Singapore OpenStreetMap data wrangling with MongoDb

1. Problem in Data set:

There are three problems encountered in the data set, they are as follows

- Abbreviated Street Names and street names ends with number
- Names given in different Languages
- Cities from nearby Singapore are included

Abbreviated Street names and street names ends with number:

The street names given in the XML data are abbreviated such as St, Rd, Dr and Ave are used instead of Street, Road, Drive and Avenue respectively. Furthermore Jl and Jln are used to represent word Jalan, Jalan in malay means street, for instance Jl Asoka is used instead of Jalan Asoka. Unlike other common street names, Jalan will be present at the start of the street name. While creating json file I replaced all the abbreviations with their full form. Some street names ended with word Cresent, which was replaced with word Crescent.

In addition some street names ends with number and I need to look for word before that number for street name auditing. For instance street name is give as Jurong west street 91, I will ignore 91 while auditing street names.

Names given in different Languages

Since Singapore is a country with multiple official languages, I was expecting names gives in different languages. When I audited data I found names are given in English, Mandarin and sometimes in Tamil and other languages. While creating json file I took only English names and names in other languages are omitted.

Cities from nearby Singapore are included

In Singapore all the postal codes should have six digits, but I found some postal codes had differ number of digits. I found this problem during data auditing phase using python. When I investigated further I found Singapore openstreet map contains data from nearby cities such as Bintan

(Indonesia), Batam(Indonesia), Johor Bahru (Malaysia), etc. Especially Johor Bahru had around 13000 entries compared to 10000 entries in Singapore.

Moreover city fields also contained area name such as Holland Village, Sembawang, Ang Mo Kio, Changi Village etc. These area values should be replaced with Singapore.

#sort cities count in descending order

>{"$match":{"address.city":{"$exists":1}}},{"$group":{"_id":"$address.city","count":{"$sum":1}}},{"$sort":{"count":-1}}

2. Overview of the Data

In this section I have discussed briefly about size of the openstreet data and basic statistics about the data using MongoDB queries

Size of the Data:

XML File size = 218MB

Json File size = 431MB

Statistics of the Data:

#total number of nodes

> db_query(db,{"type":"node"}).count()

1024434

#total number of way

> db_query(db,{"type":"way"}).count()

149684

#total number of distinct users

>len(db.city.distinct("created.user"))

1349

Top contributor

>{"$group":{"_id":"$created.user","count":{"$sum":1}}},{"$sort":{"count":-1}},{"$limit":1}

u'_id': u'JaLooNz', u'count': 316746

Users with single contribution

>{"$group":{"_id":"$created.user","count":{"$sum":1}}},{"$group":{"_id":"$count","number":{"$sum":1}}},{"$sort":{"_id": 1}},{"$limit":1}

u'_id': 1, u'number': 312

3. Additional Ideas

Statistics of the User contribution

Statistics of the User contribution is quiet skewed, with top user contributing nearly 26.97% and top 5 users contributing 53.107%. Top 10 users contributing 64.5% and top 50 users contributing 87%. There are 312 users with single contribution to this open street map data. This clearly shows only some users were actively contributing to the openstreet map data. Code for detailed analysis is available in pymongo at MongoDbClient.py

Contribution to the map data can be improved by giving some incentives to top contributor every month such as giving coupons to popular local shops. The problem with this method is that only few top users will be keep contributing and many others may not aware that they can contribute. This will again skew user contribution.

Another method is to give incentives to referral, each user can be given referral code and he can invite his friends to contribute using his referral code. The top people who referred the most can be given incentives in the form coupons or by running their name in open street map website. By this method more people will start contributing as well as create awareness about openstreet map data.

#find contribution of top n users based on value of n

```
> {"$group":{"_id":"$created.user","count":{"$sum":1}}},
   {"$sort":{"count":-1}},{"$limit":n}
```

Analysis on Restaurants and Fast food

Chinese cuisine is popular and there is no surprise in this as Singapore population is dominated by Chinese ethnic people. There are around 103 Chinese restaurants, followed by Italian and Korean with count of 41 and 35 respectively.

#restaurants in descending order based on their cuisine

```
>{"$match":{"$and"
:[{"amenity":"restaurant"},{"cuisine":{"$exists":1}}]}},
 {"$group":{"_id":"$cuisine","count":{"$sum":1}}},
  {"$sort":{"count":-1}}
```

Fast food category is dominated by western fast food chains such as McDonald, KFC and Burger King.

fast food chain count in descending order

```
>{"$match":{"$and" :[{"amenity":"fast_food"},{"name":{"$exists":1}}]}},
   {"$group":{"_id":"$name","count":{"$sum":1}}},
   {"$sort":{"count":-1}}
```

Analysis on Place of Worship

I was quiet surprised to find there are about 486 place of worship for Muslims compared to 195 for Christian and 74 for Buddhist. Since Singapore has Buddhism as major religion I was expecting more Buddhist place of worship. One possible reason for this surprise finding is that this data also includes nearby cities from Malaysia, Indonesia (both these countries has islam as their major religion). Code for the analysis is available in pymongo at MongoDbClient.py.

religion based on number of place of worrship in descending order

```
>{"$match":{"amenity":"place_of_worship"}},
  {"$group":{"_id":"$religion","count":{"$sum":1}}},
   {"$sort":{"count":-1}}
```

Conclusion

Openstreet map data for Singapore needs some major modification, as it also includes details of nearby cities. Furthermore certain critical values such as city name are provided with wrong values. At the current state openstreet map data for Singapore has to be cleaned carefully before doing any major analysis on it.