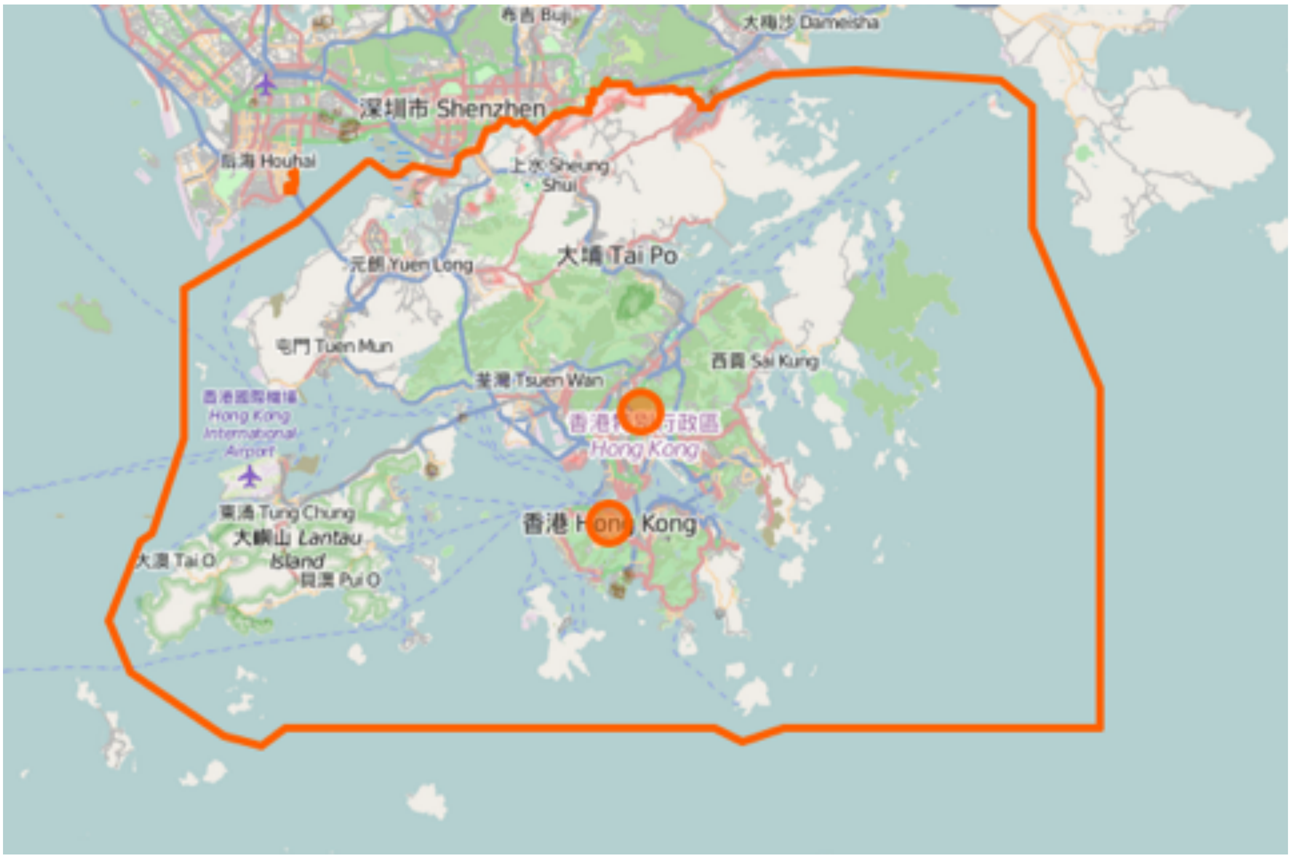
Data Wrangling Project

OpenStreetMap Sample Project
Data Wrangling with MongoDB

Lin Rui

Map Area: Hong Kong, China



https://www.openstreetmap.org/relation/913110

1. Problems Encountered in the Map

After downloading the osm data of the Hong Kong area and running it against sample_hongkong.py script, I got a smaller file sample_hongkong.osm. By checking this osm file, I notice two main problems with the data:

- Wrong located elements (some elements from another two China cities Shenzhen and Macau improperly appear in this osm file)
- Over-abbreviated street names (Taikoo Shing Rd)

Wrong located elements

Shenzhen and Macau are another two China cities which is adjacent to Hong Kong. I found some elements in the osm file from Hongkong area are actual in Shenzhen or Macau instead. For example, places like *Shenzhen Bao'an International Airport, Shenzhen Meteorological Bureau,* and *Macau International Airport* are apparently not locate in Hong kong.

