OpenStreetMap Sample Project Data Wrangling with MongoDB
Shalva Usubov

Choosing map area

I live in Krasnodar, Russia so I decide take Krasnodar city for analysis.

Map Area: Krasnodar, Krasnodar Krai, Russia
https://www.openstreetmap.org/relation/1477036

The XML file which I downloaded is 127M (11M as .zip file). I pushed on github only .zip file, because GitHub have limit 100M per file.

1. Problems Encountered in the Map

After downloading the Krasnodar area and running analysis, I noticed three main problems with the data, which I will discuss in the following order:

- Wrong street names ("Северная улица" instead "улица Северная")
- Wrong postal codes ("35001" instead "350001")
- "Incorrect" phone code format ("2671657" instead "+7 861 267-16-57")

Wrong street names

Russia has next common street types:
"улица", "бульвар", "проспект", "площадь", "переулок", "шоссе", "набережная", "проезд", "обход", "трасса", "аллея"

The first street type "улица" the most common and address should start fro this type ('улица Северная') but other street types can be in the end of street names.

And next mapping:
mapping = {
        "ул.": "улица",
        "пр.": "проспект",
        "пл." : "площадь"
      }

Example of transformation:

- Вишневая улица => улица Вишневая
- Осторожный переулок => Осторожный переулок (leave as is)
- 40-лет Победы => улица 40-лет Победы

Wrong postal codes

In general postal codes looks good, I found only 4 broken codes. For Krasnodar city postal code range from 350000 to 359999, but .osm contain additional postal codes from Adygea region https://en.wikipedia.org/wiki/Adygea . That because Adygea region neighbor of the city of Krasnodar, two regions divided by the river (with a width of 500 meters).

Example of transformation:

- 35001 converted to 350001
- 14 converted to 350000
- 305901 converted to 350000

"Incorrect" phone code format

The most common issue was with missing code of city before Krasnodar local 7 digit number.

Example of transformation:

- 2617234 converted to +7 861 261-72-34
- 786122530738 converted to +7 861 253-07-38. Here issue with double '22' after Krasnodar migrate from 6 digit to 7 digit number at 30.11.2004 http://www.yuga.ru/articles/society/3540.html
- 78612553006 to +7 861 255-30-06

2. Overview of the Data

This section contains basic statistics about the dataset and the MongoDB queries used to gather them.

File sizes

krasnodar.osm   127 MB
krasnodar.osm.json   163 MB

```
# Number of documents
> db.krasnodar.find().count()
604682

# Number of nodes
> db.krasnodar.find({"type":"node"}).count()
504100

# Number of ways
> db.krasnodar.find({"type":"way"}).count()
100562

# Number of unique users
> db.krasnodar.distinct("created.user").length
414

# Number of kindergarten
> db.krasnodar.find({"amenity":"kindergarten"}).count()
196

# Number of banks
> db.krasnodar.find({"amenity":"bank"}).count()
169

# Number of cinemas
> db.krasnodar.find({"amenity":"cinema"}).count()
7

# Top 3 contributing user
> db.krasnodar.aggregate([{"$group":{"_id":"$created.user", "count":{"$sum":1}}},
{"$sort":{"count":-1}}, {"$limit":3}])
{ "_id" : "Kotelnikov", "count" : 188197 }
{ "_id" : "Vadimi", "count" : 85564 }
{ "_id" : "Nikolya", "count" : 64927 }

# Number of users appearing only once (having 1 post)
> db.krasnodar.aggregate([{"$group":{"_id":"$created.user", "count":{"$sum":1}}},
{"$group":{"_id":"$count", "num_users":{"$sum":1}}}, {"$sort":{"_id":1}}, {"$limit":1}])
{ "_id" : 1, "num_users" : 66 }
```

3. Additional Data Overview

Biggest religion
> db.krasnodar.aggregate([{"$match":{"amenity":"place_of_worship"}},
{"$group":{"_id":"$religion","count":{"$sum":1}}}, {"$sort":{"count":-1}},{"$limit":10}])
{ "_id" : "christian", "count" : 31 }
{ "_id" : null, "count" : 5 }
{ "_id" : "muslim", "count" : 3 }

It is expected for the Krasnodar region.

Top 10 amenities

> db.krasnodar.aggregate([{"$match":{"amenity":{"$exists":1}}}, {"$group":{"_id":"$amenity",
"count":{"$sum":1}}}, {"$sort" : {"count" : -1}}, {"$limit" : 10}])
{ "_id" : "parking", "count" : 708 }
{ "_id" : "atm", "count" : 233 }
{ "_id" : "pharmacy", "count" : 223 }
{ "_id" : "cafe", "count" : 216 }
{ "_id" : "kindergarten", "count" : 196 }
{ "_id" : "bank", "count" : 169 }
{ "_id" : "fuel", "count" : 139 }
{ "_id" : "car_wash", "count" : 131 }
{ "_id" : "restaurant", "count" : 116 }
{ "_id" : "school", "count" : 115 }

It looks strange that the top 1 amenity is parking, probably that because in Krasnodar open street maps is very popular for car drivers, as example Google maps much worse than open street maps for Krasnodar and probably car drivers more active contributors.

Will be good stimulate restaurant and cafe fill their contacts to OSM, I am sure in Krasnodar we have much more restaurant/cafe than parking. Probably will can fix this issue by export data from TripAdvisor, usually restaurant/cafe care about put their contacts to TripAdvisor. As example TripAdvisor shows that in Krasnodar we have 851 restaurants
http://www.tripadvisor.ru/Tourism-g298532-Krasnodar_Krasnodar_Krai_Southern_District-Vacations.html

Conclusion

After this review of the data it's obvious that the Krasnodar area is incomplete, though I believe it has been well cleaned for the purposes of this exercise. I not found the big issues in data, all issues can be fixed with simple scripts. I cleaned street names, postal codes and phone numbers. That was great experience about find and fix issue in OSM data, I will join to OSM contributor team member. I will try use data from current research to improve Krasnodar city OSM data and then will check TripAdvisor API for able import data to OSM.

List of Resources

1. Openstreetmap http://www.openstreetmap.org
2. MongoDB Documentation http://docs.mongodb.org/manual/
3. Krasnodar change phone number format http://www.yuga.ru/articles/society/3540.html
4. Krasnodar city http://en.wikipedia.org/wiki/Krasnodar