OpenStreetMap Data Wrangling using MongoDB
Kevin H(Y) Shu
Map Area: Yangtze River Delta (Shanghai, parts of Jiangsu and Zhejiang), PRC

1 Problems encountered in map

After downloading the OSM for the Yangtze River Delta, I encountered the following main problems with the data set, in this order:

- Inconsistency with naming vis-a-vis English vs Chinese names
- Street suffixes and postal codes

1.1 Inconsistency with naming vis-a-vis English vs Chinese names

Since the data set is dealing with mainland China, most of the documents on places with names (has `tag k="name"`) have a Chinese name. Many also have English names, and occasionally the "name" tag will have an idiosyncratic mixing of the Chinese and English in the corresponding value (`v`), such as the following: "玛雅酒吧 Maya Pub."

Less commonly, but as with the case with the preceding example, the field `name:en` may have a shorter name than the English part of the `name`; in any case, the longer of the two is inserted into the `name.name` field in MongoDB.

1.2 Street suffixes and postal codes

Before the data was imported to MongoDB, I checked the street names for consistency in suffixes. While the bulk of the non-compliant suffixes are due to Chinese forming the entirety of the street name (the `addr:street` tag), a good deal of the remainder are abbreviations of common street names: "Rd" or "Rd." instead of "Road". Some are the pinyin Romanisation of the Chinese name for the street type, as in "Lu" (路) for "Road". Examples of such suffix errors are "Wukang Rd" and "Gaoyou Lu". All postal codes in mainland China comprise of six digits. The overwhelming majority of the entries (under the tag `postal_code` or `addr:postcode`) are compliant; those that are not at least six digits are discarded since it is inappropriate to second-guess the last digits, while those that consist of six digits followed by a space and non-numeric characters are trimmed to the digits themselves.

2 Data overview

2.1 File sizes

Table 1: My caption

File name	File size
shanghai_cn.osm	433 MB
shanghai_cn.json	926 MB

2.2 Number of documents

```
> db.cities.find().count()
2397382
```

2.3 Document type breakdown

Number of nodes:

```
> db.cities.find({"type": "node"}).count()
2144933
```

Number of ways:

```
> db.cities.find({"type": "way"}).count()
252137
```

2.4 Number of unique contributors

```
> query_users_desc_contrib = [{"$group": {"_id": "$created.user",
                                          "contribs": {"$sum": 1}}},
                              {"$sort": {"contribs": -1}}]
> result_users_desc_contrib = aggregate(db, query_users_desc_contrib)
> len(result_users_desc_contrib)
1376
```

3 Other ideas on the data set

3.1 Additional data exploration using MongoDB queries

Combined share of contributions from top 10 users:

```
> def retrieve_contribs(d):
>     return d['contribs']
> contribs = map(retrieve_contribs, result_users_desc_contrib)
> total_contribs = sum(contribs)
> 100.0 * float(sum(contribs[:10])) / float(total_contribs)
45.5796364534
```

Combined share of contributions from bottom 90% of users:

```
> pct_90 = len(contribs)*1/10 + 1
> 100.0 * float(sum(contribs[pct_90:])) / float(total_contribs)
5.6665145563
```

Number of one-time contributors:

```
> def is_one(x):
>     return x==1
> len(filter(is_one, contribs))
218
```

Number of unique amenities:

```
> query_amenities_desc = [{"$match": {"amenity": {"$exists": 1}}},
                          {"$group": {"_id": "$amenity",
                                      "count": {"$sum": 1}}},
                          {"$sort": {"count": -1}}]
> result_amenities_desc = aggregate(db, query_amenities_desc)
> len(result_amenities_desc)
92
```

Top five amenities: As the topography in the core urban areas of the Yangtze River Delta is largely flat, it is little surprise that bicycle rentals are the top amenity, at 2060 documents.

```
> result_amenities_desc[:5]
[{u'_id': u'bicycle_rental', u'count': 2060}
{u'_id': u'parking', u'count': 972},
{u'_id': u'school', u'count': 867},
{u'_id': u'restaurant', u'count': 727},
{u'_id': u'bank', u'count': 374}]
```

3.2 Conclusion

The main outstanding issue with the data set is the lack of English names for certain parameters in documents. Chinese names are processed poorly by MongoDB (making for difficult comparisons in queries), with output such as the following: `u'\u4e0a\u6d77\u5e02'`. A second issue is the inconsistent adherence to pinyin transcription rules; some entries use "Camel case", which capitalises the first letter of each syllable, e.g. "KunShan" as opposed to "Kunshan", while others separate the syllables when they are part of one "word", e.g. "Kun Shan" as opposed to "Kunshan".

1. The conversion of pages with only Chinese names could be done by writing additional code to "feed" the Chinese characters to a website that converts them to pinyin. This carries the clear advantage of ensuring greater completeness of information on the data, but is subject to imperfections as are all machine translations: by imperfections, converting to never-used pinyin for a particular character is not meant, but instead the rare failure to take into account local variations, such as "Luhe" as opposed to the expected "Liuhe" for that district (六合区) of Nanjing. Also, as was mentioned earlier, pinyin has grouping rules that require some human intervention beyond machine results. Another downside would be the additional computational work (on the order of $\mathcal{O}(n)$) required to complete this task.

2. The data could also be cross-checked and imputed with a non open-source API such as the map services Google or Baidu offer. The costs of doing so are A) the challenges of implementing that solution, as differing APIs may have differing structures to their data and in some cases may require "beginning from scratch" B) employee (human) input is still involved, and Google will occasionally supply English translations that are *literal* translations of the Chinese when *transcription* is more appropriate. C) Matching will likely be done on coordinates, but any map will necessarily have its precision limitations, and, in the case of Google, the apparent discrepancy between satellite imagery of some expressways in mainland China and their labels amounts to tens of metres.