OpenStreetMap Data Wrangle Project with MongoDB

Meixian Chen

Map area

Zurich, Switzerland
Data source: https://mapzen.com/data/metro-extracts/metro/zurich_switzerland/

I live in Switzerland now and visit Zurich often. I would like to explore a bit more about it during working on this project.

1.Problem Encountered in the Map:

Postcode:

The postcode code in Switzerland is 4 digital, with the country code CH. For example: 8700 CH. The country codes in the map are all correct, “CH”. The postcode in the map is of the right format, but considering it is the Canton Zurich area, postcode should be 8xxx, I eliminate the records with postcode not starting with 8.

Name tag:

Switzerland has four official languages, and many people from different countries living in Zurich. The name tags from the data are many different language.
However, for some records, the names are stored in multiple name language tags even if they are the same, here are two examples:
<tag k="name:de" v="Il Palazzetto"/>
<tag k="name:en" v="Il Palazzetto"/>
<tag k="name:fr" v="Il Palazzetto"/>

<tag k="name:fr" v="Zurich"/>
<tag k="name:en" v="Zurich"/>
<tag k="name:cn" v="苏黎世"/>

I propose to transform them into a more concise format:
for the above examples, the new representation is respectively:
{“name”: “IL Palazzeto”} and {“name”:[“zurich”,"苏黎世"]}

Beside, “name:source”, “name:botanical”, and “name:left” should be dealed differently with “name:language”.
I transform
<tag k="name:botanical" v="Tilia cordata"/>
to {“botanical”:”Tilia cordata”}

Recycling:

Recycling is an important part of swiss lift.
The recycling records is stored as:
<tag k="recycling:books" v="no"/>
<tag k="recycling:glass" v="yes"/>
<tag k="recycling:paper" v="no"/>
<tag k="recycling:scrap_metal" v="yes"/>

I propose to focus on possible recycling goods, and transform the record into following way, which is easier for further analysis:
{“recycling”:[“glass”,”scrap_metal”]}

2.Overview of the map

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

File size:

zurich_switzerland.osm : 590MB
clean.json: 766MB

Number of documents:
# sulishi is the Chinese pingyin of Zurich
db.sulishi.find().count()
3146330

Number of nodes:
db.sulishi.find({"type":"node"}).count()
2718046

#Number of ways:
db.sulishi.find({"type":"way"}).count()
427559

Number of unique users
2340

Top 5 users with the most contribution to the map:
db.sulishi.aggregate([{"$match":{"created.user":{"$exists":1}}},
{"$group":{"_id":"$created.user","count":{"$sum":1}}},

```
    {"$sort":{"count":-1}},{"$limit":5}])

[{u'_id': u'mdk', u'count': 493724},
 {u'_id': u'SimonPoole', u'count': 292844},
 {u'_id': u'Sarob', u'count': 129066},
 {u'_id': u'hecktor', u'count': 96271},
 {u'_id': u'feuerstein', u'count': 93636}]
```

The total 5 users contribute to 35% of the data of the map.

3.Data of daily life:

This section I only present the result, the mongoDB queries can be found on the code file.

Amenities

Number of amenities
26914

Top 20 amenities

```
[{u'_id': u'parking', u'count': 5503},
 {u'_id': u'bench', u'count': 3736},
 {u'_id': u'restaurant', u'count': 2199},
 {u'_id': u'drinking_water', u'count': 1574},
 {u'_id': u'school', u'count': 1110},
 {u'_id': u'post_box', u'count': 1021},
 {u'_id': u'waste_basket', u'count': 972},
 {u'_id': u'vending_machine', u'count': 892},
 {u'_id': u'bicycle_parking', u'count': 864},
 {u'_id': u'parking_entrance', u'count': 596},
 {u'_id': u'recycling', u'count': 590},
 {u'_id': u'place_of_worship', u'count': 486},
 {u'_id': u'kindergarten', u'count': 443},
 {u'_id': u'fuel', u'count': 432},
 {u'_id': u'toilets', u'count': 421},
 {u'_id': u'cafe', u'count': 421},
 {u'_id': u'fast_food', u'count': 415},
 {u'_id': u'fountain', u'count': 406},
 {u'_id': u'car_sharing', u'count': 400},
 {u'_id': u'atm', u'count': 386}]
```

