Data Wrangling with MongoDB

OpenStreetMap Project

Jamie Brand - brandjamie@hotmail.com

Map Area: Colchester, England.
https://www.openstreetmap.org/relation/76493
https://s3.amazonaws.com/metro-extracts.mapzen.com/colchester_england.osm.bz2

1. Problems Encountered in the Map

Using the functions in the 'investigate.py' script it was clear there were a number of problems with the data. I concentrated on the most common tags as identified in the 'count_tags' function. For each tag I used a regular expression to check for appropriate values and where appropriate I aggregated the data to identify any unusual values. There were a number of problems specific to different tags.
All data was converted to json using the 'clean.py' python script and imported into MongoDB using the 'mongo.py' python script (note: python scripts are using python 3).

1.1) Street names:

The first problem I found with the street names was the usage of a mixture of abbreviations or capitalizations to refer to the same thing. i.e. st. Street etc. In the cleaning process the last words of each street name were changed to full spelling with capital letters. The same was done for words preceeding street names that ended with 'North', 'South', 'West' and 'East'.
There were a also few unusual place names that I had to check manually but did not need to change. Such as, "Tally Ho" and "The Street With No Name". There was also a with "The Centre, The Crescent, Colchester Business Part", in which, after a little research, the last words were changed to 'Business Park'. In addition, a number of streets were named 'different'. As this information is of no use these tags were removed. There was also one street name value that contained the house number as well as the street name (25 Culver St W). For this tag, the street was changed to 'Culver Street West' and a housenumber tag with a value of 25 was added.

1.2) House numbers, house names and street names:

Looking at house numbers there are a lot of house names which need to be reassigned to the housename tag. Of course some number should have alphabetic components i.e. 22a, Flat 2, Unit 7. There were also housenumbers which included the street name.
The housenames were also problematic for the same reason. The tags often contained housenumber or streetname values. There were a number of housenames given as 'LB1','LB2' etc. These housenames were not changed it was unclear what the data should be. Another unusual value was 'A12 Southbound' which was removed as this is the name of a road, not a building.
When cleaning the street, housenumber and housename tags each tag was checked for what information it contained. It was usually possible to reassign the correct values to the correct tag. In some cases there were conflicting values for a tag. In those situations the tags were not changed beyond spelling and capitalization.

1.3) Postcodes:

Looking at postcodes, I found a postcode with an exclamation mark - which is not possible in the uk. I checked the address and it seems the exclamation mark should have been a '1' (possibly caused by OCR reading?). The postcode was 'CO!6 7BJ'. A couple of individual postcodes were not formatted well and were corrected.

1.4) Buildings:

The 'building' tag had some unsusual results for instance 'bing'. There was one 'no' and 9921 'yes' tags. The wiki says that 'yes' tags are acceptable but the 'no' tags had no information and so were removed. 'bing' tags were also removed.

There was at least one tag which didn't fit the apparant rule of the tag of using lower case characters with an underscore instead of a space.
There were a surprising number of 'pig_ark' and 'pigl_ark' tags which, as per the suggestion in the wiki can be merged with 'sty'.

1.5) Other tags:

The 'source:building' and 'source' tags do not contain important information for our purposes however the most common tags were 'bing' and 'Bing' so these tags were merged. All spaces were replaced with underscores.
There were no issues with the 'addr:city' tag.
There were two incorrect values in the 'natural' tag. "Bently Childrens Play Area", "Hollands Farm". These were changed to the 'housename' tag.
For the 'maxspeed' tag:
Should be in the format '80 mph' or '20 mph'. Some are in the format '80' or '20mph'. These were changed to the correct tag. An unsual value of 161 was checked but appears to be correct as it was for a railway, not a street.
The 'landuse', 'amenity','barrier' and 'service' tags were all corrected to use only lower case and underscores instead of spaces.

2. Data Overview

This section contains basic statistics about the dataset and the mongoDB queries used to gather them.

