OpenStreetMap Project
Data Wrangling with MongoDB
Map Area: Hong Kong Island
https://www.openstreetmap.org/export#map=12/22.2467/114.2092&layers=T

1. Problems Encountered in the Map

(1) Different expression way of "network"

When I tried to count the number of network points, I found that the values of some network were "MTR", while others were "港鐵 MTR ". Therefore I updated all value to "MTR", subtracting Chinese characters from the values.

(2) Multiple languages for one address

Some addresses have expression in English, Chinese and Vietnamese, e.g.

```
<tag k="name:en" v="Central"/>
<tag k="name:vi" v="Trung Hoàn"/>
<tag k="name:zh" v="中環"/>
```

while some have expression in English and Chinese, e.g.

```
<tag k="name:en" v="Admiralty Station"/>
<tag k="name:zh" v="金鐘站"/>
```

while others only have English expression, e.g.

```
<tag k="name:en" v="Ramada Kowloon"/>
```

I ignored all the expression in Non-English expression address, and left noly one expression in English for every address.

(3) Different expression ways for the same thing

e.g.: taoism - taoist
buddhism - Buddhist
McDonald's - McDonalds

I changed "taoist" into "taoism", and changed "buddhist" into "buddhism", and changed "McDonalds" into "McDonald's".

2. Data overview

(1) File Sizes

map_HK.osm 80.7MB
map_HK.osm.json 105.1MB

(2) number of tags:

{'bounds': 1, 'member': 32039, 'meta': 1, 'nd': 423245, 'node': 357046, 'note': 1, 'osm': 1, 'relation': 2529, 'tag': 189949, 'way': 41717}

(3) Number of unique users

364

(4) Number of documents

- `db.hk.find().count()`

- 398763

(5) Number of nodes

- db.hk.find({"type":"node"}).count()
- 357036

(6) Number of ways

- db.hk.find({"type":"way"}).count()
- 41717

(7) Number of guideposts for tourism

- db.hk. find({"information":"guidepost"}).count()
- 166

(8) Number of supermarkets

- db.hk. find({"shop":"supermarket"}).count()
- 106

(9) Number of McDonald's

- db.hk. find({"name":"McDonald's"}).count()
- 41

(10) Top 1 contributing user

- ```
  db.hk.aggregate(
  [{"$group": {"_id": "$created.user",
  "count": {"$sum": 1}}},
  {"$sort": {"count": -1}},
  {"$limit":1}]
  )
  ```
- [{u'_id': u'hlaw', u'count': 231573}]

(11) Number of users appearing only once (having 1 post)

- ```
  db.hk.aggregate(
  [{"$group": {"_id": "$created.user",
                "count": {"$sum": 1}}},
                {"$group":{"_id":"$count",
                "num_users":{"$sum":1}}},
                {"$sort": {"_id": 1}},
                {"$limit":1}]
  )
  ```
- [{u'_id': 1, u'num_users': 97}]
- # "_id" represents postcount

3. Additional Ideas

(1) Contributor statistics

- Top user contribution percentage ("hlaw") – 58.07%
- Combined top 2 users' contribution ("hlaw" and "KX675") – 67.48%
- Combined Top 10 users contribution – 88.05%
- Combined number of users making up only 1% of posts - 192

(about 52.74% of all users)

According to the user percentage, the top user contributes, the same as the sample project, more than half of total documents. And the second top user makes much less contribution than the top one. About half of all the users make up only 1% of the post. From the result, I guess maybe the outstanding top user("hlaw") may be a professional group in charge of Hong Kong map maintenance, which is the reason why it takes such a large percentage.

(2) Improvement suggestions

- Improvement on multi-lingual entries

  As I mentioned before, there are different languages for different addresses. I count the numbers of English names, Chinese names and Vietnamese names. They are 3723, 3443, and 55, respectively. Obviously English and Chinese are most commonly used in Hong Kong, while Vietnamese occupies more than 1%. Considering the merit of multi-lingual entries, it is worth to make every address show all the three languages using Google translate. The possible problem could be inaccuracy of translation. It may required people to check manually.

