ATLANTA- OpenStreetMap Data Wrangling

Atlanta is the capital of and the most populous city in the U.S. state of Georgia, with an estimated 2015 population of 463,878. Atlanta is the cultural and economic center of the Atlanta metropolitan area, home to 5,522,942 people and the ninth largest metropolitan area in the United States. Atlanta is the county seat of Fulton County, and a small portion of the city extends eastward into DeKalb County.

1. Problems Encountered in Data

After loading data, we did some auditing using the code from ‘audit.py’ module. We found few problems in data which are summarized below.

Function used: audit.audit(DATAFILE)

1.1 Wrong Street Names

We found out that there are some street names which do not make any sense.

a. Some street names only had number e.g. 8343 and 814
b. Some street names include the house number in the address e.g. Williams Dr #1012

1.2 Inconsistent Capitalization

We also found that that street names did not have uniform capitalization. E.g. Keeneland blvd, Concourse parkway
This was solved by using *name.title()* from python.

1.3 Parenthesis in name

We also saw only one address which had parenthesis in its name. i.e. US 41 (GA).

1.4 Wrong Address

We also found one entry which had the whole address as street address i.e. Lawrenceville Suwanee Road Northwest suite 109
It was fixed manually by breaking into address components.

```
if child.attrib['v'] == '9700 Medlock Bridge Rd.,Suite 186, Johns Creek, GA, 30097':
            node['address']['street']=update_name('9700 Medlock Bridge Rd.')
            node['address']['postcode'] ='30097'
            node['address']['housenumber'] ='186'
            node['address']['city']='Johns Creek'
            node['address']['state'] ='GA'
```

Now the result is

```
u'address': {u'city': u'Johns Creek',
u'housenumber': u'186',
u'postcode': u'30097',
u'state': u'GA',
u'street': u'9700 Medlock Bridge Road'},
```

1.5 Issues with Attribute 'k'

Function used: audit.process_map(DATAFILE)
We used above function to analyze issues with the 'k' attribute values and found following results. There are 6 values which have char problems.

```
{'lower': 2880491, 'lower_colon': 2445222, 'other': 911771, 'problemchars': 6}
```

1.6 Inconsistent Street Name Abbreviations

In our data, we found that there are too many inconsistenecies the way street names end e.g Avenue exists as => Ave, Av, Ave. .
So before converting the data into json, we updated the names of streets for consistency.
Function used: audit.update_name()

```
Bridge Mill Ave => Bridge Mill Avenue
Westchester Club Dr => Westchester Club Drive
```

1.7 ZipCode in issues

We used following query to get total records which don't have a zipcode in them

We also checked all existing zipcodes and found following issues
Query:

a. *Query: db.atlanta.find({'address.postcode':{'$exists':0},'address':{'$exists':1}}).count()*
   We found that those 23887 records who have address element, do no have zipcode in them e.g.
   ```
   u'address': {u'county': u'Fulton', u'housenumber': u'950', u'state': u'GA'},
   ```
b. Query:
   db.atlanta.aggregate([{'$match':{'address.postcode':{'$exists':1}}},{'$group':{'_id':'$address.postcode','count':{'$sum':1}}},{'$sort':{'count':-1}}])

   We also found some zipcodes don't have zipcode. They either have city or state. E.g.
   ```
   {u'_id': u'Atlanta,', u'count': 1}
   {u'_id': u'Georgia', u'count': 1}
   ```
c. Few Zipcodes didn't even belong to Atlanta e.g There are 9 records with code
   {u'_id': u'80083', u'count': 9}
   These zipcodes belong to Albert Lea, MN.

2. Overview of Data

We will give you overall size of data. It was huge dataset which took a lot of time to process. Just by looking at the number of tags found for only node, we can see that there are more thatn 25 Million records.

Function: audit.count_tags(DATAFILE)

```
{'bounds': 1,
 'member': 37313,
 'nd': 13250389,
 'node': 11701075,
 'osm': 1,
 'relation': 4181,
 'tag': 6237490,
 'way': 839060}
```

We ran queries from the MongoDB data also which matched for both ‘way’ and ‘node’

2.1 File Sizes

atlanta_georgia.osm: 2.51 GB
atlanta_georgia.osm.json: 2.71 GB

2.2 Total Counts

```
>db.atlanta.count()
12540135
```

2.3 Counts of Nodes

```
>db.atlanta.find({'type':'node'}).count()
11701075
```

2.4 Count of ways

```
>db.atlanta.find({'type':'way'}).count()
839060
```

2.5 Total Contributing Users

```
>len(db.atlanta.distinct('created.user'))
1966
```

3. Additional Insights

Let us explore bit more results by creating some pipelines.

