OpenStreetMap Data Case Study

Jungmin Park (Jungmin.j.park@gmail.com)

1. Map Area

Boston, United States (https://mapzen.com/data/metro-extracts/metro/boston_massachusetts/)
The OSM XMS file is 414 MB.
I chose this area since I want to visit this area in a few months for a tour.

2. Problem Encountered in the Map

Unnecessary abbreviated street names

: After I parsed the data, I figured out street names are abbreviated, such as “Brentwood St“,“Western Ave“, and “Soldiers Field Rd“. I think we need to remove abbreviations to make easily recognize what they mean and to make them same format. Therefore, I changed those abbreviations into the actual words such as “Street“, “Avenue“, and “Road“.
Also, I found that some street names are wrong, such as “First Street, Suite 1100“ and “Franklin Street, Suite 1702“, so I changed them into ‘First Street’ and ‘Franklin Street’ respectively.

Incorrect postal codes

: Boston area zip codes all begin with “02” however some portion of zip codes were outside this region. However, there are zip codes starting with “01”. In this case, I changed those zip codes into '00000' to distinguish from other zip codes when I need to analyze dataset.
Also, there are several zip codes whose length is over 5. However, the length of zip codes should be 5. So I updated them, too.

3. Overview statistics

Data file size

: The raw data which is XML format is 414MB.

```
os.path.getsize(OSMFILE)/1024/1024
```

I parsed it into Json format and the size of that Json file is 640MB.

```
os.path.getsize(os.path.join(path, "boston_massachusetts.osm.json"))/1024/1024
```

Number of unique users: 1,187

```
len(collection.group(["created.uid"], {}, {"count":0}, "function(o, p){p.count++}"))
```

Number of nodes: 1,932,545

```
collection.find({"type":"node"}).count()
```

Number of ways: 308,568

```
collection.find({"type":"way"}).count()
```

4. Additional exploration

Top three users contribute most of dataset. It is supposed that they were forced to insert the initial dataset. Especially, 'crschmidt' contributed approximately 54% of map.

```
# Top three users with most contributions

pipeline = [{"$group":{"_id": "$created.user",
                       "count": {"$sum": 1}}},
            {"$sort": {"count": -1}},
            {"$limit": 3}]
result = collection.aggregate(pipeline)

for x in xrange(3):
    get_record = result.next()
    print get_record
```

```
{u'count': 1204304, u'_id': u'crschmidt'}
{u'count': 430624, u'_id': u'jremillard-massgis'}
{u'count': 92178, u'_id': u'OceanVortex'}
```

```
# Proportion of the top three users' contributions
pipeline = [{"$group":{"_id": "$created.user",
                       "count": {"$sum": 1}}},
            {"$project": {"proportion": {"$divide" :["$count",collection.find().count()]}}},
            {"$sort": {"proportion": -1}},
            {"$limit": 3}]

result = collection.aggregate(pipeline)

for x in xrange(3):
    get_record = result.next()
    print get_record
```

```
{u'_id': u'crschmidt', u'proportion': 0.5372625087216202}
{u'_id': u'jremillard-massgis', u'proportion': 0.19210940971360965}
{u'_id': u'OceanVortex', u'proportion': 0.041122327526057795}
```

5. Other ideas about the datasets

Because the open street map dataset is a human edited dataset, there were some errors as we expected. I tried to update data but actually the decision to handle with wrong data has to be changed depending on how to analyze it. The best way is finding the true data and updating the wrong data into the true data.

To improve data quality, next time we could use the DOT Address Validation API from USPS to correct error and to add more detailed address information into the dataset. The following

address is the link of the DOT Address Validation API service.
: https://www.usps.com/business/web-tools-apis/welcome.htm

By cross validating using this API service, we can get much better data quality.

Also, I think if 'OpenStreetMap.org' offers the input screen to constrain what we can put the correct data, such as 5 digit Zipcode, the dataset could be more precise.