ENSAI

Année scolaire 2016-2017

Laura DUPUIS

Projet de Technologies NoSQL

Filière SID

Quel est le meilleur endroit où habiter à New York ?

Qu'est-ce que le "meilleur endroit" ?

Après trois années à Rennes, je m'envole enfin pour New York pour mon stage de fin d'étude. Les valises sont prêtes, les billets sont réservés et le passeport est à jour. Il ne reste plus qu'à trouver un appartement pour ces six mois. Mais parmi tous les appartements disponibles, le choix est compliqué. Où chercher ? Quel quartier ? Selon quels critères ?

Tout d'abord, j'aimerais être proche d'un métro. Si je vais à New York, c'est pour mon stage bien sûr, mais aussi pour en profiter et visiter. Il serait donc dommage de passer du temps dans les transports chaque jour. Je veux donc me trouver proche d'un métro. Je ne suis pas contre un peu de marche, mais dans des limites raisonnables.

Ensuite, j'ai toujours habité en campagne, avec des champs, beaucoup d'arbres et proche de la nature. Le campus de Ker Lann, sur lequel j'ai étudié pendant trois ans, n'en manque pas. Alors j'aimerais bien retrouver cette verdure à New York. Mon deuxième critère sera donc de trouver un appartement proche d'un parc. Et suffisament grand pour pouvoir s'y promener plusieurs fois sans prendre à chaque fois le même chemin. D'autant plus que j'aimerais adopter un chien à mon arrivée. Dans un petit appartement, il vaut mieux avoir des espaces verts à côté de chez soi pour pouvoir le promener tous les jours.

Et enfin, j'aimerais être proche d'un théâtre. J'aime beaucoup aller au cinéma, sortir au restaurant ou visiter des musées. Si je suis suffisamment proche d'un arrêt de métro, je pourrais m'y rendre facilement. Mais j'aime par dessus tout aller voir une pièce de théâtre ou un one man show une à deux fois par semaine. Je ne peux donc pas faire de concession sur le théâtre. Je souhaite alors en être vraiment proche et pouvoir rentrer chez moi très rapidement en marchant. J'aimerais ainsi en trouver un proche de chez moi et donc à moins de 250 mètres.

Je vais donc effectuer mes recherches principalement sur ces trois critères.

Comment trouver cet endroit idéal ? : Préparation de la base de données

Le projet est disponible sous github. Le projet NoSQL_LauraDupuis doit être placé dans un répertoire sous Ubuntu 17.4.

0 - Installation de R

Pour télécharger et préparer les données, nous allons utiliser R. Pour cela, il faut dans un premier temps installer ce logiciel en exécutant les commandes suivantes sous un terminal :

— se placer sous le dossier NoSQL_LauraDupuis .

— exécuter la commande : `chmod a+x 00_installR.sh` pour rendre le script d'installation de R exécutable

— exécuter le script : `./00_installR.sh`

1 - Téléchargement et préparation des bases de données

Pour disposer des données, nous allons télécharger les bases de données choisies depuis le site https ://opendata.cityofnewyork.us/ . Pour cela, il faut rester sous le terminal, et toujours sous le même répertoire NoSQL_LauraDupuis, exécuter la commande `sudo RScript 01_dataExtract.R` .

Ce script va générer un répertoire "data" sous le répertoire courant, qui contient cinq jeux de données utilisés au format geojson. Ce format permet d'ajouter au format json le stockage de l'information géographique. Ainsi, nous arriverons à trouver un endroit précis pour localiser le meilleur endroit de New York pour y habiter.

— Le premier contient la liste des blocs de recensements de New-York. La ville est découpée en bloc qui permet de faire un maillage très fin. (censusBlock)
— Le deuxième contient une autre liste de bloc de recensements de la ville, mais avec un maillage un peu moins fin (donc il contient mois de lignes que le premier). (censusTracts)
— Le troisième correspond à la liste des stations de métro de la ville. (subways)
— Le quatrième contient la liste des parcs de New York. (parks)
— Le cinquième correspond à la liste des théâtres. (theaters)

Chacune de ces bases détient des coordonnées géographiques sous différentes formes. Les stations de métros et les théâtres sont sous forme de point (correspond à la latitude et la longitude). Les parcs et les blocs sont des polygones, c'est-à-dire une liste de points. D'autres attributs sont conservés dans les bases. Par exemple, pour les stations de métro, nous avons bien sûr retenu un identifiant, le nom des stations et les lignes qui y passent. Pour les parcs, nous conservons l'identifiant, le nom et la surface du parc. Pour les théâtres, nous gardons l'identifiant, le nom et l'adresse. Et pour les deux bases de bloc, nous ne retenons que les coordonnées et les identifiants.

