

# 地图区域

中国,上海

<https://www.openstreetmap.org/node/778910398>

[https://mapzen.com/data/metro-extracts/metro/shanghai\\_china/](https://mapzen.com/data/metro-extracts/metro/shanghai_china/)

该城市是我能找到满足这次练习的一个比较好的选择,就是他了.说不定哪天去哪工作了,虽然我在深圳,  
可惜深圳的数据量不够.

## 数据处理

### key value 的提取规则

key: 如果没有冒号, 则是完整的标记“k”属性值, 如果有冒号, 则是冒号后面的字符。

type: 标记“k”值中冒号前面的字符, 或者如果没有冒号的话, 则是“regular”。

此外,

- 如果标记“k”值包含存在问题的字符, 则应该忽略该标记
- 如果标记“k”值包含“:”, 则“:”前面的字符应该设为标记类型, “:”后面的字符应该设为标记键
- 如果“k”值中包含其他“:”, 则应该忽略这些“:”并保留为标记键的一部分

```
if ':' in elem.attrib['k']:
    type = re.split(':', elem.attrib['k'])[0]
    tag['key'] = elem.attrib['k'][len(type)+1:]
    tag['type'] = type
else:
    tag['key'] = elem.attrib['k']
    tag['type'] = 'regular'
if LOWER_COLON.match(tag['key']) and PROBLEMCHARS.match(tag['key']):
    print tag['key']
else:
    tags.append(tag)
```

### 清理 nodes\_tags 中 name 中英混杂的问题

## 清理前

```
id,key,value,type
26466690,name,玛雅酒吧 Maya Pub,regular
26466690,amenity,pub,regular
26466690,en,Maya,name
26466690,zh,玛雅酒吧,name
26466690,street,白沙泉,addr
26466690,zh_pinyin,Mǎyǎ Jiǔbā,name
26466690,housenumber,94,addr
```

清理思路就是如果该 id对应的有 zh 属性或者 zh\_pinyin 属性,就用两者中的一个替换 name 字段,优先

使用 zh, 如果都没有,就舍去该条记录.

```
chinese_name=None
pinyin_name=None
for tag in tags:
    if tag['key']=='zh':
        chinese_name = tag['value']
    elif tag['key']=='zh_pinyin':
        pinyin_name = tag['value']

for tag in tags:
    if tag['key']=='name':
        if chinese_name!=None:
            tag['value']=chinese_name
        elif pinyin_name!=None:
            tag['value']=pinyin_name
```

# 数据基本情况和建议

## 数据大小

```
shanghai.osm ..... 762 MB
shanghai.db ..... 531 MB
nodes.csv ..... 294 MB
nodes_tags.csv ..... 9.5 MB
ways.csv ..... 26 MB
ways_tags.csv ..... 31 MB
ways_nodes.cv ..... 103 MB
```

# nodes节点数量

```
select count(*) from nodes;
```

3752991

# ways 节点数量

```
select count(*) from ways;
```

464372

# 独一无二的用户数量

```
SELECT COUNT(e.uid)  
FROM (SELECT uid FROM nodes UNION SELECT uid FROM ways) e;
```

2476

# 贡献最多的用户前十位

```
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;
```

```
"Chen Jia" "704808"  
"Austin Zhu" "191291"  
"aighes" "188950"  
"xiaotu" "176367"  
"katpatuka" "142099"  
"XBear" "125480"  
"Peng-Chung" "111460"  
"yangfl" "109888"  
"Holywindon" "102497"  
"dkt" "94491"
```

# 只贡献过一次内容的用户数量

```
SELECT COUNT(*)  
FROM  
(SELECT e.user, COUNT(*) as num  
FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e  
GROUP BY e.user  
HAVING num=1) u;
```

527

## 总结

```
select distinct nodes_tags.value  
from nodes_tags  
where nodes_tags.key='city'
```

执行上面 sql 会发现很多不是上海市,说明地图中有很多数据都不是上海市的数据.  
下面截取部分结果

```
"南汇路"  
"碧桂园凤凰城大凤凰"  
"南汇新城镇"  
"1111"  
"Shanghai"  
"宜兴市湖父镇"  
"浙江省舟山市定海区临城街道"  
"浙江省舟山市普陀区东港街道"  
"萧山市"  
"杭州市余杭区闲林街道"
```

会发现有很多上海周边城市的节点.说明我们这个地图数据不是很准确

## 建议

建议和国内专门做地图服务的机构合作.拿到他们的数据API,比如高德地图.

好处:

由于中国城市进程发展迅速,城市结构剧烈变化.而高德地图不断更新,正好可以满足我们得到精确数据的需求

而且,由于高德地图有专门的地理勘探团队,在数据准确性方面也大大高于普通用户上传.

## 问题:

高德地图提供的数据格式和 `osm` 可能不太一样,所以需要花时间去整合.这部分可能会很耗时,需要去详细

了解两边的数据格式定义.