OpenStreetMap 数据分析报告

一、 区域

我选择的城市是上海,因为上海是中国的经济中心,发展速度非常迅猛,所以我对搜集到的数据很感兴趣,想通过数据了解下上海的地理位置信息和更多的发展和未来的发展潜力

二、 初次查看文件所遇到的问题

在初次检验文件时候发现有些 tags 的节点会有异常字符

```
<tag k="name" v="Metro Line 12: 金海路 =&#62; 七莘路"/>
<tag k="name:en" v="Metro Line 12: Jinhai Road =&#62; Qixin Road"/>
<tag k="name" v="Metro Line 9: 杨高中路 =&#62; 松江南站"/>
```

三、 问题数据处理

我的做法是把节点中的异常数据删除

```
def update(k,v):
    if '=&#' in v:
        return v.replace('=&#','')
```

其他部分的数据未发现异常,所以在下一步我将数据导入到数据库

四、 导入数据到数据库并分析数据

准备导入数据的时候发现导入的数据有空行的存在,所以在写文件的时候要将 w 改成 wb 可以解决这个问题

1. 文件大小

shanghai.osm 源文件	278MB
shanghai.db 数据库文件	178MB
nodes.csv	96.4MB
nodes_tags.csv	3.34MB
ways.csv	9.52MB
ways_nodes.csv	34.4MB
ways_tags.csv	13.3MB

2. 节点数目

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

1230544

3. 路径数目

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

169557

4. 唯一的用户数量

```
select count(distinct(e.uid))
from (select uid from nodes union all select uid from ways) e;
```

1686

5. 排名前 10 的贡献者

```
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;
```

"aighes" "121974"

"zzcolin" "82496"

"xiaotu" "82016"

"Koalberry" "73657"

"Xylem" "70030"

"duxxa" "67485"

"yangfl" "61822"

"Austin Zhu" "59172"

"alberth2" "45348"

"HWST" "41550"

6. 只贡献一次的用户

```
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)
```

430

7. 人们喜爱的前 10 类饮食类型

```
select value,count(*) as num
from nodes_tags
where key = 'amenity'
group by value
order by num desc
limit 10;
```

```
"restaurant"  "878"
"bank" "371"
"cafe" "317"
"toilets"  "249"
"fast_food"   "248"
"bicycle_rental"  "212"
"bar"  "110"
"fuel" "109"
"atm"  "103"
"parking"  "92"
```

8. 人们喜爱的前 10 类食物种类

```
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
limit 10
```

```
"chinese"  "91"
"burger"   "14"
"japanese" "14"
```

"italian" "13"

"noodles" "10"

"asian" "9"

"international" "7"

"mexican" "7"

"pizza" "7"

"regional""7"

9. 需要进一步完善的数据

select e.id,e.key,e.value,count(*) as num

from (select id,key,value from nodes_tags union all select id,key,value from ways_tags) e

where e.key = 'fixme'

group by e.value

order by num desc;

返回有 121 行数据需要进行修正,相当于总量来看还是不多的根据返回前 4 的值来看

"414252807" "fixme" "fixme!" "39"

"525376138" "fixme" "extend" "16"

"403495123" "fixme" "I cannot determine if this road is named ""长康路"" (Changkang Road) or ""康文路"" (Kangwen Road)." "10"

"442045971" "fixme" "continue""8"

fixme 表示该数据需要修改,一共有 39 条需要修改

extend 表示该数据需要进一步完善,目前有 16 条需要进一步补充

I can not….表示该数据信息不明,目前有 10 条,需要进一步确定加以修改

continue 表示已解决的的点,目前有 8 条,等发现者补充上即可

10. 本地最大的宗教

select nodes_tags.value,count(*) as num

```
from nodes_tags join
(select distinct(id) from nodes_tags where value =
'place_of_worship') i
on nodes_tags.id = i.id
where nodes_tags.key = 'religion'
group by nodes_tags.value
order by num desc
limit 10;
```

"christian" "8"

"buddhist" "5"