2 - Choix et installation de la base de données

Nous avons choisi d'utiliser la base MongoDB. Après quelques recherches sur internet, elle permet de traiter efficacement les requêtes spatiales de format geojson. De plus, une aide en ligne est disponible pour utiliser des opérateurs sur ces données spatiales :
https ://docs.mongodb.com/manual/reference/operator/query-geospatial/
Cette base permet de traiter simplement les polygones et les points au format geojson.

Voici les étapes nécessaires à l'installation de MongoDB, toujours dans un terminal :

- Vérifier que vous êtes toujours dans le même répertoire courant
- Rendre le script d'installation exécutable avec `chmod a+x 02_mongoRoboInstall.sh`
- Exécuter le script d'installation : `./02_mongoRoboInstall`. Ce dernier installe mongoDB, mais aussi Robomongo, qui permet d'avoir une interface plus pratique à utiliser. Il lance ensuite MongoDB et Robomongo (ce qui peut nécessiter que votre mot de passe soit demandé).

Cependant, nous avons remarqué que cette manipulation n'est pas stable. Il arrive que ces commandes ne fonctionnent pas et que MongoDB ne s'installe pas en local. En ne sachant pas d'où vient le problème, il est aussi possible d'installer MongoDB depuis une VirtualBox. Pour cela, il faut installer une VirtualBox en amont. Puis, pour installer le client mongo il faut utiliser cette commande :

wget http ://91.121.220.23/mongo.vdi.gz

Il faut ensuite s'assurer que le MD5 du client mongo soit identique à :
" 1b0f6ea99737480259e1815f7b490e5e mongo.vdi.gz"

avec la commande suivante :
`md5sum mongo.vdi.gz`

De cette même façon, il est possible d'installer Robomongo depuis le terminal selon la commande suivante

wget https ://download.robomongo.org/1.1.1/linux/robo3t-1.1.1-linux-x86_64-c93c6b0.tar.gz

Il faut ensuite dézipper le fichier téléchargé et lancer Robomongo depuis le répertoire bin.

Pour pouvoir utiliser Robomongo, il faut lancer MongoDB sous la VirtualBox avec le login "root" et le mot de passe "root". Puis dans le shell mongo il faut saisir la commande `service mongod start`. Cette commande va activer la base de données. Robomongo pourra ensuite être lancé.

Dans le fenêtre de Robomongo qui vient de s'ouvrir, cliquer sur create pour créer une nouvelle connexion. Entrez les paramètres new_york dans le champ Name, et localhost :27017 dans le champ Address, puis cliquez sur Save. Sélectionnez alors la connection new_york que nous venons de créer, puis cliquez sur connect.

3 - Insertion en base des jeux de données

Les données au format geojson sont prêtes, il suffit de les insérer sous MongoDB. Pour cela, il faut exécuter les commandes suivantes sous le terminal, toujours sous le répertoire NoSQL_LauraDupuis (cd NoSQL_LauraDupuis) :

— `chmod a+x 03_importBDD.sh` , pour rendre le script d'installation des jeux de données en base exécutable

— `./03_importBDD.sh` . Les données sont chargées sous 5 collections dans la base new_york. Les cinq collections correspondent aux différents types de données enregistrées sous le répertoire data.

Comment trouver cet endroit idéal ? : Interrogation de la base de données

4 - Requêtes sur la base de données pour trouver l'endroit idéal

Toutes les requêtes pour trouver le meilleur endroit de la ville où habiter figurent dans le document `04_requete`. Le code doit être copier intégralement pour l'exécuter sous Robomongo.

En amont des requêtes, nous avons choisi de créer des index sur les cinq collections afin de pouvoir utiliser des requêtes géospatiales. Pour cela, les collections avec des points (subways et theaters) utilisent l'opérateur 2dsphere localisation. Les trois autres collections sous forme de polygone (parks, censusBlock, censusTracts) utilisent l'opérateur 2dsphere geometry.

On choisit tout d'abord de commencer par ne garder que les blocs (de censusTracts) qui contiennent des parcs grâce à l'opérateur $geoWithin. Tous ces blocs sont placés dans une collection temporaire, nommée tractsWithSubways.

Ensuite, nous conservons seulement parmi ces blocs, ceux qui intersectent un parc, puisque nous aimerions vivre à proximité d'un parc. Parmi ces blocs, nous conservons ceux dont l'intersection avec un parc n'est pas nulle grâce à l'opétarateur $geoIntersect. Les blocs correspondant sont recensés sous la collection tractsWithSubwaysAndParks. Ainsi, ces blocs sont suffisamment proches d'un arrêt de métro et d'un parc. Le nom et la surface du parc correspondant sont à chaque fois enregistrés sous cette collection.

