

## OpenStreetMap Data Case Study

### Map Area

地图来源 [https://mapzen.com/data/metro-extracts/metro/calgary\\_canada/](https://mapzen.com/data/metro-extracts/metro/calgary_canada/)

### Calgary\_canada

Calgary, 中文名卡尔加里。是一座位于加拿大阿尔伯塔省南部洛矶山脉的城市。多次被评为“世界上最干净的城市”。1988 年的冬奥会在这里举行。我对这座城市感兴趣，所有选取了这份地图。

## 在地图中遇到的问题

在选取 Calgary\_canada 地图的十分之一的样本地图后，我注意到了以下问题：

1. 街道名称简写的情况较多（如“1 Avenue NE”）
2. 电话的格式不统一，加拿大的代码编号为+1。大部分电话前有“+1”，少部分没有。
3. 加拿大的邮编有 6 个字符，第 1, 3, 5 为字母，第 2, 4, 6 位为数字。中间用空格分开。该样本中少量邮编没有用空格分开。

### 街道名称简写的情况较多

在审查了街道名称后，发现街道名称简写的情况较多，给不熟悉简写的用户阅读带来不便。所以将它们更正为使用以下函数在审计中的各自映射：

```
def update_name(name, mapping):  
    for word in mapping:  
        if name.endswith(word):  
            return name.replace(word, mapping[word])  
    return name
```

着更新了有问题的街道名称中的字符串，例如：

“1 Avenue NE” 变成 “1 Avenue Northwest”

### 整理电话格式

电话的格式不统一，加拿大的代码编号为+1。大部分电话前有“+1”，少部分没有。所有给没有区号的电话号码加了“+1”。

### 整理邮编号码

加拿大的邮编有 6 个字符，第 1, 3, 5 为字母，第 2, 4, 6 位为数字。中间用空格分开。该样本中少量邮编没有用空格分开。统一了邮编格式，给少量没有的加了空格。

## 数据概述

本节包含数据集的基本统计信息，使用 SQL 查询得出，为额外的想法做准备。

文件大小

```
calg.osm-----20217KB
calgary_canada.osm-----199649KB
nodes.csv-----73380KB
nodes_tags.csv-----3198KB
ways.csv-----7124KB
ways_nodes.csv-----25655KB
ways_tags.csv-----14797KB
```

*nodes 的数量*

```
4 SELECT COUNT(*) FROM nodes
5
6
```

	COUNT(*)
1	904170

*ways 的数量*

```
1 SELECT COUNT(*) FROM ways
```

	COUNT(*)
1	122436

*对地图做贡献的唯一用户的数量*

```
1 SELECT COUNT(DISTINCT(e.uid))
2 FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways) e;
```

	COUNT(DISTINCT(e.uid))
1	1840

*前十位贡献最大的用户*

```
1 SELECT e.user,COUNT(*) as num
2 FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e
3 GROUP BY e.user
4 ORDER BY num desc
5 LIMIT 10
```

	user	num
1	abDoug	278619
2	sbrown	160699
3	Zippanova	56281
4	markbegbie	38879
5	JamesBadger	32991
6	dbo-osm	32137
7	hoserab	31257
8	kor	20859
9	Sundance	17118

仅出现过 1 次编辑的用户

```

1 SELECT COUNT(*)
2 FROM
3 (SELECT e.user,COUNT(*) as num
4  FROM (SELECT user FROM nodes UNION ALL SELECT user
5  GROUP BY e.user
6  HAVING num=1) u

```

	COUNT(*)
1	236

## 额外的想法

1. 贡献最大的用户（abDoug）做出了 27.14%的编辑工作。
2. 前三名用户做出了 48.28%的编辑工作。
3. 有 12.83%的用户仅有一次编辑记录。

考虑到这样的情况，可以用游戏化的方式激励参与者。通过目标，规则，反馈系统，自愿参加等游戏化的方式，将编辑地图任务变成一个多人参与的，正反馈及时，有荣誉奖励的任务。

建议 1：使用数据时，较少使用只有一次编辑记录的用户上传的数据。

好处：仅有一次编辑记录的用户上传的数据，可能包含错误和不规范数据。

预期的问题：可能会降低用户的积极性和活跃度。

建议 2: 对于通用格式的数据, 进行一定的书写规则引导

好处： 规范数据，提高数据一致性。

预期的问题：各地通用格式的书写规则较多，可能不容易一一举例和规范。

建议 3: 游戏化的方式鼓励用户参与到地图编辑工作来。如论坛, 荣誉奖励等。

好处： 地图覆盖面更广，使用户更活跃

预期的问题： 可能需要投入人力维护论坛活跃性。

## 关于数据集的其他想法

### 前8名的休闲地方

在 Calgary, 最受欢迎的前三的休闲场所, 分别为: 操场, 野餐地, 运动中心。

```
1 SELECT value,COUNT(*) as num
2 FROM nodes_tags
3 WHERE key='leisure'
4 GROUP BY value
5 ORDER BY num desc
6 LIMIT 8
```

	value	num
1	playground	298
2	picnic_table	87
3	sports_centre	48
4	pitch	25
5	fitness_centre	11
6	park	10
7	slipway	5
8	swimming_pool	5

Calgary 里, 90%以上是阔叶树, 少于 10%为针叶树。

```
1 SELECT nodes.tags.value COUNT(*) AS num
```

在 Calgary 里，最多的建筑物是 detached，即以家庭成员为单位居住的独立住宅。其次才

```

1 SELECT value,COUNT(*) as num
2 FROM nodes_tags
3 WHERE key='building'
4 GROUP BY value
5 ORDER BY num desc
6 LIMIT 10

```

	value	num
1	detached	1843
2	house	309
3	yes	8
4	apartments	2
5	office	2
6	retail	2
7	university	2
8	pavilion	1

注：SQL 截图来源——使用了 DB.Browser.for.SQLite

参考资料：

1.SQL 示例项目

[https://s3.cn-north-1.amazonaws.com.cn/static-documents/nd002/sample\\_project\\_en.pdf](https://s3.cn-north-1.amazonaws.com.cn/static-documents/nd002/sample_project_en.pdf)

4. 优达学城论坛

5. OSM Map features

[https://wiki.openstreetmap.org/wiki/Map\\_Features#Medical\\_Rescue](https://wiki.openstreetmap.org/wiki/Map_Features#Medical_Rescue)