OpenStreetMap Data Case Study

地图区域

Wuhan, China
https://www.openstreetmap.org/relation/3076268

这张地图是我家乡的省会，从东北回来后都会去武汉玩，所以探索这份地图对我来说相对更容易，在 OpenStreetMap.org 上做出的修改也可能更准确。

地图中所遇到的问题

下载文件并处理后，我发现了一下几个明显的问题：

大量街道名称是中文的	中华路
错误的街道命名	<tag k="addr:street" v="永安堂站对面"/>
csv 文件中文字符乱码	涓冨儚鑾晴

中文名街道

针对大量街道是中文名的问题，考虑到中文名称音译成拼音后很难识别，而在数据处理中我们需要的往往只是对其的分类，我将其进行了混合命名：名称+街道类型，如中华路>>中华 Road。这样既方便了分类查询，又保留的中文的特殊性（另一方面，将中文转化为拼音很困难）。具体代码如下：

```
mapping = { "Rd" : "Road",
                    #特殊替换处理,直接替换，没有空格，最后加上空格
                    "街":" Street",
                    "大道":" Avenue",
                    "路":" Road",
                    "广场":" Square",
                    "园":" Parkway",
                    "巷":" Lane"
                    }
def update_name(name, mapping):
    """
    直接替换掉不想要的文字
    """
    m = street_type_re.search(name)
    if m and (m.group() in expected):
        pass
    else:
```

```
        for key,value in mapping.items():
            if key in name:
                name = name.replace(key,value)
    return name
```

错误的街道命名

Tag 中有许多街道命名并不是正确的街道名，如 :<tag k="addr:street" v="永安堂站对面"/>。对于这种情况，我通过 Google 搜索直接将其替换成了正确的命名，汉阳 Avenue。类似还有其他一些特殊的不规则命名情况，比较容易操作的我直接进行了替换。

csv 文件中文字符乱码

课程中给出的生成 csv 文件的代码输出的中文字符乱码，我将其用 csv.Dictwriter 改写了，另外利用了自己工作中写的 toolkit（见附件）工具包转换成 pandas dataframe 进行了筛查，部分实例代码如下：

```
with open(NODES_PATH, 'w',encoding=encoding) as nodes_file
    ways_tags_writer = csv.DictWriter(way_tags_file, fieldnames=WAY_TAGS_FIELDS )
    ways_tags_writer.writeheader()
    for i,element in enumerate(get_element(file_in, tags=('node', 'way'))):
        #tag 装进字典里
        el = shape_element(element)
        for item in el['way_tags']:
            ways_tags_writer.writerow(item)
import toolkit as tk
df_ways_tags = tk.read_csv_in_str('ways_tags.csv',sep=',',encoding='utf8')
```

数据总览和发现

文件大小

wuhan_china.osm	63.8 MB
OpenStreetMapWuhan.db	34.4MB
nodes.csv	25.4MB
nodes_tags.csv	0.54MB
ways.csv	2.0MB
ways_nodes.csv	9.2MB
ways_tags.csv	2.49MB

节点数量

```
sqlite> SELECT COUNT(*) FROM nodes
322152
```

道路数量

```
sqlite> SELECT COUNT(*) FROM ways;
35018
```

道路的平均节点数

```
SELECT COUNT(DISTINCT node_id)/COUNT(DISTINCT id ) FROM ways_nodes;
```

9.1

编辑的日期跨度

```
SELECT MIN(A.timestamp),MAX(A.timestamp)
FROM (SELECT timestamp FROM nodes UNION ALL SELECT timestamp FROM ways) AS A;
2008-9-25       2017-12-23
```

参与的用户数量

```
sqlite> SELECT COUNT(DISTINCT(A.uid))
FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways) AS A;
540
```

贡献最多的前十用户

```
sqlite> SELECT A.user, COUNT(*) as num FROM (SELECT user FROM nodes UNION ALL
SELECT user FROM ways) AS A GROUP BY A.user ORDER BY num DESC LIMIT 10;
```