2.1) File sizes:

colchester_england.osm 229MB
colchester_england.osm.json 347MB

2.2) Additional Statistics:

# Number of documents

```
    > db.colchester.find().count()
    1198581
```

# Number of nodes

```
> db.colchester.find({"type":"node"}).count()
1091773
```

# Number of ways

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

106798
```

# Number of users

```
> db.colchester.distinct("created.user").length

436
```

# Top one contributing user:

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

{ "_id" : "EdLoach", "count" : 586030 }
```

# Number of users with one contribution

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

{ "_id" : 1, "num_users" : 87 }
```

# Number of pubs

```
> db.colchester.find({"amenity":"pub"}).count()

258
```

# Number of places of worship

```
> db.colchester.find({"amenity":"places_of_worship"}).count()

238
```

3. Additional Ideas

3.1) Improving the dataset:

The biggest problem in cleaning this data was the inconsitency in dealing with certain addresses. In particular I found addresses with a Unit or Block number difficult to catagorize. I choose to tag these as house numbers. However, this often did not fit the data.

For example, here are two documents which contain a unit number as part of the address. One with unit number as a housename, one as a housenumber :

```
> db.colchester.aggregate([{"$match":{"address.housenumber":
... {"$exists":true}}},{"$match":{"address.housenumber":/Unit/}},
... {"$match":{"address.housename":{"$exists":true}}},{"$limit":1}])

{ "_id" : ObjectId("55c75ae5eab6a825c818bb9b"), "website" :
"www.archimedeslogistics.com", "address" : { "housenumber" : "Unit 5",
"postcode" : "CO11 1AL", "housename" : "No 1 The Maltings The Quayside
Maltings" }, "id" : "1422150649", "pos" : [ 51.9442634, 1.0812359 ],
"type" : "node", "created" : { "uid" : "251236", "changeset" :
"9201273", "timestamp" : "2011-09-03T15:10:27Z", "user" : "percivjr",
"version" : "1" }, "name" : "Archimedes Logistics" }
```

```
> db.colchester.aggregate([{"$match":{"address.housename":
... {"$exists":true}}},{"$match":{"address.housename":/Unit/}},
... {"$match":{"address.housenumber":{"$exists":true}}},{"$limit":1}])

{ "_id" : ObjectId("55c75b0deab6a825c81fe0b6"), "id" : "125763292",
"website" : "www.steponsafety.co.uk", "address" : { "housenumber" :
"122", "postcode" : "CO11 2LH", "housename" : "Unit 3-4" }, "building" :
"yes", "pos" : [ 0, 0 ], "type" : "way", "created" : { "uid" : "251236",
"changeset" : "8996583", "timestamp" : "2011-08-12T14:36:24Z", "user" :
"percivjr", "version" : "1" }, "name" : "Step On Safety Ltd" }
```

It is unclear how these addresses should be categorized and I feel additional tags would have been usefull here along with more specific instructions regarding how the data should be entered. While I could find information on all of the tags in the openstreetmap docmentation, I could not find anything on best practices for the format of addresses. While having a more rigid address format may prove troublesome to some of the contributors, especially when importing from other sources, it would make the data much easier to organise and search.

3.2) Additional data exploration:

# Years data was added.

```
> db.colchester.aggregate([{"$group":{"_id":
... {"$substr": "$created.timestamp",0,4]}, "count": {"$sum":1}}},
... {"$sort" : {"count":-1}}])

{ "_id" : "2010", "count" : 404780 }
{ "_id" : "2011", "count" : 246271 }
{ "_id" : "2012", "count" : 168189 }
{ "_id" : "2014", "count" : 142004 }
{ "_id" : "2013", "count" : 95009 }
{ "_id" : "2015", "count" : 94812 }
{ "_id" : "2009", "count" : 39958 }
{ "_id" : "2008", "count" : 6486 }
{ "_id" : "2007", "count" : 538 }
{ "_id" : "2006", "count" : 534 }
```

# Biggest religion:
# In fact the only religion represented. I can only assume the data is incomplete.

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

    { "_id" : "christian", "count" : 237 }
```

