

OpenStreetMap 数据案例研究

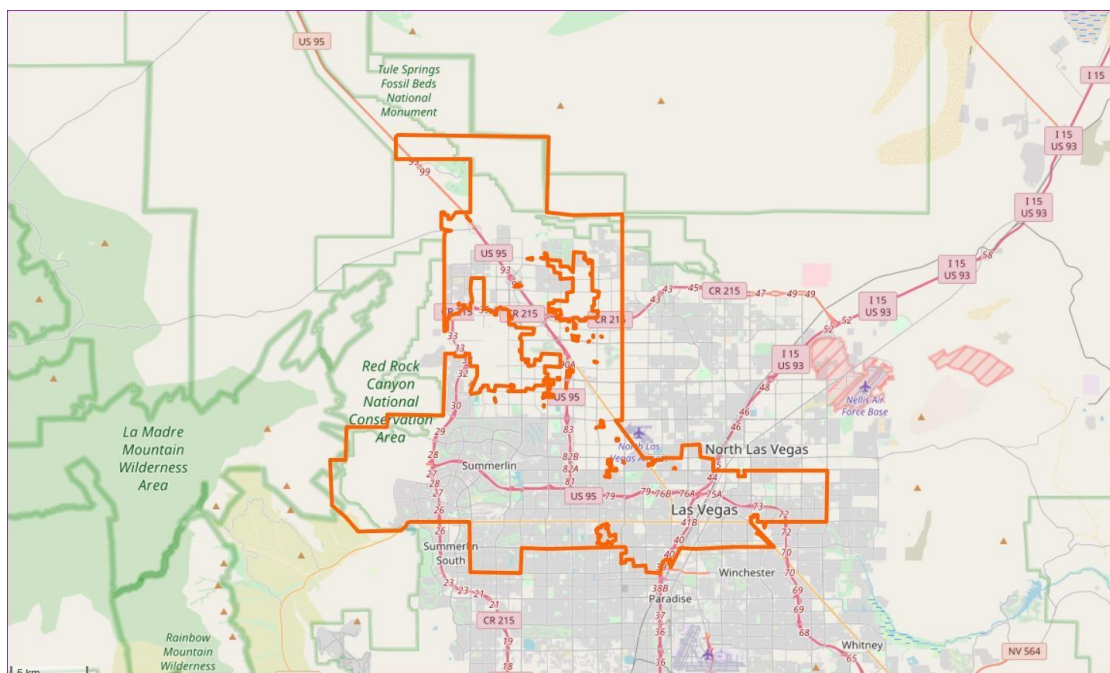
地图区域

拉斯维加斯，内华达州

<http://www.openstreetmap.org/relation/170117>

<https://mapzen.com/data/metro->

extracts/metro/lasvegas_nevada/



拉斯维加斯是美国内华达州的最大城市，以赌博业为中心的庞大的旅游、购物、度假产业而著名，是世界知名的度假圣地之一。

地图中遇到的问题

- 1、防止出现中文字符问题，遂采用美国地图，文化差异比较大。
- 2、地图中街道名称不完善，很多是缩写形式：

```
In [13]: #查看所有缩写类型
st_types = audit(OSMFILE)
pprint.pprint(dict(st_types))

'Ave': set(['Chandler Ave',
            'E Cheyenne Ave',
            'E Sahara Ave',
            'E Tropicana Ave',
            'East Tropicana Ave',
            'S Eastern Ave',
            'S. Eastern Ave',
            'Sunset River Ave',
            'W Cactus Ave',
            'W Sahara Ave',
            'W Twain Ave',
            'W Washington Ave',
            'W. Sahara Ave']),
'Ave.': set(['200 Hoover Ave.',
             '6601 W. Twain Ave.',
             'East Twain Ave.',
             'Glendale Ave.',
             'Hoover Ave.',
             'W. Arby Ave.',
             'West Sahara Ave'])
```

使用 mapping 字典配合 update_name 函数来转换街道名称：

```
#缩写与常用街道名称转换表
mapping = {
    "Ave": "Avenue",
    "Ave.": "Avenue",
    "AVE": "Avenue",
    "AVE.": "Avenue",
    "ave": "Avenue",
    "Blvd": "Boulevard",
    "Blvd.": "Boulevard",
    "blvd": "Boulevard",
```

```
#将缩写类型替换为完整类型
def update_name(name, mapping):
    #print mapping.values()
    shortname = mapping.keys()
    for word in shortname:
        if word in name:
            name = name.replace(word,mapping[word])
    return name
```

转换效果：

West Sahara Ave.	=>	West Squareahara Avenuenue
East Twain Ave.	=>	East Twain Avenuenue
Hoover Ave.	=>	Hoover Avenuenue
Glendale Ave.	=>	Glendale Avenuenue
200 Hoover Ave.	=>	200 Hoover Avenuenue

3、地图中电话号码格式很不规范，有空格和特殊

符号：

1-702-836-3278', '17022511600', '(702) 240-6366', '702-331-4466', '702-5000', '+1 (702) 714-0586', '7024572541', '7024572706', '7024325869', '563804', '17028399221', '17028780044', '17026582112', '7026582144', '1727', '(702) 750-2329', '702-320-0703', '702-441-0719', '702-565-9727', '2-222-0907', '17023850838', '17022278826', '17023627283', '1702282563', '240', '+1 702 5882404', '702-868-9096', '+1 702 388 2220', '+1 702 3800', '(702) 732-7440', '(702) 671-0001', '(702) 385-2662', '(702) 3992) 949-2583', '(702) 388-8338', '(866) 983-4279', '(702) 776-7770', '

