Problem Set 3

The map I used for this project is the map of Hawaii islands.

I downloaded from Geofabrik

http://download.geofabrik.de/north-america/us/hawaii.html

1. Data Wrangling Proces

After initial running some initial inquiries to the dataset. There are several issues with the data.

1. Abbreviation of road names
2. Use of indigenous language in street names.
3. Zipcode with dash and state code.
4. K values with colon other than addresses.

1. Abbreviation of road names:

   I ran a query against the dataset to check the last word of the street names. I also ran another query against the dataset to check what are the names that ended with unexpected last word.

```
context=ET.iterparse(filename)
streettype=dict()
oddstreetname=set()
for event, elem in context:
        if elem.tag=="node" or elem.tag=="way":
            for tag in elem.iter("tag"):
                if tag.attrib['k']=="addr:street":
                    lastword=tag.attrib['v'].split()[-1]
                    if lastword not in expected:
                            oddstreetname.add(tag.attrib['v'])
                    if lastword in streettype:
                            streettype[lastword]=streettype[lastword]+1
                    else:
                            streettype[lastword]=1
```

   It turned out that there are many streets that are ended with something like “St.”, “Blvd”, “Av” etc. I solved this problem by substituting the last

word of the street name into its complete form.

2. Use of indigenous language:
   There are multiple street names that are not English like “Ala Ike” and “Ala Kalanikaumaka”. I did not do any adjustments to them.

3. Zip code with dash and state code.
   There are 17 problematic zip codes with dash and state code, such as “96712-999” and “HI 96732”. I changed the one with dash to only five digits during the auditing period. I deleted the one started with HI after uploaded the data to MongoDB using the following command:
   db.openstreet.update({"address.postcode":{"$regex":"HI"}},{"$unset":{"address.postcode":""}}, multi=True)
   Other than this issue, the zip codes are pretty clean. All the zip codes fall within the range of 96701 and 96898, which is the range of Hawaiian zip code.

4. K values with colon other than addresses:

```
tag=dict()
context=ET.iterparse(filename)
for event, elem in context:
    if elem.tag=="node" or elem.tag=="way":
        for el in elem.iter("tag"):
            if el.attrib['k'] not in tag:
                tag[el.attrib['k']]=1
            else:
                tag[el.attrib['k']]=tag[el.attrib['k']]+1
```

   I also ran a query against all the k values in the tags. There are 3 values set that are not addresses, such as “telescope:diameter”, “telescope:spectrum”, “telescope:type”. I decided to add 3 additional dictionaries (telescope, roof and building_details) into the Json file.

2. Data Analysis

The followings are data overview and some interesting findings that I found using MongoDB.

1. File size:
   Hawaii-latest.osm:222.8MB
   Hawaii-latest.osm.json:217.4MB

2. Number of Nodes and Ways:
   db.opens.find({“type”:”node”}).count()
   db.opens.find({“type”:”way”}).count()
   There are 1088442 nodes and 80175 ways.

3. Number of Documents:
   db.opens.find().count()
   1168617

4. Number of Unique Users:
   users=db.openstreet.distinct("created.user")
   len(users)
   866

5. The List of User Contribution:
   most=db.openstreet.aggregate([{"$group":{"_id":"$created.user","count":{"$sum":1}}}, {"$sort":{"count":1}}])
   The user that contributed the most is Tom_Holland. He contributed 451057 times. The second one is ksamples with the contribution of 147724.
   The top 4 contributors contributed 72.57% of the entries, which is surprisingly similar to the wealth distribution of the world.

6. Numbers of Restaurants Providing Different Cuisines:

number=db.opens.find({"amenity":"restaurant"}).count()

restaurants=db.opens.aggregate([{"$match":{"amenity":"restaurant", "cuisine":{"$exists":1}}}, {"$group":{"_id":"$cuisine", "count":{"$sum":1}}}, {"$sort":{"count":1}}])

There are total 488 restaurants on the map.

The most common cuisine is “regional”. There are 36 of them. The second and third most common are “pizza” and “thai”, which has 19 and 17 restaurants respectively. According the result, we can tell that the restaurants at Hawaii are fairly diverse.

7. Incompleteness of Restaurant Data:

   The list of restaurants on the map is incomplete. There are 16 Japanese restaurants in the dataset, but according to Trip Advisor there are at least 53 Japanese restaurants. There are also 11 Vietnamese restaurants on Trip Advisor, but there is only 1 in the dataset.

8. Number of Restaurants in the Same Zip Code:

   restaurants=db.opens.aggregate([{"$match":{"amenity":"restaurant", "address.postcode":{"$exists":1}}}, {"$group":{"_id":"$address.postcode", "count":{"$sum":1}}}, {"$sort":{"count":1}}])

   There are 20 restaurants at the area code 96815, 6 restaurants at 96714. Zip code 96815 is at the Waikiki Bay of Honolulu, which is the downtown area. Area code 96714 is Hanalei Bay of Kauai Island, which is a small touristy town.

3. Additional Ideas:

1.Improvement of Coverage:

After running some queries against the data set it is quite obvious that the Hawaii map is incomplete and the sites are highly concentrated in Honolulu around Waikiki Bay. 28% of the sites are concentrated in single zip code area (96815, Waikiki Bay area). Top 5 zip codes account for 58.2% of the sites on the map. Those 5 zip codes are all around Waikiki

Bay.

The reason why the map is so highly concentrated in a single area might be because of the lack of diverse contributors and the scattering islands of Hawaii. As stated in the previous sections, top 4 contributors contributed 72.57% of the entire data. For a place with so many islands, it is hard to cover every corner by only a few people.

It might be a good idea to use a ranking system to motivate people to update the information. It can also collaborate with local stores to update their own information as a way to promote their businesses.

2.Address should be Mandatory:

There are many sites that do not have an address. As the result, it is very hard for users to locate where the locations are. But in some areas in the world, especially in some developing countries where address system is not in place, it will be hard to implement this improvement.

3.Other Interesting Findings:

There are 22 telescopes in Hawaii. Those telescopes are owned by research institutes from all over the world (ex, US, Japan, Canada, France etc.)

db.opens.find({"telescope":{"$exists":1}})

The most common building color is beige (33%).

num=db.opens.find({"building_details.color":{"$exists":1}}).count()

mat=db.opens.aggregate([{"$match":{"building_details.color":{"$exists":1}}}, {"$group":{"_id":"$building_details.color", "count":{"$sum":1}}}])