OpenStreetMap Sample Project
Data Wrangling with MongoDB

Prasad Pagade, 1/6/2017

Map Area: Santa Clara, CA, United States

Map Link - https://www.openstreetmap.org/relation/2221647#map=12/37.3875/-121.9232

1 Problems Encountered in the Map

After downloading the dataset and running it, I noticed two problems with the data.

1. Over-abbreviated street names
2. Inconsistent postal codes

Over-abbreviated street names:

After parsing the raw xml data file, I found some street names are abbreviated, such as "Wolfe St", “wilcox ave” and “Homestead Rd”. I updated these abbreviations with the actual words like “Street”, “Avenue” and “Road”.

Inconsistent postal codes:

I cleaned up some inconsistent postal codes like “CA 95054” to “95054”

2 Data Overview

Data file size The original downloaded OpenStreetMap in XML format is of size 58 MB. I parsed it into JSON format with the street types corrected, and the resulting JSON file is of size 62MB. The attached sample “.osm” file is a sample of the original XML file, with first hundred elements been selected.

Summary statistics of the dataset

Below, are some summary statistics of the dataset as well as the python code that are used to generate these results.

Number of documents: 2180736

```
collection.find().count()
```

Number of unique users: 534

```
len(collection.distinct('user'))
```

Number of nodes: 263052

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

Number of ways: 39339

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

3 Some more exploration on the data set

Top 5 contributers:

```
result = collection.aggregate( [    { "$group" : {"_id" : "$user",
                                    "count" : { "$sum" : 1} } },
                                    { "$sort" : {"count" : -1} },
                                    { "$limit" : 5 } ] )
print(list(result))
```

```
[{u'count': 58912, u'_id': u'samely'}, {u'count': 34481, u'_id': u'dannyka
th'}, {u'count': 23884, u'_id': u'RichRico'}, {u'count': 17966, u'_id': u'
karitotp'}, {u'count': 15622, u'_id': u'matthieun'}]
```

Top 10 amenities in Santa Clara

```
amenity = collection.aggregate([{'$match': {'amenity': {'$exists': 1}}}, \
                               {'$group': {'_id': '$amenity', \
                                           'count': {'$sum': 1}}}, \
                               {'$sort': {'count': -1}}, \
                               {'$limit': 10}])
print(list(amenity))
```

```
[{u'count': 354, u'_id': u'parking'}, {u'count': 237, u'_id': u'restaurant'},
{u'count': 113, u'_id': u'fast_food'}, {u'count': 71, u'_id': u'school'}, {u'
count': 61, u'_id': u'cafe'}, {u'count': 58, u'_id': u'place_of_worship'}, {u
'count': 48, u'_id': u'fuel'}, {u'count': 42, u'_id': u'parking_entrance'}, {
u'count': 31, u'_id': u'toilets'}, {u'count': 27, u'_id': u'bench'}]
```

Top 10 cuisines in Santa Clara:

```
cuisine = collection.aggregate([{"$match":{"amenity":{"$exists":1},
                        "amenity":"restaurant",}},
                {"$group":{"_id":{"Food":"$cuisine"},
                        "count":{"$sum":1}}},
                {"$project":{"_id":0,
                         "Food":"$_id.Food",
                         "Count":"$count"}},
                {"$sort":{"Count":-1}},
                {"$limit":10}])
print(list(cuisine))
```

```
[{u'Food': None, u'Count': 77}, {u'Food': u'mexican', u'Count':
17}, {u'Food': u'pizza', u'Count': 15}, {u'Food': u'indian', u'C
ount': 15}, {u'Food': u'chinese', u'Count': 11}, {u'Food': u'san
dwich', u'Count': 9}, {u'Food': u'sushi', u'Count': 9}, {u'Food'
: u'thai', u'Count': 8}, {u'Food': u'american', u'Count': 8}, {u
'Food': u'japanese', u'Count': 8}]
```

One thing that caught my attention was that Indian and Chinese cuisines are quite popular in Santa Clara. This is true for the fact that there are lot of Indian and Chinese engineers residing in Santa Clara give that it is in the heart of the Silicon Valley.

4. Conclusion

Summary statistics

- size of the file --> The original OSM file is **58.944542 MB** --> The JSON file is **61.66255 MB**
- number of unique users --> **534**
- number of nodes and ways --> nodes: **263052** ways: **263052**

Ideas to improve the data

I found that the Santa Clara OSM data was fairly clean. There are 534 unique users contributing to the data. Every person has their own way of addressing street names. We can eliminate these types of human errors by keeping a data entry form which has pre-defined naming conventions. Also, this dataset can be better understood by visualizations using a visualization tool like Tableau. This would helps us compare different nodes in the data visually.

- Cost of implementation: We need a team and structure that can maintain the above suggestions. Once the data entry form is set up, only minimal supervision would be required to make any changes periodically.