OpenStreetMap Data Case Study Of Nanjing

Problems Encountered In The Map ¶

After using python code to make the original map smaller and running it against some codes, I noticed three problems with the data, which I will discuss in the following order:

1. Different ways to show the same meaning when expressing number of gate
2. Different phone number format
3. The traditional Chinese expression is hard to understand its meaning from the name
4. Chinese is the biggest problem for me during my queries.

Different ways to show the same meaning when expressing number of gate

Problem Review

When I look through the names of nodes,I found some interesting names and some problems which will lead to misundersatnding of the map.

- 街口#4
- 新街口#11 The number above represent the order of entrances.
- 奥体东门5
- 下马坊2
- 马群1
- The number above also represent the order of entrances.But they have different ways to show,which may cause some problems when doing some queries.

```
import re
number=re.compile(’\d’)
for event,elem in ET.iterparse(’sample.osm’):
    if elem.tag==’node’:
        for child in elem:
            if child.attrib[’k’]==’name’:
                if number.search(child.attrib[’v’]):
                    print child.attrib[’v’]
```

Above is the code I use to find the problems,and I have used a function to solve this problem

Problem Solution

I use regular expression to find the '#' in the file and replace it with' '

```
def eli_diff(string):
    ”””This function is used to clear wrong expressions of gate”””
    DIFFERENCE=re.compile(r’#’)
    if DIFFERENCE.search(string):
        move_str=string.repalce(’#’,’ ’)
    else:move_str
    return move_str
```

Different phone number format

I found some phone numbers with different format,which will cause error in practice

Problem Review

1. +86 25 86647225...........This is the standard format.
2. 025-52774818..........This one do not have +86,since openstreetmap is an international website,it is better to add +86.
3. +86-25-86812222
4. +86 553 584 4888.........This one lack one number,and dial this number will lead you to nothing.

Problem Solution

```
def clear_number(number):
    """This function is used to clear wrong number format according to the situation sho
wed above"""
    is86=re.compile(r'^\W86')
    if len(number.split())>3:
        return "wrong"
    if not is86.search(number):
        number='+86-'+number
    number=number.replace(' ','-')
    return number
```

The traditional Chinese expression is hard to understand its meaning from the name

Problem Review

Chinese has various meaning when express the same thing so there are many misundersandings when I see the map

- 高阳
- 竹园
- 双岗
- 管头 ###### Those names will not indicate what it is just from the name.
- 禄口山大巴 ###### The name above will cause misunderstanding

Problem Solution

For this problem is hard to find what the pattern is,I have not figured out a way to solve it programatically yet.The reason for the error is that

Chinese is the biggest problem for me during my queries.

Problem Review

This actually exalts me.Because when I use the code to divide the osm file into several csv files,I use Excel to open it,but I found some lines lost its format

- 1042174613 32.082303 118.8904146 sinopitt 253683 3 13930721 2012-11-19T11:58:23Z............This is the right format

- 1042174614 1042174614 118.8892711 鏉庨殕闃?4071947 3 39987055 2016-06-13T08:31:00Z..........This is the wrong format

This also reflects in my queries,everytime I use my queries to get a table with Chinese,it will get wrong .

- babribeiro|157.........This is the line without Chinese,it is alright
- 甯傚寇銈嗐倞瀛您108..........This is the line with Chinese,it is wrong,and you can see there is no '|'to seperate the string and number,and if I use this table to do some queries,it will definitely get wrong.

Problem Solution

- For the first one,if I open it with notepad it will be alright,which means my files are not wrong,later I found if I use some settings in the Excel,it will be right in Excel,that's because Excel won't automatically use"utf-8"

- For the second problem, I have not found a solution yet.

Data Overview

This section contains basic statistics about the dataset, the SQL queries used to gather them, and some additional ideas about the data in context.

