深圳地图探索报告

报告结构

本报告包含以下内容的分析结果,具体代码请见 github 中另一个 ipynb 文件

- 1.清洗数据
- 2.用 SQL 自由查询及探索
- 3.根据探索结果提出地图改进建议

项目目标

探索深圳地图数据集,并得出有建设性的见解。

项目流程:

清洗深圳地图上的街道名,将 XML 格式的数据写入 CSV,再将 CSV 导入数据库以便能用 SQL 进行查询和探索,最后根据探索结果为改进地图提出建议。

一、清洗数据

本项目仅针对地图数据里的街道类型进行清洗,首先观察数据,在我浏览街道类型的过程中,发现了几个问题,归纳如下:

- 1.街道类型不统一,对于同一种街道类型有不同的写法,如: Av, Ave, Avenue 为同一意思,只不过前两个是简写,为了便于研究,应将其统一
 - 2.街道名变量当中有些并非街道名,而是商场或建筑名称,如 CocoPark
 - 3.街道名既有中文,也有英文,也有中英混杂,该如何处理?

下面针对以上问题一一解答

问题 1: 我继续观察了所有的街道类型,并建立一个规范化模板,比如将 Av,Ave,Avenue 统一为 Avenue,将 Rd,Road,road,Lu 统一 Road,这里我用了 python 中的 replace 方法来实现。变更后的街道名举例如下:

正风路 Zhengfeng Rd => 正风路 Zhengfeng Road Hongbao Lu => Hongbao Road 坂雪岗大道 Bǎnxuě Gǎng Av => 坂雪岗大道 Bǎnxuě Gǎng Avenue

问题 2: 这种情况应该是数据没有按类别录入,我浏览了整个数据集,发现只有几个个例, 考虑到对研究影响不大,故不作处理。

问题 3:

首先,考虑到实施难度,这里不进行中英文统一,即允许街道名既有中文,也有英文。

- 如果街道名为纯英文,则按问题1的方式处理。
- 如果街道名为纯中文,则不存在街道类型为简写的问题,不作处理。
- 如果街道名为中英文夹杂,观察得知英文部分一般为中文街道的翻译,且跟在中文后面, 因为表达的是同一个意思,故对研究影响不大,同样只需要按问题1的方式来处理。

二、探索数据集

这里是自由探索环节,包括该区域数据的一些基本概况:维护人数、节点数、宾馆数等等。

维护人数

```
sqlite> SELECT COUNT(DISTINCT(uid))
...> FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways) AS n;
515
```

节点数

```
sqlite> SELECT COUNT(*)
...> FROM nodes;
251925
```

途径数

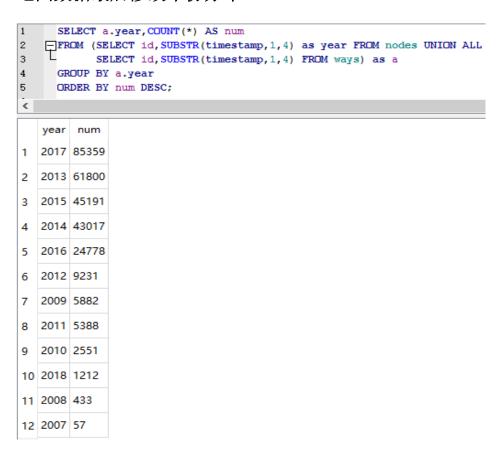
```
sqlite> SELECT COUNT(*)
...> FROM ways;
32974
```

宾馆数

```
1 SELECT count(distinct id)
2 FROM nodes_tags
3 WHERE value like '%酒店%' OR value like '%hotel%';
4
5
6
7
count(distinct id)
1 102
```

所选区域纬度范围

地图数据最后修改年份分布



可以看到最后修改年份越近,数据记录大致也越多,2018年才开始,所以记录数排在后面。

三、改进建议

建议 1:

唯一用户数相对地图的节点数和途径数还是太少,可能会导致维护压力大,容易产生局限性或**忽略一些细节的地方,所以建议引进更多专业的人参与到地图的维护中,**不同背景的人加入有助于提升地图数据的全面性和准确性。

潜在问题:

- 调动参与积极性有赖于建立一套良好的激励制度,因此可能增加一定运营成本。
- 需要审核新进维护人员的素质,否则可能导致地图数据质量下降。

建议 2:

对于最后修改年份较早的记录,存在老旧数据的可能性更高,且由于数量相对最近修改的少很多,因此有必要增加审查次数,以便及时发现数据中需要更新的地方。

潜在问题:

审查过程中可能存在因年代差异导致的表述差异(如饭店改名),需要进一步考证,增加了审查成本,因此要设定合理的审查频率。