open street map 项目报告

邓剑波 2018-3-17

项目简介

Open street map （https://www.openstreetmap.org） 是一个可以由用户自由修改和标记的开放地图项目，用户可以通过对地图进行修改和审核，来完善地图的标记和功能，通过用户共同的贡献，创造出一个可自由编辑的世界地。而对open street map 特定区域的探索，可以通过数据分析的方式，调查该地的状况，从而得出对生活和工作有帮助的结论。

一、数据说明

地图来源：https://www.openstreetmap.org

源文件下载来源：http://overpass-api.de/query_form.html

地图区域：min_lon = 111.0304336 min_lat = 22.0586758 max_lon = 114.8118653 max_lat = 23.7075167

区域说明：所选中的地图区域包含了大部分的广东省珠三角的城市数据，选择数据的原因是珠三角是我主要的生活和工作的区域，所以希望通过对该区域地图的探索，能够对该区域有一个更深刻的了解。

二、数据清理

1、数据问题

通过对数据的审查，发现以下问题：

● 首先是街道名字不统一，本应填写街道的名字会有多余的信息，如<tag k="addr:street" v="广州市越秀区北站路168号" />；

● 街道名字会有中英混合的状况，如 <tag k="addr:street" v="湖贝路 Hubei Rd" />；

● 城市名信息不一致，简繁体同时存在，拼音和中文同时存在，如 <tag k="addr:city" v="深圳市 Shenzhen"/>， <tag k="addr:city" v="廣州" />；

● 城市名信息错乱，有的是区域名，如<tag k="addr:city" v="广州市天河区珠江新城"/>。

2、清理数据

街道名称清理

通过审查数据，找出不符合规范的街道名称，然后通过update_name对数据进行清理更新，更新具体方式如下

In [ ]:

```
st_name_zh = re.compile(ur".*?([\u4E00-\u9FA5]+)") #匹配中文字符
mapping = {u'白云大道南':u'白云大道',
           u'三元里大道中':u'三元里大道',
           u'福中路市民中心':u'福中路',
           u'中山大道西':u'中山大道',
           u'广州市南沙区环市大道南':u'环市大道',
           u'布吉街道大芬社區':u'布吉街道',
           u'合新路九號':u'合新路',
           u'冈州大道东':u'冈州大道',
           u'江门市新会区冈州大道':u'冈州大道',
           u'芳村大道中':u'芳村大道',
           u'中康路卓越城北區':u'中康路',
           u'江南大道中':u'江南大道',
           u'广州大道中':u'广州大道',
           u'中山大道西':u'中山大道',
           u'广州大道北':u'广州大道',
           u'白云大道南':u'白云大道',
           u'黄埔大道西':u'黄埔大道',
           u'芳村大道中':u'芳村大道',
           u'高新科技園高新南':u'高新路',
           u'佛平':u'佛平二路'
           }
def update_name(name,mapping):
    zh_name = st_name_zh.match(name)
    problemname = mapping.keys()
    #先对名字进行判断，如果街道名字包含中英文，先去除英文
    if zh_name:
        name = zh_name.group(0)
        for key in problemname:
            if key in name:
                name = name.replace(key, mapping[key])
    else:
        name = None    #对于不符合中文格式的字段，返回None值
    return name
```

