OpenStreetMap Data Wrangling with MongoDB of Manila, Philippines

Map Area: Manila, Philippines

Problems Encountered in the Map

Since the data is entered by users, there are various problems with entry and formatting.

Street Names

Once example is "**De Venecia St.**" Some streets end in "Street, St., Ave., Ave, etc. I had to do some formatting so the street would be updated to "**De Venecia Street**."

Many Filipinos love to use abbreviations in normal conversation. This is similar to using LOL (laughing out loud) in social media.

Postal Codes

The postal codes in Philippines are in a four digit format. However, there were 3 problem postal codes "NCR 1007", "58001" and "`12001". These are all in Manila so they should just be "1007", "5800" and "1200" so I updated the postal code formatting to check for these cases and fix.

Abbreviations

Some of the abbreviations of the amenities are self-explanatory like "bbq" (Barbeque) & "atm" (Automated teller machine).

However, it's difficult to tell what some of the <tag = k values are for example in the following (with no further information).
Ele
Fee
Hgv
Lcn

Incomplete data & some potential improvements

For a population of Manila, 1.652 million (2010) according to US Census Bureau, the number of nodes & ways seems low. I compare to the sample report of Charlotte, NC and their population of 792,862 (2013) according to US Census Bureau. The charlotte data had just over 1.5 million documents compared to almost 400,000 for the Manila dataset. I'm not sure how recent this data is and perhaps in time, more data will contributed.

I notice that Openstreetmaps has facebook & twitter pages. I am a big facebook user and so are people in Philippines. Additional, technology is growing fast there since many companies are moving their call centers there. However, I have never heard of Openstreetmaps until this project. Perhaps, there can be better marketing through social media. Openstreetmap can work with those social media giants to

get the word out about Openstreetmaps. Then people will start to add more data to the dataset. An online form that forced more complete data could help drive usage.

According to Openstreetmaps, there are one million mappers. According to the following article, http://geoawesomeness.com/the-us-mobile-app-report-google-maps-app-64-5m-users-apple-maps-42m/
many map applications have multi-millions of users. It's not a fair comparison but it just gives an idea of the usage. If Openstreetmaps usage increases or explodes even a little, the amount of data auditing and cleaning tasks could be immense before any analysis could be done.

For the most part, I was able to pull interesting data statistics on the map of my mom's hometown.

Data Overview

This section contains basic statistics about the dataset and the MongoDB queries used to gather them.

File sizes

manila.osm . . . . 71 MB
manila.osm.json . . . 107 MB

Number of documents

```
db.manila_main.find().count()
385104
```

Number of nodes

```
db.manila_main.find({"type" : "node"}).count()
314713
```

Number of ways

```
db.manila_main.find({"type" : "way"}).count()
70361
```

Number of unique users

```
len(db.manila_main.distinct("created.user"))
754
```

Top 1 contributing user

```
unique_user = [{"$group": {"_id":"$created.user","count":{"$sum":1} } },
     {"$sort":{"count":-1} },
     {"$limit": 10}]
Unique user
[{u'_id': u'jmbangate', u'count': 284865}
```

Number of users appearing only once (having 1 post)

```
once_user = [{"$group": {"_id":"$created.user","count":{"$sum":1} } },
     {"$group": {"_id":"$count","num_users":{"$sum":1} } },
     {"$sort":{"count":-1} },
     {"$limit": 10}]
```

```
One-time user
[{u'_id': 156, u'num_users': 1},
 {u'_id': 42, u'num_users': 1},
 {u'_id': 58, u'num_users': 1}
```

…

Data Exploration

I was particular interested the following data. Filipinos love to eat and it's a big part of the culture. When you go to a house, the welcoming question is not "How are you?" It's "Did you eat already?" So I was not surprised to see restaurant as the top amenity.

Top 10 amenities

```
amenity = [{ "$match" : { "amenity" : {"$exists" : 1 } } },
      { "$group" : { "_id" : "$amenity", "count" : { "$sum" : 1 } } },
      { "$sort" : { "count" : -1 } },
      { "$limit" : 10}]
```

```
 [{u'_id': u'restaurant', u'count': 595},
 {u'_id': u'bank', u'count': 473},
 {u'_id': u'fast_food', u'count': 436},
 {u'_id': u'school', u'count': 267},
 {u'_id': u'parking', u'count': 265},
 {u'_id': u'place_of_worship', u'count': 182},
 {u'_id': u'cafe', u'count': 145},
 {u'_id': u'pharmacy', u'count': 135},
 {u'_id': u'fuel', u'count': 112},
 {u'_id': u'bar', u'count': 89}]
```

Since the country is predominantly a Christian county, I expected this result.

