Project - Wrangle OpenStreetMap Data

by:Harish Garg(harish.garg@gmail.com)

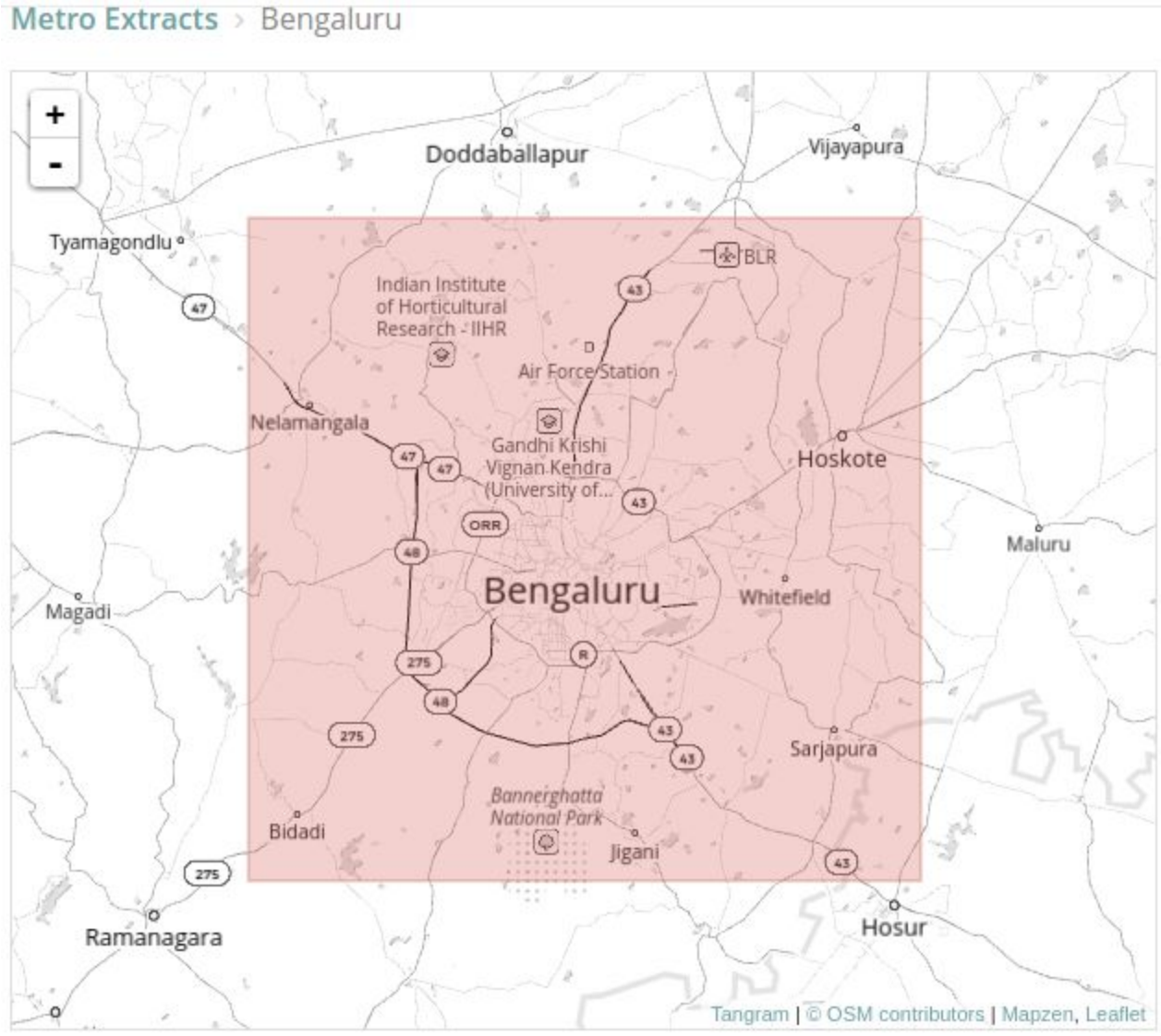
Table of Contents

Project Overview

The purpose of this project is to analyze OpenStreepMap Data for a chosen map area. This project is created for Udacity Data Analyst NanoDegree

About the Map Area

I chose city of Bengaluru, India from https://mapzen.com/data/metro-extracts/



The reason I chose this area is - I am a resident of Bengaluru from last few years. However, I it still keep discovering new areas in this city and thought choosing this area for this project will be a dual learning experience for me - know my current city more as well as finish the project.

Problems encountered in the map

I started with downloading the pre-extracted metro file for the Bengaluru City(full) from https://mapzen.com/data/metro-extracts/metro/bengaluru_india/.

The OSM XML file was 610 MB. However, I got stuck at the very 1st script i.e. mappraser.py. Over multiple runs, my system got hung and was never able to finish. So I used the create sample script to choose every 10th node. The resulting OSM XML file ,came about to be around 62 MB.

