Introduction

Map Area/ Data:

City: Chennai, Tamil Nadu, India

The dataset for the city of Chennai was obtained from metro extracts of the open street maps available at https://www.nextzen.org/metro-extracts/index.html#chennai_india. I'm interested in this part of the world because it is my hometown and I would like to explore the contributions to the open street map and get an idea of all the changes that can be made to improve this data.

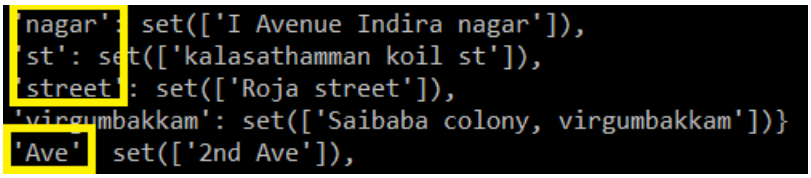
Identifying problems in the Chennai city area of the open street maps

To get a general overview of the data and to get an idea of the problems present, a sample of the entire data was taken using the sample.py code and then the Unix command less was used to look at the layout of the data. With the help of audit.py code, unusual street names and postal code was printed out.

Problems encountered in the map

Street Type

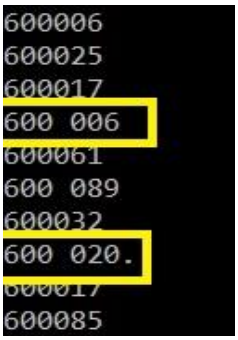
In Chennai, the major categories of street will be addressed by one of the following: Street, Road, Avenue, Nagar. When looking at the data, there were inconsistencies in these type like shown in the below figure.



In most cases street, avenue and roads were in unabbreviated forms namely – St, Ave or Rd. In order to make this consistent, generic mapping for the street type was added to the code.

Postal code

In Chennai, the postal codes are continuous six-digit numbers starting with the digit six. However, as seen above, there are inconsistencies that must be fixed.



Analysis

Data Overview

Following is the statistical summary of the open street map data of city of Chennai data set. After loading data into MongoDB, I used a MongoDB GUI called Robomongo for querying the database.

File Size:

Chennai_india.osm: 396 MB

Chennai_india.osm.json: 471 MB

Data auditing code before importing it to Mongo db : project.py

Import statement for importing the JSON file generated from the project.py code.

mongoimport --db udacity --collection map --drop --file chennai_india.osm.json

Statistical summary from MongoDb:

db.stats()


```
{
    "db" : "udacity",
    "collections" : 1.0,
    "views" : 0.0,
    "objects" : 2277535.0,
    "avgObjSize" : 237.26497507173326,
    "dataSize" : 540379285.0,
    "storageSize" : 162471936.0,
    "numExtents" : 0.0,
    "indexes" : 1.0,
    "indexSize" : 22843392.0,
    "fsUsedSize" : 135417294848.0,
    "fsTotalSize" : 987211096064.0,
    "ok" : 1.0
}
// ----------------------------------------
```


Number of documents: 2277535

```
db.getCollection('map').find().count()
```

Number of nodes: 1861817

```
db.getCollection('map').find({"type":"node"}).count()
```

Number of ways: 415680

```
db.getCollection('map').find({"type":"way"}).count()
```

Number of unique users: 1282

```
db.getCollection('map').distinct("created.user").length
```

Top 3 contributors:

```
{
    "_id" : "maheshrkm",
    "count" : 164818.0
}
// ------------------------------------------
{
    "_id" : "PlaneMad",
    "count" : 102285.0
}
// ------------------------------------------
{
    "_id" : "venkatkotha",
    "count" : 93227.0
}
// ------------------------------------------
```

```
db.map.aggregate([

{$match: {'created.user': {$exists: 1} }},
 {$group: {_id: '$created.user','count':{$sum: 1}}},
 {$sort: {'count': -1}},
 {$limit: 3}
])
```

Top 10 amenities:

_id	count
place_of_worship	704.0
school	536.0
restaurant	424.0
atm	267.0
hospital	241.0
bank	240.0
college	189.0
fuel	147.0
pharmacy	145.0
bus_station	111.0

```
db.map.aggregate([

{$match: {'amenity': {$exists: 1} }},
 {$group: {_id: '$amenity', count: {$sum: 1}}},
 {$sort: {'count': -1}},
 {$limit: 10}
])
```

Top 10 restaurants:

Adyar Ananda Bhavan	21.0
Adyar Anand Bhavan	9.0
Hotel Saravana Bhavan	8.0
Hot Chips	5.0
KFC	5.0
Murugan Idli Shop	4.0
Subway	3.0
Bombay Halwa House	3.0
Wangs Kitchen	3.0

```
db.map.aggregate([

{$match: {'amenity': "restaurant"}},
 {$group: {_id: '$name', count: {$sum: 1}}},
 {$sort: {'count': -1}},
 {$limit: 10}
])
```

Top 3 Banks:

State Bank of India	28.0
HDFC Bank	18.0
Indian Overseas Bank	16.0

```
db.map.aggregate([

{$match: {'amenity': "bank"}},
 {$group: {_id: '$name', count: {$sum: 1}}},
 {$sort: {'count': -1}},
 {$limit: 3}
])
```

Other ideas about the dataset:

The open street maps data is missing newly opened buildings and amenities that I know of. The city of Chennai has developed a lot lately with the opening of the metro rail system and this data lacks station information of the metro train. Also, there is no information of the new malls and coffee places. One

way of improving the content is by encouraging the people of Chennai to contribute to the open data when they do a social media check in. After all it is their city data and one should take pride in contributing to it as a result of which will be beneficial of analysts using this open data from bringing in good changes and improvements in the city system.

One of the anticipated problems in such update of information is when people try to update neighboring locations instead of the current location they are checking in from. The accuracy and validity have to be ensured and taken charge by the contributor.

Code File:

project.py