Files Size

```
nanjing_china.osm..........72.0Mb
xnanjing.db...........52.8Mb
nodes.csv..........28.8Mb
nodes_tags.csv..........0.7Mb
ways.csv..........2.5Mb
ways_tags.csv..........2.9Mb
ways_nodes.csv..........10.0Mb
```

Number Of Nodes

```
SELECT COUNT(*) FROM nodes;
```

352356

Number Of Ways

```
SELECT COUNT(*) FROM ways;
```

43721

Number Of Unique Users

```
sqlite> SELECT COUNT(DISTINCT(e.uid))
FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways) e;
```

366

Number Of Banks

```
sqlite> SELECT COUNT(nodes.id)
   ...> FROM nodes JOIN nodes_tags
   ...> ON nodes.id=nodes_tags.id
   ...> WHERE value==’bank’
```

89

The Top Ten Contributor in Specied Field

Since I am typically interested in bridges in Nanjing,so I decide to make some queries through the contributors have made contribution in auditing bridges:

```
sqlite> SELECT user,count(*)
   ...> FROM ways JOIN ways_tags
   ...> ON ways.id=ways_tags.id
   ...> WHERE key=='bridge'
   ...> GROUP BY user
   ...> ORDER BY count(*) desc
   ...> LIMIT 10;
```

1. Chen Jia,275
2. sinopitt,259
3. jamesks,246
4. yangfl,238
5. greecemapper,199
6. babribeiro,157
7. 甯傚寇鍏嗗悍瀛恷,108
8. Dmitry2013,97
9. bhw98,79
10. aighes,76

Top Ten Contributors

```
sqlite> SELECT e.user,COUNT(*) as num
   ...> FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e
   ...> GROUP BY e.user
   ...> ORDER BY num DESC
   ...> LIMIT 10;
```

1. Chen Jia,89952
2. sinopitt,37557
3. 鍥涘浗鍐涙□鍐涢暱,32259
4. bhw98,19026
5. jerryhappy,17710
6. 甯傚寇鍏嗗悍瀛恷,16572
7. Sunng,15867
8. METRO2333,14602
9. jamesks,13104
10. katpatuka,8098

Additional Ideas

As I have found the top two contributors in bridge field and overall are the same, I guess maybe the top contributors cover most of the posts. So I decide to make some queries of this part and try to get some statistics.

Here are some user percentage statistics:

- Total posts:400077
- Top user contribution percentage ("Chen Jia") 22.48% (Total posts:89952)
- Combined top 3 users' contribution ("Chen Jia" and "sinopitt"and"鍥涘浗鍐涙□鍐涢儴") 39.93% (Total posts:159768)
- Combined Top 10 users contribution 66.67% (Total posts:264747)
- Total user:366 ##### Here are the codes I use to get the statics:

```
import sqlite3
file='xnanjing.db'
con=sqlite3.connect(file)
c=con.cursor()
QUERY='SELECT e.user,COUNT(*) as num\
     FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e\
     GROUP BY e.user\
     ORDER BY num DESC\
     LIMIT 1;'
c.execute(QUERY)
all_rows = c.fetchall()
print(all_rows)
```

```
import sqlite3
file='xnanjing.db'
con=sqlite3.connect(file)
c=con.cursor()
QUERY='SELECT COUNT(*) as num\
     FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e\
     ;'
c.execute(QUERY)
all_rows = c.fetchall()
print(all_rows)
```

```
import sqlite3
file='xnanjing.db'
con=sqlite3.connect(file)
c=con.cursor()
QUERY='SELECT SUM(num) FROM\
     (SELECT e.user,COUNT(*) as num\
     FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e\
     GROUP BY e.user\
     ORDER BY num DESC\
     LIMIT 10);'
c.execute(QUERY)
all_rows = c.fetchall()
print(all_rows)
```

There are 366 unique users participating in editing this map,but top ten contributors have made 66.67% of the total posts,which is to say most of people who participated do not edit the map frequently.So why?I think that is because there is not a clear way to encourage people to participate.If we give those who participate frequently some rewards ,like badges or leaderboard that may spur other people's passion in creating the map.Basically I think the benefit of implementing this is attracting more participates,and more participates usually means more accurate of the map,thus it will attract more people and enter a virtuous circle.But somehow there exists difficulty of implementing this measure.Since making a reward system is not a easy job,it needs the developers of this website to make more investment both in time and money,besides,we are not so sure whether those rewards will attract people or not.So we need to try and try in order to figure out the best way to reward.

Conclusion

Through the wrangling,I have learned a lot.Especially,I realize the difficulties use deal with Chinese encode data.And I have also found the map of nanjing in OPENSTREET MAP is incompleted.I think seldom people will use this map in China because the barrier of language,so there is a reason that why many districtions have been misundersood.I think this website can offer us a version of Chinese,let more Chinese take part in processing the map,after all Chinese will have a better understanding of Chinese and sites named in Chinese.