可以看出在上海市宗教信仰的数量不是很多,其中基督教和佛教是主流宗教

11. 废弃的地点

```
select e.id,e.key,e.value,count(*) as num
from (select id,key,value from nodes_tags union all select
id,key,value from ways_tags) e
where e.key = 'description'
group by e.id
```

一共返回 90 条数据,通过对这些数据进行查看,发现这些地点都是临时成立的地区,在活动或者节目结束之后没有改成原先的名字,导致了数据有冗余

12. 生活设施

```
select e.id,e.key,e.value,count(*) as num
from (select id,key,value from nodes_tags union all select
id,key,value from ways_tags) e
where e.key = 'name' and e.value like '%医院'
group by value;
```

这里我查询了医院,小学,中学,大学的生活设施,,只需要更改 e.value 即可,具体结果如下:

医院	84 所

小学	119 所
中学	171 所
大学	19 所

从生活设施的数量上来看,上海市的生活设施还是非常富足的

13. 文化设施

图书馆:

```
select e.id,e.key,e.value,count(*) as num
from (select id,key,value from nodes_tags union all select
id,key,value from ways_tags) e
where e.key = 'name' and e.value like '%图书馆'
group by value
```

电影院:

```
select e.id,e.key,e.value,count(*) as num
from (select id,key,value from nodes_tags union all select
id,key,value from ways_tags) e
where e.key = 'name' and
(e.value like '%影院' or e.value like '%影城')
group by value
```

这里我查询了图书馆和电影院的数据,结果如下

图书馆	33
电影院	24

从文化设施的数量上来看,上海市的精神文明建设属于全国的前列

14. 长度前 10 的地名

```
select e.id ,e.value,length(e.value)
from (select id,key,value from nodes_tags union all select
id,key,value from ways_tags) e
where e.key = 'alt_name'
group by value
order by length(e.value) desc
limit 2
```

从查询的结果来看上海最长的地名只有 35 个字符,说明上海市的地点命名有着很好的规范,没有出现过长的地名

"4762962773" "The Lion King; Mickey Film Festival" "35"

"114518336" "Internation Students Dormitory" "30"

五、 建议与结语

建议:

1. 修改标签中的特殊字符,同时删除 k=address 中的省市和国家

好处:

1) 可以维护数据的一致性

2) 因为省,市数据大量的重复,减少数据的冗余

预期的问题:

1) 联系当时的记录员进行更改,可能会花费较多的时间去核对

2) 在修改的过程中,地址,地名还有可能会发生变动

预期结论:

通过联系记录员并将特殊的字符进行更改,可以使数据更具有实时性和一致性,虽然会花费联络,修改的成本但风险远远小于收益

2. 将标签中 k=’fixme’的值发布到网上，让网友知道有哪些点是需要修改的并且让其附近的网友修改。

好处:

1) 可以维护数据的完整性

2) 可以让附近的人员将这些数据更新到网上,更具有时效性

预期的问题:

1) 如何去激励修改数据的人员

2) 如何审核数据的正确性

预期结论:

通过调动网友的积极性来填充待修改的数据,我认为可以将贡献度高的网友的放在首页上表示感谢,同时要注意数据的格式,否则有可能变成脏数据,但是修改付出的成本还是小于收益

结语:

1. 在对这些数据进行处理之后,发现部分数据还有有重复,缺失的情况,这是手工填写错误导致的,有些数据格式异常可以在代码中进行修改,有些异常数据只能手动进行修改,这使得我的工作难度增加
2. 此次练习也很好的锻炼了我数据清洗和分析的能力,将我从一个新手慢慢引入了数据分析的大门,我知道如何取从网上下载数据,如何将数据导入数据库,然后根据 SQL 语句来找出我想知道的结果
3. 本次实验中,也存在这很多不足之处,比如重复的数据没有清理,清洗的过程过于简单,在解析 xml 文件->csv 文件时会出现过于缓慢的情况,这是由于 SAX 解析导致的,有没有更快速的解析方法呢?