经过更新后的数据，街道名称将符合规范，更新状况例子如下：
广州大道北 Guangzhou Ave N > 广州大道
江门市新会区冈州大道东 > 冈州大道

城市名清理

城市名称的清理方式和街道名称类似，不同的地方在于可以通过中国人书写地名的习惯来减少部分的工作量；

另外，由于城市名中有很多用户录入的是区域的名字，我利用open street map在线地图，查找了该区域所属城市，也列入更新范围。

具体方式如下：

In [ ]:

```
#所选区域包含城市仅有以下几个，非下列城市的都属于超出范围
expected_city = [u'深圳市',u'广州市',  u'东莞市',  u'佛山市',  u'中山市',  u'江门市']
#匹配中文字符
city_name_zh = re.compile(ur".*?([\u4E00-\u9FA5]+)")

city_mapping = {u'深圳':u'深圳市',
           u'龙岗中心城':u'深圳市',
           u'龙岗中':u'深圳市',
           u'沙頭角':u'深圳市',
           u'大运新城':u'深圳',
           u'体育新城':u'深圳市',
           u'六榕街道':u'广州',
           u'大塘街道':u'广州',
           u'白云区':u'广州',
           u'廣州':u'广州',
           u'廣州市':u'广州市',
           u'北京街道':u'广州',
           u'广州荔湾区解放中路街道':u'广州',
           u'广州':u'广州市',
           u'佛山':u'佛山市',
           u'佛山市南海區':u'佛山市',
           u'龙江镇':u'佛山',
           u'東莞市長安鎮':u'东莞',
           u'常平镇':u'东莞市',
           u'东莞生态园':u'东莞',
           u'东莞':u'东莞市',
           u'东东莞':u'东莞',
           u'新会区':u'江门市',
           u'沙坪市':u'江门市',
           u'石岐区街道':u'中山市',
           u'沙溪镇':u'中山市'
            }


def update_city_name(name,mapping):
    problemname = mapping.keys()
    zh_name = city_name_zh.match(name)
    if zh_name:
        if name[:3] in expected_city:    #这里的[:3]和下面的[-3:]使用的原因是中国地名的书写习惯，通过提取前三个字或者后三个字，能够减少工作量
            name = name[:3]
        elif name[-3:] in expected_city:
            name = name[-3:]
        else:
            for key in problemname:
                if key in name:
                    name = name.replace(key, mapping[key])[:3]
    else:
        name = None
    return name
```

经过更新后的城市名称例子如下：
广州荔湾区解放中路街道 > 广州市
广东省深圳市 > 深圳市
沙頭角 Sha Tau Kok > 深圳市

三、数据探索(SQL)

通过清理后的数据，写入CSV文件，得到如下CSV文件：
nodes.csv : 132 MB
nodes_tags.csv : 3.67 MB
ways.csv : 10.4 MB
ways_tags.csv : 45.6 MB
ways_nodes.cv : 13.5 MB
以下通过sqlite3进行数据探索

用户修改时间

nodes.csv文件中，timestamp描述的是用户修改地图的时间戳，通过百度百科的查询，open street map项目开始于2006年，我想通过统计根据年份修改时间戳的数量，来查看珠三角区域用户每年录入修改的状况。

In [ ]:

```
sqlite> SELECT strftime('%Y', nodes.timestamp) AS time,
   ...> COUNT(*) FROM nodes
   ...> GROUP BY time
   ...> ORDER BY COUNT(*)
   ...> DESC
   ...> ;
```

2017,417299
2016,294588
2015,255496
2014,227407
2013,176972
2012,105156
2018,74388
2011,26303
2010,19164
2007,15662
2009,10259
2008,1453

通过对用户修改时间戳的统计，可以发现，珠三角区域用户开始参与标注修改open street map项目开始于2007年，除去今年2018年，从2010年开始，用户参与的热情逐年提高，在2017年达到峰值，总共标注修改此处超过41万次。

但是有趣的是，2007年的标注修改数量却高于2008年和2009年，特别是2008年，用户参与次数相较2007年下降了90.6%，2009年回升，但是也仅为2007年数量的65.5%，之后的年份数量又开始大幅提升。我试图通过查找2008年open street map的新闻，试图证实是否是项目出现了大的困难或者危机，但是并没有找到相关新闻，只能假设是因为金融危机，严重的影响了用户参与免费公开项目的热情。

城市名数量探索

nodes_tags.csv 和 ways_tags.csv文件中包含了大量的城市名，通过对城市名数量的统计，希望找出每个城市被标注的程度。

In [ ]:

```
sqlite> SELECT tags.value, count(*) FROM
   ...> (SELECT value, key FROM nodes_tags
   ...>  UNION ALL
   ...>  SELECT value, key FROM ways_tags)
   ...> tags
   ...> WHERE tags.key = 'city' AND tags.value IS NOT ""
   ...> GROUP BY tags.value
   ...> ORDER BY count(*) DESC;
```

