Project 3: OpenStreetMap Data Case Study

Map Area

Bengaluru, Karnataka, INDIA

- https://mapzen.com/data/metro-extracts/

Bengaluru is considered as the IT capital of India, The city saw tremendous growth and industrial revolution in the last decade. It also hosts many fortune 500 companies; this made me pick this city's map to do my case study.

Problems Encountered In The Map

I used the provided code, To make a OSM file for a smaller area (Using a larger value of k). In the initial inspection, I found the following problem with the map while auditing the data.

- Incorrect postal codes
- Abbreviated street names
- Errors in street tag
- Full address of the city in street tag
- Inconsistency in city tag

Errors in Street Tag:

I removed some errors while auditing the street tag. Here are some examples:

1. Comma in the end
2. Comma in the beginning
3. Full stop in the end

Incorrect Postal Code:

I found much discrepancy while auditing the postal code. Some examples are mentioned below:

1. 560 001
2. – 560094
3. Bengaluru
4. 5600011
5. 560027"
6. 56006
7. 560001ph
8. iam in bang

I corrected the postal code where it was possible, but where the postal code was incomplete or had a text in place, I removed it while updating the database.

Abbreviated Street Names:

I found some street names and words to be abbreviated in street tag. Here are some of the examples:

1. "st": "Street"
2. "rd": "Road"
3. "blk": "Block"

I updated these abbreviated words respectively.

Full address instead of street name:

While auditing the street tag, I noticed that most values in the tag contain full address instead of just the street name. Here are some examples:

1. `I Main, I Phase, V Stage, BEML Layout, RajaRajeshwari Nagar`
2. `1st Cross, 1st Main, 1st Phase, 5th Stage, BEML Layout`
3. `13th cross, R.A. Road, Ejipura`

I found this problem in most of the tags to be corrected manually. To get the best of our map data, where ever I found comma in the street tag, I changed “street” to “full” type while updating the database.

Inconsistency in city tag

While auditing the city tag, I noticed some inconsistency:

1. Bangalore
2. bengaluru
3. Whitefield, Bangalore
4. Bellary
5. Whitefield

Bangalore was the old city name which now has been named Bengaluru, Hence both of these are correct and used popularly. I also found some tags to have area information attached. I decided not remove this information as Bengaluru is a big metro city. For consistency purposes I updated city name to be “Bengaluru” in most of the cases and ignored others.

Data Overview and Additional ideas

This section contains basic statistics about the dataset, the SQL queries used to gather them, and some additional ideas about the data in context.

FILE SIZES

```
charlotte.osm ..... 294 MB
charlotte.db ..... 129 MB
nodes.csv ..... 144 MB
nodes_tags.csv ..... 0.64 MB
ways.csv ..... 4.7 MB
ways_tags.csv ..... 20 MB
ways_nodes.cv ..... 35 MB
```

NUMBER OF NODES

```
sqlite> SELECT COUNT(*) FROM nodes;
```

2855613

NUMBER OF WAYS

```
sqlite> SELECT COUNT(*) FROM ways;
```

654455

NUMBER OF UNIQUE USERS

```
sqlite> SELECT COUNT(DISTINCT(e.uid))
FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways) e;
```

1747

FILE SIZES

```
charlotte.osm ..... 294 MB
charlotte.db ..... 129 MB
nodes.csv ..... 144 MB
nodes_tags.csv ..... 0.64 MB
ways.csv ..... 4.7 MB
ways_tags.csv ..... 20 MB
ways_nodes.cv ..... 35 MB
```

TOP 10 CONTRIBUTING USERS

```
sqlite> 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;
```

jasvinderkaur,126140
akhilsai,119332
premkumar,116184
saikumar,115138
shekarn,100001
vamshikrishna,94386
PlaneMad,93977
himalay,88514
himabindhu,87410
sdivya,85202

Number of users appearing only once (having 1 post)

```
sqlite> SELECT COUNT(*)
FROM
    (SELECT e.user, COUNT(*) as num
```

```
    FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e
    GROUP BY e.user
    HAVING num=1)  u;
```

426

References:

1. Stroop Effect - Wikipedia
2. P-Value Calculator