OpenStreetMap Project
Data Wrangling with MongoDB
Baba Mahmudov

Map Area: Niamey, NE, Niger
http://www.openstreetmap.org:
Niamey

1. Problems Encountered in the Map

After downloading a sample size of the Niamey area and skimming it, I noticed three main problems in the data:

- The names of type of Holy Trinity schools were in french;
- There were underscore characters in some tag values;
- Some postal codes were incorrect.

French names of school types

As I noticed in the data all the names of Holy Trinity school types were in French language: Secondaire, primaire and kindergarten. I changed them to English versions: Secondary, primary and kindergarden.

Underscore characters in some tag values

Strings of all tag values consisted of two or more words (like "place_of_worship") had underscore ("-") character in them. I removed underscore characters.

Incorrect postal codes

Some postal codes had words, letters and dot notations. I checked on internet that B.P means boite postale which is translated into english as mailbox and PO box means post-office box, which is relevant in Africa. I deleted them and made postcodes consist of only numbers.

Also, I modified values starting with "addr:" and created tag values "address" and those "address" values included a list of subtags consisted of second part of those strings after "addr:" part.

2. Data Overview

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

Fill sizes

niamey_niger.osm	62 mb
niamey_niger_new.osm.json	72 mb

```
# Number of documents
> client.dbname3.coll3.find().count()
334382
```

```
# Number of nodes
> client.dbname3.coll3.find({"type":"node"}).count()
270572

# Number of ways
> client.dbname3.coll3.find({"type":"way"}).count()
63671

# Number of unique users
> len(client.dbname3.coll3.distinct('created.user'))
151

# Top 10 contributing users
> client.dbname3.coll3.aggregate([{"$group":{"_id": "$created.user",
"count":{"$sum":1}}},
                               {"$sort":{"count":-1}},
                               {"$limit":10}])
[{u'_id': u'mahaman ali douka', u'count': 168794},
 {u'_id': u'AlioSamaila', u'count': 49869},
 {u'_id': u'fatimanalher', u'count': 46566},
 {u'_id': u'Mahamane Abdoulkader', u'count': 26236},
 {u'_id': u'dbusse', u'count': 11652},
 {u'_id': u'Hassane Amadou Daouda', u'count': 9831},
 {u'_id': u'datalogg', u'count': 3470},
 {u'_id': u'Dan Wood', u'count': 2815},
 {u'_id': u'nomisonson', u'count': 1533},
 {u'_id': u'zenfunk', u'count': 1273}]

# Number of users having 1 post
> client.dbname3.coll3.aggregate([{"$group": {"_id": "$created.user", "count":
{"$sum":1}}}, {"$group": {"_id":"$count", "num_users":{"$sum":1}}},
                               {"$sort": {"_id":1}}, {"$limit": 1}])
[{u'_id': 1, u'num_users': 22}]
```

3. Additional Ideas

Contributor statistics

Top user contribution percentage (“mahaman ali douka”) - 50.48%
Combined top 2 users' contribution (“mahaman ali douka” and “AlioSamaila”) - 65.39%
Combined Top 10 users contribution - 96.31%

There are many OpenStreetMap groups on facebook based in different countries. If all countries had such a facebook group and gave awards to top contributers, it would encourage many people to make edits on the map. However, this in turn may bring many incorrect information. Moreover OpenStreetMap needs more elaboration. There are very few street names in Niamey map. More contributors would mean more street names.

Additional data exploration using MongoDB queries

```
# Biggest religion in Niamey
> client.dbname3.coll3.aggregate([{"$match":{"amenity":{"$exists":1},
"amenity":"place of worship"}},
                               {"$group":{"_id":"$religion", "count":{"$sum":1}}},
                               {"$sort":{"count":-1}},
                               {"$limit":1}])

[{u'_id': u'muslim', u'count': 116}]
```

It is not surprising, as 80% of Niger's population are muslim.

```
# Number of hotels in Niamey
> len(list(client.dbname3.coll3.find({"tourism":"hotel"})))
22

# Top 3 places in Niamey
> client.dbname3.coll3.aggregate([{"$match":{"place":{"$exists":1}}},
                               {"$group":{"_id":"$place", "count":{"$sum":1}}},
                               {"$sort":{"count":-1}},
                               {"$limit":3}])

[{u'_id': u'village', u'count': 64},
 {u'_id': u'suburb', u'count': 60},
 {u'_id': u'hamlet', u'count': 4}]

# Number of cemeteries in Niamey
> client.dbname3.coll3.find({"landuse":"cemetery"}).count()
4
```

Lets see now, how many of them belong to muslims or christians

```
> client.dbname3.coll3.aggregate([{"$match": {"$and":[{"religion":{"$exists":1}},
{"landuse":"cemetery"}]}},
                               {"$group":{"_id":"$religion", "count":{"$sum":1}}},
                               {"$sort":{"count":-1}}])

[{u'_id': u'muslim', u'count': 2}, {u'_id': u'christian', u'count': 1}]
```

# Most edited item in the data set

In order to find it, I change 'version' number strings into integers:

```
>for doc in client.dbname3.coll3.find({"created.version":{"$not":{"$type":15}}}):
    doc["created"]["version"] = int(doc["created"]["version"])
    client.dbname3.coll3.find_one_and_update({'_id':doc['_id']},
{"$set":{'created':{'version':doc['created']['version']}}})
```

Then I make query to find the most edited item:

```
> client.dbname3.coll3.aggregate([
                               {"$group":{"_id": "$_id", "maxVersion":{"$max":
"$created.version"}}},
                               {"$sort":{"maxVersion":-1}},
                                {"$limit":1}
                                ])

[{u'_id': ObjectId('57c81a798c8a3fcf53dfc119'), u'maxVersion': 238}]
```

Conclusion
Although I cleaned some parts of the data, Niamey area still has some problems. Firstly, it has many french words and when we process the file some french letters turn into problematic characters. For example, "Avenue du Général de Gaulle" turned into "Avenue du G\u00e9n\u00e9ral de Gaulle". The data needs more and more cleaning. During the data wrangling I solved some such problems and gave some basic descriptions about the data of Niamey area.