I filter all these Shenzhen elements by using regular expression functions in my Python script clean_to_json.py.
The major part of codes like this:

```
wrong_el_re = re.compile(u'深圳|Shen Zhen|澳門|Macau', re.IGNORECASE)
list_wrong = set()
with codecs.open(filein, 'r') as fin:
    for _ , el in ET.iterparse(fin):
        if el.find("tag") != None:
            sub_tag = el.find("tag[@k='name']")
            if sub_tag != None:
                m = wrong_el_re.search(sub_tag.attrib['v'])
                if m != None:
                    list_wrong.add((el.tag, sub_tag.attrib['v']))
```

Then I just drop these elements when generating the JSON file.

Over-abbreviated street names

When I list all the highway types by running the Python script clean_to_json.py, I found some over-abbreveated street names, like “Rd”, “St”, “str”, “St.”, “Ave”, “Ave.”, “Bd”, and then I update all this problematic substrings.

2. Data Overview

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

File sizes

HongKong.osm …... 271.1 MB
Hongkong-clean.json …... 439.2 MB

```
# Number of documents
> db.hongkong.find().count()
1379491

#  Number of nodes
> db.hongkong.find({"type": "node"}).count()
1243325

#  Number of ways
> db.hongkong.find({"type": "way"}).count()
131027
```

```
#  Number of relations
> db.hongkong.find({"type": "relation"}).count()
5139

#  Number of unique users
> db.hongkong.distinct("created.user").length
911

# Top 1 contributing user
> db.hongkong.aggregate([{"$group": {"_id":"$created.user", "count":
{"$sum":1}}},{"$sort": {"count": -1}}, {"$limit":1}])
{ "_id" : "hlaw", "count" : 521440 }

# Number of users appearing only once (having 1 post)
> db.hongkong.aggregate([{"$group":{"_id":"$created.user", "count":
{"$sum":1}}}, {"$group":{"_id":"$count", "num_users":{"$sum":1}}},
{"$sort":{"_id":1}}, {"$limit":1}])
{ "_id" : 1, "num_users" : 170 }
```

3. Additional Ideas

English and Chinese Names of the elements

Hong Kong is an international metropolis whose official languages are Chinese and English. So every named elements should at least have both Chinese and English names, and having more other language names is appreciated. Here are some numbers of named elements:

```
# Number of elements having an English name

> db.hongkong.find({"tags.k":"name:en"}).count()
302418

# Number of elements having an Chinese name
> db.hongkong.find({"tags.k":"name:zh"}).count()
308270

# Number of elements having both Chinese and English names
> db.hongkong.aggregate([
...                         {"$match":{"tags.k":"name:en"}},
...                         {"$match":{"tags.k":"name:zh"}},
...                         {"$group":{"_id":null, "count":{"$sum":1}}}
...                       ])
{ "_id" : null, "count" : 279686 }
```

As we can see above, some of the named elements are lack of either Chinese name or English name. I am thinking about extracting these name-missing elements, then post them on the web and encourage volunteers to fill in the missing part. Since it’s just a translating work, it is easy for the volunteers to do it.

Additional data exploration using MongoDB queries

```
# Number of public-transport facilities, grouped by ‘node’, ‘way’ and
‘relationship’
> db.hongkong.aggregate([{"$match": {"tags":{"$elemMatch": {"k":
"public_transport"}}}},{"$group":{"_id":"$type", "count":{"$sum":1}}},
{"$sort":{"_id":1}}])
{ "_id" : "node", "count" : 16749 }
{ "_id" : "relation", "count" : 5 }
{ "_id" : "way", "count" : 377 }

# Number of tourism facilities
> db.hongkong.find({"tags":{"$elemMatch": {"k": "tourism"}}}).count()
51113

# Top 10 tourism facilities
> db.hongkong.aggregate([
...                         {"$match": {"tags.k": "tourism"}},
...                         {"$unwind": "$tags"},
...                         {"$match":{"tags.k": "tourism"}},
...                         {"$group":
...                            {"_id":"$tags.v", "count":{"$sum":1}}
...                         },
...                         {"$sort":{"count":-1}},
...                         {"$limit": 10}
...                       ])
{ "_id" : "information", "count" : 23471 }
{ "_id" : "viewpoint", "count" : 9479 }
{ "_id" : "hotel", "count" : 7134 }
{ "_id" : "attraction", "count" : 3017 }
{ "_id" : "artwork", "count" : 2330 }
{ "_id" : "picnic_site", "count" : 1642 }
{ "_id" : "zoo", "count" : 1269 }
{ "_id" : "camp_site", "count" : 1054 }
{ "_id" : "hostel", "count" : 862 }
{ "_id" : "guest_house", "count" : 595 }
```

Conclusion

Since Hong Kong is an international metropolis which attracts many foreign visitors every year, it’s important to correctly name the elements in multilingual way, especially for those public-transport facilities and tourism facilities. It’s possible to invent a mechanism to attracts the map viewers to do a little work to help doing the multilingual naming work each time they browse the Hong Kong map, especially for Hong Kong residents.