Data Wrangle OpenStreetMaps Data

by Eduardo Eidelwein Berlitz in fulfillment of Udacity's Data Analyst Nanodegree, Project 3

Map Area: Porto Alegre, Brazil
https://www.openstreetmap.org/relation/242397
https://s3.amazonaws.com/metro-extracts.mapzen.com/porto-alegre_brazil.osm.bz2

Section 1: Problems Encountered in the Map

Street Abbreviations

After running the audit script from lesson 6 against my chosen area I realized that I should change the regular expression to find the first word of the street name, and not the last. That's because the language in this map area is Portuguese. After that it was possible to list all sorts of unusual streets.

Looking at the names and street types I discovered that there are some abbreviated names. So I modified the *shape_element* function to correct these abbreviations. Basically all instances were corrected from "**Av**" or "**Av.**" to "**Avenida**", which is the Portuguese for "**Avenue**" and also "**R.**" to "**Rua**", which means "**Street**".

Incorrect Postal Codes

After listing all map postcodes using the script **print_postcodes.py** I have found that almost all of them follow the postal code format `**00000-000**` but some of them were in the wrong format as you can see in the sample below:

```
...
90030-010
93216-120
93510-310 ÔÇÄ
90230091
92500-000
92410350
92020-970
...
```

So using regular expressions I manage to correct these postal codes to the correct format by replacing the original value by the first 5 digits followed by "-" and the others 3 digit.

After that the osm.json file was generated and imported into a MongoDB collection using the following command:

```
> mongoimport -d poa -c poa --file porto-alegre_brazil.osm.json
```

I also noticed that the osm data extracted with the *mapzen metro extracts* tool contains other cities besides the chosen one. In fact that's because this tool extracts all the data within a box.

So it was found nodes and ways of other cities around Porto Alegre while executing the MongoDB queries.

Section 2: Data Overview

File sizes

- porto-alegre_brazil.osm (101.8MB)
- porto-alegre_brazil.osm.json (104.6MB)

Number of documents

```
> db.poa.find().count()
530665
```

Number of nodes

```
> db.poa.find({"type":"node"}).count()
459399
```

Number of ways

```
> db.poa.find({"type":"way"}).count()
71250
```

Number of unique users

```
> db.poa.distinct("created.user").length
432
```

Number of cafes

```
> db.poa.find({amenity:"cafe",type:"node"}).count()
41
```

Number of restaurants

```
> db.poa.find({amenity:"restaurant", type:"node"}).count()
339
```

Top 10 appearing amenities

```
> db.poa.aggregate([
    {$match:{"amenity":{"$exists":1},"type":"node"}},
    {"$group":{"_id":"$amenity","count":{"$sum":1}}},
    {$sort:{"count":-1}},
    {"$limit":10}
])
[{"_id": "telephone","count": 696},
{ "_id": "bench", "count": 375},
{ "_id": "restaurant", "count": 339},
{ "_id": "fuel", "count": 246},
{ "_id": "waste_basket", "count": 245},
{ "_id": "taxi", "count": 236},
{ "_id": "bank", "count": 166},
{ "_id": "school", "count": 148},
{ "_id": "pharmacy", "count": 147},
{ "_id": "place_of_worship","count": 146}]
```

Sort cities by count, descending

```
> db.poa.aggregate([
    {$match:{"address.city":{"$exists":1}}},
    {"$group":{"_id":"$address.city","count":{"$sum":1}}},
    {$sort:{"count":-1}},
    {"$limit":10}
])
[{ "_id": "São Leopoldo", "count":745 },
{ "_id": "Novo Hamburgo", "count":279 },
{ "_id": "Porto Alegre", "count":134 },
{ "_id": "Montenegro", "count":43 },
{ "_id": "Ivoti", "count":33 },
{ "_id": "Brochier", "count":33 },
{ "_id": "Canoas", "count":26 },
{ "_id": "Estrela", "count":20 },
{ "_id": "Taquara", "count":20 },
{ "_id": "Campo Bom", "count":13 }]
```

