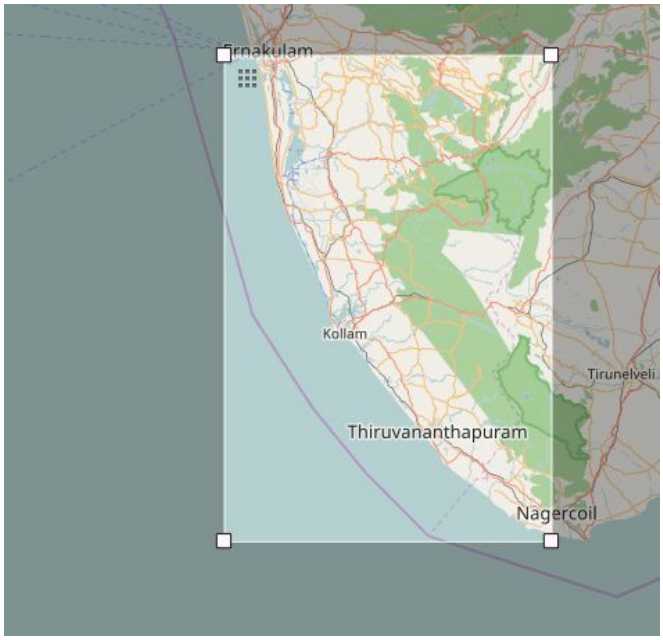
P3-Wrangling OpenStreetMap Data

OpenStreetMap Data Case Study

Map area

I chose South Kerala,India as my region of study since I was born and brought up in Thiruvananthapuram,Kerala also known as God's own country :)
So,it will help me validate the data using my personal knowledge of the area.
Openstreetmap URL:http://www.openstreetmap.org/export#map=8/9.636/77.657
Here is the screenshot showing the region I chose



Data Audit

Different tag types in osm file

Basically,there are three data primitives(nodes,ways and relations). A node is one of the core elements in the OpenStreetMap data model. It consists of a single point in space defined by its latitude, longitude and node id. A way is an ordered list of nodes which normally also has at least one tag or is included within a Relation. A relation is one of the core data elements that consists of one or more tags and also an ordered list of one or more nodes, ways and/or relations as members which is used to define logical or geographic relationships between other elements.

Problematic tags

Each of the tags inside a node,way or a relation has a key value pair.

For eg.<tag k="name" v="Evanchalikkal Church Road"/>

Next I checked for the k values for any problematic tags based on four categories:

1. Tags that contain only lowercase letters and are valid

eg.<tag k="cycleway" v="no"/>

2.For otherwise valid tags with colon in their names

eg.<tag k="maxspeed:motorcycle" v="50"/>

3.Tags with problematic characters

eg.<tag k="_OBJECTID_1_" v="17523"/>

We would like to change the data model and expand the "addr:street" type of keys to a dictionary like this: {"address": {"street": "Some value"}}

1.Problems encountered in the map

After downloading a portion of south kerala and checking the xml data,I noticed two main problems:

1. Street address inconsistencies
2. City and district name inconsistencies

Using the audit_street_type() function,I was able to find out the street types that are not of the standard form

Street address inconsistencies

- Different types of abbreviations are used
  Eg.for the word *Road,*the different abbreviations are *Rd.,rd*
- Spelling mistakes
  Eg.For *junction*,it is misspelled as *junctin*
- Lower case
  Eg.*road* instead of *Road*

Using the update_name function,the following changes were made.

road->Road
rd->Road
Rd->Road
Jn->Junction
jun->Junction

City and district name inconsistencies

- The main problem I encountered with the district name was the different misspellings.
  Example:for the district Alappuzha the different misspellings were 'Allapuzha','Allapura'
- Improper abbreviations
  Eg:*TV Puram* for *Thiruvananthapuram*

Some instances of changes made:

Allapura->Alapuzha
Allapuzha->Alapuzha
TV Puram->Thiruvananthapuram

- Most of the inconsistencies solved by using update_name function

Inconsistencies in postal code

Example:the different formats like *685 509,686613*

- Using the update_zip function,all the pincodes were converted into standard format with a space in the middle of the six digits

2.Data Overview

This section contains basic statistics about the dataset and the MongoDB queries used to gather them.

