OpenStreetMap 案例研究

区域

中国北京
地图数据文件下载链接
选择原因：北京是中国首都，地图编辑者多，地图数据信息较为详细。

审查数据发现问题

- **部分道路（highway）的英文名（name:en）不规范。**

  例如：
  jie（街）、Str 应改为 Street
  lu（路）、Rd. 应改为 Road
  Hutong（胡同） 应改为 Alley 等。
  注：

  1 只修改英文名（name:en），不修改国际名（int_name），因为数据文件中国际名普遍为 jie（街），lu（路）的形式。

  2 此处选取有 highway 属性的道路进行清洗，这里的 highway 并非仅指高速公路，根据 OSM 命名的中文标准（与英文标准不同，链接：Zh-hans:Map Features），highway 是用于街道的最主要标签和唯一的标记，通过对数据集的审查也可以发现，不仅是高速公路、一级二级公路、街道、胡同全部都有 highway 属性，只不过属性值各有不同，详细可见上面的标准。

- **银行名字（name）不一致，大部分为中文，但有的为英文（银行的英文名应放在 name:en 属性）**

  如 Bank of Communications 应改为交通银行等

修正道路（highway）的英文名(name:en)

用 Python 构造一个修正函数：

```
def update_name(name, mapping):
    wrong_name_list = mapping.keys()
    for wrong_name in wrong_name_list:
        if name.endswith(wrong_name):
            name = name.replace(wrong_name,mapping[wrong_name])
    return name

mapping = {"jie":"Street","Jie":"Street",'JIE':"Street","St": "Street",
           "St.": "Street", "Str":"Street",'Hwy':'Highway', "Ave": "Avenue",
           "lu":"Road","Lu":"Road","Rd.": "Road","road":"Road",
           "hutong":"Alley","Hutong":"Alley","xiang":"Alley","Xiang":"Alley"}
```

完整代码见文件 clean_shape_convert.py

一致化银行名字（name）

```
def consistent_name(name, mapping_bank):
    wrong_name_list = mapping_bank.keys()
    for wrong_name in wrong_name_list:
        if name == wrong_name:
            name = name.replace(wrong_name,mapping[wrong_name])
    return name
```

数据集概述

注：数据库软件为 MySQL 而非 sqlite.

文件大小

```
beijing_china.osm ..... 182 MB
beijing_db.sql ..... 122 MB
nodes.csv ..... 68.9 MB
nodes_tags.csv ..... 3.06 MB
ways.csv ..... 7.36 MB
ways_tags.csv ..... 8.5 MB
ways_nodes.cv ..... 24.6 MB
```

nodes 数量

```
SELECT count(*)
FROM nodes;
```
863664

ways 数量

```
SELECT count(*)
FROM ways;
```
128218

unique users 数量

```
SELECT count(distinct subq.uid)
FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways) as
subq;
```
1846

有 highway 标签的 ways 数量

```
SELECT count(*)
FROM ways_tags
WHERE key = 'highway';
```
64186

有 highway 标签且有英文名（name:en）的 ways 数量

```
SELECT count(distinct subq.id)
FROM ways_tags JOIN (
SELECT id
FROM ways_tags
WHERE key = 'highway') as subq
ON ways_tags.id = subq.id
WHERE ways_tags.key = 'en' and ways_tags.type = 'name';
```
4152

有名字的银行网点的数量

```
SELECT count(*)
FROM nodes_tags
WHERE `key` = 'amenity' and value = 'bank';
```
456

关于数据集的其他探索和想法

道路（highway）的各种别名的数量

```
SELECT ways_tags.key, count(distinct subq.id) as count
FROM ways_tags JOIN (
SELECT id
FROM ways_tags
WHERE key = 'highway') as subq
ON ways_tags.id = subq.id
WHERE ways_tags.type = 'name'
GROUP BY ways_tags.key
ORDER BY count DESC
LIMIT 5;
```

查询结果：

```
en, 4152
zh, 949
zh_pinyin, 896
fr, 353
de, 71
```

可以看到，在道路的各种别名中：
道路的英文名（en）数量最多，但也仅占总数的 6.5%，有些道路有英文名甚至汉语拼音、法文名等，但有些道路完全没有别名，道路别名信息比较混乱，没有统一格式。

额外的想法

在探索数据集的过程中，发现的一个问题是道路或节点的属性值命名比较混乱，缺乏统一的标准，另外有一些数据未严格按 OSM 命名的中文标准（链接：Zh-hans:Map Features）进行命名，以至于出现很多诸如 expressway 等不规范的命名，建议在地图编辑器中添加一个功能，时刻提醒编辑者或贡献者们在编辑地图时注意命名标准，这样在之后的数据分析时会方便很多。

改进的益处及潜在的问题

如果贡献者能够在编辑数据时遵守一个统一的标准，那么别人在下载数据用于数据分析时就能够避免很多的清洗工作。但是这个标准可能是相对复杂的，要求贡献者去严格遵守，无形中增加了编辑数据的门槛，另外，贡献者可能来自世界各地，而各国的道路、设施命名的标准不同，这也是此措施的一个潜在问题。