I ran mapparser.py on the above data file. The OSM file was parsed successfully and it it showed 284261 Nodes.

```
{'member': 524,
 'nd': 351633,
 'node': 284261,
 'osm': 1,
 'relation': 94,
 'tag': 77656,
 'way': 65260}
```

Issue # 1
Finding out how many entries are with pronlem chars.
Next, I ran the tags.py script on the OSM file. It showed below data.
{'lower': 75284, 'lower_colon': 2224, 'other': 148, 'problemchars': 0}

There are 0 k tags with problem chars.

Issue # 2
Next, I decided to audit the postal codes. All bangalore codes start with 5 and should be of length of 6, all numbers. 63 unique postal codes (or PIN codes as they are called in India). The ones with issues and their correct format are they have a space char in the middle.

'560 001' -> 560001
'560 068' -> 560068

This was fixed before generating the JSON. and the output from the script is…

```
harish:~/data_analysis_nanodegree_projects/Project3_Wrangle_OpenStreetMap_Data$ python final_parser.py
'Changed 560 001 to 560001'
'Changed 560 068 to 560068'
```

Issue # 3

Running audit.py showed lot of issues in street names. Main issues were badly capitalized names or with special characters. For example...

```
"ROad": "Road",
"road": "Road",
"stage": "Stage",
"cross": "Cross",
"main": "Main",
"street": "Street",
"vijayanagar": "Vijayanagar",
"road\)": "Road",
"Colony\)": "Colony",
"Road\)": "Road"
```

Here is the output from the script…

```
harish:~/data_analysis_nanodegree_projects/Project3_Wrangle_OpenStreetMap_Data$ python final_parser.py
'Changed 560 001 to 560001'
'West of Chord Road (LIC Colony)=>>West of Chord Road (LIC Colony'
'Changed 560 068 to 560068'
'27 main, sector-1, hsr layout=>>27 Main, sector-1, hsr layout'
'Hosur road=>>Hosur Road'
'Hosur road=>>Hosur Road'
'100 Feet Road(S K Karim Khan Road)=>>100 Feet Road(S K Karim Khan Road'
'seegehalli, whitefield road=>>seegehalli, whitefield Road'
'24th main=>>24th Main'
'80 Feet Road(Sir C.V. Raman Hospital Road)=>>80 Feet Road(Sir C.V. Raman Hospital Road'
'jnguva street=>>jnguva Street'
'Swamy Vivekananda Road (Old Madras Road)=>>Swamy Vivekananda Road (Old Madras Road'
'Swamy Vivekananda Road (Old Madras Road)=>>Swamy Vivekananda Road (Old Madras Road'
'10th cross=>>10th Cross'
'vijayanagar=>>Vijayanagar'
'bannerughatta road=>>bannerughatta Road'
'Indiranagar Double Road (Paramahansa Yogananda road)=>>Indiranagar Double Road (Paramahansa Yogananda Road)'
'Indiranagar Double Road (Paramahansa Yogananda Road)=>>Indiranagar Double Road (Paramahansa Yogananda Road'
'Indiranagar Double Road (Paramahansa Yogananda road)=>>Indiranagar Double Road (Paramahansa Yogananda Road)'
'Indiranagar Double Road (Paramahansa Yogananda Road)=>>Indiranagar Double Road (Paramahansa Yogananda Road'
'kodigehalli main road, 1st main, 4th cross=>>kodigehalli main Road, 1st main, 4th cross'
'kodigehalli main Road, 1st main, 4th cross=>>kodigehalli Main Road, 1st Main, 4th cross'
'kodigehalli Main Road, 1st Main, 4th cross=>>kodigehalli Main Road, 1st Main, 4th Cross'
'100 Feet Road(S K Karim Khan Road)=>>100 Feet Road(S K Karim Khan Road'
'100 Feet Road(S K Karim Khan Road)=>>100 Feet Road(S K Karim Khan Road'
'100 Feet Road(S K Karim Khan Road)=>>100 Feet Road(S K Karim Khan Road'
'100 Feet Road(S K Karim Khan Road)=>>100 Feet Road(S K Karim Khan Road'
'100 Feet Road(S K Karim Khan Road)=>>100 Feet Road(S K Karim Khan Road'
'100 Feet Road( S K Karim Khan Road)=>>100 Feet Road( S K Karim Khan Road'
'100 Feet Road(S K Karim Khan Road)=>>100 Feet Road(S K Karim Khan Road'
'100 Feet Road( S K Karim Khan Road)=>>100 Feet Road( S K Karim Khan Road'
'100 Feet Road( S K Karim Khan Road)=>>100 Feet Road( S K Karim Khan Road'
'100 Feet Road(S K Karim Khan Road)=>>100 Feet Road(S K Karim Khan Road'
'80 Feet Road(Sir C.V. Raman Hospital Road)=>>80 Feet Road(Sir C.V. Raman Hospital Road'
'80 Feet Road(Sir C.V. Raman Hospital Road)=>>80 Feet Road(Sir C.V. Raman Hospital Road'
```

