OpenStreetMap Project
Data Wrangling with MongoDB

By Mengyao Gao

Problems Encountered in the Map

Two main problems were encountered in the original map dataset. I will discuss them in the following order.

- Over-abbreviated Street Names ('East Bidwell St.')
- Inconsistent Postcodes ('CA 95826', '95832-1447')
- Problematic characters ('Segment #')

Over-abbreviated Street Names

Some street names are over-abbreviated. Different formats of street names may cause confusion. I updated all the abbreviations to full format for unity. For example, 'East Bidwell St.' is updated to 'East Bidwell Street'.

Postcodes

The original dataset presents postcodes in three formats: 5-digit ('95832'), 5-digit with 4-digit extension ('95382-1447') and state characters followed 5-digit ('CA 95832'). To benefit MongoDB aggregation calls on postcodes, I stripped the state characters and 4-digit extensions, standardizing postcodes.

Problematic Characters

I observed some problematic characters in this dataset. I deleted these documents when cleaning and transforming the XML dataset to JSON one.

Overview of the Data

Basic statistics about the dataset are calculated by the MongoDB queries below.

File sizes

sacramento_california.osm: 217.7MB
sacramento_california.json: 221.8MB

Number of documents

```
> db.sacramento.find().count()
1019239
```

Number of unique users

```
> db.sacramento.distinct("created.user").length
609
```

Number of nodes

```
> db.sacramento.find({'type':'node'}).count()
929573
```

Number of ways

```
> db.sacramento.find({'type':'way'}).count()
89600
```

Number of shops

```
> db.sacramento.find({'amenity':'shop'}).count()
22
```

Number of banks

```
> db.sacramento.find({'amenity':'bank'}).count()
94
```

Number of restaurants

```
> db.sacramento.find({'amenity':'restaurant'}).count()
393
```

Other ideas about the dataset

More user activeness needed

```
> db.sacramento.find({'amenity':{'$exists':1}}).count()
6691
```

Only 6691 amenities are labeled in this dataset, compared to total 929573 nodes. More amenities need to be flag by users. Further, flags are not united. For example, 'fuel' and 'gas' are presented to one thing. But it is very time-consuming and high costly to get more information to fix these issues. OpenStreetMap could design more attractive policies to activate users.

Other data exploration

Top 3 brands of fuel

```
>
db.sacramento.aggregate([{'$match':{'amenity':{'$exists':1},'name
':{'$exists':1},'amenity':'fuel'}}, {'$group': {'_id':
'$name','count':{'$sum':1}}},{'$sort':{'count':-
1}},{'$limit':3}])
{ "_id" : "Chevron", "count" : 34 }
{ "_id" : "Shell", "count" : 29 }
{ "_id" : "Arco", "count" : 18 }
```

Top 3 brands of fast food

```
>
db.sacramento.aggregate([{'$match':{'amenity':{'$exists':1},'name
':{'$exists':1},'amenity':'fast_food'}}, {'$group': {'_id':
'$name','count':{'$sum':1}}},{'$sort':{'count':-
1}},{'$limit':3}])
{ "_id" : "Taco Bell", "count" : 23 }
{ "_id" : "Subway", "count" : 22 }
{ "_id" : "McDonald's", "count" : 21 }
```