OpenStreetMap Project

Data Wrangling with MongoDB

Sheng Weng

Map Area: Austin, TX, United States, and San Antonio, TX, United States
OSM XML file source: https://mapzen.com/metro-extracts/

1. Problems Encountered in the Map of Austin

1) I first worked on the data of Austin area. After running austin_texas.osm file in audit.py, I found the main problem with the data is the inadequate use of street names. I updated all problematic address strings, for example,
2) I replaced "IH-35", "IH35", "I-35", "I H 35" with "IH 35". This is one of the main highways that lie across Austin.
3) I replaced "brigadoon lane" with "Brigadoon Lane".
4) I replaced "US" with "U.S.", and "TX" with "Texas".
5) At first, I did not know what "FM" or "F.M." stands for. After going through the data it turned out they represent "Farm-to-Market Road". So I was able to make the correction after careful auditing the data.
6) Other detailed changes can be found in the mapping dictionary in data.py file.

2. Data Overview

1) File sizes

   austin_texas.osm…… 179.1 MB
   austin_texas.osm.json…… 196.7 MB

2) Number of documents

```
> db.austin.find().count()
861767
```

3) Number of nodes

```
> db.austin.find({"type": "node"}).count()
780753
```

4) Number of ways

```
> db.austin.find({"type": "way"}).count()
81004
```

5) Number of unique users

```
> db.austin.distinct("created.user").length
897
```

6) Sort postcodes by count, descending

```
>db.austin.aggregate([{"$match":{"address.postcode":{"$e
xists":1}}}, {"$group": {"_id":"$address.postcode",
"count":{"$sum":1}}}, {"$sort":{"count":-1}}, {"$limit": 3}])
{ "_id" : "78704", "count" : 61 }
{ "_id" : "78705", "count" : 54 }
{ "_id" : "78757", "count" : 47 }
```

7) Top 3 popular cuisines

```
> db.austin.aggregate([{"$match": {"cuisine": {"$exists":
1}, "amenity": "restaurant"}}, {"$group": {"_id": "$cuisine",
"count": {"$sum": 1}}}, {"$sort": {"count": -1}}, {"$limit": 3}])
{ "_id" : "mexican", "count" : 67 }
{ "_id" : "american", "count" : 25 }
{ "_id" : "pizza", "count" : 24 }
```

3. Other Ideas

1) Sort road surface types by count

```
> db.austin.aggregate([{"$match": {"surface": {"$exists":
1}}}, {"$group": {"_id": "$surface", "count": {"$sum": 1}}},
{"$sort": {"count": -1}}, {"$limit": 3}])
{ "_id" : "asphalt", "count" : 3408 }
{ "_id" : "paved", "count" : 531 }
{ "_id" : "concrete:plates", "count" : 197 }
```

From this result we know that most of the roads are asphalt. It is an instructive information especially for the government budgeting new roads.

2) Number of parking place

I compared the number of parking place in Austin and San Antonio.

```
> db.austin.find({"amenity": "parking"}).count()
1854

> db.antonio.find({"amenity": "parking"}).count()
1089
```

After that, I calculated the areas of the two cities covered in the data sets based on the “bounds” information.
For Austin,
<bounds minlon="-98.21200" minlat="29.93100" maxlon="-97.23400" maxlat="30.67000".
For San Antonio,
<bounds minlon="-98.95100" minlat="29.10900" maxlon="-97.88000" maxlat="29.95100".

Therefore, the map area of San Antonio is about 1.248 times larger than that of Austin. However, the number of parking amenity in San Antonio is only about 0.587 times that of in Austin. This is consistent with my feeling that Austin is more of a crowded city as compared with San Antonio.