P3 OpenStreetMap 数据整理及探索——hawli

1.选择地图

数据来源如下：

- openstreetmap地址
- 选取上海区域下载

地图描述及选择理由：

该地图为上海地区，地处长江入海口，是长江经济带的龙头城市，隔东中国海与日本九州岛相望，南濒杭州湾，北、西与江苏、浙江两省相接。

上海是我曾经工作的区域之一，也是考虑定居的城市之一，之前在那里半年并没有对这个城市有深入的了解，希望通过这次整理能有进一步的了解。

2.地图数据存在的问题：

- 该地区为下载时为上海数据，但因上海与江苏和浙江相接，地图实际上包含着大量江苏和浙江城市的数据，如整理，需剔除上海以外的城市数据，或改为长三角地区数据。

```
<tag k="network" v="南通市区公交"/>
<tag k="is_in" v="China; Zhejiang; Hangzhou"/>
<tag k="name" v="杭州江南实验学校"/>
```

- 该地图为中国地区，很多地名都非汉字显示，而有同样种类的名字，比如哈根达斯，有的名字用中文显示有的用英文显示，有的连名字都没有，而且存在错误，未体现一致性，建议名字统一为中文，其他语言名字用其它标签。

```
<tag k="name" v="杭州联合银行 Hangzhou United Bank"/>
<tag k="name" v="KFC(东门店)"/>
<tag k="name" v="Ningbo Women & Children Hospital"/>
<tag k="name" v="Starbucks"/>
```

- 该地区标记的名字中拼音类型有部分未字母，无声调，应统一改为拼音格式。

```
<tag k="name:zh_pinyin" v="Liang'an Kafei"/>
<tag k="name:zh_pinyin" v="Yinzhou Qu"/>
<tag k="name:zh_pinyin" v="Xinqiao Zhen"/>
```

`<tag k="name:zh_pinyin" v="Dongba Zhen"/>`

- 该地区标记的种类较多，但数据不齐全，每一个用户贡献的数据都非统一结构，建议贡献数据的时候有统一的标签结构和标准，未在架构内的信息采用其它标签。

3.数据概述

3.1 文件大小

```
shanghai_china.osm     769.5MB
shanghai4.osm          11.1MB
shanghai.db            5.8MB
nodes.csv              4.3MB
nodes_tags.csv         141KB
ways.csv               382KB
ways_tags.csv          461KB
ways_nodes.csv         1.5MB
```

3.2 唯一用户的数量

```
sqlite> select count(distinct(e.uid))
from (select uid from nodes union all select uid from ways) e;
889
```

3.3 节点和途径的数量

```
sqlite> SELECT COUNT(*) FROM nodes;
51893

sqlite> SELECT COUNT(*) FROM ways;
6481
```

3.4 节点中前十个分类的数量

```
sqlite> select value,count(*)as num
from nodes_tags where key ='amenity'
group by value order by num desc limit 10;
bicycle_rental|35
restaurant|15
bank|8
fast_food|7
fuel|7
parking|7
cafe|6
school|6
toilets|6
atm|4
```

3.5 前十名贡献者的名字

```
sqlite> 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;
Chen Jia|10011
aighes|2630
xiaotu|2529
Austin Zhu|2336
katpatuka|2040
XBear|1801
Peng-Chung|1611
yangfl|1607
Holywindon|1443
dkt|1376
```

3.6 公路附近的标签类型及数量

```
sqlite> select tags.value,count(*)as count
from(select*from nodes_tags union all
select*from ways_tags)tags
where tags.key = 'highway'
group by tags.value
order by count desc;
residential|636
service|451
tertiary|450
unclassified|331
secondary|321
primary|234
motorway|177
footway|152
motorway_link|122
bus_stop|87
crossing|66
traffic_signals|60
trunk|52
primary_link|48
trunk_link|48
track|37
path|34
secondary_link|26
pedestrian|22
construction|20
cycleway|19
motorway_junction|14
road|14
living_street|13
steps|13
tertiary_link|10
platform|3
turning_circle|2
```

3.7 地方的类型及数量

```
sqlite> select tags.value,count(*)as count
from(select*from nodes_tags union all select*from ways_tags)tags where tags.key='p
lace'
group by tags.value
order by count desc;
village|18
town|14
suburb|7
hamlet|2
island|2
islet|1
locality|1
```

3.8 旅游相关类型及数量

```
sqlite> select tags.value,count(*)as count
from(select*from nodes_tags union all select*from ways_tags)tags
where tags.key='tourism'
group by tags.value
order by count desc;
hotel|12
attraction|6
museum|5
viewpoint|3
chalet|1
hostel|1
information|1
```

4.关于数据集的其它想法

4.1 改进建议：

- 定制地图边界标准，所有区域的数据可以统一分类，减少临界区域数据混淆的情况。
- 定制统一标签标准，以国家为对象，定制标签架构，使用者添加数据的时候根据标签来进行数据增添。
- 为了鼓励更多的使用者贡献数据，可组织数据集活动与交友，比方在上海区域，让使用者集中一起，分成n组，每组选择一个种类为目标，比方旅游，酒店，餐饮等，然后选择某一片区域范围进行为期一天的采集活动，边采集边上传数据，最后根据数据的数量和质量进行评比，选出最佳贡献组，给予“地图之王”称号，可增加大家的积极性。

4.2 获得收益及风险：

通过前期结构性的调整，可提高数据录入的质量，减少后期整理的时间，有利于保持数据的一致性，也方便贡献者有指导有目的地提供数据，但提前确定结构化框架，会降低灵活性，也容易限制贡献者的想象空间，

妨碍提供更加完善的数据及解决方案。 通过活动的形式，既能增加贡献者的积极性，贡献者在这个过程中可以选择自己贡献的区域及活动类型，充分体现了其主观能动性，而采集的过程与队友的协同作战，既满足了贡献者的社交需求，又感受到团队的力量，最后在贡献数据和后期评比的过程中，感受到成就感与自信，让贡献者感受到做这件事情的意义和使命，从而进一步为数据集的发展做贡献。然而采用活动的方式也有其弊端，首先需要花费一定的人力及物力，而对于活动目的来说，我们采取这种形式是为了增加贡献者的积极性，但是也可能带来其他的问题，比方评比和排名，容易让一部分靠前的人更有积极性，而靠后的人反而打消了积极性，存粹提供数据的内部动机转为外部动机驱使，反而违反了我们做活动的目的。 因此，在做调整之前，应充分考虑其所带来的收益及风险，以及收益和风险对我们所要达到的目标的贡献程度，然后采取适当的改进方式。

5.参考资料

参考网站书籍，论坛等：

中文转拼音

中文转unicode：

在 Python 中解析并修改XML内容的方法

正则表达式在线转换工具

schema

拼音也要转unicode