OpenStreetMap Data Case Study

MAP area:
Taipei, Taiwan
https://drive.google.com/open?id=0B2B1W8xYjS33RUFqc1JXQ1BPT0U
The file is 145 MB
Taipei is my hometown, I was born here so I'm interested to see what will the data be and what improvement could I make!

Problems Encountered in the Map

After play around of my dataset, I found some problems:

- different city street names(It should only contain Taipei city, but I see many other city inside)
- Inconsistent postal codes("226", "22606")
- Inconsistent street name("台北市中山區玉門街 1 號","迪化街 116 號". First one also contains the city name and district name, second one only contains street name)

Inconsistent street name and different city street names:
This problem makes the map inconsistent, so I wrote a function, try to replace abbreviations with the full name to standardize.

```
def update_name(street_name):
    wrong_street = [u"台北市",u"南港區",u"大安區",u"中山區",u"大同區",u"淡水區",u"淡水区",u"內湖區"]
    for wrong in wrong_street:
        street_name = re.sub(wrong,"", street_name)
    return street_name
```

Here are my results:

```
台北市內湖區陽光街 383 號 => 陽光街 383 號
台北市內湖區瑞湖街 80 號 => 瑞湖街 80 號
台北市內湖區瑞湖街 80 號 => 瑞湖街 80 號
```

Inconsistent postal code:
I found that there are several postal code contains more than 3 digit number. To make the map more consistent, I wrote a function to fix the number to only 3 digit number.

```
def update_post(post):
    if len(post)>3:
        postal_code = post[:3]
        return postal_code
    else:
        return post
```

Here are my results:

```
10056 => 100
11675 => 116
11071 => 110
```

Update_name function return the clean street name without city name and district name.
Update_post function return the postal code only contain 3 number

Data overview and Additional idea

- **number of nodes**

```
QUERY = """
SELECT COUNT(*)
FROM nodes
"""
```

```
665642
```

- **number of ways**

```
QUERY = """
SELECT COUNT(*)
FROM ways
"""
```

```
85815
```

- **To check which postcode is most popular in taipei city map**

```
QUERY = """
SELECT tags.value, COUNT(*) as count
FROM (SELECT * FROM nodes_tags
    UNION ALL
    SELECT * FROM ways_tags) tags
WHERE tags.key = 'postcode'
GROUP BY tags.value
ORDER BY count DESC
LIMIT 10;
```

```
[(u'106', 340), (u'104', 149), (u'220', 142), (u'114', 107), (u'238', 87),
(u'100', 78), (u'111', 75), (u'231', 71), (u'112', 69), (u'116', 66)]
```

106 is the post code of Da'an district, where my university located. Da'an also offers some of Taipei's most expensive residential real estate. And it has the highest population in Taipei city.

- **To check how many unique user contribute to the map**

```
QUERY = """
SELECT  COUNT (DISTINCT nodes.user)
FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) nodes;
"""
```

```
1435
```

- **Top 5 contributor in Taipei map**

```
QUERY = """
SELECT nodes.user, COUNT (*) as num
FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) nodes
GROUP BY nodes.user
ORDER BY num DESC
LIMIT 5;
"""
```

```
[(u'Supaplex', 178571), (u'Littlebtc', 77375), (u'siaoyo', 46792),
(u'\u611b\u53f0\u73a9', 44876), (u'Vintagejhan', 41001)]
```

(this one is number 4 in the most contributing user. This Chinese means "Love Taiwan!"

print u'\u611b\u53f0\u73a9'

愛台玩)

- **To check how many karaoke in Taipei**

```
QUERY = """
SELECT nodes_tags.key, nodes_tags.value
FROM nodes_tags
WHERE nodes_tags.key = "karaoke";
"""
```

```
[(u'karaoke', u'yes'), (u'karaoke', u'yes')]
```

Only two?? That's definitely a mistake, if you want to go karaoke in Taipei, you will have plenty of options!!

- **To check the TOP 10 shop in Taipei**

```
QUERY = """
SELECT nodes_tags.value, COUNT(*) as num
FROM nodes_tags
WHERE nodes_tags.key = "shop"
GROUP BY nodes_tags.value
ORDER BY num desc
LIMIT 10;
"""
```

```
[(u'convenience', 1906), (u'clothes', 523), (u'supermarket', 386),
(u'hairdresser', 310), (u'bakery', 276), (u'yes', 191), (u'beverages', 162),
(u'books', 161), (u'motorcycle', 144), (u'chemist', 117)]
```

- **To check which food Taipei residents most likes**

```
QUERY = """
SELECT nodes_tags.value, COUNT(*) as num
FROM nodes_tags
  JOIN (SELECT DISTINCT(id) FROM nodes_tags WHERE value = 'restaurant') i
  ON nodes_tags.id = i.id
WHERE nodes_tags.key = 'cuisine'
GROUP BY nodes_tags.value
ORDER BY num DESC
LIMIT 5;
"""
```

```
[(u'chinese', 269), (u'japanese', 136), (u'regional', 54), (u'italian', 53),
(u'breakfast', 51)]
```

Other ideas about the datasets

When I saw the list of unique key, I found that there are so many different keys but same values, like currency, bike or bicycle, etc. I think it will be a good idea to make a list of unique key, each unique key should contains different value. When people try to make some revises on the map, they can choose correspond key. Only if there are no correspond key, they may use their own key. This can reduce many duplicate key, making the map more efficient. One thing is that when creating the unique key list, people will lost their freedom to compile the map. Since the open street map is like Wiki, everyone has the right to revise it, it will be a dilemma when facing the free of compiling and consistency of map. But I prefer the consistency of map, so I would suggest to use the unique key list, but people can still make their own key once the key they want is not in the list.

We can make the rank of top contributor, just like the stars in GitHub. If someone's contribution is well, people can give him a star. The rank system will show the one who has most stars.

Conclusion

This is really a tough project, cost me almost three weeks to finish it. This Taipei city map is incomplete and many mistakes need to be fixed. Like there are lots of key meaing the same value('NTD',"TWD" all means Taiwan dollars). And the street name are not consistent. After using sql to analyze this database, I found that the content are not complete. Like the karaoke in Taipei are way more than 2. But, the number of convenience stores are correct! There are almost 2,000 convenience stores in Taipei and the number convenience stores from database is 1906, which is almost the same. The food that taipei residents enjoy is Chinese food, which is not suprised to me.