Wrangle OpenStreetMap Data

Author: Esha Chaudhary

Map Area: Austin, TX, United States

Location: http://www.openstreetmap.org/relation/113314

Data Extract Link : https://s3.amazonaws.com/metro-extracts.mapzen.com/austin_texas.osm.bz2

Area description and reason for choice:
Austin is the capital of the U.S. state of Texas. It is the 11th-most populous city in the U.S. and the 4th-most populous in Texas. It is the fastest growing large city in the United States. I chose this area because this is where I currently live and city is new to me.

Problems encountered and Data cleaning

Below are the steps taken to investigate and clean the dataset before loading it into the database.

I. A sample file was created from the original extract.
Original file size: 1.4GB
Sample file size: 143.2MB

II. Script mapparser.py was run on sample dataset to count occurrence of each tag,.
{'member': 2466,
'nd':701474,
'node':639940,
'osm':1,
'relation':241,
'tag':239116,
'way':67065}

III. Script tags.py was run to see if there are any problems with the "k" value of each "<tag>".
No problems were found with the tag names.
{'lower': 131004, 'lower_colon': 106900, 'other': 1212, 'problemchars': 0}

IV. Script audit.py was used to audit the data. Below are the fields that were investigated.

- **Street Names**
  There were some inconsistencies in the street names abbreviation. For example, some street names were named properly as "ABC Avenue" while some were "ABC Ave". For clarity purpose all abbreviated street names were translated to their long forms. Below are some street names from the dataset and their corresponding translations.

  Claro Vista Ct → Claro Vista Court
  Hilltop Canyon Cv → Hilltop Canyon Curve
  Bill Baker Dr → Bill Baker Drive

- **Phone numbers**
  It was observed that the phone numbers were not in a specific format. Some had country code while some had it missing. Some numbers had spaces between numbers while some didn't.

  Below are some examples of numbers from the dataset.
  (512) 443-1057
  +1 512 368 1818
  5124780098
  512-459-2300
  +15125706300

  to avoid confusion all numbers were converted to '+1 999-999-9999'. Also the numbers with insufficient digits were replaced by null.

- **Direction**
  While auditing fields with direction data, no problems were found. But all direction abbreviations were translated to their proper names before loading the data in db.
  e.g E → East
  N:S → North:South

- **Postcodes/Zip codes**
  Postcode data was quite clean so no transformations were performed. Austin zip codes are of 5 digits, I expected to see some alphanumeric codes but no problems were found.

- **Speed**
  Maxspeed column had some values with missing speed unit. So "mph" was added wherever unit was missing.

- **State Code**
  state code was inconsistent in the dataset so all records were replaced with common state code "TX".

V. After auditing was completed, cleandata.py script was run to clean the data. Script also created 5 csv files, which were then loaded to SQLite database.

Data Overview

This section contains some SQL queries performed on the dataset to gather some basic statistics and to explore the area.

File size:
nodes.csv – 60.2MB
nodes-tags.csv – 1.1MB
ways.csv – 4.8MB
ways_nodes.csv – 16.9MB
ways_tags.csv – 6.9MB

Unique users:
sqlite> SELECT COUNT(DISTINCT(e.uid))
FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways) e;
798

Number of nodes:
nodes: 639940

Number of ways:
ways: 67065

Number of nodes tags:
nodes_tags: 31791

Number of ways tags:
ways_tags : 206053

Number of ways nodes:
ways_nodes: 701474

Top 10 contributing users

sqlite>SELECT e.user, COUNT(*) as num
FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e
GROUP BY e.user
ORDER BY num DESC
LIMIT 10;

patisilva_atxbuildings|274259
ccjjmartin_atxbuildings|129979
ccjjmartin__atxbuildings|94018
wilsaj_atxbuildings|35898
jseppi_atxbuildings|30062
woodpeck_fixbot|22098
kkt_atxbuildings|15788
lyzidiamond_atxbuildings|15640
richlv|4997
johnclary_axtbuildings|4827

Number of users appearing only once (having 1 post)
sqlite> SELECT COUNT(*)
FROM
(SELECT e.user, COUNT(*) as num
 FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e
 GROUP BY e.user
 HAVING num=1) u;
197

Top 20 amenities
sqlite> SELECT value,count(*) as total
FROM node_tags

```
WHERE  key = 'amenity'
GROUP BY value
ORDER BY  total DESC
LIMIT 20;
```

```
waste_basket|58
restaurant|50
place_of_worship|45
fast_food|39
bench|34
fuel|26
school|20
bar|16
bank|11
cafe|10
pharmacy|10
parking|9
post_box|9
drinking_water|7
bicycle_parking|6
pub|6
toilets|6
atm|5
fire_station|4
grave_yard|4
```

Hospital Names:

```
sqlite> SELECT value
FROM  node_tags
WHERE  key='name' AND  id in(SELECT distinct id FROM node_tags WHERE key = 'amenity' AND
value = 'hospital');
```

```
Georgetown Hospital
Austin Women's Hospital
```

Types of 'restaurant':

```
sqlite> SELECT nt.value, COUNT(*) as num
FROM node_tags nt
JOIN (SELECT DISTINCT(id) FROM node_tags WHERE value='restaurant') i
ON nt.id=i.id
WHERE nt.key='cuisine'
GROUP BY nt.value
ORDER BY num DESC;
```

```
mexican|6
pizza|5
american|2
cajun|2
indian|2
BBQ|1
Jamaican,_Cuban|1
```

barbecue|1
burger|1
chinese|1
italian|1
peruvian|1
regional|1
salad|1
sandwich|1
thai|1
vietnamese|1

Tourist spots in city:

```
sqlite> SELECT *
FROM node_tags where key='name'
AND id in(SELECT id from node_tags WHERE  key = 'tourism' and value not in('hotel','motel'));
```

368165070|name|French Legation Museum|regular
447647000|name|Trail Head for Cypress Creek Nature Preserve|regular
2449728191|name|512|regular
2449734965|name|446|regular
4460646698|name|The Austin Visitor Center|regular

ATM Opeartors:

```
sqlite> SELECT value
FROM node_tags nt
JOIN(select id from node_tags where key='amenity' and value = 'atm')i
ON nt.id = i.id
WHERE  nt.key in('operator','name');
```

Wells Fargo
BBVA Compass
Chase
Wells Fargo
Bank of America

Additional Ideas for improving dataset

The quality of the dataset was not very bad. But quality can be more improved by enforcing some standards to certain fields. For example standardizing street names abbreviations or providing country code to phone numbers. This process will not completely clean the data but quality of data will be better. I am sure this will also help people who are exploring a new area.

The problem with this solution is that contributors might not like this approach as this is an open source project and contributors focus is more on adding information rather than standardizing it. Even if a dedicated team is set to handle such issues, the amount of effort and cost for maintaining such activities might be expensive.

References:

https://wiki.openstreetmap.org/wiki/OSM_XML
https://www.tutorialspoint.com/sqlite/sqlite_select_query.html