Data Wrangle Openstreetmaps Data

Udacity Nanodegree Project 2

Coursework:

- Data Wrangling with MongoDB

By : Joe Nyzio

Data Wrangle Openstreetmaps Data

Udacity Nanodegree

By Joe Nyzio

Project Report

Problems encountered with my map

Up until auditing my data I was still pretty naive about what “dirty data” meant. It seems pretty ridiculous but I had the idea that “it couldn’t happen to me” like it was a sickness or something. I did the audit and what I think is probably a relatively small amount of problems came up within the data. I did a few things to tidy it up.

- **Add expected results**

The first thing I did was add more expected results like “Alley”, “Circle”, “Pike”, “Terrace”, and “Ridge”.

```
expected = ["Street", "Avenue", "Boulevard", "Drive", "Court", "Place", "Square", "Lane", "Road", "Trail", "Parkway", "Commons", "Alley", "Circle", "Pike", "Terrace", "Ridge"]
```

- **Remap variables**

I remapped a few variables like Blvd and Blvd. to Boulevard, and Ln to Lane.

```
mapping = { "St": "Street",
          "St.": "Street",
          "street": "Street",
          "Atreet" : "Street", #Fix Arch Street typo
          "Ave": "Avenue",
          "Ave." : "Avenue",
          "Rd": "Road",
          "Rd.": "Road",
          "Blvd" : "Boulevard",
          "Blvd." : "Boulevard",
          "Ln" : "Lane",
          "PA" : "",
          "446-1234" : "",
          "BlvdHavertown" : "Bouldevard" #Fix Brookline
```

```
    "Orgontz" : "Ogontz" #No r in Ogontz
    }
```

- **Spelling Errors**

A few other results I remapped because of spelling errors I caught. “Arch Atreet” was probably meant to be “Arch Street”, “Orgontz” might exist somewhere in the city I don’t know about but I assumed they were going for “Ogontz”.

```
'Atreet': set(['Arch Atreet']),
'72nd Ave and Orgontz Ave',
```

- **Bad Members**

Whoever typed “'446-1234': set(['1 Brookline BlvdHavertown, PA 19083(610) 446-1234'])” into the address box should be banned from the openstreetmaps community.

```
'446-1234': set(['1 Brookline BlvdHavertown, PA 19083(610) 446-1234']),
```

- **Numbered streets**

Some streets came up only having their number like ‘37th’ and ‘43rd’. I left them this way because it is not wrong, just incomplete.

```
{'37th': set(['N 37th']),
 '39th': set(['N 39th']),
 '43rd': set(['N 43rd']),
```

- **NEWS**

Many different ways to signify directions were used. ‘N 9th’, ‘N. 9th’, ‘North 9th’, are examples of the most common ways of representing direction.

```
'E. Mt Airy Ave',
'East Wadsworth Ave',
```

Overview of the Data

A collection of data from Philadelphia Area from the OpenStreetMaps dataset.

```
<bounds minlat="39.8858" minlon="-75.3391" maxlat="40.1022" maxlon="-75.1362"/>
```

- **Size of the file**

philly_map 57.9 MB.

philly_map.json 71.9 MB.

```
> db.philly.dataSize()
71943008
```

- **Collections**

There are 248,106 collections within the file.

```
> db.philly.count()
248106
```

- **Format**

```
> db.philly.findOne()
{
        "_id" : ObjectId("55302a3830d782052b0569cf"),
        "id" : "27148343",
        "type" : "node",
        "pos" : [
                39.8824957,
                -75.1948014
        ],
        "created" : {
                "changeset" : "14253878",
                "user" : "bigwebguy",
                "version" : "2",
                "uid" : "635896",
                "timestamp" : "2012-12-12T21:23:52Z"
        }
}
```

- **Number of unique users**

There are 613 unique users who have contributed to this dataset.

```
> db.philly.distinct( ‘created.user’ ).length
613
```

Top contributing user

A member named “dchiles” is the top contributing users with 51,831 posts.

```
>db.philly.aggregate( [ { '$group' : { “_id” : “$created.user” , “count” : {  “$sum” : 1 } } } , {
“$sort” : { “count” : -1 } } , { “$limit” : 1 } ] )

{ "_id" : "dchiles", "count" : 51831 }
```