Most common building types:

```
db.poa.aggregate([
    {'$match': {'building': {'$exists': 1}}},
    {'$group': { '_id': '$building','count': {'$sum': 1}}},
    {'$sort': {'count': -1}}, {'$limit': 10}
])
[{"_id" : "yes", "count" : 5174 },
{ "_id" : "house", "count" : 811 },
{ "_id" : "industrial", "count" : 384 },
{ "_id" : "residential", "count" : 362 },
{ "_id" : "university", "count" : 215 },
{ "_id" : "commercial", "count" : 207 },
{ "_id" : "school", "count" : 130 },
{ "_id" : "apartments", "count" : 128 },
{ "_id" : "office", "count" : 108 },
{ "_id" : "roof", "count" : 63 }]
```

Top 1 contributing user

```
> db.poa.aggregate([
    {"$match":{"type":"node"}},
    {"$group":{"_id":"$created.user","count":{"$sum":1}}},
    {"$sort":{"count":-1}},
    {"$limit":1}
])
[{ "_id" : "SergioAJV", "count" : 67429 }]
```

Number of users appearing only once (having 1 post)

```
> db.poa.aggregate([
    {"$group":{"_id":"$created.user","count":{"$sum":1}}},
    {"$group":{"_id":{"postcount":"$count"},"num_users":{"$sum":1}}},
    {"$project":{"_id":0,"postcount":"$_id.postcount","num_users":1}},
    {"$sort":{"postcount":1}},
    {"$limit":1}
])
[{ "num_users": 64,"postcount": 1 }]
```

Biggest religion

```
> db.poa.aggregate([
    {"$match":{"amenity":"place_of_worship","type":"node"}},
    {"$group":{"_id":"$religion","count":{"$sum":1}}},
    {"$sort":{"count":-1}},
    {"$limit":1}
])


[ { "_id" : "christian", "count" : 125 } ]
```

Most popular cuisines in fast foods

```
db.poa.aggregate([
    {"$match":{"cuisine":{"$exists":1},"amenity":"fast_food"}},
    {"$group":{"_id":"$cuisine","count":{"$sum":1}}},
    {"$sort":{"count":-1}},
    {"$limit":5}
])
[{"_id" : "burger", "count" : 29 },
{ "_id" : "sandwich", "count" : 10 },
{ "_id" : "regional", "count" : 3 },
{ "_id" : "pizza", "count" : 3 },
{ "_id" : "hotdog", "count" : 2 }]
```

Top 10 gas stations brands

```
db.poa.aggregate([
    {"$match":{"brand":{"$exists":1},"amenity":"fuel"}},
    {"$group":{"_id":"$brand","count":{"$sum":1}}},
    {"$sort":{"count":-1}},
    {"$limit":5}
])
[{ "_id" : "Ipiranga", "count" : 101 },
{ "_id" : "BR", "count" : 54 },
{ "_id" : "Shell", "count" : 53 },
{ "_id" : "Charrua", "count" : 13 },
{ "_id" : "Petrobras", "count" : 7 },
{ "_id" : "Esso", "count" : 6 },
{ "_id" : "Megapetro", "count" : 6 },
```

```
{ "_id" : "Vale", "count" : 3 },
{ "_id" : "Latina", "count" : 2 },
{ "_id" : "Race Trac", "count" : 2 }]
```

Section 3: Additional Ideas

Analyzing the data it is possible to realize that the city of Porto Alegre (which is the state capital) have fewer addresses associated with this data set than the city of São Leopoldo. But not all nodes or ways include this information since its geographical position is represented within regions of a city. What could be done in this case, is check if each node or way belongs to a city based on the latitude and longitude and ensure that the property "address.city" is properly informed. By doing so, we could get statistics related to cities in a much more reliable way. In fact, I think this is the biggest benefit to anticipate problems and implement improvements to the data you want to analyze. Real world data are very susceptible to being incomplete, noisy and inconsistent which means that if you have low-quality of data the results of their analysis will also be of poor quality.

Another alternative to help in the absence of information in the region would be the use of gamification to make more people help in the map contribution. Something like Waze app already does today.

Section 4: Conclusion

This review of the data is cursory, but it's obvious that the Porto Alegre area is incomplete, though I believe it has been well cleaned for the purposes of this exercise.

Section 5: References

- OpenStreetMap Wiki Page
- OpenStreetMap Wiki Page - OSM XML
- OpenStreetMap - Map Features
- Python Regular Expressions
- MongoDB Operators
- Metro Extracts - City-sized portions of OpenStreetMap
- Udacity course, "Data Wrangling with MongoDB"