipeline =[
    {
        "$group": {
            "_id": "$Boro",
            "nbrCases": { "$sum": 1 },
            "yesCases": {
                "$sum": {
                  "$cond": [ { "$eq": [ "$CaseJudgement", "YES" ] }, 1, 0 ]
                }
            },
            "noCases": {
                "$sum": {
                  "$cond": [ { "$eq": [ "$CaseJudgement", "NO" ] }, 1, 0 ]
                }
            }
        }
    },
    {"$project" : {"_id":1,"nbrCases" : 1,"yesCases":1,"noCases":1,
    "ratioY":{ "$divide": [ "$yesCases", "$nbrCases" ] },
    "ratioN":{ "$divide": [ "$noCases", "$nbrCases" ] }}},
    {"$sort":  {"ratioY" : -1} }]
```

On obtient la sortie de cette requête ci-dessous :


```
[{'_id': 'STATEN ISLAND',
  'nbrCases': 1223,
  'noCases': 1146,
  'ratioN': 0.9370400654129191,
  'ratioY': 0.06295993458708095,
  'yesCases': 77},
 {'_id': 'BROOKLYN',
  'nbrCases': 21964,
  'noCases': 21201,
  'ratioN': 0.9652613367328355,
  'ratioY': 0.03473866326716445,
  'yesCases': 763},
 {'_id': 'QUEENS',
  'nbrCases': 10058,
  'noCases': 9726,
  'ratioN': 0.9669914495923643,
  'ratioY': 0.033008550407635714,
  'yesCases': 332},
 {'_id': 'BRONX',
  'nbrCases': 21329,
  'noCases': 20890,
  'ratioN': 0.9794176942191383,
  'ratioY': 0.02058230578086174,
  'yesCases': 439},
 {'_id': 'MANHATTAN',
  'nbrCases': 12808,
  'noCases': 12695,
  'ratioN': 0.9911773891317927,
  'ratioY': 0.00882261086820737,
  'yesCases': 113}]
```


On peut voir que Staten Island là encore a le meilleur ratio de cas jugés. De plus il a 20 fois moins de cas que les autres quartiers (liés à la population plus faible tout de même). Ainsi, la justice paraît plus à l'écoute des citoyens du quartier de Staten Island.

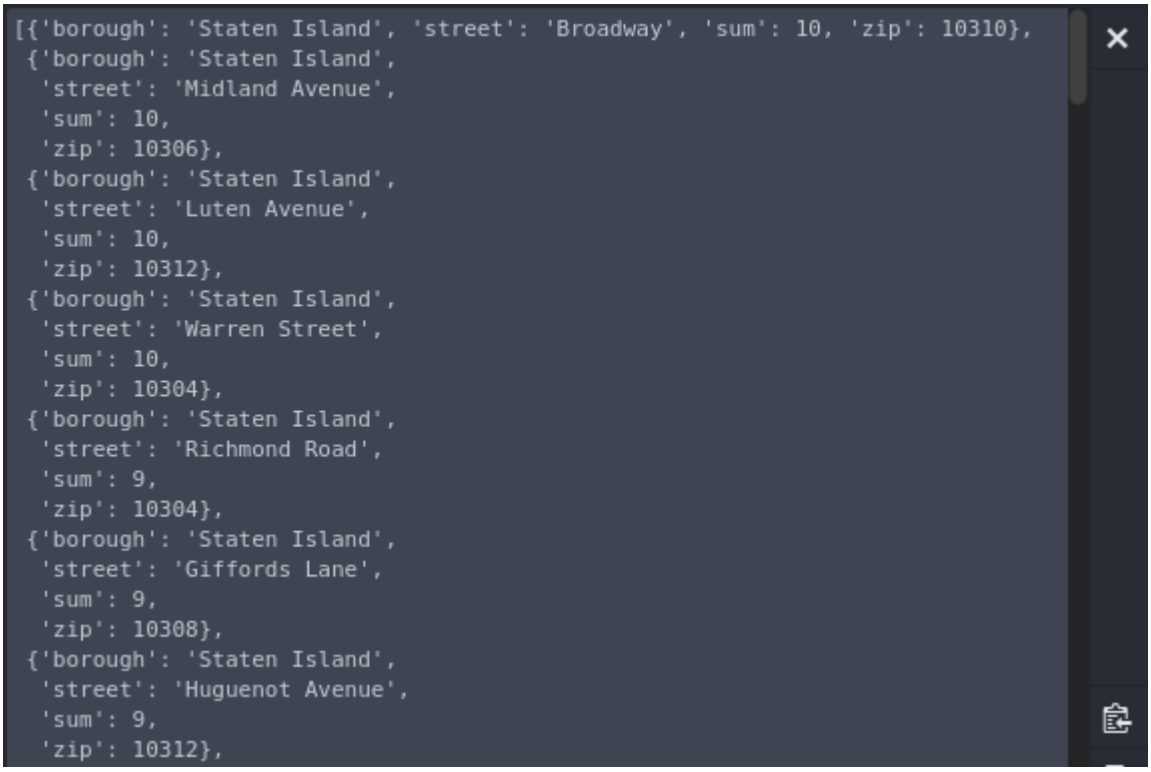
Après ces trois différents datasets, on peut vraiment affirmer que le meilleur quartier pour vivre selon mes critères est le Staten Island.

Maintenant il s'agit de trouver la meilleure location pour établir son logement.

Activités à New York

On choisit un dataset sur les activités extra-scolaires. En effet, il n'y malheuresement pas d'autres données sur des activités. On a tout de même au final, des données sportives et également des données sur les librairies. Donc ce dataset restera quand même représentatif des locations les plus vivantes et animées.

On va sommer par location les activités pour chaque rue du dataset. On sélectionne ensuite le quartier de Staten Island et on trie par nombre d'activités le plus important ce qui nous donne comme sortie :