Number of users with 1 post

There are 149 users that have only made 1 post.

```
> db.philly.aggregate ( [ { ‘$group’ : { ‘_id’ : ‘$created.user’ , ‘count’ : {‘$sum’:1} } } , { ‘$group’ :
{ ‘_id’ : ‘$count’ , ‘num_users’ : { ‘$sum’ : 1 } } } , { ‘$sort’ : { ‘_id’ :1 } } , { ‘$limit’ : 1 } ] )

{ "_id" : 1, "num_users" : 149 }
```

Number of nodes and ways

There are 227,927 nodes in the dataset.

```
> db.philly.find( { type : ‘node’ } ).length()
227927
```

There are 20,176 ways in the dataset.

```
> db.philly.find( { type : ‘way’ } ).length()
20176
```

Number of chosen type of nodes, like cafes, shops etc

In this area of Philadelphia there are...

```
> db.philly.find( { shop: { $exists : true } } ).count()
```

507 shops.

```
> db.philly.find( { amenity : ‘cafe’ } ).length()
```

109 cafes.

```
> db.philly.find( { amenity : ‘bar’ } ).length()
```

54 bars.

```
> db.philly.find( { amenity : ‘library’ } ).length()
```

65 libraries.

```
> db.philly.find( { amenity : ‘university’ } ).length()
```

21 universities.

```
> db.philly.find( { amenity : ‘school’ } ).length()
```

436 schools.

```
> db.philly.find( { amenity : ‘parking’ } ).length()
```

295 parking lots. If thats true why did I always have to park in some weird alley and walk a mile to get to school?

Other Searches

- ### Most common ‘building types’

This made me laugh. I’m not sure why everyone is putting ‘yes’ as the building type. I might be misunderstanding the query. Either that or the website design is making people believe it’s asking whether the address is a building or not.

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

```
{ "_id" : "yes", "count" : 745 }
{ "_id" : "university", "count" : 45 }
{ "_id" : "residential", "count" : 14 }
{ "_id" : "dormitory", "count" : 12 }
{ "_id" : "apartments", "count" : 7 }
{ "_id" : "hospital", "count" : 7 }
{ "_id" : "tank", "count" : 6 }
{ "_id" : "house", "count" : 6 }
{ "_id" : "commercial", "count" : 6 }
{ "_id" : "entrance", "count" : 4 }
```

- **Most common ‘amenity’**

There are tons of schools in this area. I’m not sure how this is true. There has to be things being classified as schools that aren’t. Maybe nurseries, preschools, and daycares classify themselves as schools. I could be wrong but that seems like an enormous amount of school. A few on this list came up earlier when I did individual searches.

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

{ "_id" : "school", "count" : 436 }
{ "_id" : "restaurant", "count" : 341 }
{ "_id" : "parking", "count" : 295 }
{ "_id" : "social_facility", "count" : 225 }
{ "_id" : "place_of_worship", "count" : 175 }
{ "_id" : "car_sharing", "count" : 164 }
{ "_id" : "fast_food", "count" : 123 }
{ "_id" : "cafe", "count" : 109 }
{ "_id" : "fire_station", "count" : 88 }
{ "_id" : "post_box", "count" : 74 }
```

Other ideas about the dataset

It would be interesting if open street maps was an app that people could allow to access their location. The collected data could be used to build real time traffic reports based on the accumulation of data derived from vehicle speeds of people using the app. I haven't really looked into how google does it but they’re probably doing this already. I think being an open sourced non commercial entity would give people more faith in allowing the app to access their information.

A cross between https://photosynth.net/ and openstreetmaps that could access user photos by and upload them based on the lat/long they were taken could be some sort of open sourced google earth.

Conclusion

This dataset seems haphazard. It may be worth it for the site to attempt to set a standard for how data is put into their system. It's probably easier to set and attempt to hold a standard input than to constantly have to clean up after everyone. The people working on this project and making updates are likely already trying to help so I'm sure they would listen to some guidance for the sake of making what they do even more useful.