Author: Roxy Lu

Map Area: Shanghai, CN https://www.openstreetmap.org/relation/913067

Problems encountered in your map

Problems:

1. Inconsistency in translation:

```
# Chinese
<tag k="addr:street" v="江西中路“/>
# English
<tag k="addr:street" v=" Pudong Avenue" />
<tag k="addr:street" v=" Wuzhong Rd." />
# Mixed
<tag k="addr:street" v=" 仙霞西路 (Xianxia West Rd)" />
# Pinyin
<tag k="addr:street" v=" wukang lu" />
```

2. More detailed location info after street name:

```
<tag k="addr:street" v=" 上海市长宁区中山西路1277号" />
<tag k="addr:street" v=" 中山西路 1245 弄" />
```

3. Empty name

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

4. Duplicated address info

```
<tag k=" address"  v=" 张东路2281弄" />
<tag k="addr:street" v=" 上海" />
```

Solution:

1. Use Baidu Translation API to translate any English / Pinyin street name into Chinese.
2. Remove anything behind “路”.

Algorithm:

```
1. If string contains Chinese characters
    1.1 remove all English characters from string
    1.2 remove anything behind EXPECTED word “路”
    1.3 return result.
2. Else
    2.1 If string in cache
        2.1.1 get translated word from cache
    2.2 Else
        2.2.1 translate into Chinese using Baidu Translate API
        2.2.2 store translated word in cache
    2.3 remove anything behind EXPECTED word “路”
    2.4 return result
3. do nothing if the name is empty
```

Explanation:

1. Check if there is any Chinese character (CJK UNIFIED) in the string. If so, I would remove all the ascii letters, punctuation, digits and whitespaces from the string.

```
仙霞西路 (Xianxia West Rd) => 仙霞西路
```

2. Then remove all the detailed location info after “路”.

```
中山西路 1245 弄 => 中山西路
```

3. In case 2, we have a string that contains no Chinese characters. So this would be either Enligsh or Pinyin. Here I use Baidu Translation API to process it to Chinese.

```
Wuzhong Rd => 吴中路
wukang lu => 武康路
```

4. MongoDB is used for translation result caching. This would help improve the performance, and make the translation result editable, in case we are not satisfied with the result by Baidu Translation API.

Overview of the data

File Sizes:

- 133M shanghai.osm
- 193M shanghai.osm.json

Import to MongoDB:

```
mongoimport --jsonArray -d osm -c shanghai --file shanghai.osm.json
```

Mongo Shell:

```
# Number of documents:
> db.shanghai.count()
720904

# Number of Nodes:
> db.shanghai.find({type:"node"}).count()
630008

# Number of Ways:
> db.shanghai.find({type:"way"}).count()
90892
```

More Data Exploration With MongoDB:

The following result is the output of query.py:
https://github.com/roxylu/Wrangle-OpenStreetMap-Data/blob/master/osm_code/query.py

```
Number of contributing users:
1266
==========================================
Top contributing user as a percentage of total documents:
10.77 [%]
==========================================
Top 10% of contributing user as a percentage of total documents:
94.72 [%]
==========================================
Contributions by year:
```

Year: Contributions

2007: 3825

2008: 765

2009: 9148

2010: 19395

2011: 69698

2012: 136454

2013: 43556

2014: 91579

2015: 127185

2016: 166675

2017: 52624

==========

Number of top 5 shops:

supermarket: 256

convenience: 238

clothes: 90

mall: 71

bakery: 58

==========

Number of top 10 convenience:

全家FamilyMart: 20

全家: 18

快客: 17

好德: 14

Family Mart: 10

Alldays: 10

罗森LAWSON: 9

Lawson: 5

良友: 5

FamilyMart: 4

==========

Number of nodes without addresses:

Nodes_without_address: 628534

==========

Number of top 5 street addresses:

沪南路: 1018

```
鹤庆路: 116
安宁路: 75
松花一村: 61
延吉东路: 53
==========================================
Number of top 10 amenities:
restaurant: 786
parking: 563
school: 402
bank: 324
toilets: 264
cafe: 251
fast_food: 191
hospital: 129
bar: 97
fuel: 94
```

Other ideas about the dataset

Additional Improvements:

Inconsistent Names

Inconsistency in the use of names makes the analysis on the dataset no more accurate and reliable.

Taking top 10 convenience as an example:

```
全家FamilyMart: 20
全家: 18
快客: 17
好德: 14
Family Mart: 10
Alldays: 10
罗森LAWSON: 9
Lawson: 5
良友: 5
FamilyMart: 4
```

4 of the 10 results are the same convenience: ”FamilyMart”, and ”好德” equals to ”Alldays”, and there are two duplicated ”Lawsons”.

So it would be better if we could have datalist autocomplete dropdown on user input, to see if it is similar to an existing word, and make suggestions.

For example:

```
<input type="text" list="convenience">
<datalist id="convenience">
  <option value="全家">
  <option value="FamilyMart">
  <option value="快客">
  <option value="好德">
  <option value="罗森">
  <option value="Lawson">
  <option value="良友">
</datalist>
```

By checking and suggesting input, we could control typo errors and improve data quality.

However, the datalist does not support translations. So “全家” and “FamilyMart“
would still be two entries.

Multiple Sources

Currently all the dataset is populated only from one source: OpenStreetMaps. While this crowdsourced repository pulls from multiple sources ,some of data is potentially outdated, especially those within China. There are quite a few nodes without address.

```
Nodes_without_address: 628534
```

It would have been an interesting exercise to validate and / or pull missing information from Baidu Maps API, since every node has latitude-longitude coordinates.

Conclusion

During this project, I used data munging techniques, such as assessing the quality of the data for validity, accuracy, completeness, consistency and uniformity, to clean the OpenStreetMap data for Shanghai, CN. Thank you for reading, and please refer to the Python files to see implementations with more details.