People complains about food in German speaking area, actually, there are many different cuisine in Zurich.
Top 10 cuisine in Zurich area:

[{u'_id': u'regional', u'count': 271},
 {u'_id': u'italian', u'count': 211},
 {u'_id': u'pizza', u'count': 133},
 {u'_id': u'asian', u'count': 71},
 {u'_id': u'thai', u'count': 67},
 {u'_id': u'burger', u'count': 66},
 {u'_id': u'kebab', u'count': 60},
 {u'_id': u'chinese', u'count': 46},
 {u'_id': u'coffee_shop', u'count': 45},
 {u'_id': u'international', u'count': 43}]

Zurich is a famous trourist city:
Top 10 tourist place
[{u'_id': u'information', u'count': 1397},
 {u'_id': u'picnic_site', u'count': 559},
 {u'_id': u'hotel', u'count': 248},
 {u'_id': u'artwork', u'count': 234},
 {u'_id': u'viewpoint', u'count': 207},
 {u'_id': u'museum', u'count': 139},
 {u'_id': u'attraction', u'count': 116},
 {u'_id': u'guest_house', u'count': 22},
 {u'_id': u'camp_site', u'count': 15},
 {u'_id': u'hostel', u'count': 10}]

Recycling is an important part of swiss life. (You might not know a 35L trash bag costs around 2 CHF (about 2 US dollar) in Zurich. Recycling protects the earth and also your wallet.)

Recycling spot:
[{u'_id': u'underground', u'count': 1},
 {u'_id': u'dump', u'count': 1},
 {u'_id': u'underfloor-container', u'count': 1},
 {u'_id': u'centre', u'count': 41},
 {u'_id': u'container', u'count': 204}]

top 5 items with the most recycling spot
[{u'_id': u'glass', u'count': 258},
 {u'_id': u'cans', u'count': 234},
 {u'_id': u'clothes', u'count': 161},

{u'_id': u'glass_bottles', u'count': 113},
{u'_id': u'scrap_metal', u'count': 103}]

top 5 items with the fewest recycling spot
[{u'_id': u'tin', u'count': 1},
{u'_id': u'hardcore', u'count': 1},
{u'_id': u'compost', u'count': 1},
{u'_id': u'aerosol_cans', u'count': 1},
{u'_id': u'dvds', u'count': 1}]

4.Suggestion of improvement

I found out the records about traffic are not organized in a clear way, for example
<tag k="highway" v="crossing"/>
<tag k="crossing" v="traffic_signals"/>
<tag k="traffic_signals" v="signal"/>
<tag k="traffic_sign" v="city_limit"/>

There are different key names related to traffic information, and the values sometime could be the key of other tag.
A better representation to organize the data tags could be:
<tag k="traffic:crossing" v="traffic_signals"/>
<tag k="traffic:signals" v="signal"/>
<tag k="traffic:sign" v="city_limit"/>

The main benefit of the new representation is, when searching about traffic-related information, we would not miss some records due to the different naming of keys. The new key names enable us to search the traffic signs by looking up all keys name starting with “traffic”. Moreover, we could nest different types of travel information under the “traffic” field in the json format.

However, the new implementation still consist of a lot of redundant information (eg. crossing: traffic_signals, traffic_signals:signal). Another issue about traffic is that, even standing on the same location, a traffic sign facing different direction might give different instructions. We should also consider these when creating and working on the data.

5.Some thoughts

It is very interesting to see a city from data.
Query with MongoDB(with pymongo library) is so much faster than analysis text with python.