Project 3: OpenStreetMap Data Wrangling with SQL

地图区域:格拉斯哥。
选择理由：我喜欢苏格兰，所以选择苏格兰的城市。

数据审计

独立标签数量：
使用 cElementTree 分析 osm 文件后得出的标签数量：
'nd': 562843,
'node': 493780,
'tag': 320815,
'way': 80094,
'member': 24036,
'relation': 1010,
'bounds': 1,
'osm': 1

检查标签 k 值：
对每个 k 值使用正则，将结果分成 4 组：
'lower': 276700 仅包含小写字母且有效的标记
'lower_colon': 23397 名称中有冒号的其他有效标记
'problemchars': 0 字符存在问题的标记。
'other': 20718 不属于其他 3 类。

地图中遇到的问题

街道地址不符：
1. 使用缩写
2. 首字母大小写混用
3. 拼写错误

有以下修正：
```
Springbank Sreet => Springbank Street
Ardgay St => Ardgay Street
downi => Downy
Garfield Strret => Garfield Street
pollokshaws road => Pollokshaws Road
Marihill road => Marihill Road
canal bank => Canal Bank
```

数据概述

文件大小：

Glasgow.osm 107mb

Glasgow.db 62mb

nodes.csv : 41 MB

nodes_tags.csv 6.1 MB

ways.csv : 4.6 MB

ways_nodes.csv 13.2 MB

ways_tags.csv 6.5 MB

节点数量：

sqlite> SELECT count(*) from nodes;

493779

道路数量：

sqlite> select count(*) from ways;

80094

用户数量：

sqlite> select count(distinct(uid)) from

...> (select uid from nodes union all

...> select uid from ways);

806

贡献地图最多的 10 名用户：

sqlite>select user,count(user) from

...> (select user from nodes union all select user from ways)

...> group by user order by count(user) desc

...> limit 10

drnoble|167694

cupofcoffee|62723

addavies|62252

crossmyloof|34533

Central America|29846

Ossian Lore|23158

Hawkeye|21078

i am tiz|19596

c3pol|13797

andypreece|10305

只提交过 1 次的用户数量：

```
sqlite> 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) u;
195
```

10 大便利设施：

```
sqlite> select value,count() from nodes_tags
...> where key='amenity' group by value
...> order by count() desc limit 10;
bicycle_parking|535
post_box|380
fast_food|294
restaurant|273
cafe|244
pub|238
telephone|157
bench|130
atm|118
bar|96
```

关于数据集的其他想法

从数据上可以看出，大部分地图数据都是由少数人提供的。贡献地图最多的 drnoble 提供了接近 30%的数据量，同时也比 2-10 名提供者的总和都要多。这带来了数据不准确以及更新滞后的问题。同时由于也缺乏一些景点和热门地点等有用信息。同时由于手工输入，不可避免的也带来一些错误。

所以可以做出以下改进：

建议 1：想办法提升用户活跃性，使用用户激励手段，比如积分，贡献排行榜，用户特权等。

好处：

增加数据量，丰富地图数据

提升纠错性，不会因为个人失误造成数据错误。

问题：

过多的参与者导致数据格式不统一

过程比较复杂，成本较高

建议 2：增加地图信息，比如景点或热点

好处：

吸引更多用户。

问题：

专业性下降，和其他地图有同质化倾向。