Project - Wrangle OpenStreetMap Data

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

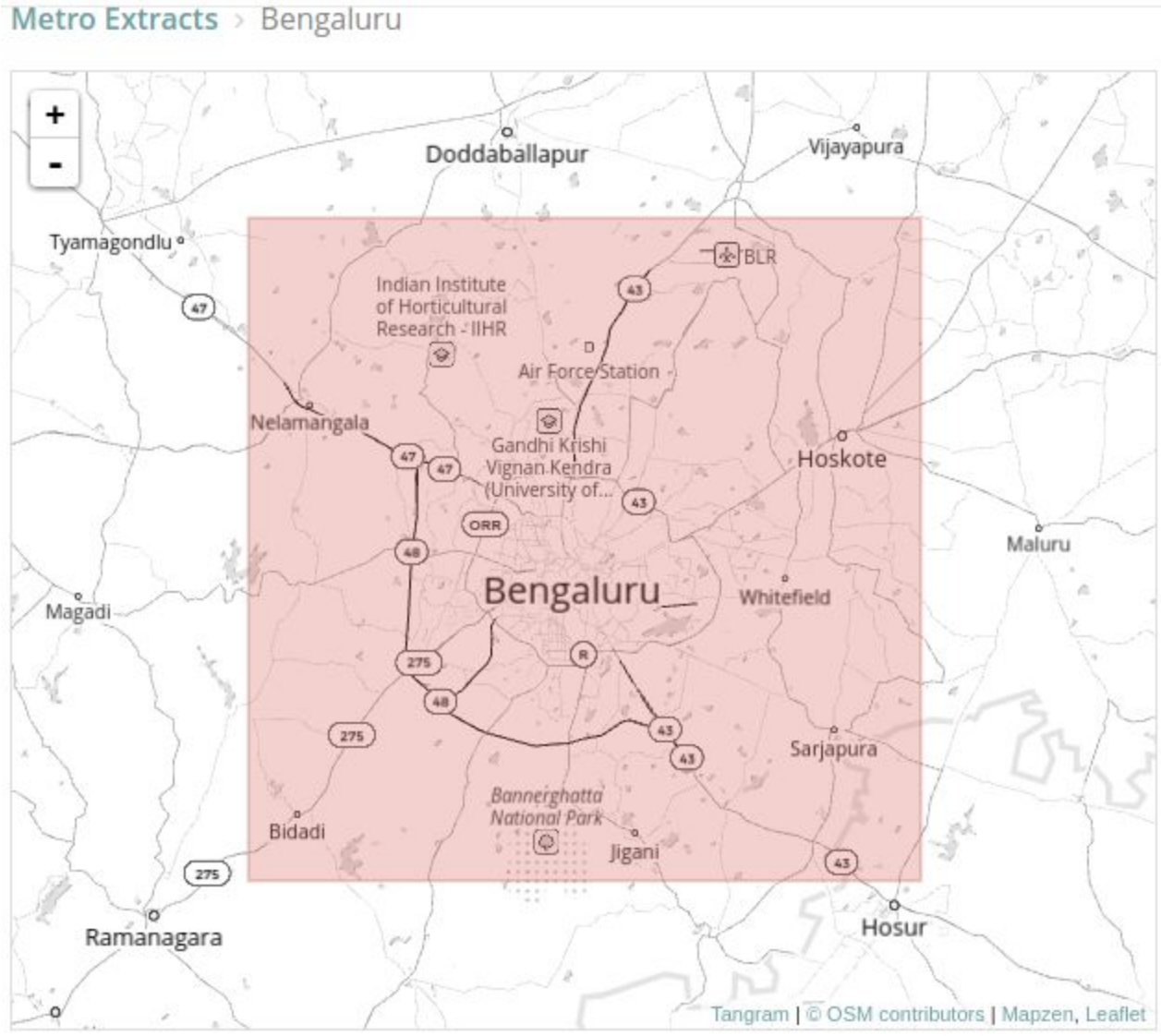
Table of Contents

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}

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.

Users.py shows there are 878 unique users who contributed.

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 are...

'560 001'
'560 068’

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"

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

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 }

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.