```
[{'borough': 'Staten Island', 'street': 'Broadway', 'sum': 10, 'zip': 10310},
 {'borough': 'Staten Island',
  'street': 'Midland Avenue',
  'sum': 10,
  'zip': 10306},
 {'borough': 'Staten Island',
  'street': 'Luten Avenue',
  'sum': 10,
  'zip': 10312},
 {'borough': 'Staten Island',
  'street': 'Warren Street',
  'sum': 10,
  'zip': 10304},
 {'borough': 'Staten Island',
  'street': 'Richmond Road',
  'sum': 9,
  'zip': 10304},
 {'borough': 'Staten Island',
  'street': 'Giffords Lane',
  'sum': 9,
  'zip': 10308},
 {'borough': 'Staten Island',
  'street': 'Huguenot Avenue',
  'sum': 9,
  'zip': 10312},
```

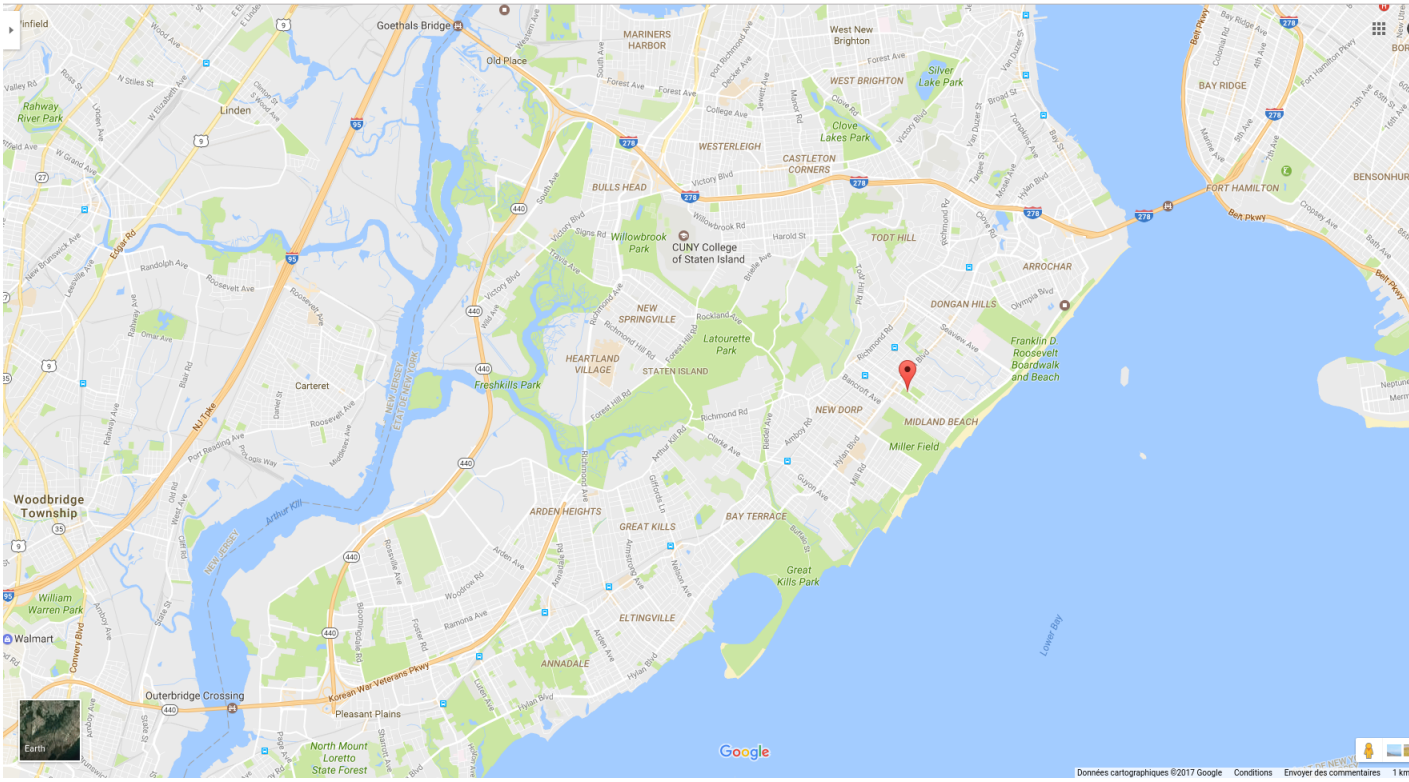
Conclusion

Ensuite si on revient sur la sortie de la qualité de l'air, on peut remarquer que la meilleure zone dans Staten Island se situe entre Tottenville et South Beach.

Par rapport aux rues qui ont le plus d'activités, on peut voir que Midland Avenue est celle qui a donc :

- la meilleure qualité de l'air,
- le moins de complaintes de logements et une justice clémente,
- proche d'un parc où la crimininalité est quasiment nulle
- et enfin le plus d'activités

Ainsi j'habiterais à **Midland Avenue, Staten Island, 10306, New York** et qui est proche de la plage en plus !



A Boucle de requêtes MongoDB en Python + Scaling des données de 0 à 100