3.1 Most Common Speed and Highest Speed

```
>db.atlanta.aggregate([{'$match':{'maxspeed':{'$exists':1}}},{'$group':{'_id':'$maxspeed','count':{'$sum':1}}},{'$sort':{'count':-1}},{'$limit':10}])
{u'_id': u'45 mph', u'count': 2350}
{u'_id': u'55 mph', u'count': 1699}
{u'_id': u'35 mph', u'count': 1691}
{u'_id': u'25 mph', u'count': 1171}
{u'_id': u'65 mph', u'count': 1050}
{u'_id': u'70 mph', u'count': 603}
{u'_id': u'40 mph', u'count': 409}
```

```
     {u'_id': u'30 mph', u'count': 358}
     {u'_id': u'15 mph', u'count': 285}
     {u'_id': u'50 mph', u'count': 158}
```

Majority of places have speed of 45 mph. Highest speed allowed is 70 mph.

3.2 Smoothness of Highways

```
     >db.atlanta.aggregate([{'$match':{'smoothness':{'$exists':1}}},{'$group':{'_id':'$smoothness',
     'count':{'$sum':1}}},{'$sort':{'count':-1}},{'$limit':10}])
```

```
{u'_id': u'good', u'count': 1387}
{u'_id': u'excellent', u'count': 1111}
{u'_id': u'intermediate', u'count': 153}
{u'_id': u'bad', u'count': 124}
{u'_id': u'very_bad', u'count': 8}
{u'_id': u'horrible', u'count': 6}
{u'_id': u'goo', u'count': 1}
     {u'_id': u'exc', u'count': 1}
```

That is good that majority of the roads in Atlanta have good or excellent smoothnes.

3.3 Total Streets named containing "Peach"

As Georgia is called Peach State, so we should find out how many streets have name Peach i

```
     >db.atlanta.find({'address.street':{'$regex' : ".*[Pp]each.*"}}).count()
     1221
```

3.4 Top Contributing Users

```
     >result=db.atlanta.aggregate([{'$group':  { '_id':  "$created.user", 'count':
     { '$sum':  1 }}}, {'$sort':  { "count":  -1}}, {'$limit':  10}])
```

```
{u'_id': u'Liber', u'count': 5378291}
{u'_id': u'Saikrishna_FultonCountyImport', u'count': 2410657}
{u'_id': u'woodpeck_fixbot', u'count': 1508201}
{u'_id': u'afonit', u'count': 336149}
{u'_id': u'Jack the Ripper', u'count': 322658}
{u'_id': u'rjhale1971', u'count': 281143}
{u'_id': u'Jack Kittle Buildings', u'count': 247159}
{u'_id': u'maven149', u'count': 196132}
{u'_id': u'Chris Lawrence', u'count': 124373}
     {u'_id': u'macon_cfa', u'count': 100104}
```

3.5 Top Amenities

```
     db.atlanta.aggregate([{'$group':  { '_id':  "$amenity", 'count':
     { '$sum':  1 }}}, {'$sort':  { "count":  -1}}, {'$limit':  10}])
```

```
{u'_id': None, u'count': 12512622}
{u'_id': u'parking', u'count': 8739}
{u'_id': u'place_of_worship', u'count': 5742}
{u'_id': u'school', u'count': 2389}
{u'_id': u'grave_yard', u'count': 2186}
{u'_id': u'restaurant', u'count': 1376}
{u'_id': u'fuel', u'count': 865}
{u'_id': u'fast_food', u'count': 833}
{u'_id': u'parking_space', u'count': 817}
    {u'_id': u'bank', u'count': 342}
```

4. Suggestions to Improve Data

a. OSM should enforce some stricter checks when accepting the users entry. E.g. give errors if zip code is missing, notification for incomplete street ending etc.
b. There should be ranking system based on quality of contribution from each user. And at the end best contributors should be rewarded with points etc.
c. Probably use the USPS data base to verify the addresses entered and also fix the missing zip codes intelligently.

Stricter checks might decrease the contribution from users, because users tend to like easier methods. Also rewarding would cost. Probably OSM can start generating some income from ads, and then share the profits with contributors too.

5. Conclusion

Data wasn't clean as we expected. We found problems in dataset as we had seen in exercises. Some street names were not correct, some address were totally wrong.

We found out also that majority of the roads in Atlanta are in good, rather excellent condition. We also have majority of roads with speed limit of 45 mph.

User 'Liber' made the most contributions to the. In other reports we have seen that user 'woodpeck_fixbot' has most entries, but here he is at 3rd place.

It would be interesting to investigate data more for existing running track, as I am more of a runner, and finding new places would be fun.