OpenStreetMap Project

Data Wrangling with MongoDB

Swain Tseng

Map area: Dublin, Ireland

https://www.openstreetmap.org/relation/282800

https://mapzen.com/data/metro-extracts/

1.The Reason Why is Dublin, Ireland

I am a Taiwanese and I am a navy so far. In our country, we have to be in military for one year. There is no freedom and excuse only the command of naval officer. My girlfriend have studied in Dublin, Ireland for eight months and she usually shared her life so I want to realize more about Dublin,Ireland.

2.Problems Encountered in the Map

After downloading the area file of Dublin, Ireland, I was trapped in how to use mongoimport to access file. I thought the osm file can be imported directly into mongodb and it didn't work. So, I asked my instructor and he told me that I use preparing_for_database.py to transfer it to Json file. Then, I found some problems with the data.

- Over-abbreviated street names("lord edward st")
- Post codes show up at city(Dublin 7)
- Post codes aren't standardized

Over-abbreviated Street Names

When I import dublin_ireland.osm.json to MongoDB, I found that some some street name is abbreviation such as, Ave, Rd, Rd. , Roafd, St, st so I created Audit.py to transfer “lord edward st” to “lord edward street”

Post Code also Show up at City

The problem is that the post codes also are written in city. Considering this problem, we can remove the post code and the file size will be smaller. We can get post code from postcode column so it is not necessary to put post code in city.

#Sort city by count, descending

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

{ "_id" : "Dublin", "count" : 7741 }
{ "_id" : "Lucan", "count" : 1322 }
{ "_id" : "Dublin 6", "count" : 994 }
{ "_id" : "Blanchardstown", "count" : 955 }
{ "_id" : "Dublin 1", "count" : 392 }
{ "_id" : "Dublin 8", "count" : 322 }
{ "_id" : "Dublin 7", "count" : 314 }
{ "_id" : "Dublin 2", "count" : 298 }
{ "_id" : "Dublin 3", "count" : 242 }
```

Post Code aren’t standardized

The standard post code of Ireland should be like this(D01 to D24 and D06W) but a few data does not like that. The data have another 4 character behind the post code. I think that it is not correct. Then, I searched on the internet and found it is open post code which is calculated by a free algorithm from latitude and longitude but most people only use D01 to D24 and D06W.

#Sort postcode by count, acesending

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

```
{ "_id" : "D11 K254", "count" : 1 }
{ "_id" : "D15 N5H6", "count" : 1 }
{ "_id" : "D15 C9K2", "count" : 1 }
{ "_id" : "D15 V2X4", "count" : 1 }
{ "_id" : "D15 E8Y7", "count" : 1 }
{ "_id" : "25443", "count" : 1 }
{ "_id" : "D15 C9Y5", "count" : 1 }
{ "_id" : "D15 A3A8", "count" : 1 }
{ "_id" : "D15 WD5P", "count" : 1 }
{ "_id" : "D15 DX0H", "count" : 1 }
{ "_id" : "D15 TRX7", "count" : 1 }
```

The 3 and 5 column should be changed to D03 and D05.

#Sort postcode by count, descending

```
db.ireland.aggregate([{"$match":{"address.postcode":{"$exists":1}}},
{"$group":{"_id":"$address.postcode", "count":{"$sum":1}}}, {"$sort":
{"count":-1}}])
{ "_id" : "D02", "count" : 350 }
{ "_id" : "D07", "count" : 190 }
{ "_id" : "D01", "count" : 63 }
{ "_id" : "5", "count" : 50 }
{ "_id" : "D06", "count" : 33 }
{ "_id" : "3", "count" : 30 }
{ "_id" : "D08", "count" : 24 }
{ "_id" : "D11", "count" : 20 }
{ "_id" : "D12", "count" : 16 }
{ "_id" : "D04", "count" : 15 }
```

3.Data Overview

This section shows the basic statistics about the dataset and I use Mongodb queries to display it.

File size:

dublin_ireland.osm 259.8MB

dublin_ireland.osm.json 365.4MB

#Number of documents

```
> db.ireland.find().count()
1282393
```

#Number of nodes

```
> db.ireland.find({"type":"node"}).count()
1089072
```

#Number of ways

```
> db.ireland.find({“type":"way"}).count()
193289
```

#Number of unique users