Top 10 religions

```
religion = [{ "$match" : {"amenity" : {"$exists" : 1}, "amenity" : "place_of_worship" } },
       { "$group" : { "_id" : "$religion", "count" : { "$sum" : 1 } } },
       { "$sort" : {"count" : -1 } },
       {"$limit":10 }]
```

```
 [{u'_id': u'christian', u'count': 144},
 {u'_id': None, u'count': 28},
 {u'_id': u'buddhist', u'count': 3},
 {u'_id': u'muslim', u'count': 3},
 {u'_id': u'taoist', u'count': 2},
 {u'_id': u'hindu', u'count': 1},
 {u'_id': u'jewish', u'count': 1}]
```

I wanted to take another look at the eating culture. Looking at the data, they were a lot of restaurants or fast food places that had no further information on cuisine. I was not surprised about the burger joints because they love McDonalds, Burger King and other local burger restaurants like Jolibee. I didn't realize there were a lot of Chinese restaurants but it's not surprising. A lot of Filipinos would prefer to eat Filipino food at home.

Top 10 restaurants

```
restaurant = [{ "$match" : { "amenity" : {"$exists" : 1 },
                "amenity" : { "$in" : ["fast_food", "restaurant"] } } },
        { "$group" : { "_id" : "$cuisine", "count" : { "$sum" : 1 } } },
        { "$sort" : { "count" : -1 } },
        { "$limit" : 10}]
```

```
[{u'_id': None, u'count': 554},
{u'_id': u'burger', u'count': 78},
{u'_id': u'chinese', u'count': 66},
{u'_id': u'filipino', u'count': 43},
{u'_id': u'asian', u'count': 36},
{u'_id': u'chicken', u'count': 36},
{u'_id': u'pizza', u'count': 35},
{u'_id': u'japanese', u'count': 23},
{u'_id': u'korean', u'count': 16},
{u'_id': u'regional', u'count': 11}]
```

It's good to see there are a lot of parks and pitches (fields). Since the people love to eat, at least there are places to work off the food.

Top 10 leisure locations

```
leisure = [{ "$match" : { "leisure" : {"$exists" :1 } } },
      { "$group" : { "_id" : "$leisure", "count" : { "$sum" : 1 } } },
      { "$sort" : { "count" : -1 } },
      { "$limit" : 10}]
```

```
[{u'_id': u'pitch', u'count': 111},
{u'_id': u'swimming_pool', u'count': 94},
{u'_id': u'park', u'count': 67},
{u'_id': u'sports_centre', u'count': 39},
{u'_id': u'garden', u'count': 36},
{u'_id': u'playground', u'count': 12},
{u'_id': u'fitness_centre', u'count': 11},
{u'_id': u'golf_course', u'count': 7},
{u'_id': u'stadium', u'count': 6},
{u'_id': u'track', u'count': 5}]
```

One of the favorite sports in the Philippines is basketball because of the popularity of the PBA (Philippine Basketball Association). It's so hot and tropical there and many people play basketball in their slippers!

Top 10 sport locations

```
sport = [{ "$match" : { "sport" : {"$exists" :1 } } },
      { "$group" : { "_id" : "$sport", "count" : { "$sum" : 1 } } },
      { "$sort" : { "count" : -1 } },
      { "$limit" : 10}]
```

```
[{u'_id': u'basketball', u'count': 69},
{u'_id': u'tennis', u'count': 22},
{u'_id': u'swimming', u'count': 16},
{u'_id': u'multi', u'count': 15},
{u'_id': u'golf', u'count': 10},
{u'_id': u'soccer', u'count': 9},
{u'_id': u'badminton', u'count': 4},
{u'_id': u'pool', u'count': 4},
{u'_id': u'bowling', u'count': 4},
{u'_id': u'skateboard', u'count': 3}]
```

Conclusion

So, we saw with the Manila dataset that there are problems with the data just like any data that relies on human editing and entering. It's also incomplete but cleaned enough to be able to do basic statistics reporting. As I stated, since this is my mother's hometown and I've visited twice before, I had some knowledge of the culture. For the most part, I was not surprised from the food culture but it was extremely interesting to dig into the data.