OpenstreetMap Project Data Wrangling with MongoDB

# Problems encountered in the map

After initial Downloading the file for Dubai- Abu Dhabi Area , and running against provisional data.py File. I have noticed following issue

## Names are in Arabic

The data contains “name” attribute value in Arabic, the correct “name” is in attribute of K tag with the name “Name:en”

<tag k="name" v="کیش"/>

<tag k="name:en" v="Kish"/>

### Solution for Names in Arabic

While creating the Json file , we first check for existence of “name:en" and the same equivalent values are now passed to json attribute “name”

## Data has In:country from iran

The Data contains tag element with attribute “is\_in:country” with text value of “Iran” which is different country than United Arab Emirates, which we are interested in, these are errors in the data.

The Kish island is not part of United Arab Emirates , it is part of iran, but somehow it is part of this Openstreet Map data, which is to be corrected.

<tag k="is\_in:country" v="Iran"/>

### Solution for data contains IRAN

While creating the Json file , we first check for existence of "is\_in:country" and the same equivalent values of V attribute if it is Iran, we simply skip the row

## Incorrect data for amenity

On querying dubai collection, we found the distinct value of Amenity using following clause

mongo --quiet local --eval 'printjson(db.dubai.distinct("amenity"))' > output.json

we figured out that there is an incorrect value of Amenity as “Jebel Ali Station”, the location is correct and it is Jebel Ali Station , so we corrected the Amenity as bureau\_de\_change",

### Updating Incorrect data for amenity

using update\_amenity.py

> db.dubai.find({"amenity":"Jebel Ali Station"}).count()

1

> db.dubai.find({"amenity":"Jebel Ali Station"})

{ "\_id" : ObjectId("572a225072edaabb5f02fb92"), "amenity" : "Jebel Ali Station", "network" : "RTA", "created" :

{ "changeset" : "37140259", "user" : "ConsEbt", "version" : "8", "uid" : "313448", "timestamp" : "2016-02-11T07:

03:44Z" }, "pos" : [ 55.0910256, 24.9774785 ], "station" : "subway", "address" : { "city" : "Dubai", "country" :

"AE" }, "railway" : "station", "type" : "node", "id" : "1379661746", "name" : "UAE Exchange" }

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

176

After executing update\_amenity.py

$ sudo python3 update\_amenity.py

$ mongo

> use local

switched to db local

> db.dubai.find({"amenity":"Jebel Ali Station"})

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

177

> db.dubai.find({"amenity":"Jebel Ali Station"}).count()

0

> db.dubai.find({"name" : "UAE Exchange"})

{ "\_id" : ObjectId("572a225072edaabb5f02fb92"), "amenity" : "bureau\_de\_change", "type" : "node", "railway" : "st

ation", "name" : "UAE Exchange", "id" : "1379661746", "address" : { "city" : "Dubai", "country" : "AE" }, "pos"

: [ 55.0910256, 24.9774785 ], "created" : { "uid" : "313448", "changeset" : "37140259", "timestamp" : "2016-02-1

1T07:03:44Z", "version" : "8", "user" : "ConsEbt" }, "network" : "RTA", "station" : "subway" }

{ "\_id" : ObjectId("572a225072edaabb5f032d59"), "amenity" : "bureau\_de\_change", "name" : "UAE Exchange", "create

d" : { "changeset" : "9881953", "user" : "MathewJ", "version" : "1", "uid" : "451206", "timestamp" : "2011-11-20

T10:25:22Z" }, "pos" : [ 55.3384777, 25.2772196 ], "type" : "node", "id" : "1509578887" }

{ "\_id" : ObjectId("572a225172edaabb5f0337f7"), "amenity" : "bureau\_de\_change", "name" : "UAE Exchange", "create

d" : { "changeset" : "10410357", "user" : "Walter Schlögl", "version" : "2", "uid" : "78656", "timestamp" : "201

2-01-16T17:51:39Z" }, "pos" : [ 55.3476672, 25.2785833 ], "type" : "node", "id" : "1536329967" }

This confirms the update clause works fine.

# Data Overview

The section contains basic statistics about the dataset and the MongoDB Queries used to gather them

## File sizes

dubai\_abu-dhabi.osm 353.5 MB

dubai\_abu-dhabi.osm.json 157.9 MB

## Number of Documents

> use local

switched to db local

> db.dubai.find().count()

655515

## Number of Documents of type node

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

655509

## Number of Documents of type way

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

0

## Number of distinct users

> db.dubai.distinct("created.user").length

612

## Top contributing user

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

{ "\_id" : "eXmajor", "count" : 158029 }

## Number of contributing user with one post

> db.dubai.aggregate([{$group:{\_id:"$created.user","count":{$sum:1}}},{$match:{"count": 1}},{$group:{\_id:null,"count1":{$sum:1}}}, {"$limit":1}])

{ "\_id" : null, "count1" : 119 }