

Openstreetmap 项目使用 mongodb 进行清洗

地图 url: https://mapzen.com/data/metro-extracts/metro/shanghai_china/

上海: 我读书, 工作, 生活的地方

1、 地图中遇到的问题	1
路名中英文混杂	1
邮编错误	1
2、 数据查看	2
3、 其他	3
数据集的问题和建议	3
地图数据太大	3
安装 mongodb	3

1、地图中遇到的问题

路名中英文混杂

如下:

```
mapping = { "#1": "",
             "Lane 30 of West Dalian Road": "",
             "Lane 82 of East Yanji Road": "",
             "Xiuyan Road": "秀沿路",
             "Songhua Community": "",
             "yindu road": "银都路",
             "Jiashan Market": "嘉善老市"
             ... }
```

查看了一下发现直接改掉也不行, 就先不动了。

邮编错误

```
mapping = {
    '201315 上海': '201315',
    '2000080': '200080',
    '21500': '215027',
    '21351': '213353',
}
```

使用 mongodb 修改邮编:

```
for i in range(len(mapping)):
```

```
db.autos.update({'address.postcode':list(mapping.keys())[i]},{'$set':{'address.postcode':list(mapping.values())[i]}},multi=True)
```

2、数据查看

文本大小

shanghai_china.osm: 779M

shanghai_china.osm.json: 984M

几个 node

```
db.autos.find("type": "node")
3869506
```

几个 way

```
db.autos.find("type": "way")
471827
```

几个餐厅

```
db.autos.find("amenity": "restaurant")
1496
```

数量前 5 的餐馆

```
aggregate(db, [{"$match":{"amenity": "restaurant"}},
{"$group":{"_id":"$name","count":{"$sum":1}}},
{"$sort":{"count":-1}}
,{"$limit":5}])
```

结果: [{'count': 454, '_id': None}, {'count': 12, '_id': '食堂'}, {'count': 7, '_id': '外婆家'}, {'count': 6, '_id': '必胜客'}, {'count': 6, '_id': '肯德基'}]

发现: 上海最多的餐馆是食堂。。。

哪条街上餐馆最多

```
pipeline = [{"$match":{"amenity": "restaurant"}},
{"$group":{"_id":"$address.street","count":{"$sum":1}}},
{"$sort":{"count":-1}}
,{"$limit":5}]
```

结果: [{'_id': None, 'count': 1203}, {'_id': '解放路', 'count': 15}, {'_id': '淮海中路', 'count': 8}, {'_id': '南京西路', 'count': 7}, {'_id': '天钥桥路', 'count': 6}]

发现: 上海最多的餐馆是解放路, 下次去试试看

前 5 个创建了餐馆标签的创建人

```
pipeline = [{"$match":{"amenity": "restaurant"}},
{"$group":{"_id":"$created_by","count":{"$sum":1}}}, {"$sort":{"count":-1}}, {"$limit":5}]
```

结果: [{'count': 1480, '_id': None}, {'count': 14, '_id': 'JOSM'}, {'count': 1, '_id': 'iLOE 1.9'}, {'count': 1, '_id': 'Potlatch 0.4a'}]

发现: 大部分餐馆标签创建者都没有署名, 有署名中创建最多的人是 JOSM。

3、其他

数据集的问题和建议

有些餐厅有时属于 **restaurant**，有时属于 **fast food**，对最后的统计会造成一定的困扰。

有两个解决方案：

- 1: 如果一个餐厅大部分时候是 **fast food**，那就把其他地址的同名餐厅都统一成 **fast food**
- 2: 先限定 **fast food** 对菜品的标准，例如菜单中含有 **burger** 的是快餐店，将菜单中出现 **burger** 的餐厅一律改成 **fast food**。

可能造成的问题：第二种方式对菜品的界定很难穷举，就算穷举了也不能排除，出现快餐食品的餐厅是正式餐厅的可能性

地图数据太大

转成 json 文件后有将近 1G，电脑跑死机好几次，后来换了公司的电脑，跑了半小时。。。

安装 mongodb

之前一直用的是 mysql 和 Oracle，第一次接触 mongodb，下面是经验总结：

安装 mongodb：安装，在安装根目录下新建 data 文件夹，在里面新建文件夹 db 和 log，在安装目录 MongoDB 下新建配置文件 mongodb.cfg，里面写入

systemLog:

destination: file

path: D:\data\log\mongod.log #刚刚建的 log 路径

storage:

dbPath: D:\data\db #刚刚建的 db 路径

连接：打开 mongod.exe，使用 pymongo 连接