Huy Do

3/13/16

P3: Wrangle OpenStreetMap Data

Link to map:

https://www.openstreetmap.org/export#map=10/36.1201/-95.8626

1. Problems Encountered in the Map

For this project, I chose the Tulsa area as I know it well. To meet the 50MB requirement of this project, I extended the map area to include a few surrounding cities. After downloading the map and running the data, I noticed the following problems:

Street type is abbreviated:

Sometimes street type (Street, Avenue, etc…) is abbreviated (St., Ave., etc…). To maintain consistency in address, street type is updated to full-length words. In addition, typo such as Avenu is corrected.

Street type is omitted:

Some street address does not have the street type and/or is not properly capitalized. Since this is a small dataset, I have chosen to correct this using the mapping to update street name. For example, sheridan is changed to Sheridan Street.

Street address is not always labeled with a preceding addr:

Street address is not always labeled with a preceding addr. Sometimes it is stored within name in tag attribute k. To separate the street address from other names stored in the name tag (e.g name of business, name of building, etc…), a heuristic is used. If a string in the name tag contains a number and a street type (e.g. avenue, street, station, etc…), it is assumed to be a street address and is shaped as other street address in the tag addr. This is an imperfect solution that may include values that are not street address (e.g. 25th fire station). However, the exceptions are few and can be easily fixed manually.

Incorrect zip code:

Occasionally, postal codes in the data set are incorrect. To address this problem, I use a function to convert the postal code to the simple digit format by splitting it by the hyphen. Afterward, the function

checks for the proper length (5 digits) and the leading 2 digits (starting with 74). If the postal code passes the test, it is stored in the address dictionary. Otherwise, it is dropped.

2. Data Overview

File sizes:

Tulsa map.osm……61,816 KB

Cleaned Tulsa map.osm.json……..66,006 KB

Number of documents:

```
>db.Tulsa1.find().count()

273278
```

Number of nodes

```
> db.Tulsa1.find({"type":"node"}).count()

233375
```

Number of ways

```
> db.Tulsa1.find({"type":"way"}).count()

39903
```

Number of unique users:

```
> pipeline = ([{"$group":{"_id": "$created.user","count":{"$sum" : 1}}},

            {"$sort" : {"count" : -1 }} ])

288
```

Top 3 contributing users:

```
> pipeline = ([{"$group":{"_id": "$created.user","count":{"$sum" : 1}}},

            {"$sort" : {"count" : -1 }} ])
```

Paul Johnson

Daniel Jeffries

woodpeck_fixbot

Top 10 amenities:

```
> pipeline = ([{ "$match" : {"address.amenity": {"$exists": True} } },
            {"$group":{"_id": "$address.amenity"}},
            {"$sort" : {"count" : -1 }},
            {"$limit" : 10}]) [{u'_id': u'pub'},
[{u'_id': u'pub'},
 {u'_id': u'kindergarten'},
 {u'_id': u'fire_station'},
 {u'_id': u'community_centre'},
 {u'_id': u'recycling'},
 {u'_id': u'school'},
 {u'_id': u'restaurant'},
 {u'_id': u'post_box'},
 {u'_id': u'place_of_worship'},
 {u'_id': u'fast_food'}]
```

Top 3 cuisines:

```
> pipeline = ([{ "$match" : {"address.cuisine": {"$exists": True} } },
            {"$group":{"_id": "$address.cuisine"}},
            {"$sort" : {"count" : -1 }},
            {"$limit" : 3}])
[{u'_id': u'burger'}, {u'_id': u'Pub'}, {u'_id': u'deli'}]
```

3. Additional Ideas

When reviewing the data, I notice that there are many instances whereas tag k has “tiger:” in its attribute such as tiger:name, tiger:zip, tiger:county. According to the openstreetmap wiki page, TIGER data (Topologically Integrated Geographic Encoding and Referencing system) is produced by the US Census Bureau. It is a public domain data source with many features. Thus, it was used as a primary

source of populating the OpenStreetMap's United States map in 2007 and 2008. Since OpenStreetMap has been actively updated by users since then, it would be interesting to see the difference between the current OpenStreetMap and TIGER data. The comparison may also provide helpful insights when cleaning up the map data.

While the TIGER data was useful in updating the initial United States map for OpenStreetMap project, it has many problems. Hence, bots are used to map automated edit to correct the data. The woodpeck_fixbot, the 3rd most contributing user for my Tulsa dataset, is an example of such bot. One of the primary function of this bot, according to its wiki page, is to remove the TIGER tags from the OpenStreetMap data. As there are still many instances of TIGER tags in my dataset, a review of the bot functions is desirable to improve the quality of OpenStreetMap data in the future.

4. Reference

https://docs.mongodb.org/manual/reference/mongo-shell/

https://docs.mongodb.org/manual/meta/aggregation-quick-reference/

https://www.openstreetmap.org/export#map=10/36.1201/-95.8626

http://wiki.openstreetmap.org/wiki/TIGER

http://wiki.openstreetmap.org/wiki/Fixbot