OpenStreetMap Sample Project
Data Wrangling with MongoDB

Tim Lindsey

1. Problems Encountered in the Map

2. Data Overview

3. Additional Ideas

1. Problems Encountered in the Map

The data set downloaded had two primary issues after investigating a sample of the downloaded set from Map Zen

- Minimal address information
- Unique encoding requirements for the document

Minimal Address information

The data in the data set in general is pretty well formed, there are no significant lapses in data or inappropriate data polluting the set. For that reason I was comfortable injecting all “nd” and “tag” information into the json output (as compared to lesson 6 where a good bit of data was instructed to be ignored). However I notice as compared to much of the US data and data from the lesson there is not a ton of address info in the set. This just appears to be a deficiency of the users that were posting this data.

Unique Encoding Requirements For The Document

```
}
```

Because of some of the unique characters involved in the language from the data set (it’s all in icelandic), I had to before some character encoding to move the data to Unicode then into the proper encoding before output into the json output file. After research it appears this would have been slightly easier in Python 3+ as it defaults to Unicode rather than UTF-8. Once example of a troublesome line:

<tag k="addr:street" v="Hljóðalind"/>

2. Data Overview

Here is some general info on the data used in this lesson.

File Size:

reykjavik_iceland.osm - 181.6MB

reykjavik_iceland.json - 176.8

Example of Data Imported Into Mongo:

Data structure of single user

> *db.collection1.findOne()*

```
{
        "_id" : ObjectId("56b05556398a2ff21ff6b7e6"),
        "id" : "12885854",
        "pos" : [
                "-22.5434936",
                "63.9721673"
        ],
        "created" : {
                "timestamp" : "2013-09-07T14:43:51Z",
                "changeset" : "17719085",
                "user" : "MdMax",
                "version" : "5",
                "uid" : "136345"
        }
}
```

Submissions by an active user “MdMax”

```
db.collection1.find({},{"user": "MdMax"})
{ "_id" : ObjectId("56b05556398a2ff21ff6b7e6") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7e7") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7e8") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7e9") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7ea") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7eb") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7ec") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7ed") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7ee") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7ef") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7f0") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7f1") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7f2") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7f3") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7f4") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7f5") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7f6") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7f7") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7f8") }
{ "_id" : ObjectId("56b05556398a2ff21ff6b7f9") }
```

Unique Users

```
db.collection1.distinct("created.user").length
559
```

Total Records in The Document

865913

3. Additional Ideas

Some consideration of ideas to add to help make the dataset more complete:

Motivate users to complete parameters completely and correctly:

Overall it’s clear the users have taken good care to keep the data set clear and pretty consistent. As compared to the data set from lesson 6 it seems that there are far fewer anomalies to account for. Also, a nice bonus was that there were no street names that needed to be changed from abbreviation to the full name. However, there are a few instances where there appear to be inconsistencies in data loading. Here is one example:

```
<tag k="addr:housenumber" v="2"/>
<tag k="addr:housenumber_1" v="4"/
```

These are 2 tags side by side in one of the datasets. It does not appear as though this would be a valid address input and the issue was either introduced by human error or by some sort of bug in the system that compiles this data. It would appear enforcing stricter parameters around these inputs would help clean the data up front.

Increase datapoints

While the data is clean and a good effort it appears by the users, it seems like the available data is a little light. It may be possible to help motivate new or existing users to provide additional data in the future by offering local discounts, perhaps announcing “winners” and offering prized through some sort of locally help competition, or encouraging locales to get the word out about the program and drive their own activity in openmaps.

Conclusion

Getting data from a place like iceland provides some interesting insights into how a project like this translates overseas. Considering the project is largely volunteer run the data itself is extremely clean and practical to access and load into the database. However it is a somewhat light dataset as opposed to some of the other regions and I believe that to be able to fully compare and find use of this data the users will have to populate this region with more useful data points or new users will have to get interested in the project to help fill the Reykjavik dataset in more completely.