使用正则表达式配合判断语句进行清洗：

```
#正则表达式
phone_num_re = re.compile(r'\+1\s\d{3}\s\d{3}\s\d{4}')
#完善电话号码
def update_phone(phone_num):
    m = phone_num_re.match(phone_num)

    if m is None:

        if "-" in phone_num:
            phone_num = re.sub("-", " ", phone_num)
        if "." in phone_num:
            phone_num = re.sub(".", "", phone_num)
        if "(" in phone_num or ")" in phone_num:
            phone_num = re.sub("[()]", "", phone_num)

        elif re.match(r'\d{10}', phone_num) is not None:
            phone_num = phone_num[:3] + " " + phone_num[3:6] + " " + phone_num[6:]
        elif re.match(r'\d{11}', phone_num) is not None:
            phone_num = phone_num[:1] + " " + phone_num[1:4] + " " + phone_num[4:7] + " " + phone_num[7:]
        elif re.match(r'\d{3}\d{3}\d{4}', phone_num) is not None:
            phone_num = "+1 " + phone_num

        elif re.match(r'1\s\d{3}\s\d{3}\s\d{4}', phone_num) is not None:
            phone_num = "+" + phone_num
        elif re.match(r'\+1\s\d{3}\s\d{7}', phone_num) is not None:
            phone_num = phone_num[:8] + phone_num[8:11] + " " + phone_num[11:]

        elif sum(c.isdigit() for c in phone_num) < 10:
            return None








    return phone_num
```

清洗效果：

```
702 435 6758
702 462 9500
702 685 9595
170 286 97440
702 574 2113
702 455 7522
702 451 0100
180 078 61000
702 778 4444
702 933 0775
702 690 9000
702 641 7590
702 383 2859
170 268 57712
```

数据概述和其他想法

文件大小

	las-vegas_nevada.osm	2017/12/30 11:41	OSM 文件	231,728 KB
	las-vegas_nevada.db	2017/12/31 11:02	Data Base File	153,225 KB
	nodes	2017/12/30 14:44	Microsoft Excel ...	91,001 KB
	ways_nodes	2017/12/30 14:44	Microsoft Excel ...	32,093 KB
	ways_tags	2017/12/30 14:44	Microsoft Excel ...	15,195 KB
	ways	2017/12/30 14:44	Microsoft Excel ...	7,027 KB
	nodes_tags	2017/12/30 14:44	Microsoft Excel ...	2,393 KB

节点数

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

1099352

道路数

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

118370

唯一用户数

```
select count(e.uid)
```

```
from (select uid from nodes union select uid from ways) as  
e;
```

1134

前 10 贡献用户

```
select e.user, count(*) as num
from (select user from nodes union all select user from
ways) as e group by e.user order by num desc
limit 10;
```

	user	num
1	alimamo	251121
2	tomthepom	121046
3	woodpeck_fixbot	69687
4	alecdhuse	66429
5	abellao	55041
6	Tom_Holland	51525
7	gMitchelID	44485
8	robgeb	40624
9	TheDutchMan13	38620
10	MojaveNC	27397

仅出现一次的用户数

```
select count(*)
from (select e.user, count(*) as num
      from(select user from nodes union all select user from
ways) as e group by e.user having num = 1) u;
```

关于数据集的其他想法

截至 2017 年 12 月 30 日，拉斯维加斯地区数据更新于 9 天之前，杭州地区数据更新于 2 个月之前，三四线城市数据更加稀少。对于普通老百姓甚至是业内人士来说，大家的积极性都不算很高。

建议一：官方利用地图数据做一些好的产品。

好处：给社会带来便捷的同时，也传播了自身的价值，吸引更多贡献者参与分享地理信息，进入良性循环。

坏处：投入较多的人力物力，增加公司成本。

建议二：降低利用和贡献门槛。

好处：扩大潜在贡献者人数，有利于地图社区繁荣。

坏处：需要与广大贡献者进行深入沟通来改善产品体验，增加时间成本

其他数据调查

前 10 名出现的设施

```
select value, count(*) as num from
```

```
nodes_tags
```

where key='amenity' group

by value

order by num desc limit

10;

	value	num
1	restaurant	487
2	place_of_worship	292
3	fuel	284
4	fast_food	281
5	fountain	266
6	school	204
7	shelter	122
8	toilets	87
9	cafe	82
10	bar	80

最大的宗教

select nodes_tags.value, count(*) as num from

nodes_tags

join (select distinct(id) from nodes_tags where

value='place_of_worship') as i on

nodes_tags.id=i.id where


```
nodes_tags.key='religion' group by
```

```
nodes_tags.value order by num desc limit 1;
```

```
christian
```

最受欢迎的菜肴

```
select nodes_tags.value, count(*) as num from
```

```
nodes_tags
```

```
    join(select      distinct(id)from      nodes_tags      where  
value='restaurant') as i    on nodes_tags.id=i.id where  
nodes_tags.key='cuisine' group by nodes_tags.value order by  
num desc;
```

	value	num
1	mexican	41
2	pizza	31
3	american	21
4	italian	18
5	steak_house	16
6	burger	15
7	chinese	13
8	asian	9
9	japanese	9
10	buffet	7

结论

OSM 是一款由网络大众共同打造的免费开源、可编辑的地图服务，但是往往由少数人贡献了大部分数据，本次数据集有 1134 名用户参与贡献，这对于一个现代化城市人口来说，还远远不够。当然由于只分析了一个旅游城市，理解会有片面性。

另外，度假胜地拉斯维加斯，公共设施中酒店、教堂、加油站排前三位，大家最爱吃墨西哥菜。