Data Wrangle OpenStreetMaps Data

Mohammed Awad-Allah

Map Area: Cairo, Egypt

Downloaded from: https://mapzen.com/data/metro-extracts
https://s3.amazonaws.com/metro-extracts.mapzen.com/cairo_egypt.osm.bz2

Problems Encountered in the Map

During the inspection of the data, I encountered the following problems:

1. Street Names are sometimes written in Arabic and sometimes in English.
2. Street type (Street, Road ... etc) is sometimes omitted altogether.

1. Arabic Street Names

In order to inspect Street names when written in arabic, the regular expression needed to be changed. As the word Street in Arabic (□□□□) would come at the beginning of the street name instead of the end. The regular expression would also need to incorporate Unicode characters in the match.

```
ar_street_type_re = re.compile(r'^\S+\b\.?', re.UNICODE)
```

I also updated the list of expected street types to add Arabic ones.

```
expected = ["Street", "Avenue", "Boulevard",
               "Drive", "Court", "Place",
               "Square", "Lane", "Road",
               "Trail", "Parkway", "Commons",
              u"شارع", u"ميدان", u"طريق",
              u"محور", u"المحور", u"كورنيش",
              u"مجاورة", u"امتداد"]
```

2. Omitted Street types

For Arabic Streets, all the ones that has an omitted type were streets, so prepending the word □□□□ (Street in Arabic) was easy.

However, for street names in English, that wasn't the case, and further data cleaning was not possible.

Data Overview

Here, I present some basic statistics about the data.

- **Original Data file Size:** 68 MB.
- **JSON file size:** 78 MB

- **Number of Documents:** 377430

```
db.nodes.count()
```

- **Number of Nodes:** 331607

```
db.nodes.find({"type" : "node"}).count()
```

- **Number of Ways:** 45814

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

- **Number of Unique Users:** 539

```
len(db.nodes.distinct("created.user"))
```

- **Number of Historic Places:** 402

```
db.nodes.find({"historic": {"$exists": True} }).count()
```

Additional Ideas

Arabic and English Names of Nodes

It appears that the nodes don't always have names in both Arabic and English, With English appearing more than Arabic.

- **Number of Nodes with a name attribute:** 2761
- **Number of Nodes with a name:en attribute:** 2032
- **Number of Nodes with a name:ar attribute:** 1723
- **Number of Nodes with a both name:ar and name:en attributes:** 1703

```
with_name = db.nodes.find({ "type": "node", "name" : { "$exists": True}}).count()
with_en_name = db.nodes.find({ "type": "node", "name:en" : { "$exists":
True}}).count()
with_ar_name = db.nodes.find({ "type": "node", "name:ar" : { "$exists":
True}}).count()
with_both_name = db.nodes.find({ "type": "node", "name:en" : { "$exists": True},
"name:ar" : {"$exists": True}}).count()
```

Top 5 Amenities

1. place_of_worship: 407
2. parking: 336
3. restaurant: 209
4. school: 178
5. cafe: 150

```
top_5_amenities = db.nodes.aggregate(
                        [ {"$match":{"amenity":{"$exists":1}}},
                          {"$group":{"_id":"$amenity","count":{"$sum":1}}},
                          {"$sort":{"count":-1}},
                          {"$limit": 5}])
```

Conclusion

The data for Cairo Map appears to need more cleaning, specifically in the areas of street names with no types and the availability of Arabic and English names.