

# OpenStreetMap 数据案例研究

## 在地图中遇到的问题

下载好地图数据，然后将地图数据缩小 5 倍，即 k=5 代码环境：py 3

- 语言问题：中英文合并在一起，处理起来比较乱
- 下载广州的数据里面有香港的数据
- Way 的 key 值除了 name，还有 zh，zh\_pinyin 等
- 地址中路的名称不统一，比如 Rd，road)，road 等不规范的后缀

## 规范道路类型

`#期待的道路类型`

```
expected = ["Street", "Avenue", "Boulevard", "Drive", "Court", "Place", "Square", "Lane",  
"Road", "Trail", "Parkway", "Commons"]
```

`# 数据中需更新的道路类型`

```
mapping = { "St.": "Street",  
            "Ave": "Avenue",  
            "Bou": "Boulevard",  
            "Dr": "Drive",  
            "CT": "Court",  
            "Pl": "Place",  
            "Sq": "Square",  
            "La": "Lane",  
            "Rd.": "Road",  
            "Rd": "Road",  
            "Tr": "Trail",  
            "PARKWAY": "Parkway",  
            "Co": "Commons"  
}
```

`#矫正不规范的道路名称#`

```
def update_name(name, mapping):  
    shortname = mapping.keys()  
    for word in shortname:  
        if word in name:  
            name = name.replace(word,mapping[word])  
    return name
```

## 数据概述

OSM XML 是否足够大？  
-OSM XML 未压缩时超过 50 MB，有 562MB  
使用了数据库查询来提供数据库的统计概述  
-直接在文件查看其大小为 561520kb

map_sample.osm	11MB
map-广州-v2.py	12KB
OSM.db	87.4MB
nodes.csv	42.7 MB
nodes_tags.csv	1.07 MB
ways.csv	3.03 MB
ways_tags.csv	14.3 MB
ways_nodes.csv	3.92 MB

## 唯一用户的数量

```
select count(*) from nodes group by id  
531667
```

## 节点和途径的数量

```
节点：select count(*) from nodes_tags group by id  
15880  
途径：select count(*) from ways group by id  
52510
```

## 所选节点类型（如：咖啡店、商店等）的数量

```
select count(*) from nodes_tags where key='highway'
3111
select count(*) from nodes_tags where key='street';
104
select count(*) from ways_tags where key='zh';
路的名字是中文的有 2819
```

## 是店铺的数据

```
slect count(*) from nodes_tags where key='shop';
246
```

## 店铺类型数据最多的五家连锁

```
SELECT nodes_tags.value,COUNT(*) as num
FROM nodes_tags JOIN
(SELECT DISTINCT(id) FROM nodes_tags WHERE key='shop')as i
ON nodes_tags.id=i.id
WHERE nodes_tags.key='name'
GROUP BY nodes_tags.value
ORDER BY num DESC
Limit=5;
```

	value	num
1	7-Eleven	17
2	国美电器	3
3	新华书店	3
4	Circle K	2
5	Mannings	2

广州该样本数据中店铺最多的是 7-11 便利店，从经验来说是合理的一个结果。

## 关于数据集其他想法

好处：

1. 数据多，覆盖面广
2. 可以轻松获得数据

预期的问题：

1. 因为是国外的网站，所以对语言方面比较担心
2. 数据是否是新的，因为中国发展比较快，很多零售店铺都一直在更新
3. Open street map 上的数据个人认为作为练习或者初步了解一个城市是很有用的，但是如果准确地了解一个城市，数据不够新和本土化的话，可信度将不高。