File sizes

kerala_south.osm.....111 MB
kerala_south.db.....457 MB
nodes.csv.....46.4 MB
nodes_tags.csv.....1.14 MB
ways.csv.....2.10 MB
ways_tags.csv.....2.97 MB
ways_nodes.csv.....15.4 MB

Number of nodes

```
cur.execute('''
    SELECT COUNT(*) FROM nodes;
''')
```

Output

3480870

Number of ways

```
cur.execute('''
    SELECT COUNT(*) FROM ways;
''')
```

Output

180060

Number of unique users

```
cur.execute('''
SELECT COUNT(DISTINCT(e.uid))
FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways) e;
''')
```

Output

1163

Top 10 contributing users

```
cur.execute('''
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;
''')
```

Output

```
[(u'Prasanth Rajan', 336310), (u'Arun Dhilogics', 280957), (u'premkumar',
222417), (u'matthayichen', 201603), (u'sk_trav', 187143), (u'PlaneMad',
186024), (u'uttelamiak', 172750), (u'manuvarkey', 167038), (u'SM@Edit',
105037), (u'Akhilan', 104318)]
```

3.Data Exploration

Biggest Religion

```
cur.execute('''
SELECT nodes_tags.value, COUNT(*) as num
FROM nodes_tags
    JOIN (SELECT DISTINCT(id) FROM nodes_tags WHERE value='place_of_worship')
i
    ON nodes_tags.id=i.id
WHERE nodes_tags.key='religion'
GROUP BY nodes_tags.value
ORDER BY num DESC
LIMIT 1;

''')
```

Output

[(u'christian', 2058)]

Interestingly,according to 2011 Census of India figures, 54.73% of Kerala's residents are **Hindus**, 26.56% are **Muslims**, 18.38% are **Christians**, and the remaining 0.32% follows another religion or no religion.

Top 10 amenities

```
cur.execute('''

SELECT value, COUNT(*) as num
FROM nodes_tags
WHERE key='amenity'
GROUP BY value
```

```
ORDER BY num DESC
LIMIT 10;

''')
```

Output

```
[(u'place_of_worship', 4968), (u'restaurant', 2238), (u'school', 2070),
(u'fuel', 1932), (u'bank', 1752), (u'hospital', 1218), (u'atm', 1170),
(u'bus_station', 594), (u'post_office', 426), (u'cafe', 420)]
```

Most popular cuisines

```
cur.execute('''

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

''')
```

Output

```
[(u'indian', 216), (u'regional', 138), (u'vegetarian', 48), (u'international',
12), (u'other', 12), (u'Arabian,_Malabar,_Chicken', 6), (u'Fried_Chicken', 6),
(u'Indian', 6), (u'North_indian', 6), (u'Regional', 6), (u'SOUTH_INDIAN', 6),
(u'South_Indian_Meals', 6),
(u'Thai,_Japanese,_Singaporean,_Malaysian,_Chinese,_Vietnamese,_Indonesian',
6), (u'arab;chinese;north_indian;south_indian', 6),
(u'breakfast;vegetarian;local', 6), (u'chayakada + meals', 6), (u'chicken',
6), (u'italian', 6), (u'north_indian', 6), (u'north_indian,_punjabi', 6),
(u'pizza', 6), (u'seafood', 6), (u'south_indian', 6), (u'tibetan', 6)]
```

4.Additional Ideas for improvement

1.The quality of openstreetmap data can be assessed by comparing it with proprietary data or data of governmental map agencies.
2..Reviews and ratings as in a Google map can be included for the tourist attractions and restaurants.This will greatly help both the local people and tourists alike.Here also,the reviews may depend from one person to another.

3.Automated auditing tools can be brought out for the community to use during the uploading process of data.
4.A heat map layer could be overlaid on the map showing how frequently or how recently certain regions of the map have been updated. These map layers could help guide users towards areas of the map that need attention in order to help more fully complete the data set.

5.Conclusion

The openstreetmap data of South Kerala is of fairly good quality.Most of the problems encountered are due to the usage of unstandardised abbreviations,spelling mistakes and other typos.Also,there is the use of unconventional postal code types.These errors can be avoided by using quality assurance tools and validation.
There's plenty of potential to extend OpenStreetMap to include user reviews of establishments, subjective areas of what classifies a good vs bad neighborhood, housing price data, school reviews, walkable paths, quality of mass transit like bus transport,metro rail,, and a bunch of other metrics that could form a solid foundation for robust recommender systems. These recommender systems could help users in deciding where to live or what restaurants to check out.The data is far too incomplete to be able to implement such recommender systems as it stands now.