```
def scale100(data):
    daata = data.drop(data.columns[[0,1]], axis=1)
    maxn = 100
    minn = 0
    maxo = daata.max()
    mino = daata.min()
    new = maxn-((maxn-minn)/(maxo-mino)*(daata-maxo)+maxn)
    return(pd.concat([data[[0,1]],new],axis = 1))

scores = pd.DataFrame(columns = ['_id','quartier_id'])
for i in [639,640,641,642,643,644,645,646,647,657]:
        pipeline =[
        { "$match": { "$and": [ {"geo_type_name":"UHF42"},
                                {"indicator_id":i}] } },
                {"$group": {
                  "_id" : "$geo_entity_name",
                          "quartier_id" : {"$first":"$geo_entity_id"},
                  "total" : {"$avg":"$data_valuemessage"}
                }
            }
            ,{"$project" : {"total" : 1,"quartier_id":1 }}
            ]
        tt = pd.DataFrame(list(airMG.aggregate(pipeline)))
        tt.rename(columns={'total':'ind'+str(i)}, inplace=True)
        scores = pd.merge(scores,tt,how='outer',on=['_id','quartier_id'])

scores100 = scale100(scores)
scores100['quartier_id'] = scores100['quartier_id'].apply(str)
scores100['average'] = scores100.mean(numeric_only=True, axis=1)
scores100[['_id','quartier_id','average']].sort('average',ascending = 0)
```