GeoSUN	111526
Soub	47736
jamesks,	24376

"Gao xioix"	17894
katpatuka	17225
"samsung galaxy s6"	13781
dword1511	13527
flierfy	5473
hanchao	5283
keepcalmandmapon	4936

只出现一次的用户数（只有一条记录）

```
sqlite> SELECT COUNT(*) FROM
(SELECT A.user, COUNT(*) as num FROM (SELECT user FROM nodes UNION ALL SELECT user
FROM ways) AS A GROUP BY A.user HAVING num = 1);
```

101

其他发现

出现次数最多的前十个公共设施

```
sqlite> SELECT value, COUNT(*) as num FROM nodes_tags WHERE key='amenity' GROUP BY
value ORDER BY num DESC LIMIT 10;
```

restaurant	165
school	150
bank	131
townhall	75
parking	74
fast_food	60
fuel	57
bicycle_parking	35
hospital	32
atm	27

节点数量生成最多的前十个月份

```
SELECT date,COUNT(*) AS num FROM (SELECT SUBSTR(timestamp,1,7) AS date FROM
```

nodes) GROUP BY date ORDER BY num DESC LIMIT 10;

2011-03	25803
2013-07	21768
2011-04	17665
2013-08	15964
2011-02	13229
2016-12	12937
2012-08	11968
2011-06	8824
2017-03	7247
2012-04	7245

道路数量生成最多的前十个月份

SELECT date,COUNT(*) AS num FROM (SELECT SUBSTR(timestamp,1,7) AS date FROM ways) GROUP BY date ORDER BY num DESC LIMIT 10;

2016-12	3138
2013-07	2766
2011-03	2161
2016-11|	1526
2017-03	1391
2017-05	1192
2017-10	1162
2012-08	1119
2011-04	946
2016-03	928

关于数据集的其他想法

整体来书数据集不详细 。相比纽约大都会数据集动辄上 G，本数据集在 100Mb 以内，很多新奇地点多的信息并不包含在内，使得很多分析和应用无法进行。
建议 1：大学给大学生开展一些地理信息系统相关的课程。
好处：这样可以让学生把学业和实际结合和起来，是一个非常好的锻炼机会，同时也能帮助丰富数据。
预期的问题：可能需要大量资金的投入，而且不是每个学生都想学，可能缺乏动力。

建议 2：让企业和政府开展合作，互相在 OpenStreetmap 上共享数据资源。
好处：这样应该能很快的丰富数据，同时也节约资金。由于时商业参与，投入能较快的变现，动力比较足。
预期的问题：由于企业间存在竞争关系，如果涉及到商业机密和资产分配可能会遇到阻力，政府可能需要发挥协调作用。

很多地方中文和英文地名混合出现，而且存在地名缺失。多种语言名称混合出现时造成数据处理非常麻烦，中文可以较好的保留原意，但有些分析工具并不支持中文；英文可以应用在大多数软件上，但是音译过来又容易使人产生误解。
建议：公开城市交通公共运输信息（比如公交、地铁），各地的站名设置，给路标同时设置英文和中文名。
好处：这些数据是已经人工整理好的，质量相对较高，可以直接转换使用，非常便捷和节省资金。
预期的问题：相关部分或利益相关方不一定愿意公开数据，和业务实际结合的商业公司可以发挥一定推动作用。

结论

在数据探索和审查之后，武汉的 Openstreetmap 数据明显是不完整的，相比深圳（166M）和香港（720M）的数据，拥有较小的数据量，未来有很多的补充空间，我所做的贡献主要在于去除一些明显的数据错误和不规整的命名。由于在中国很多地名很难用英文来表示出来，所以中英文混合可能是更好的表示方式。工作中我一直将地理信息和车辆的调度情况结合起来，Openstreetmap 是我接触到的第一个地理信息数据集，其中有很多地方可以利用到工作中。未来随着武汉的发展，信息化程度的提高，这些数据将有更广泛的用途。