OpenStreetMap Project

Data Wrangling with MongoDB

Map Area: Austin, TX, United States

Contents

Introduction 1

1. Problems Encountered in the Map 1

  1.1 Over-abbreviated street types 2

  1.2 Abbreviated direction 2

  1.3 Direction at the end of a street name 2

  1.4 Formats of Interstate Highway 35 3

  1.5 Phone Number 3

  1.6 Zip Code 3

  1.7 City Name 3

  1.8 State name 3

2. Data Overview 4

3. Additional Ideas 6

4. Conclusions 9

References 9

Introduction

The purpose of this project is to practice data wrangling. Since it requires the dataset to be larger than 50 MB and the data for the city I live is about 7 MB, I chose the OpenStreetMap data for the city of Austin, Texas.

1. Problems Encountered in the Map

After downloading this dataset, I briefly went over it and observed the following problems in the data.

- Over-abbreviated street types with or without punctuation, such as “Rd”, “St”,”St.”,”Ave”,“Ct”,”Blvd”,”Blvd.”,”Cv”,”Dr”,”Expwy”,”Ln”, “Hwy”
- Abbreviated direction including “S”, “S.”, “N”, “N.”, “E”, “E.”, “W”, “W.”
- I 35, IH 35, Ih 35, IH-35
- Phone number format.
- Zipcode: “TX 78759-3504”, “Tx”
- City: “Taylor, TX”, “Bee Cave, TX”, “Dripping Springs, Tx”, “N Austin”, “Austin, TX”
- State: “Tx”, “Texas”, ” Metric Boulevard”

1.1 Over-abbreviated street types

In this dataset, there are many kinds of over-abbreviated street types. Here is a list of corrections I made.

Abbreviation	*Change*
Rd/Rd.	Road
St/St.	Street
Ave/Ave.	Avenue
Ct/Ct.	Court
Blvd/Blvd.	Boulevard
Cv/Cv.	Cove
Dr/Dr.	Drive
Expwy/Expwy.	Expressway
Ln/Ln.	Lane
Hwy/Hwy.	Highway

1.2 Abbreviated direction

In this dataset, many of the directions are abbreviated using “N”, “S”, “E” or “W”. Through the data cleaning process, I changed them to full-name as “North”, “South”, “East” and “West”, respectively.

Together with the step in 1.1, now the street names in the dataset are in a uniform way. For instance, “S 1st St” now would become “South 1st Street”.

1.3 Direction at the end of a street name

There are a few street with the direction at the end of the name, e.g. “Hwy 290 W”, “Capital of TX Hwy N”. In the data cleaning process, I firstly put the direction at the front of the name, and the do other data wrangling steps.

1.4 Formats of Interstate Highway 35

In this dataset, interstate highway 35 is abbreviated in several ways as: “I 35”, “IH 35”, “Ih 35”, “IH-35”. In the data cleaning process, I have replaced all of them using “Interstate Highway 35”.

1.5 Phone Number

There are several formats of phone number in this dataset, e.g. “+1 512 836 7644”, “5123947041”, “+1 512 322-9168”, “(512) 996-8900”, “512-301-1007”, “+1-512-821-9472”,” 1-512-322-9977”,” 15124209035”,” 512 328 2114”.

Moreover there are some ill-formatted phone numbers, such as: “512) 719-5575”.

Since someone from foreign countries may want to call these phone numbers, it is better to update these number with the country code, area code and numbers in a formatted way. Therefore, I have updated all the phone number in the following format: +country code (area code)-other numbers. For instance, “5123947041” would now be “+1(512)394-7041”.

1.6 Zip Code

In this dataset, some postcodes are of the form “Tx 78759-3504” or “TX 76574”. Since the state information is redundant, I have removed the state information. So now all the postcode should be just numbers. To further facilitate later analysis, I also removed the latter four digits and only kept the first five digits.

1.7 City Name

In this dataset, some of the name fields contain the state name “TX”, some others don’t. Since this map is for the city of Austin in TX, we do not need the state name in the city field. Therefore, I have removed them. Now, “Taylor, TX”, “Bee Cave, TX”, “Dripping Springs, Tx”, “Austin, TX” would simply be “Taylor”, “Bee”, “Dripping Springs”, “Austin”.

1.8 State name

Several kinds of state name exist in this dataset, such as “Tx”, “Texas”, “TX”. And there are some state fields with the entry “Metric Boulevard”, which is incorrect. I have replace them to be “TX” for uniformity.

2. Data Overview

File size:

austin_texas.osm 170,908 KB

austin_texas.osm.json 255,506 KB

Number of documents:

db.austinmapfull.find().count()

856626

Number of nodes with contributor info:

db.austinmapfull.find({"created.user":{"$exists":1}}).count()

856626

Number of ways

db.austinmapfull.find({"type":"way"}).count()

80276

Number of nodes

db.austinmapfull.find({"type":"node"}).count()

775045

Number of unique users

db.austinmapfull.distinct("created.user").length

928

Top 10 contributing user

```
db.austinmapfull.aggregate([{"$group":{"_id":"$created.user","count":{"$sum":1}}},
{"$sort":{"count":-1}},{"$limit":10}])
```

```
{ "_id" : "woodpeck_fixbot", "count" : 239487 }
{ "_id" : "varmint", "count" : 38453 }
{ "_id" : "richlv", "count" : 36847 }
{ "_id" : "Clorox", "count" : 36304 }
{ "_id" : "Iowa Kid", "count" : 34288 }
{ "_id" : "HJD", "count" : 29017 }
{ "_id" : "afdreher", "count" : 26302 }
{ "_id" : "Chris Lawrence", "count" : 18542 }
{ "_id" : "TexasNHD", "count" : 18079 }
{ "_id" : "Longhorn256", "count" : 17951 }
```

