Project 3: OpenStreetMap Data Case Study

Map Area

Beijing, China

- https://mapzen.com/data/metro-extracts/metro/beijing_china/

Beijing is the capital city of my home country, so I would like to take this chance to learn more about this city and of course I will be very happy if I could make some contribution to this map on OpenStreetMap.org.

Problems Encountered in the Map

I use the code provided to generate a sample OSM file from the original OSM file. After using pandas to explore the CSV files generated by running data.py, I noticed several problems with the data, which I will explain in details in the following order:

- Incorrect English translation of road names ("Jiuzong Lu")
- Inconsistent English translation of street and road names ("Dexian Road", "Lugu Rd.", "5th Ring road", "SHANYUAN St", "Xizhimen North Street", "Gulouwai Str", "Beijingzhan Str.", "Shangye west street")
- Some node tags with `"k=addr:street"` have a mapping `"v"` value which is not a street name:

  ```
  1 <tag k="addr:street" v="Sanlitun SOHO" ></tag>
  ```

  "Sanlitun SOHO" is not a street name.

Incorrect English translation of road names

After using pandas to explore the CSV files generated by running data.py, the first problem captures my eyes is the incorrect English translation of road names, for example "Jiuzong Lu", "Lu" actually is phonetic symbol of Chinese character "路", which means "Road". Therefore the correct translation should be "Jiuzong Road". I will modify the audit.py file to solve this problem together with next problem.

Inconsistent English translation of street and road names

Also I discovered the problem of Inconsistent English translation of street and road names.

Street name of English	Road name of English
SHANYUAN St	Dexian Road
Xizhimen North Street	Lugu Rd.
Gulouwai Str	5th Ring road
Beijingzhan Str.	
Shangye west street	

I modified the audit.py to solve the problems mentioned above, below are some of the functions:

```
def audit_street_type(street_types, street_name):
    m = street_type_re.search(street_name)
    if m:
        street_type = m.group()
        if street_type in unexpected:
            street_types[street_type].add(street_name)
```

```
def is_street_name(elem):
    return (elem.attrib['k'] == "name:en")
```

```
def audit(osmfile):
    osm_file = open(osmfile, "r")
    street_types = defaultdict(set)
    for event, elem in ET.iterparse(osm_file, events=("start",)):

        if elem.tag == "way":
            for tag in elem.iter("tag"):
                if is_street_name(tag):
                    audit_street_type(street_types, tag.attrib['v'])
    osm_file.close()
    return street_types
```

```
def update_name(name, mapping):
    m = street_type_re.search(name)
    if m:
        weird_street_type = m.group()
        print weird_street_type
        name = name.replace(weird_street_type, mapping[weird_street_type])
    return name
```

This updated all the problematic names, for example:

"Jiuzong Lu" becomes: "Jiuzong Road"

"Gulouwai Str" becomes: "Gulouwai Street"

"Lugu Rd." becomes: "Lugu Road"

Overview of the data

This section presents the overview statistics of the dataset.

File sizes

```
Beijing_China.osm ..................... 186.7 MB
Beijing_openstreetmap.db ............. 99 MB
Beijing_nodes_cleaned.csv ............ 69.8 MB
Beijing_nodes_tags_cleaned.csv .......  3 MB
Beijing_ways_cleaned.csv ............. 6.8 MB
Beijing_ways_tags_cleaned.csv ........ 8.3 MB
Beijing_ways_nodes_cleaned.csv ....... 24.2MB
```

Number of nodes

```
sqlite> SELECT COUNT(*) FROM node;
851757
```

Number of ways

```
sqlite> SELECT COUNT(*) FROM way;
116605
```

Number of unique users

```
SELECT COUNT(e.uid)
   ...> FROM (SELECT uid FROM node UNION SELECT uid FROM way) as e;
1657
```

Major contributor (users who contributing more 5,000 times)

```
sqlite> SELECT e.user, COUNT(*) as num
   ...> FROM (SELECT user FROM node UNION ALL SELECT user FROM way) e
   ...> GROUP BY e.user
   ...> HAVING num >= 10000
   ...> ORDER BY num DESC;
```

```
"Chen Jia"            210797
R438                  152434
ij_                   85385
hanchao               65723
katpatuka             41590
m17design             21808
Esperanza36           19046
nuklearerWintersturm  17151
RationalTangle        14585
u_kubota              11217
```

Data Exploration

Top 5 appearing brand

```
sqlite> SELECT value, COUNT(*) as num
   ...> FROM node_tags
   ...> WHERE key = "brand"
   ...> GROUP BY value
   ...> ORDER BY num DESC
   ...> LIMIT 5;
```

```
"中国石化",17
"中国石油",8
"Bank of China",2
Shell,2
Apple,1            ps: how to have pretty print results in terminal?
```

Top 5 popular cuisines of restaurant

```
SELECT value, count(*) as num
   ...> FROM node_tags
   ...> WHERE key='cuisine'
   ...> GROUP by value
   ...> ORDER BY num DESC
   ...> LIMIT 5;
chinese,182
coffee_shop,60
american;burger,57
chicken,55
japanese,28
```

It's not surprise that Chinese restaurant dominates the market. In addition to that, it's safe to infer that MacDonald and KFC are also very popular in Beijing, China.

Additional Ideas

"FIXME" key, value statistics and suggestion

I notice there are some tags in which the key is "FIXME", and the value of such tag did not state clearly what is the problem with the tag, what need pay attention to, what is the situation now. Detailed statistics are as follows:

- Total number of tags with `key='FIXME'` is 754.

```
SELECT count(*)
   ...> FROM (SELECT key, value From node_tags WHERE key = 'FIXME' UNION ALL select key, value from way_tags WHERE key = 'FIXME') e;
754
```

- Top five values of such tags are as follows.

```
SELECT value, count(*)
   ...> FROM (SELECT key, value From node_tags WHERE key = 'FIXME' UNION ALL select key, value from way_tags WHERE key = 'FIXME') e
   ...> group by value
   ...> order by count(*) desc
   ...> limit 5;
"Possibly a tower",223
"Or no?",102
"请看一看铁塔的号码和电线的名字。有一张牌子在每个铁塔的前面。",79
"Is here wall?",76
"Is here a tower?",10
```

The 'value' of 'FIXME' tags should provide some useful information for later contributors, however, looking at the top five values and their counts, they are probably generated automatically by bots and the information they provided may be misleading. To facilitate collaboration among contributors, the community of Openstreetmap should set certain rules, for example: 1. try to edit the value manually and explain the problem clearly; 2. if you cannot make the First rule and want to generate by bots, please use default value,like '0', in order to avoid misleading.

References:

1.https://gist.github.com/carlward/54ec1c91b62a5f911c42#file-sample_project-md

2.https://gist.github.com/swwelch/f1144229848b407e0a5d13fcb7fbbd6f