- Improvement on recycling sites

  According to my statistic on amenities, "recycling" is among top 10. When I checked the details in every recycling sites, I found that in some recycling sites there were information about what kind of items could be recycled here, e.g:

  ```
  <tag k="amenity" v="recycling"/>
   <tag k="recycling:cartons" v="yes"/>
  <tag k="recycling:magazines" v="yes"/>
  <tag k="recycling:newspaper" v="yes"/>
  <tag k="recycling:paper" v="yes"/>
  <tag k="recycling:paper_packaging" v="yes"/>
  ```

  However, most recycling sites only had name information and were lack of recycling items information. E.g:

  ```
  <tag k="amenity" v="recycling"/>
  <tag k="building" v="yes"/>
  <tag k="building:levels" v="1"/>
  <tag k="name" v="基業街垃圾收集站 Kei Yip Street Refuse Collection Point"/>
  <tag k="name:en" v="Kei Yip Street Refuse Collection Point"/>
  <tag k="name:zh" v="基業街垃圾收集站"/>
   <tag k="recycling_type" v="centre"/>
  ```

  I think the information of recycling materials is very useful for both residents and recycling company. Residents know where to throw

recycling materials, and recycling companies know where to get what they need. This information could be extracted at the website Environmental Protection Apartment of Hong Kong.

(3) Additional data exploration using MongoDB queries

- Top 10 appearing amenities (ignore"None")

```
db.hk.aggregate([
{"$match":{"amenity":{"$exists":1}}},
{"$group":{"_id":"$amenity","count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":10}
])
```

{u'_id': u'parking', u'count': 77},
{u'_id': u'place_of_worship', u'count': 44},
{u'_id': u'community_centre', u'count': 35},
{u'_id': u'public_building', u'count': 30},
{u'_id': u'toilets', u'count': 30},
{u'_id': u'school', u'count': 23},
{u'_id': u'marketplace', u'count': 23},
{u'_id': u'recycling', u'count': 19},
{u'_id': u'ferry_terminal', u'count': 15},
{u'_id': u'taxi', u'count': 88}]}

- Top 4 religions

```
db.hk.aggregate([
{"$match":{"amenity":{"$exists":1},            "amenity":
"place_of_worship"}},
{"$group":{"_id":"$religion",
"count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":4}
])
```

[{u'_id': u'christian', u'count': 76},
{u'_id': u'taoism', u'count': 26},
{u'_id': u'buddhist', u'count': 8},
{u'_id': u'muslim', u'count': 3}]

From the result above, I can tell that people in Hong Kong have multiple religions, among which Christian makes up most.

- Top 4 banks

```
db.hk.aggregate([
{"$match":{"amenity":{"$exists":1}, "amenity": "bank"}},
{"$group":{"_id":"$name",
"count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":5}
```

```
])
```

```
[{u'_id': u'HSBC', u'count': 26},
{u'_id': u'BEA', u'count': 13},
{u'_id': u'BOC', u'count': 12},
{u'_id': u'Standard Chartered', u'count': 11},
{u'_id': u'Hang Seng Bank', u'count': 11},
```

- Most popular fast food

```
db.hk.aggregate([
{"$match":{"amenity":{"$exists":1},              “amenity”:
“fast_food”}},
{"$group":{"_id":"$name",
"count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":5}
])
```

```
[{u'_id': u"McDonald's", u'count': 41},
{u'_id': u'KFC', u'count': 6},
{u'_id': u'Maxim MX', u'count': 5},
{u'_id': u'Burger King', u'count': 5},
{u'_id': u'Yoshinoya', u'count': 4},
```

4. conclusion

After this review of the data, I found that OpenStreetMap in Hong Kong has been relatively well maintained. There is rarely different expression for address words like “street”, “road”, etc. However, it does exist small differences in words like “McDonald’s” and “Taoism”. They are too tedious if I try to fix everyone of them at the beginning. But when I made statistics for detail information, the flaws stood out and I can fix them and got exact result I was looking for.

Besides, my HK map file is only about 100MB, while producing the json file took me almost ten minutes. If I want to deal with bigger data, I think map reduce must be necessary to increase the program efficiency.