DATA WRANGLING

Wrangle OpenStreetMap data with MongoDB

APRIL 23, 2017

FAHAD ALHAJJAJ

Contents

OBTAINED DATA 2

PROBLEMS ENCOUNTERED IN YOUR MAP 2

STREET TYPES AND STREET NAMES INCONSISTENCY 2

POSTCODE INCONSISTENCY & ACCURACY 2

NON-ASCII CHARACTERS AND UNICODE TYPE CHARACTERS 2

OVERVIEW OF THE DATA 3

NUMBER OF DOCUMENTS IN BROOKLYN COLLECTION 3

NUMBER OF DIFFERENT NODE TYPES IN BROOKLYN COLLECTION 3

NUMBER OF UNIQUE USERS 3

TOP CONTRIBUTOR 3

TOP 10 CONTRIBUTORS 3

NUMBER OF USERS WHO CONTRIBUTED ONLY ONCE 3

TOP 10 POSTCODES RECORDED IN BROOKLYN COLLECTION 4

TOP 10 CITIES RECORDED IN BROOKLYN COLLECTION 4

TOP 5 AMENITIES RECORDED IN BROOKLYN COLLECTION 4

TOP 5 CUISINES RECORDED IN BROOKLYN COLLECTION 4

OTHER IDEAS ABOUT THE DATASET 5

CONCLUSION 5

CITATION 6

Obtained Data

Map Area: Brooklyn, NY, USA. Mapzen, Download

Zip code database from unitedstateszipcodes.org

Problems encountered in your map

Street types and street names inconsistency

Street types were inconsistence, for example, 'Street' type had more than 5 different cases and misspelling. These are some of them, 'St', 'St.', 'St.,', 'St,', 'st', 'ST' and 'Streeet'. Also some street names had multiple street name or more information than a street name, adding a suit number.

Different street types was handled By using a python Dictionary. These multiple types were replaced to one type 'Street'. And that happened to all different type cases like 'Ave' was replaced by Avenue and so on.

Grand St → Grand Street

More than one name and having multiple name at once was handled by removing any thing after a street type.

Smith St & Bergen St → Smith Street

Main St., Suite 500 → Main Street

Some name did not fit in any of the above category, so they were ignored and not included in the database. Examples of street names that were ignored:

218650358, Bowery, The Bowery, Macdougal.

Postcode inconsistency & accuracy

Postcodes had some inconsistencies in them. Some had a leading 'NY' and some had two postcodes separated by a ';' and some had an addition 4 digits. To fix all of that, the state initials were deleted, Only one postcode is accepted, And the additional 4 digits were deleted:

NY 10003 → 10003

11214;11223 → 11214

11224-4003 → 11224

There was only one case were it was obvious that it wasn't a postcode and it was ignored:

(718) 778-0140

To make sure all postcodes were valid, a zip code database was obtained from unitedstateszipcodes.org and all postcode from the osm file were validated.

Non-ASCII characters and Unicode type characters

All non-ASCII characters that were found in the osm file were ignored and their tag was not included in the database. Example ' Сундучок'.

Overview of the data

The OSM file uncompressed size: 638 MB

The JSON file size: 955 MB

Number of documents in Brooklyn collection

```
> db.brooklyn.find().count()
```

```
2991649
```

Number of different node types in Brooklyn collection

```
> db.brooklyn.aggregate([{$group:{_id: "$node_type", total: {$sum : 1}}}])
```

```
{ "_id" : "way", "total" : 492999 }
{ "_id" : "relation", "total" : 2070 }
{ "_id" : "node", "total" : 2496580 }
```

Number of unique users

```
db.brooklyn.distinct("created.user").length
```

```
1692
```

Top contributor

```
> db.brooklyn.aggregate([{$group:{_id: "$created.user", total: {$sum : 1}}}, { $sort : { total : -1}}, { $limit: 1}])
```

```
{ "_id" : "Rub21_nycbuildings", "total" : 1739725 }
```

Top 10 contributors

```
> db.brooklyn.aggregate([{$group:{_id: "$created.user", total: {$sum : 1}}}, { $sort : { total : -1}}, { $limit: 10}])
```

```
{ "_id" : "Rub21_nycbuildings", "total" : 1739725 }
{ "_id" : "ingalls_nycbuildings", "total" : 373576 }
{ "_id" : "ediyes_nycbuildings", "total" : 189681 }
{ "_id" : "celosia_nycbuildings", "total" : 117294 }
{ "_id" : "ingalls", "total" : 105079 }
{ "_id" : "lxbarth_nycbuildings", "total" : 79596 }
{ "_id" : "aaron_nycbuildings", "total" : 41996 }
{ "_id" : "ewedistrict_nycbuildings", "total" : 35016 }
{ "_id" : "smlevine", "total" : 25017 }
{ "_id" : "robgeb", "total" : 24400 }
```