"深圳市",269
"广州市",142
"东莞市",12
"中山市",10
"佛山市",8
"江门市",7

通过对提取数据中城市被标注的次数，可以看出珠三角中最热心于open street map项目的的是来自深圳市的用户。

对比深圳市和广州市的状况，通过搜索，我查询了广州市的总人口为1449.84万 (2017年)，深圳市的总人口为1252.83万 (2017年)，深圳市的人口数量小于广州市人口数量，但是深圳市用户在城市名上的编辑数量确实广州市用户的两倍。考虑到两个城市的状况，并且会参与open street map项目的用户群体应该大体是IT行业人员，果然深圳的“码农”比广州的多。

生活状况探索

通过三个小探索来从侧面观察该地区的生活环境和习惯，首先统计下该区域的设施状况。

In [ ]:

```
sqlite> SELECT tags.value, count(*) FROM
   ...> (SELECT value, key FROM nodes_tags
   ...>  UNION ALL
   ...>  SELECT value, key FROM ways_tags)
   ...> tags
   ...> WHERE tags.key = 'amenity'
   ...> GROUP BY tags.value
   ...> ORDER BY count(*) DESC
   ...> LIMIT 10;
```

```
school,767
parking,693
restaurant,633
bank,319
toilets,275
bus_station,250
fast_food,231
hospital,216
fuel,212
cafe,155
```

可以看到，该区域设施被标注最多的市学校，首先，该地区的教育资源相对丰富，并且学龄人口较多；另外，排第二位的设施是停车场，证明该区域的汽车数量比较多，大量的设施建设都需要考虑停车问题；由以上两个数据，可以推测，该区域的经济状况较好。

另外，通过统计cuisine（风味）项对该地区居民的饮食做一个调查。

In [ ]:

```
sqlite> SELECT tags.value, count(*) FROM
   ...> (SELECT value, key FROM nodes_tags
   ...>  UNION ALL
   ...>  SELECT value, key FROM ways_tags)
   ...> tags
   ...> WHERE tags.key = 'cuisine'
   ...> GROUP BY tags.value
   ...> ORDER BY count(*) DESC
   ...> LIMIT 10;
```

```
chinese,123
burger,62
chicken,32
coffee_shop,27
regional,12
japanese,9
pizza,8
sandwich,7
cantonese,6
american,5
```

在中国，中餐口味的餐馆最多可以理解，令我意外的是burger占据了第二的位置，看来，一方面归功于麦当劳和肯德基等的西式快餐的扩张，另外一方面也可以看出该区域居民相对开放，能够普遍接受西式饮食。

最后查看下生活的休闲设施状况。

In [ ]:

```
sqlite> SELECT tags.value, COUNT(*) FROM
   ...> (SELECT value, key FROM nodes_tags
   ...>  UNION ALL
   ...>  SELECT value, key FROM ways_tags)
   ...> tags
   ...> WHERE tags.key = 'leisure'
   ...> GROUP BY tags.value
   ...> ORDER BY COUNT(*) DESC
   ...> LIMIT 10;
```

```
pitch,1697
park,813
swimming_pool,201
garden,111
sports_centre,103
stadium,74
track,52
golf_course,37
playground,27
fitness_centre,17
```

看来该地区居民对球类运动比较感兴趣，也有可能是因为球场（pitch）建设相对简单导致球场设施最多。另外，排名前十的休闲设施中，和运动相关的就有六个（pitch ， swimming_pool ， sports_centre ， stadium ， golf_coures ， fitness_centre），该地区居民对运动的热情比较高，生活习惯比较健康。