'8th cross=>>8th Cross'
'lakshmipuram 1st main road=>>lakshmipuram 1st main Road'
'lakshmipuram 1st main Road=>>lakshmipuram 1st Main Road'
'Budigere road=>>Budigere Road'
'M G ROad=>>M G Road'
'Kodichikkanahalli main road=>>Kodichikkanahalli main Road'
'Kodichikkanahalli main Road=>>Kodichikkanahalli Main Road'
'Abbigere Main road=>>Abbigere Main Road'
'13th cross=>>13th Cross'
'19th D cross=>>19th D Cross'
'3rd cross=>>3rd Cross'
'100 Feet Road(S K Karim Khan Road)=>>100 Feet Road(S K Karim Khan Road'
'1st C cross, 2nd phase 4th block, BSK 3rd stage=>>1st C Cross, 2nd phase 4th block, BSK 3rd stage'
'1st C Cross, 2nd phase 4th block, BSK 3rd stage=>>1st C Cross, 2nd phase 4th block, BSK 3rd Stage'
'100 Feet Road(S K Karim Khan Road)=>>100 Feet Road(S K Karim Khan Road'
'7th cross=>>7th Cross'

Overview of the Data

size of the file: 62M

I ran the final_parser.py on the OSM XML file, which produced a JSON file. The I imported the JSON file into a local instance of mongodb by using below command.

mongoimport --db openstreetmap --collection bengaluru --drop --file ~/datasets/sampleK10.osm.json

After that< i launched the mongo shell in the Linux terminal to run few more stats…

Number of unique users:
>db.bengaluru.distinct("created.user").length
877

Number of Nodes
> db.bengaluru.find({"type": "node"}).count()
284260

Number of Ways
>db.bengaluru.find({"type":"way"}).count()
65258

Number of chosen type of nodes, like cafes, shops etc.

We get the list of amenities and their counts by using this query.
>db.bengaluru.aggregate([{$match: {'amenity': {$exists: 1} }}, {$group: {_id: '$amenity', count: {$sum: 1}}}, {$sort: {'count': -1}}])
Few Top ones are…
{ "_id" : "restaurant", "count" : 127 }

{ "_id" : "place_of_worship", "count" : 88 }
{ "_id" : "bank", "count" : 65 }
{ "_id" : "school", "count" : 64 }
{ "_id" : "atm", "count" : 59 }
{ "_id" : "hospital", "count" : 51 }
{ "_id" : "fast_food", "count" : 39 }
{ "_id" : "college", "count" : 30 }
{ "_id" : "parking", "count" : 29 }
{ "_id" : "cafe", "count" : 28 }

Other ideas about the datasets

Problem - Incomplete Dataset
One of the biggest feeling you get when you look at the OSM dataset for Bangalore is that it's very incomplete. For example, look at the list of amenities and their counts by using this query.
>db.bengaluru.aggregate([{$match: {'amenity': {$exists: 1} }}, {$group: {_id: '$amenity', count: {$sum: 1}}}, {$sort: {'count': -1}}])
Few Top ones are…
{ "_id" : "restaurant", "count" : 127 }
{ "_id" : "place_of_worship", "count" : 88 }
{ "_id" : "bank", "count" : 65 }
{ "_id" : "school", "count" : 64 }
{ "_id" : "atm", "count" : 59 }
{ "_id" : "hospital", "count" : 51 }
{ "_id" : "fast_food", "count" : 39 }
{ "_id" : "college", "count" : 30 }
{ "_id" : "parking", "count" : 29 }
{ "_id" : "cafe", "count" : 28 }

Now, if one has lived in Bengaluru for more than a day, you can see that the no. of restaurants and place_of_worhsip numbers in this dataset are way less than what they actually should be. Seems like the contributions to the openstreetmap are not very complete as far as Bengaluru is concerned.

Solution Proposed
OpenStreetMap is a crowdsource project. And that is good because maps are generally data driven and the people contributing about where live / work / play are best contributors. And that's the key. You need to get more people contributing, more data about their local area. And the best way to do that would be build a mobile map app from which people can contribute directly, from wherever they are. One can build gamification in this where the users can be rewarded with badges like Local Guide Level 1 and so on.

Benefits

- Since people carry and use mobile devices everywhere they go, they are more likely to submit more data
- More people submit data about a particular place, with GPS co-ordinates, higher the accuracy of the data.
- Data will be more up to date.

Issues

- More users connecting means more load on the server resources, means more costs.
- OpenStreetMap would have to invest in building and maintaining these apps.