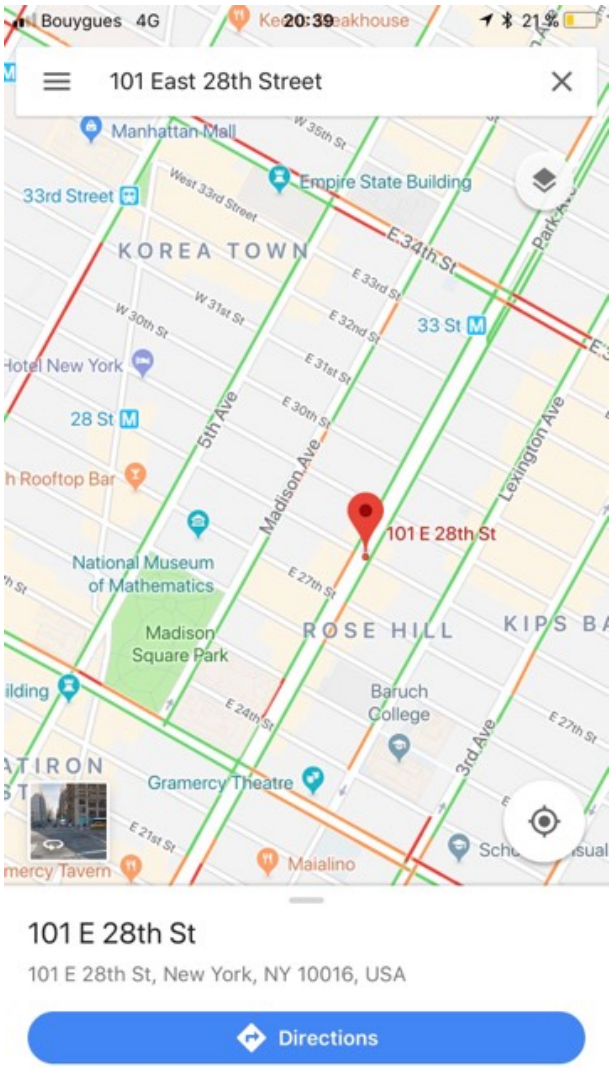
Puis, pour être à proximité d'un théâtre (à moins de 250 mètres), nous gardons parmi les blocs sélectionnés précedemment, les théatres situés à moins de 250 mètres. Pour cela, nous utilisons les opérateurs $nearSphere et $maxDistance. Cela permet d'avoir dans chaque bloc, le ou les théâtres dans un rayon de 250 mètres maximum.

Après cette requête, nous obtenons 5 blocs qui correspondent à nos attentes. Comme nous souhaitons être proche d'un assez grand parc, nous décidons de trier ces blocs selon la taille du parc associé. Le bloc retenu sera celui du parc le plus grand.

5 - L'endroit idéal trouvé

Le bloc trouvé correspond à un bloc de Manhattan proche du parc Madison Square Park de superficie de 6.234 à proximité de la station de métro 28th St où les lignes 4-6-6 Express circulent, à proximité du théâtre Gramercy Arts Theatre à l'adresse 138 E 27th St. Toutes les coordonnées de ce bloc peuvent donc nous convenir.

Toutes les adresses de ce bloc peuvent donc nous convenir. Nous avons choisi de prendre la première coordonnée du bloc puisque nous l'avons utilisée dans la requête pour trouver les théâtres à moins de 250 mètres. Nous choisissons les coordonnées du point suivant : longitude =-73.9840748526015 et latitude = 40.7433247184166. En visualisant ce point sur le site https ://www.coordonnees-gps.fr/conversion-coordonnees-gps, nous pouvons trouver l'adresse suivante : 101 E 28th St, NY, New York 10016, États-Unis.



Sur l'image ci-dessus, nous observons bien que l'adresse trouvée correspond tout à fait à nos critères de recherche. Il y a bien un théatre et un parc à proximité, ainsi que plusieurs stations de métro très proches.

6 - Améliorations possibles

Pour avoir plus de choix, et avoir accès à toutes les adresses possibles, nous pouvons retourner sur le site https ://data.cityofnewyork.us/ et identifier toutes les adresses qui correspondent aux critères recherchés pour trouver l'endroit idéal où vivre à New York.

Concernant les requêtes, il est possible d'effectuer les mêmes sur les censusBlocks, c'est à dire sur un maillage plus fin et donc plus précis de New York. Nous n'avons finalement pas utilisé cette base dans les requêtes, puisque le temps d'exécution était trop long. Il serait donc possible d'introduire un MapReduce poure exécuter les requêtes souhaitées. De plus, au moment de trouver les théâtres les plus proches, nous comparons des blocs avec des points. Nous avons donc choisi de travailler avec une des coordonnées des blocs (la première : coordinates[0][0][0]). Nous pensons qu'il est possible de travailler avec le bloc entier, mais cela n'a pas abouti à une requête s'exécutant sans erreur.