1. Problems encountered in the OSM map of Mumbai

Incomplete map

First and foremost, the available map of Mumbai was a fraction of the size of other comparable metros, and hence seemed incomplete.

Mumbai: 233 sq mi, pop. 12.4 m, compressed XML 7.6 MB
Madrid: 233 sq mi, pop. 3.1 m, compressed XML 46.9 MB
Marseille: 92 sq mi, pop. <1 m, compressed XML 74.9 MB

Small number of contributors

Running 'users.py' code revealed only about 850 contributors to the map in a city of over 12 million people, which possibly explains the relatively small map available for Mumbai.

Difficult cleanup

An audit of the street addresses in the map data (audit.py) displayed 119 street types. Only 10 or so street types are generally used in Mumbai.

In India, street names typically contain landmark, locality or neighborhood information, so regular street types such as Road / Lane / Street don't always appear at the end of an address line.

Sample street addresses from the map data:

"Gulmohar Road, Juhu Scheme, Vile Parle West"
"4, Lokhandwala Complex Rd, Shastri Nagar, Andheri West"

So, cleaning up Mumbai street addresses for consistency requires a different approach and greater effort as compared to US street addresses.

2. Overview of the data

Uncompressed, the Mumbai OSM XML data was 109.9 MB in size. The XML data was processed by the 'data.py' code to generate a file 125.5 MB in size, containing JSON documents shaped as per the prescribed model.

The JSON documents were loaded into a MongoDB database named 'osm', in a collection named 'mumbai', and then summarized as follows.

Total nodes: 536044

```
> db.mumbai.find({"type": "node"}).count()
536044
```

Total ways: 47565

```
> db.mumbai.find({"type": "way"}).count()
47565
```

Total unique contributors: 852

```
> db.mumbai.distinct("created.user").length
852
```

Top 10 contributors:

```
> db.mumbai.aggregate([{$group: {_id: "$created.user", count:
{$sum: 1}}}, {$sort: {"count": -1}}, {$limit: 10}])
{ "_id" : "parambyte", "count" : 87942 }
{ "_id" : "PlaneMad", "count" : 65715 }
{ "_id" : "balaji88", "count" : 58003 }
{ "_id" : "MJL Wood", "count" : 57813 }
{ "_id" : "udaya", "count" : 43477 }
{ "_id" : "smith_dsm", "count" : 38340 }
{ "_id" : "Giyavudeen", "count" : 28981 }
{ "_id" : "Heinz_V", "count" : 21675 }
{ "_id" : "indigomc", "count" : 19185 }
{ "_id" : "singleton", "count" : 15533 }
```

Top 10 users' contribution: 436664 or 75%

```
> db.mumbai.aggregate([{$group: {_id: "$created.user", count:
{$sum: 1}}}, {$sort: {"count": -1}}, {$limit: 10}, {$group:
{_id: "", sum: {$sum: "$count"}}}])
{ "_id" : "", "sum" : 436664 }
```

Average contribution of each user: 685

```
> db.mumbai.aggregate([{$group: {_id: "$created.user", count:
{$sum: 1}}}, {$group: {_id: "", avg: {$avg: "$count"}}}])
{ "_id" : "", "avg" : 685.231220657277 }
```

Top amenities: place of worship, restaurant, school, …

```
> db.mumbai.aggregate([{$match: {"amenity": {$exists: 1}}},
{$group: {_id: "$amenity", count: {$sum: 1}}}, {$sort:
{"count": -1}}, {$limit: 10}])
{ "_id" : "place_of_worship", "count" : 354 }
{ "_id" : "restaurant", "count" : 241 }
{ "_id" : "school", "count" : 234 }
{ "_id" : "bank", "count" : 198 }
{ "_id" : "hospital", "count" : 154 }
{ "_id" : "fuel", "count" : 123 }
{ "_id" : "parking", "count" : 121 }
{ "_id" : "bus_station", "count" : 112 }
{ "_id" : "cafe", "count" : 107 }
{ "_id" : "college", "count" : 94 }
```

Top religions: Reflects religious diversity of city and country

```
> db.mumbai.aggregate([{$match: {"amenity":
"place_of_worship"}}, {$group: {_id: "$religion", count:
{$sum: 1}}}, {$sort: {"count": -1}}])
{ "_id" : "hindu", "count" : 121 }
{ "_id" : "muslim", "count" : 94 }
{ "_id" : "christian", "count" : 52 }
{ "_id" : null, "count" : 49 }
{ "_id" : "zoroastrian", "count" : 9 }
{ "_id" : "jain", "count" : 8 }
{ "_id" : "buddhist", "count" : 7 }
{ "_id" : "sikh", "count" : 6 }
{ "_id" : "jewish", "count" : 4 }
{ "_id" : "Jain", "count" : 1 }
```

Fuel type indicated: Only 31 times across 123 fuel amenities

(Important to know as many cars in Mumbai run on CNG or diesel.)

```
> db.mumbai.find({"amenity": "fuel", "fuel:cng": {$exists:
1}}).count()
7
> db.mumbai.find({"amenity": "fuel", "fuel:diesel": {$exists:
1}}).count()
24
```

Total hospitals: 154 (out of 253* actual)

* http://www.tiss.edu/tiss-attachements/downloads/list-of-hospitals-in-mumbai/view

```
> db.mumbai.find({"amenity": "hospital"}).count()
154
```

Total vegetarian restaurants: Only 10 of 241 restaurants

(Large population of vegetarians in Mumbai.)

```
> db.mumbai.find({"amenity": "restaurant", "diet:vegetarian":
{$exists: 1}}).count()
10
```

3. Final thoughts

As noted in section 1, two problems afflict the OSM data for Mumbai:

- The data seems incomplete as compared to other metros.
- The street addresses contain landmark and neighborhood information, which makes cleaning up street addresses challenging.

Raising awareness of the OSM project may make the data more complete, especially with the ubiquity of GPS-enabled smartphones, and the city's high level of pedestrian traffic.

The data.py script can be used to split street information into street, landmark and neighborhood attributes, where possible. The street information can then be cleaned up consistently. Availability of landmark and neighborhood information as part of the address would be an added bonus in a city like Mumbai, where awareness of landmarks and neighborhoods is far greater than street names.

Author: Bhavin V. Choksi

Project: Data Wrangling with MongoDB