OpenStreetMap数据分析

本次项目选择中国成都的地图进行分析，因我对成都比较熟悉，使用Python对数据进行清洗，再通过SQL进行数据分析。 使用的数OpenStreetMap的开放地图数据：https://mapzen.com/data/metro-extracts/metro/chengdu_china/ (https://mapzen.com/data/metro-extracts/metro/chengdu_china/)

一、地图中遇到的问题

存在拼音表示的街道名

地图数据中，存在拼音和英文表示的街道名，如“Tian Xian Qiao Bei Jie”等；

解决方法是用翻译后的中文街道名替换拼音街道名称。

存在以小区名代替街道名

地图数据中，存在小区名代替街道名的，如“高新区南城都汇2A期汇雅园”；

解决方法是用正确的街道名称替换。

In [ ]:

```
#audit_ways.py
#英文街道名与中文街道名的映射
english_name_mapping = { "Tian Xian Qiao Bei Jie": "天仙桥北街",
            "Long du nan Road": "龙都南路",
            "JinAn": "金安路",
            "Huadu Avenue West": "华都大道西",
            "Jinquan Street": "金泉街",
            u"高新区南城都汇2A期汇雅园": "景明路"
            }

#判断是否街道名
def is_street_name(elem):
    return (elem.attrib['k'] == "addr:street")

#审核街道名
def audit(element):
    if element.tag == "node" or element.tag == "way":
        for tag in element.iter("tag"):
            if is_street_name(tag):
                for key,value in english_name_mapping.items():
                    if key == tag.attrib['v']:
                        tag.set('v', value)
```

In [ ]:

```
#process_map.py
#在shape_element中调用audit()方法
from audit_ways import audit
def shape_element(element, node_attr_fields=NODE_FIELDS, way_attr_fields=WAY_FIELDS,
                  problem_chars=PROBLEMCHARS, default_tag_type='regular'):
    """Clean and shape node or way XML element to Python dict"""

    node_attribs = {}
    way_attribs = {}
    way_nodes = []
    tags = []  # Handle secondary tags the same way for both node and way elements

    audit(element)
    ...
```

二、数据概览

将清洗后的数据输出为csv格式，然后导入到sqlite3数据库中。

文件大小

- chengdu_china.osm 53.8MB
- chengdu.db 38.4MB
- nodes.csv 21.5MB
- nodes_tags.csv 387KB
- ways.csv 2.02MB
- ways_tags.csv 2.51MB
- ways_nodes.csv 7.56MB

nodes的数量

SELECT COUNT(*) FROM nodes;

263373

ways的数量

SELECT COUNT(*) FROM ways;

33994

用户的数量

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

556

贡献前十的用户

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 ;

巴山夜雨,48929
geodreieck4711,18739
katpatuka,18228
ff5722,14053
hanchao,12359
jamesks,10792
Ernst Poulsen,9080
guanchzhou,8662
jechterhoff,8530
AntiEntropy,7332

只贡献过一次的用户数

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;

119

贡献前十的用户占总贡献的比例

select ((select sum (num1) from (SELECT COUNT (*) as num1 FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e GROUP BY e .user ORDER BY num1 DESC LIMIT 10))1.) /(select count(*) as num2 from (SELECT user FROM nodes UNION ALL SELECT user FROM ways));

0.5269

贡献前一百的用户占总贡献的比例

select ((select sum (num1) from (SELECT COUNT (*) as num1 FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e GROUP BY e .user ORDER BY num1 DESC LIMIT 100))1.) /(select count(*) as num2 from (SELECT user FROM nodes UNION ALL SELECT user FROM ways));

0.9641

三、额外的想法

3.1 额外的想法

从上面的数据统计来看，当前的地图数据量太少，参与的用户也比较少，才五百多，所以：

建议1：增加开放地图项目的宣传，让更多的人参与进来

好处：
1，更多的人参与进来，会提供更多的数据，可使地图数据更详细；
2，地图数据达到一定量后，就会有更多的人和公司会考虑采用本地图数据，从而参与的人越多；
3，使用的人越多，发现的问题越多，从而数据会更精确。

预期的问题：
1，贡献数据的人越多，会导致数据存在更大的不一致性；
2，很多比较少使用地图的用户贡献的数据质量也会比较低，用户增多后会导致低质量的数据也会增多。

建议2：采用其它地图项目的数据进行补充

好处：
1，通过采用其它地图项目的数据，可以快速扩大本地图数据的数据量；
2，可以通过其它地图项目的数据对本地图数据进行校验，从而提交数据的准确性。

预期的问题：
1，其它地图项目的数据也可能存在错误的数据，采纳前也需要进行数据校验；
2，采用其它地图项目的数据，可能存在著作权，知识产权冲突的问题。

3.2 进行额外的数据探索

出现次数最多的便利设施

```
SELECT value, COUNT(*) as num FROM nodes_tags WHERE key='amenity' GROUP BY value ORDER BY num DESC LIMIT 10;
```

```
restaurant|166
toilets|71
parking|64
cafe|62
bank|61
school|48
hospital|43
fuel|37
bicycle_parking|30
fast_food|25
```

出现次数最多的餐厅

```
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;
```

chinese|12
asian|2
barbecue|2
japanese|2
regional|2
BBQ|1
burger|1
burger;italian_pizza;american|1
chinese;local|1
chinese;noodles|1
german|1
international|1
pizza|1
spanish|1
tea;local|1
turkish|1
燃面，姜鸭面|1

出现次数最多的商店类型

SELECT value, COUNT(*) as num FROM nodes_tags WHERE key='shop' GROUP BY value ORDER BY num DESC LIMIT 10;

clothes|46
supermarket|46
convenience|27
gift|14
yes|13
florist|12
bakery|7
department_store|7
outdoor|7
bicycle|4

四、结论

从上面的数据可以看出，出现最多的便利设施是餐厅、厕所、公园、咖啡馆，出现最次数最多的餐厅是中餐馆，出现次数最多的商店类型是服装店和超市，与成都的休闲文化相匹配。 从对成都的地图数据分析结果来看，当前的地图数据量太少，标注的并不详细，可能国内对这种开放式的地图了解的人还比较少，需要更多的人来贡献数据。 经过数据的清洗和矫正，数据的准确性和完整性有了一定的提升，但可能仍然存在一些错误数据。 通过本次练习，让我认识到开放数据地图可以让所有感兴趣的人都可以加入到地图的编辑和矫正中来，可以快速的构建地图数据，同样因为有很多人同时在进行编辑操作，就会带来数据的不一致性。所以在进行数据分析前，应对数据进行清洗和矫正。