Number of users who contributed only once

```
> db.brooklyn.aggregate([{$group:{_id: "$created.user", total :{$sum:1}}}, {$group:{_id: "$total", num_users:{$sum: 1}}}, {$sort :{_id: 1}}, { $limit: 1 }])
```

```
{ "_id" : 1, "num_users" : 524 }
```

Top 10 postcodes recorded in Brooklyn collection

```
> db.brooklyn.aggregate([{"$match": {"address.postcode": { "$exists": true}}},
{$group:{_id: "$address.postcode", total: {$sum : 1}}}, { $sort : { total : -1}}, { $limit: 10
}])
```

```
{ "_id" : "11234", "total" : 20141 }
{ "_id" : "11236", "total" : 15232 }
{ "_id" : "11385", "total" : 15101 }
{ "_id" : "11229", "total" : 12545 }
{ "_id" : "11203", "total" : 11785 }
{ "_id" : "11207", "total" : 11188 }
{ "_id" : "11208", "total" : 11142 }
{ "_id" : "11223", "total" : 10763 }
{ "_id" : "11214", "total" : 10061 }
{ "_id" : "11204", "total" : 9904 }
```

Top 10 cities recorded in Brooklyn collection

```
> db.brooklyn.aggregate([{"$match": {"address.city": { "$exists": true}}},{$group:{_id:
"$address.city", total: {$sum : 1}}}, { $sort : { total : -1}}, { $limit: 10 }])
```

```
{ "_id" : "Brooklyn", "total" : 1801 }
{ "_id" : "New York", "total" : 1433 }
{ "_id" : "New York City", "total" : 110 }
{ "_id" : "Hoboken", "total" : 91 }
{ "_id" : "Jersey City", "total" : 63 }
{ "_id" : "Corona", "total" : 48 }
{ "_id" : "Forest Hills", "total" : 46 }
{ "_id" : "Brookklyn", "total" : 22 }
{ "_id" : "Rego Park", "total" : 19 }
{ "_id" : "Elmhurst", "total" : 15 }
```

Top 5 amenities recorded in Brooklyn collection

```
> db.brooklyn.aggregate([{"$match": {"amenity": { "$exists": true}}},{$group:{_id:
"$amenity", total: {$sum : 1}}}, { $sort : { total : -1}}, { $limit: 5 }])
```

```
{ "_id" : "bicycle_parking", "total" : 2824 }
{ "_id" : "restaurant", "total" : 1188 }
{ "_id" : "place_of_worship", "total" : 937 }
{ "_id" : "parking", "total" : 769 }
{ "_id" : "school", "total" : 614 }
```

Top 5 cuisines recorded in Brooklyn collection

```
> db.brooklyn.aggregate([{"$match": {"cuisine": { "$exists": true}}},{$group:{_id:
"$cuisine", total: {$sum : 1}}}, { $sort : { total : -1}}, { $limit: 5 }])
```

```
{ "_id" : "coffee_shop", "total" : 104 }
{ "_id" : "pizza", "total" : 92 }
{ "_id" : "mexican", "total" : 79 }
{ "_id" : "american", "total" : 72 }
{ "_id" : "burger", "total" : 71 }
```

Other ideas about the dataset

The number of users who contributed in the database was small ‘1692’ even though the database was quite large. The top 10 contributors have covered 91.3% of the database and the number 1 contributor ‘Rub21_nycbuildings’, has 1739725 contributions out of 2991649 which is 58.2% of all contributions. 31% of users only contributed once.

The OpenStreetMap can encourage people to contribute more by recognizing their efforts, by giving them contribution badges, for example, contributor of the month badge. Also, OpenStreetMap can give contributors exclusive invitations to city events and new contributors can get exclusive offers when they contribute a certain number of times.

One benefit of this solution is that more people will be encourage to participate either by creating new nodes, editing existing node, or adding new tags to existing nodes. Another benefit is that people can invite and advertise for the site and the database because of what they get out of contributing on the website.

However, more people mean more human errors and mistakes, which means more time auditing data. Also, some users might add bad information because they don’t have the time to put good ones, and the reason of doing that is to get contribution points and benefits.

Conclusion

The database had a few problems. In this project, many of those problems were found and the data was cleaned. OpenStreetMap can create an automated program that can monitor, detect and fix problems when data is entered into the database.

Citation

1. Udacity.com: Working with Mongodb
2. unitedstateszipcodes.org: Zip code database
3. Udacity.com: MongoDB Sample Project