# Biggest Christian Denomination.

```
    > db.colchester.aggregate([{"$match":{"amenity":{"$exists":true}}},
    ... {"$match":{"amenity":"place_of_worship"}},
    ... {"$match":{"religion":"christian"}},
    ... {"$match":{"denomination":{"$exists":true}}},
    ... {"$group":{"_id":"$denomination","count":{"$sum":1}}},
    ... {"$sort":{"count":-1}},{"$limit":5}])

    { "_id" : "anglican", "count" : 137 }
    { "_id" : "methodist", "count" : 20 }
    { "_id" : "catholic", "count" : 10 }
    { "_id" : "baptist", "count" : 9 }
    { "_id" : "united_reformed", "count" : 7 }
```

# Most popular restaurant cuisines:

```
    > db.colchester.aggregate([{"$match":{"amenity":{"$exists":true}}},
    ... {"$match":{"amenity":"restaurant"}},{"$match":{"cuisine":
    ... {"$exists":true}}}, {"$group":{"_id":"$cuisine","count":
    ... {"$sum":1}}},{"$sort":{"count":-1}},{"$limit":5}])

    { "_id" : "indian", "count" : 19 }
    { "_id" : "italian", "count" : 9 }
    { "_id" : "chinese", "count" : 7 }
    { "_id" : "fish_and_chips", "count" : 6 }
    { "_id" : "pizza", "count" : 3 }
```

# Most popular fast food cuisines:

```
    > db.colchester.aggregate([{"$match":{"amenity":{"$exists":true}}},
    ... {"$match":{"amenity":"fast_food"}},{"$match":{"cuisine":
    ... {"$exists":true}}},{"$group":{"_id":"$cuisine","count":{"$sum":1}}},
    ... {"$sort":{"count":-1}},{"$limit":5}])

    { "_id" : "chinese", "count" : 29 }
    { "_id" : "fish_and_chips", "count" : 27 }
    { "_id" : "sandwich", "count" : 11 }
    { "_id" : "burger", "count" : 7 }
    { "_id" : "indian", "count" : 5 }
```

#Most popular sports

```
     > db.colchester.aggregate([{"$match":{"sport":{"$exists":true}}},
     ... {"$group":{"_id":"$sport",  "count":{"$sum":1}}},
     ... {"$sort":{"count":-1}},{"$limit":5}])

     { "_id" : "soccer", "count" : 48 }
     { "_id" : "tennis", "count" : 31 }
     { "_id" : "cricket", "count" : 22 }
     { "_id" : "multi", "count" : 20 }
     { "_id" : "bowls", "count" : 19 }
```

```
# Most common pub name
     > db.colchester.aggregate([{"$match":{"amenity":{"$exists":true}}},
     ... {"$match":{"amenity":"pub"}}, {"$group":{"_id":"$name",  "count":
     ... {"$sum":1}}},{"$sort":{"count":-1}},{"$limit":5}])

     { "_id" : "The Crown", "count" : 6 }
     { "_id" : "The Red Lion", "count" : 4 }
     { "_id" : "The Swan", "count" : 4 }
     { "_id" : "White Hart", "count" : 4 }
     { "_id" : "The White Hart", "count" : 3 }
```

3.3) Conclusion:

After this review of the data it’s obvious that the Charlotte area is incomplete, though I believe it has been well cleaned for the purposes of this exercise. It interests me to notice a fair amount of GPS data makes it into OpenStreetMap.org on account of users’ efforts, whether by scripting a map editing bot or otherwise. With a rough GPS data processor in place and working together with a more robust data processor similar to data.pyI think it would be possible to input a great amount of cleaned data to OpenStreetMap.org.

The OSM data for the Colchester area has a lot of information which, on the surface seems quite complete. However there are obviously some gaps in the data. For instance the dataset returned 128 christian places of worship but none for any other religion. A quick internet search shows there are both mosques and synagues in the area but these are either not present or mislabled in our data. The timestamps show the data is relatively recent with most of the information being collected in the last six years. The is a substantial amount of recent activity so I expect to see many of the gaps filled in the near future.