OpenStreetMap Data Wrangling with MongoDB

Eric Carlsen

Map Area: Centennial, Colorado (suburb of Denver)

Introduction

The area I chose to map consisted of the Centennial, Colorado region. This is the area in which I live and work. I've been a contributor to OpenStreetMap (OSM) for many years now. My focus has been completing new neighborhoods, but also fleshing out the existing cycling infrastructure. As such, I decided to investigate the state of the data in regards to cycling.

Problems Encountered in the Map

Street Names

There were standard issues with this data set, most notably streets that didn't conform to the standard of not using abbreviations. This data was corrected at the xml level, before importation into mongodb. Some useful statistics on the topic were reported, however.

```
Total number of streets processed: 261
Number of streets corrected: 50
Percent of streets corrected: 19.2%
```

City Names

The names of the cities in the database were given by the following query, with its results (db and collection preface left off for formatting):

```
aggregate([{"$match": {"address.city": {"$exists": True}}},
           {"$group": {"_id": "$address.city", "count": {"$sum": 1}}},
           {"$sort": {"count": -1}}])
```

```
City: Aurora              , Count: 375
City: Centennial          , Count: 237
City: Parker              , Count: 217
City: Denver              , Count: 82
City: Englewood           , Count: 22
City: Greenwood Village   , Count: 22
City: Foxfield            , Count: 5
City: Lone Tree           , Count: 4
City: Cherry Hills Village, Count: 3
City: Highlands Ranch     , Count: 2
City: Littleton           , Count: 1
City: Centenn             , Count: 1
```

This shows that the only error consists of a city abbreviation of Centenn, for the city Centennial. This was corrected for in a generic fashion, however, allowing for more future corrections than this one. A

dictionary of key value incorrect/correct pairs can be provided that is iterated through. The documents matching the incorrect city are then updated with the save command, as shown in Lesson 4.

```
for incorrect_city in cities_corrections:
    res = self.collection.find({"address.city": incorrect_city})
    for doc in res:
        doc["address"]["city"] = cities_corrections[incorrect_city]
        self.collection.save(doc)
```

Miscellaneous Problems

Some other areas of investigation for problems were conceived by the online tool keepright. For example, within the area of interest it was found that there were 22 documents with a FIXME tag. There were also 5 places of worship listed without a corresponding religion. These are areas which would be great for immediate editing, which I plan on doing as soon as this assignment is passed.

Overview of the Data

The downloaded osm file was 102 MiB. Some overall statistics about the number of entries are given below.

```
Number of documents in the database: 523060
Number of nodes in the database: 474952
Number of ways in the database: 48102
```

As an OSM contributor I was curious to see how much my contributions had impacted this area. My username is ecarl65, and I performed the following query (the database collection is an object variable stored in self.collection).

```
self.collection.find({"created.user": "ecarl65"}).count()
```

Which gave results of (with the percent being the percent of the total number of documents)

```
Number of edits by ecarl65: 14436 (2.8%)
Number of unique users: 370
```

In order to represent the size of the data captured, and for some density calculations performed later, I also showed the limits on the downloaded map. I used the python geopy package to calculate the extent of the map area in the east/west and north/south directions, which I used to estimate the area of the map.

```
Reported Minimum Latitude: 39.4913
Reported Maximum Latitude: 39.6678
Reported Minimum Longitude: -104.96
Reported Maximum Longitude: -104.531
Reported Distance Across Constant Latitude: 36.858 (km)
Reported Distance Across Constant Longitude: 19.596 (km)
Reported Area: 722.269 (km^2)
```

The postal codes and their counts all seemed consistent with the map area.

Additional Ideas

Cycling Ways

As mentioned previously, my particular area of interest with OSM is in the capabilities to aid in cycling and path infrastructure mapping. Multiple years ago I purchased a Garmin cycling computer and noticed, despite its high cost, that the cycling paths were incomplete. Shortly after that I realized the potential of OSM to fill this gap. After finding tools to convert the maps to usable formats I set to work on altering and filling in the local path network. But let's begin by determining how cycling friendly this particular area of Colorado is.

