Project: JOCON37 OpenStreetMap Data Case Study

Denver, Colorado - United States

https://github.com/jocon37/OpenStreetMap-Data-Wrangling
(https://github.com/jocon37/OpenStreetMap-Data-Wrangling)

Introduction

I chose this region - city for two reasons. Firstly, I had a difficult time being able to get the right size file to extract from OpenStreetMap so I took what would work, but my family use to have a small ranch in a city near Denver, so it is a place I am familiar with.

Problems Encountered in the Map

The data throughout the XML should truly follow specific formatting and should be consistent and easy to read.

- There are overabbreviated street names like "Hwy", "Ave", "St", "S", "Blvd", etc. For example, Blvd has 1720 entries, Ave has 7945, and St has 11,199 entries on the file but also has 'st'.
- Correction of misspellings
- Correction of words that have been combined incorrectly like MainStreet, which should be Main Street.

Auditing and Clearing:

The main function I used for clearning the street names is:

def audit_street_type(street_types, street_name): m = street_type_re.search(street_name) if m: street_type = m.group() if street_type not in expected: street_types[street_type].add(street_name)

Overview of the data

I extracted the OSM.XML ZIP file and ran data.py and audit.py, I created the .csv files. I then used SQLITE to create a database and load the .csv files into tables so I could query the tables.

- denver_colorado_osm.xml 714MB
- nodes.csv 271MB
- nodes_tags.csv 16MB
- ways.csv 25MB
- ways_nodes.csv 86MB
- ways_tags.csv 45MB

I'm glad I did my analysis on the full XML file. Because when I view the stats on the sample OSM file, the denver_colorado.osm, it's not as rich in information. For example, the file is 3.46MB, which is manageable. However for a research standpoint it did not give me much to go on.

- nodes.csv only had 1.84MB
- nodes_tags.csv had 85.5KB
- ways.csv had 42B (yes bytes)
- ways_nodes had 22B
- ways_tags had 20B

When I looked at the ways files, you can guess by the size, the only thing there was the headers. Immediately I thought something was wrong with my code. But it worked just fine with the full file. So after some research within the sample OSM, I realized, there are no entries. The sample file, really would not have given me much fun to work with!

To get an idea of the number of unique users in the ways.csv file we can look at the disntinct uid variable:

```
sqlite> select count(distinct(a.uid))
   ...> from(select uid from nodes union select uid from ways) as a;
1467
sqlite>
```

Returns: 1467

We can look at the basic count of nodes and ways:

```
sqlite> select count(*)
   ...> from nodes;
3132214
sqlite> select count(*)
   ...> from ways;
396041
sqlite>
```

Querying the nodes table returns a count of 3,132,214 and the ways table returns a count of 396,041.

We can look at various types of nodes out of the nodes_tags tables by looking for different items. For example, let's look at how many BBQ places there are?

```
sqlite> select count(*)
   ...> from nodes_tags
   ...> where key = 'amenity' and value = 'bbq';
28
sqlite>
```

Our query returns 28 BBQ facilities. What about bicycle_parking?

```
sqlite> select count(*)
   ...> from nodes_tags
   ...> where key = "bicycle_parking";
111
sqlite>
```

A quick querry on they key in the nodes_tags table returns 111.

If we wanted to, we could look at the data to see who the top contributing users are to OneStreetMap with our queries.

```
sqlite> select a.user, count(*) as num
   ...> from (select user from nodes union all select user from ways) as a
   ...> group by a.user
   ...> order by num desc
   ...> limit 10;
chachafish,1269527
"Your Village Maps",761766
CornCO,168736
GPS_dr,139611
woodpeck_fixbot,118067
Stevestr,97600
DavidJDBA,82366
"RustProof Labs",58097
EnigmaQuip,39968
ColoSean,34139
sqlite>
```

- chachafish 1269527
- "Your Village Maps" . 761766
- CornCO 168736
- GPS_dr 139611
- woodpeck_fixbot 118067
- Stevestr 97600
- DavidJDBA 82366
- "RustProof Labs" 58097
- EnigmaQuip 39968
- ColoSean..... 34139

In doing some research it looks like chachafish is an actual person. They have quite a bit of activity in OneStreetMap's forum (https://forum.openstreetmap.org/viewforum.php?id=67) (https://forum.openstreetmap.org/viewforum.php?id=67)). "Your Village Maps" is also an individual. I couldn't find anything on CornCo.

We can look at the top 10 categories of shops through queries as well:

```
sqlite> select count(*), value
   ...> from nodes_tags
   ...> where key = 'shop'
   ...> group by value
   ...> order by count(*) desc
   ...> limit 10;
388,yes
360,convenience
233,car_repair
230,hairdresser
190,alcohol
133,clothes
129,supermarket
127,car
126,doityourself
118,beauty
sqlite> _
```

- Yes 388
- Convenience 360
- Car_repair 233
- Hairdresser 230
- Alcohol 190
- Clothes 133
- Supermarket 129
- Car 127
- Doityourself ... 126
- Beauty 118

The "Yes" shop means that the shop is unidentified or unspecified. The Doityourself shops include stores such as Lowes, HomeDepot and even some Self Storage.

It is interesting and rather humorous to see that there are more alcohol shops than supermarkets. Not sure what that says about us as a society. I can't help but laugh a little to my self and wonder if it is like that in my area as well. I also did look up on my own the marijuana shops here out of curiosity since it is legal in Colorado. I have to admit, I was half expecting it to be in the top 10 list (I say half humorly), however, I was very surprised to see that there are only 4 shops listed. I can only think that some of those shops are listed under "yes" - the unspecifed shops.

If you want to know what do in and around Denver, Colorado? We can look up tourism in the tables with key = "tourism".

```
sqlite> select distinct(value)
   ...> from nodes_tags
   ...> where key = 'tourism';
museum
guest_house
information
picnic_site
viewpoint
hotel
attraction
motel
artwork
trail_riding_station
camp_site
caravan_site
gallery
hostel
apartment
yes
sqlite>
```

You can look for items like Museum, Picnic_site, Viewpoint, Hotel, Artwork, Trail_riding_station, Camp_site and more. There is a lot of options available to look for.

Other Information

It might be helpful, as a suggestion, for OneStreetMap, to provide some guidelines or manuals for mappers to have to follow for when they are mapping. That way, they would have consistency when mappers are coding addresses and postal codes and it might cause less confusion.

It would be easier to read and be easier to search.

The challenge would be getting eveyrone to read the guidelines, plus people have their own preferences and style for doing things.

Starting with a group of top mappers to test the idea of set guidelines and implementing it. Sometimes getting the popular group doing something the rest will follow.

Make a game out of it by creating a rank system out of the guidelines.

Create a peer review system.