```
> db.ireland.distinct("created.user").length
1421
```

#Top 5 contributing user

```
db.ireland.aggregate([
... ... {"$group":{ "_id":"$created.user", "count":{"$sum":1}}}
... ... ,{"$sort":{"count":-1}}
... ... ,{"$limit":5}])
{ "_id" : "Nick Burrett", "count" : 234952 }
{ "_id" : "mackerski", "count" : 185196 }
{ "_id" : "brianh", "count" : 155255 }
{ "_id" : "Dafo43", "count" : 153517 }
{ "_id" : "Conormap", "count" : 63929 }
```

#Number of users post 1-5 times

```
db.ireland.aggregate([{"$group":{"_id":"$created.user", "count":
{"$sum":1}}}, {"$group":{"_id":"$count", "num_users":{"$sum":1}}},
{"$sort":{"_id":1}}])
{ "_id" : 1, "num_users" : 346 }
{ "_id" : 2, "num_users" : 144 }
{ "_id" : 3, "num_users" : 89 }
{ "_id" : 4, "num_users" : 64 }
{ "_id" : 5, "num_users" : 73 }
```

The “_id” represent how many post created and the “num_users” represent how many people post “_id”

4.Additional Ideas

Contributor statistics and creative thinking

The statistics below is some contributor statistics.

- Top 1 user's percentage of all post ……………..……….. 21.5%
- Combined top 2 user's percentage of all post ………… 32.7%
- Combined top 5 user's percentage of all post ………… 61.8%

There are 1421 users and top 5 users contribute 61.8% of all post. OpenStreetMap should give them an award to thank their tremendous contribution. With this thinking, We can suggest OpenStreetMap to create a leaderboard and give top 3 contributor bonus such like, money, badge, to motivate more users or potential users in order to map edit.

#Top 10 appearing amenity

```
db.ireland.aggregate([{"$match":{amenity:{"$exists":1}}}, {"$group":
{"_id":"$amenity","count":{"$sum":1}}},{"$sort":{"count":-1}},
{"$limit":10} ])
{ "_id" : "parking", "count" : 2250 }
{ "_id" : "pub", "count" : 697 }
{ "_id" : "restaurant", "count" : 662 }
{ "_id" : "fast_food", "count" : 588 }
{ "_id" : "cafe", "count" : 575 }
{ "_id" : "school", "count" : 545 }
{ "_id" : "post_box", "count" : 443 }
{ "_id" : "place_of_worship", "count" : 386 }
{ "_id" : "bench", "count" : 377 }
{ "_id" : "bicycle_parking", "count" : 313 }
```

#How many nightclub

```
db.ireland.aggregate([{"$match":{amenity:"nightclub"}},  {"$group":
{"_id":"nightclub","count":{"$sum":1}}}])
{ "_id" : "nightclub", "count" : 20 }
```

Ireland is a country whose people love beer. As my queries below can know the second appearing amenity is pub, my friend who live in Dublin also tell me that almost people would go to pub after getting off work. There is a interesting point which is that Irish love drinking but they would go to pub rather than nightclub. Pubs are 30 times than nightclub.

Additional queries in MongoDB

#How many bunkers

```
db.ireland.aggregate([{ "$match":{military:"bunker"}}, {"$group":
{"_id":"bunker","count":{"$sum":1}}}])
{ "_id" : "bunker", "count" : 5 }
```

5.Conclusion

After doing this project, the file of Dublin, Ireland is almost completed. A small country is edited by 1282393 documents but the dataset have to be cleaned. OpenStreetMap can define a clear rule which is created by region or country in order to making data cleaner. As I mentioned above, post code standard(open post code or traditional post code) and city name format should be unified. OpenStreetMap can also use tutorial to guide users so that the data will be cleaner and data analysts or students would not spend so much time to process data.

I believe that OpenStreetMap can establish an environment where people can not only edit map but also have fun. It can post some interesting issue on the website and let people to verify or reject it. Then, give them some points which is to rank on the leaderboard and it can connect with FaceBook or Twitter so that your friends will be on the leaderboard. People will be motivated by their friends. OpenStreetMap will be not only a map but also a playground.

Reference:

http://wiki.openstreetmap.org/wiki/Map_Features#Amenity

https://docs.mongodb.org/manual/

https://en.wikipedia.org/wiki/Postal_addresses_in_the_Republic_of_Ireland