The following query was used to list the counts of each type of highway, which is shown after the query.

```
self.collection.aggregate([
    {"$match": {"type": "way", "highway": {"$exists": True}}},
    {"$group": {"_id": "$highway", "count": {"$sum": 1}}},
    {"$sort": {"count": -1}},
    {"$limit": 12}
])
```

```
Highway: residential         , Count: 9929
Highway: service             , Count: 7819
Highway: footway             , Count: 7023
Highway: path                , Count: 2097
Highway: tertiary            , Count: 1197
Highway: secondary           , Count: 839
Highway: cycleway            , Count: 791
Highway: living_street       , Count: 634
Highway: track               , Count: 575
Highway: motorway_link       , Count: 275
Highway: unclassified        , Count: 248
Highway: motorway            , Count: 161
```

From these results it can be seen that the cycleway, indicating a dedicated cycling path, is the 7th most common type of highway. However, there are also footways and paths that are more common, would any of those allow for cycling? For that another query was performed, looking at the bicycling tag on the data.

```
self.collection.aggregate([
    {"$match": {"type": "way"}},
    {"$group": {"_id": "$bicycle", "count": {"$sum": 1}}},
    {"$sort": {"count": -1}},
])
```

```
Bicycle: None                , Count: 45449
Bicycle: yes                 , Count: 1432
Bicycle: no                  , Count: 674
Bicycle: dismount            , Count: 277
Bicycle: designated          , Count: 258
Bicycle: permissive          , Count: 7
Bicycle: allowed             , Count: 5
```

This shows the overwhelming majority of ways do not have bicycling tags. Performing the following query that incorporates both sources of data for determining the allowance of cycling, and then dividing by the total number of ways, gives the following results.

```
self.collection.aggregate([
    {"$match": {"type": "way", "$or": [
        {"highway": "cycleway"},
        {"bicycle": {"$in": ["yes", "designated", "permissive", "allowed"]}}]}},
    {"$group": {"_id": None, "count": {"$sum": 1}}}
])
```

```
Number of ways in which bicycling is allowed: 2048
Percent of ways in which bicycling is allowed: 4.3%
```

A total of 2048 ways that are suitable for cycling indicates 4.3% of total ways. However, given that most ways do not have a bicycling tag it seems likely that this information is sorely incomplete.

Cycling Shops

Lastly, let's determine how many cycling shops there are in the map area, and what the density of these shops is. Sometimes the primary function of a shop is other than cycling, but searching for the shop_1 key can help broaden the potential results.

```
num_bike_shops = self.collection.find(
    {"$or": [{"shop": "bicycle"}, {"shop_1": "bicycle"}]}
).count()
```

```
Number of bike shops: 7
Number of bike shops per km^2: 0.00969168
```

For an area of over 700 square kilometers it seems that 7 shops is rather low. However, one would have to compare to other areas, such as downtown Denver, for a truly meaningful contrast.

Suggestions for Improvement

The OSM database can be improved for bicycling friendliness. The first suggestion involves an outreach effort to suggest adding a bicycling tag on all new ways (other than ways used for buildings and such). It could be added to the "howto" guides published for new OSM users. However, reaching existing users might be a challenge. Perhaps a pinned post on the popular forums, wiki, and other related sites would be useful. As an extreme measure, checks could be added to the popular editing programs (iD, Potlatch, and JOSM) to not accept a new street or path based way without a bicycling tag.

There are drawbacks to these requirements, however. An outreach program may have limited visibility and success. While a hard requirement on new ways having bicycling tags could frustrate new and existing users. Existing users will likely not be used to that approach, and may be turned off by it. New users may not like unexpected requirements such as that to impede their ability to make progress. The OSM project suggests that a user actually physically have been on a path or road before submitting it (using GPS information). However, it is possible to use some open satellite imagery to trace new roads and paths. In that situation it may be that the user doesn't know about the suitability of a road for cycling.

As far as bicycling shops, it seems part of the issue lies with the need for a primary designation for a shop amenity. The current database system doesn't handle multiple sports stores well. For instance, near my house is a Sports Authority store, which carries equipment for camping, cycling, golfing, soccer, skiing, snowboarding, shooting, fishing, weightlifting, and many other sports. Ideally one could enter these multiple categories. But currently it seems that the only way to do so is with different shop tags, labeled as shop, shop_1, shop_2, etc. But these extra tags don't seem to be well supported. For example, the keepright website referenced above produces errors on keys with an underscore, such as shop_1. But the importance of having shop="bicycle" lies in the labeling of bike shops in cycling computers and in OSM based cycling websites, such as opencyclemap.org.

Ideally a tag would be added that can accept an array of values. But that would require a prohibitively difficult re-design of the current paradigm. Most OSM tags now do not involve arrays. Perhaps a related tag could be added that indicates if a store has some cycling gear, and the main shop tag could be shop="sports." It is always difficult, both for implementation and adoption, to add a new tag to the existing file formats.

Conclusions

The OpenStreetMap data has great potential for rapid updating to dynamic circumstances. However, it seems that more critical mass is required in people contributing to the project. Especially in the area of cycling, it would be useful for more editors to contribute their knowledge of bike shops and roads or paths that are safe for cycling. As OSM is used in more projects, and on cycling computers such as Garmins, it will become even more important to have complete information on cycling infrastructure.