Number of users appearing only once, twice, or three times (having 1, 2 or 3 posts)

```
db.austinmapfull.aggregate([{"$group":{"_id":"$created.user","count":{"$sum":1}}},
{"$group":{"_id":"$count", "num_users":{"$sum":1}}},{"$sort":{"_id":1}},{"$limit":3}])
```

```
{ "_id" : 1, "num_users" : 180 }
{ "_id" : 2, "num_users" : 73 }
{ "_id" : 3, "num_users" : 38 }
```

3. Additional Ideas

Additional data exploration using MongoDB queries:

Number of restaurant and pub:

db.austinmapfull.find({"amenity":"pub"}).count()

35

db.austinmapfull.find({"amenity":"restaurant"}).count()

687

Top 10 appearing amenities:

db.austinmapfull.aggregate([{"$match":{"amenity":{"$exists":1}}},{"$group":{"_id":"$amenity","
count":{"$sum":1}}},{"$sort":{"count":-1}},{"$limit":10}])

{ "_id" : "parking", "count" : 1852 }

{ "_id" : "restaurant", "count" : 687 }

{ "_id" : "waste_basket", "count" : 591 }

{ "_id" : "school", "count" : 563 }

{ "_id" : "fast_food", "count" : 503 }

{ "_id" : "place_of_worship", "count" : 487 }

{ "_id" : "fuel", "count" : 371 }

{ "_id" : "bench", "count" : 349 }

{ "_id" : "shelter", "count" : 231 }

{ "_id" : "bank", "count" : 150 }

Biggest religion group:

db.austinmapfull.aggregate([{"$match":{"amenity":{"$exists":1},"amenity":"place_of_worship"}}
,{"$group":{"_id":"$religion","count":{"$sum":1}}},{"$sort":{"count":-1}},{"$limit":1}])

{ "_id" : "christian", "count" : 445 }

Popular cuisines:

```
db.austinmapfull.aggregate([{"$match":{"amenity":{"$exists":1},"amenity":"restaurant"}},{"$group":{"_id":"$cuisine","count":{"$sum":1}}},{"$sort":{"count":-1}},{"$limit":6}])
```

```
{ "_id" : null, "count" : 369 }
{ "_id" : "mexican", "count" : 66 }
{ "_id" : "american", "count" : 26 }
{ "_id" : "pizza", "count" : 23 }
{ "_id" : "chinese", "count" : 20 }
{ "_id" : "italian", "count" : 16 }
```

Number of posts still in the first version:

```
db.austinmapfull.find({"created.version":"1"}).count()
```

```
446217
```

Most common postal code:

```
db.austinmapfull.aggregate([{"$match":{"address.postcode":{"$exists":1}}},{"$group":{"_id":"$address.postcode","count":{"$sum":1}}},{"$sort":{"count":-1}},{"$limit":1}])
```

```
{ "_id" : "78704", "count" : 65 }
```

Number of postal code starting with ‘78’:

```
db.austinmapfull.find({"address.postcode":{"$regex":"^78.+"}}).count()
```

```
1058
```

Nodes with or without address:

```
db.austinmapfull.aggregate([{"$match":{"type":"node","address":{"$exists":1}}},{"$group":{"_id":"Nodes_w/o_Addr","count":{"$sum":1}}}])
```

```
{ "_id" : "Nodes_w/o_Addr", "count" : 1490 }
```

db.austinmapfull.aggregate([{"$match":{"type":"node","address":{"$exists":0}}},{"$group":{"_id":"Nodes_w/o_Addr","count":{"$sum":1}}}])

{ "_id" : "Nodes_w/o_Addr", "count" : 773555 }

Observations:

(1) The contribution of users is heavily skewed.

- Top user's contribution percentage is 239487/856626 = 27.96%;
- Top three users' contribution percentage is 36.75%;
- Top 10 users' contribution percentage is 57.82%;

Since there are totally 928 uses, these 1% users contribute about 60% of the data.

(2) There are totally 775045 nodes. Within them 773555 nodes do not have addresses. That is 99.81% of the nodes do not have address information.

(3) Most of the zip code in Austin starts with '78'. Actually, on this map, 1058 addresses' zip code starts with '78'.

(4) The most popular food in Austin is Mexican food. While there are totally 69 American, Pizza, and Chinese restaurants, there are 66 Mexican restaurant, almost 10% of the total 687 restaurants. This is probably due to the demographic composition in Austin.

Other observations:

(1) There are many nodes that only have trivia information, such as id, longitude, latitude, user, version, timestamp, etc. These node do not provide any useful information for this map. If we remove these nodes, we could save a lot of space and time to analyzing this dataset. The challenge is that, when removing these node, many ways and relations will be affected. So removing nodes while updating ways and relations could be very challenging.

(2) Another possible improvement to this dataset is that, for every node, we add another attribute to this node indicating which ways this node is affiliated to. It could be implemented as a dictionary in Python with "Ways Affiliated" as the key and a list of ids of ways as the value. The challenging part is the algorithm complexity, which could be *O(NM)*, where N is the number of nodes, and M is the number of ways.

4. Conclusions

Lots of data cleanings have been done for this dataset. On the other hand, several insights have been gained through data exploration of this dataset. Since the much information of Austin on this map is still missing,

References

1) MAPZEN: https://mapzen.com/metro-extracts/