Data Wrangling with MongoDB and OpenStreetMap data

André Marais

Map Area: Cape Town and surrounding area, South Africa

Map data gathered from https://mapzen.com/data/metro-extracts

1. Concerns with the Data

After subsetting the data to a smaller sample (of about 50Mb), I encountered 2 problems and also something else that might give rise to some concern

- There are quite a few non-address fields which have more than one colon, this might mean we have to create an array for this set of information.
- There are also a few CoTC (City of Cape Town) attributes which refer to municipal assets. If this is useful information, we can present these attributes in a better fashion
- Overabundance of certain postal codes

Non-address attributes with more than one colon

Upon further investigation I realized these mostly refer to ocean assets (buoys, lights, markers etc). While this data surely can be useful at some stage, I decided to omit it for now. The values contained in these fields are all *but* uniform, and cleaning it would require some deeper knowledge into ocean artifacts (what they are, what they should do, what elements are important etc).

CoTC attributes

These attributes refer to municipal assets, mainly road signs. Some of the fields have values (only speed limits at this stage), so I decided to only include attributes with non-zero values (ie if the speed limit or road number was zero, I omitted the attribute)

Overabundance of certain postal codes

While it's not unusual to see a skewed distribution of postal codes, one still needs to investigate why. At my work (insurance industry) I realized that the default (ie lazy) post code for Pretoria is 0001. We ignore those post codes, since no address has that actual post code. The most frequent post code for this data set was 7405, which is the Pinelands area. Upon further investigation I realized that the most frequent user on this data set is Adrian Frith, who lives in Pinelands. Adrian contributed more than 75% of this particular data set, and more than 80% of his work was done in Pinelands. This explains why we can be content with the high frequency of post code 7405 and we don't have to worry about cleaning it.

2. Data Overview

The original data set (after uncompressing) sits at 212 MB. Knowing that it will take a century to upload a file of that size to Github (on a line with 0.8Mbs upload speed yay), I decided to take a subset of the data of around 53MB. The JSON file is sitting at 81MB.

Document count

```
[{"$group" : {"_id" : "$type", "count" : {"$sum" : 1}}},
 {"$sort" : {"count" : -1}}]
```

shows that there are 242,587 nodes and 43,146 ways, totalling 285,733 documents. There are also 2 'palms', 5 'routes' and 1 multipolygon we can conveniently ignore.

User statistics

As stated earlier, Adrian Firth is the top contributor. Below is a list of the top 5 contributors:

```
[{"$match":{"address.city":{"$exists":1}}},
 {"$group":{"_id":"$created.user", "count":{"$sum":1}}},
 {"$sort":{"count":-1}},
 {"$limit" : 5}]
```

shows:
[{u'_id': u'Adrian Frith', u'count': 1019},
 {u'_id': u'Stefanoodle', u'count': 133},
 {u'_id': u'adjuva', u'count': 86},
 {u'_id': u'Kelerei', u'count': 44},
 {u'_id': u'Redro', u'count': 7}]

User per city

```
[{"$match":{"address.city":{"$exists":1}}},
 {"$group": {"_id": {"user" : "$created.user",
                                  "city" : "$address.city"},
                      "count":{"$sum":1}}},
 {"$sort":{"count":-1}},
 {"$limit" : 10}]
```

You can see Adrian is very active in Pineland specifically:
[{u'_id': {u'city': u'Pinelands', u'user': u'Adrian Frith'}, u'count': 819},
 {u'_id': {u'city': u'Rosebank', u'user': u'Adrian Frith'}, u'count': 99},
 {u'_id': {u'city': u'Loevenstein', u'user': u'Stefanoodle'}, u'count': 96},
 {u'_id': {u'city': u'Cape Town', u'user': u'adjuva'}, u'count': 86},
 {u'_id': {u'city': u'Mowbray', u'user': u'Adrian Frith'}, u'count': 85}]

Some street name information

This is just interesting, the street names in Pinelands are named quite aptly :)

```
[{"$match":{"address.city":'Pinelands',}},
 {"$group":{"_id": {"street name" : "$address.street"},
"count":{"$sum":1}}},
 {"$sort":{"count":-1}},
 {"$limit" : 10}]
```

```
[{u'_id': {u'street name': u'Ringwood Drive'}, u'count': 46},
 {u'_id': {u'street name': u'Forest Drive'}, u'count': 44},
 {u'_id': {u'street name': u'Victory Avenue'}, u'count': 16},
 {u'_id': {u'street name': u'Union Avenue'}, u'count': 14},
 {u'_id': {u'street name': u'Daffodil Way'}, u'count': 12},
 {u'_id': {u'street name': u'Camp Road'}, u'count': 12},
 {u'_id': {u'street name': u'Brookdale Avenue'}, u'count': 12},
 {u'_id': {u'street name': u'Meerlust'}, u'count': 11},
 {u'_id': {u'street name': u'Days Walk'}, u'count': 11},
 {u'_id': {u'street name': u'Sunny Way'}, u'count': 10}]
```

Restaurants and Fast food cuisine types

```
[{"$match":{"amenity":"fast_food"}},
 {"$group":{"_id":"$name", "count":{"$sum":1}}},
  {"$sort":{"count":-1}}]
```

Shows a decent variety of options, with the expected McDonald's topping the list

and

```
[{"$match":{"amenity":"restaurant"}},
  {"$group":{"_id":"$cuisine", "count":{"$sum":1}}},
   {"$sort":{"count":-1}}]
```

Unfortunately shows very little information, Most of the restaurants haven't been classified yet.

Other Ideas

The main issue is the small pool of contributors. There were 59 contributors, with only a very few likely being bots (having bot in username, or a having username which consisting of upper-, lowercase letter and numbers). Barely a third of these users made more than one contribution and only 7 users made more than 5 contributions. OpenStreetmap is quite unknown to most people, mostly because the benefits of using it is not too obvious.

Cape Town specifically has a lot of outdoor activities (off road cycling, mountain biking, hiking, mountain climbing etc). I think OpenStreetMap might not be the best platform for the average user to save this kind of information to, but perhaps one can look into combining data from Garmin and Strava. Quite a large portion of active people use fitness devices these days (some with GPS), so it's just a matter of getting the data and going through a similar cleaning exercise.

One major issue might be the aggregation of all of the data. If there are a few people who all walked/ cycled/ climbed the same route, how would we go about mapping the combined data? And considering that some climbing routes aren't really fixed routes, it might become messy to show all the different routes. One solution that pops up is that we can use zoning instead of having distinct paths or routes, but this would then rely on a lot more data to 'flesh out' the zones.

Apart from gathering good quality data from Google Maps using the API (the R package is quite awesome, haven't checked the Python library out yet), you can also use food ordering sites to scrape data. See it as a food location repository :) But the main problem still remains, all of this is human entered data. Whether it's from Google Maps or a food website, fingers pressed the keys. I think if OpenStreetMap starts being more useful to the average person, more people will start using it

Conclusion

Cape town has some decent clean data, but there's still a lot of work to do to make it attractive to the average user. I think the main problem is that most people might see OpenStreetMap as 'another Google Maps', while it can be more like a 'Map wikipedia'. I think the data you see on OpenStreetMap reflects the culture/ sentiment of the users. A more 'outdoor' city might have a higher frequency of outdoor activities, while a more metro city will have more traffic/street related data. It would be super interesting to see if you can combine Strava and Garming data with OpenStreetMap to show outdoor activities people can do