一些其他数据状况

nodes 数量

In [ ]:

```
sqlite> SELECT COUNT(*) FROM nodes;
1624147
```

ways 数量

In [ ]:

```
sqlite> SELECT COUNT(*) FROM ways;
178282
```

用户数量

In [ ]:

```
sqlite> SELECT COUNT(DISTINCT(f.uid))
   ...> FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways)f;
1317
```

被标记最多的道路

In [ ]:

```
sqlite> SELECT tags.value, count(*) FROM
   ...> (SELECT value, key FROM nodes_tags UNION ALL
   ...>  SELECT value, key FROM ways_tags) tags
   ...> WHERE tags.key = 'street' AND tags.value IS NOT ""
   ...> GROUP BY tags.value
   ...> ORDER BY count(*)
   ...> DESC
   ...> limit 20;
```

```
"中山二路",14
"建设路",14
"人民南路",13
"北环大道",12
"深南大道",12
"罗芳路",12
"人民北路",11
"深南东路",10
"东风西路",8
"景田北街",8
"滨河大道",8
"香梅路",8
"南山大道",7
"天河路",7
"天润路",7
"春风路",7
"景田路",7
"百花一路",7
"迎春路",7
"南湖路",6
```

四、总结与建议

总结

通过对珠三角地图的数据清理和探索，可以描述出对该地状况的一个“侧身像”：珠三角区域经济较为发达，居民的生活条件优越，基础设施例如学校、医院和运动场所等足够完善；居民的文化观念相对开放，除了传统的中餐之外，西式餐饮也广受欢迎；不同的城市有着明显的职业人群，例如同为大城市，深圳从事IT相关人数就比广州多，并且城市发展并不是太均衡，广州深圳两个城市就占据了大量的标记数据；另外，珠三角的休闲场所以运动场所居多，说明该地区居民的生活习惯相对较好，对运动场所有着较高的需求。

同时，在数据的清理阶段，也发现了open street map项目的一些遗憾，例如录入格式的不统一，导致了许多数据无法阅读，或者数据含义不明确，从而丧失了大量的数据信息；另外，在地图地区特色化上并没有得到较好的支持，例如许多场景和地点都没有用中文标识。数据的不统一导致了后期数据清理工作太过繁琐和复杂，而且没有办法很好的保留数据，我在清理过程中就不得不丢弃大量的不明所以的数据，很多也许只有录入者明白，但是对于其他人的可读性太低。

建议

- 在录入地图时提供规范格式指引，如街道名称上，规定不许使用简写等：

好处：

1、会使得用户录入修改更加的规范，提高录入效率；

2、经过规范录入的数据，会大量减少数据清理的工作，更加利于日后的数据分析；

3、规范化的数据也更加具有可读性，对其他地图使用者来说也更加容易理解。

预期的问题：

1、存在抹杀个性化的可能，例如在某些特定的地址无法用常规拼写方式表达，这样可能是的部分数据更加难以读懂；

2、对于数据分析，在源头就进行了数据筛选是方便了工作，但是实际中可能真实存在的异常值有会被忽略，降低数据分析质量。

- 对每个国家的地图录入标准引入更多的本地化，例如对于中国地区，录入中文名为必选，英文名为备选，其他国家语言类似：

好处：

1、能够进一步的规范数据，提高数据的可读性；

2、在分析数据时候能够更加的简洁，并且对于需求量较大的当地分析师来说，可以结合文化习惯等进行分析，提高效率。

预期的问题：

1、数据的复杂程度会提升；

2、对于非本国分析师而言，数据将会更加难以理解，通用程度会降低。

五、数据概况

zhusanjiao.osm : 341 MB 珠三角完整osm文件
SAMPLE_FILE.osm : 9.74 MB 珠三角采样osm文件
nodes.csv : 132 MB nodes清理后信息文件
nodes_tags.csv : 3.67 MB nodes_tags清理后信息文件
ways.csv : 10.4 MB ways清理后信息文件
ways_tags.csv : 45.6 MB ways_tags清理后信息文件
ways_nodes.cv : 13.5 MB ways_nodes清理后信息文件