整理 OpenStreetMap 数据

Map Area: Beijing

1. 地图数据中存在的问题

1.1 数据中的 type 字段可能包含除 node 和 way 之外的其他值

数据初步处理完成后，为了验证 type 字段仅取值仅为 node/way，执行查询语句如下：

```
> db.beijing.distinct('type')
[
        "node",
        "Cypress",
        "国槐",
        "way",
        "公园",
        "应急避难场所",
        "office",
        "multipolygon"
]
```

结合上述结果和原始文件 beijing_china.osm 发现类似 `<tag k="type" v="Cypress"/>` 的标签会导致 type 的取值被覆盖，因此需要对类似标签的 `k` 进行变更。具体如下：

- 对于 `国槐` / `Cypress`，其取值描述的为数的种类，因此将其 key 由 type 变更为 species。
- 通过查看 `type` 为 `公园` / `应急避难场所` 的数据发现，该数据与应急避难场所相关。故参照其他应急避难场所数据将 `type` 调整为 `应急避难场所类型`，并对该条数据的其他字段 key 值作出对应调整。
- 通过查看 `building` 字段取值发现， `office` 为 `building` 的其中一种取值，因此可以将此类标签 key 值调整为 `building`。
- 查看 `type` 为 `multipolygon` 的数据，其中还存在 `area` 字段，对比其他存在 `area` 字段的数据无法确定 `multipolygon` 的具体含义，做忽略处理。

1.2 相同含义字段 key 取值不同

继续检查 key 的取值，得到结果如下：

```
> db.runCommand({"mapreduce" : "beijing","map" : function() {for (var key in this)
{ emit(key, null); }},"reduce" : function(key, stuff) { return null; }, "out": "bei
jing" + "_keys"})
> db.beijing_keys.find().sort({_id:-1})
{ "_id" : "黄南苑小区", "value" : null }
{ "_id" : "车库", "value" : null }
{ "_id" : "疏散人数（万）", "value" : null }
{ "_id" : "疏散人数(万人)", "value" : null }
{ "_id" : "疏散人数", "value" : null }
{ "_id" : "开发商", "value" : null }
{ "_id" : "建成时间", "value" : null }
{ "_id" : "应急避难场所类型", "value" : null }
{ "_id" : "应急避难场所类别", "value" : null }
{ "_id" : "应急避难场所疏散人数万人", "value" : null }
{ "_id" : "应急避难场所疏散人口万人", "value" : null }
{ "_id" : "应急避难场所总面积万平米", "value" : null }
{ "_id" : "应急避难场所", "value" : null }
{ "_id" : "应急避难人数（万人）", "value" : null }
```

其中表述疏散人数的字段包括 `疏散人数（万）`、`疏散人数(万人)`、`疏散人数`、`应急避难场所疏散人数万人`、`应急避难场所疏散人口万人`、`应急避难人数（万人）`，需将其统一为 `capacities`。且当 key 为 `疏散人数` 时，需要将对应的 value 格式化为以 `万人` 为单位。

1.3 电话号码格式不统一

检查 phone 字段发现，改字段的格式较为混乱，应调整为统一的格式。

```
> db.beijing.find({'phone': {$exists: true}},{phone:1})
{ "_id" : ObjectId("5b17e9db03601f638ae387c8"), "phone" : "+86 10 6582 2892" }
{ "_id" : ObjectId("5b17e9db03601f638ae387d4"), "phone" : "(010)64629112" }
{ "_id" : ObjectId("5b17e9db03601f638ae427cf"), "phone" : "01051696505" }
{ "_id" : ObjectId("5b17e9dc03601f638ae498ba"), "phone" : "+86-10-60712288" }
```

这里将电话号码统一调整格式如下：`+86-10-88888888`。

2. 数据概述

2.1 文件大小

```
~/Dropbox/DAND/P3(master*) » ls -alh beijing_china.osm*
```

```
-rw-r--r--@ 1 dragonkid  staff   195M Dec 12 22:06 beijing_china.osm
-rw-r--r--@ 1 dragonkid  staff   242M Jun  6 22:03 beijing_china.osm.json
```

2.2 总记录数

```
> db.beijing.find().count()
1060380
```

2.3 贡献总人数

贡献总人数 1994 人：

```
db.beijing.distinct("created.user").length
```

贡献最多的人为 `Chen Jia`：

```
db.beijing.aggregate([{"$group":{"_id":"$created.user", "count":{"$sum":1}}},{"$sor
t":{"count":-1}}, {$limit: 1}])
```

2.4 节点和途径数

其中 `node` 节点数据 923234 条， `way` 节点数据 137132 条：

```
> db.beijing.find({'type': 'way'}).count()
137132
> db.beijing.find({'type': 'node'}).count()
923234
```

2.5 应急避难场所可容纳的总人数

应急场所共可容纳 100.1291 万人。

```
> db.beijing.aggregate({$group:{_id: '', capacities: {$sum: '$capacities' }}}, {$pr
oject: {_id: 0, count: '$capacities'}})
{ "count" : 100.1291 }
```

4. 额外想法或改进建议

在检查数据的时候发现问题主要出现在两方面：

1. key 的取值意义模糊、不统一
2. 某些字段缺乏特定的填写规范导致格式混乱

因此建议如下：

1. 通过审查当前已有字段，将使用频率较高的字段固定下来并详细描述其使用场景。
2. 对于个别字段，例如：电话号、传真号、邮编等，明确其格式，保证录入数据的规范性。

预期的问题：

1. 本次整理的数据中共有字段 311 个，实际需要固定下来并详细描述使用场景的字段可能较多，在实际操作中难以真正为用户提供参考。
2. 只提供参考建议一类的规范不足以达到使数据规范化的目的，需要找到数据入口，在入口处对字段的规范性进行校验并对不符合规范的数据做出友好的提示。

Refs

- https://s3.cn-north-1.amazonaws.com.cn/static-documents/nd002/SampleDataWranglingProject_en.pdf