OpenStreetMap Project
Data Wrangling with MongoDB

Gustavo Pereira

Map Area: Montevideo, Uruguay

1. Problems encountered in the map

The main issue I found was that in Spanish most street names do not begin with “Calle”, so you will get tons of first names as the beginning of a name. However, the data was VERY clean with respect to street, avenue, etc. names and there were very few abbreviations.

As a matter of fact, because the abbreviations were so few, I included them all in the regex expression to find and change them.

Another issue was that middle names were included as either the first letter only, eg.“A”, or as first letter and a dot, eg “A.”. To further complicate matters, there are streets that might be named “Pasaje A”, just using a letter of the alphabet. I chose to go the “easy” way and remove all dots from abbreviated names.

There was only one street with lower case name “soriano", which I fixed to “Soriano”.

The use of Unicode forced me to run in IPython, as a regular Python shell would throw errors regarding not being able to translate to ASCII.

Also, while looking at “schools”, I found another problem which I believe is beyond my current level in both Python and MongoDB. It is a consistency problem with the way the word ‘number’ is represented. In Spanish, “número” can be abbreviated as “No.”,”Nº”,”Nro” or just a blank. All public schools and high schools in Uruguay have a number… I think that the way to fix it would be very similar to the way street names were fixed.

Here are a few examples from the entry (edited to show the examples)

```
> db.montevideo_uruguay.osm.aggregate([{"$match":{"amenity":{"$exists":1},"amenity":"school"}},
{"$group":{"_id":"$name",  "count":{"$sum":1}}},{"$sort":{"count":-1}}])

{ "_id" : "Liceo 61", "count" : 1 }
{ "_id" : "Escuela nº 55 - República Federal de Alemania", "count" : 1 }
{ "_id" : "Escuela Nro 92", "count" : 1 }
{ "_id" : "Escuela Nº 130 \"Andrés Bello\" y Nº 283", "count" : 1 }
{ "_id" : "Liceo N° 30 Cagancha", "count" : 1 }
```

2. Data Overview

Some statistics on the data

montevideo_uruguay.osm file size: 132,140KB
montevideo_uruguay.osm.json file size: 98,013KB

I imported the json file into MongoDb, with collection name “montevideo_uruguay.osm”

Number of documents

>db.montevideo_uruguay.osm.find().count()
371994

Number of nodes

> db.montevideo_uruguay.osm.find({“type”:”node”}).count()
355597

Number of ways

> db. montevideo_uruguay.osm.find({“type”:”way”}).count()
15579

Distinct users

>db. montevideo_uruguay.osm.distinct(“created.user”).length
214

Top 5 contributors

>db.montevideo_uruguay.osm.aggregate([{"$group":{"_id":"$created.user","count":{"$sum":1}}},{"$sort":{"count":-1}},{“$limit”:5}])

{ "_id" : "xybot", "count" : 204045 }
{ "_id" : "muralito", "count" : 100957 }
{ "_id" : "MaxM", "count" : 36882 }
{ "_id" : "EkiZ", "count" : 9878 }
{ "_id" : "Zeroth", "count" : 5701 }

The top 3 contributors account for over 90% of the overall document. Also, the first one seems to be automated, given the name “xybot”.

3. Additional exploration and ideas

Very few people are involved in creating the OpenStreetMap repository for Montevideo. There should be a way to involve more people. Perhaps the fact that a GPS is necessary to log the data, and it being quite expensive here in Uruguay, prevents it.

My suggestion would be that institutions with the ability to purchase GPS devices at scale (city council, universities, etc.) give their employees/students/interns the ability to improve the map of given zones. This would be particularly interesting for students in Urban Planning because it would give them a chance to explore, record and then have enough material for their projects. Also, CS students might benefit from seeing all the steps involved in creating, cleaning and then do further processing of the data.

The main problem I see is that some training might be required, or that said institutions might not be willing to spend resources on the project. This could be counters by showing the advantages for the people working in the actual data gathering/wrangling/cleaning/processing as well as for society as a whole. In fact, the more and the easier the access is to information, the better choices a person can make. “Where is the nearest hospital? Pharmacy?” “I’ve run out of money, where’s the nearest ATM?” are just two simple questions that might help people under two very different situations where they could benefit from access to a good map with the information available.

Additional exploration

Top 5 amenities

```
>db.montevideo_uruguay.osm.aggregate({"$match":{"amenity":{"$exists":1}}},{"$group":{"_id":"$amenity","count":{"$sum":1}}},{"$sort":{"count":-1}},{"$limit":5})
```

```
{ "_id" : "pharmacy", "count" : 181 }
{ "_id" : "atm", "count" : 128 }
{ "_id" : "restaurant", "count" : 114 }
{ "_id" : "fuel", "count" : 110 }
{ "_id" : "bank", "count" : 87 }
```

It’s interesting that the main “amenity” are pharmacies, followed by ATMs. It seems to me that there is a “services” component skew in the data.

Pharmacy chains

```
>db.montevideo_uruguay.osm.aggregate({"$match":{"amenity":{"$exists":1},"amenity":"pharmacy"}},{"$group":{"_id":"$name","count":{"$sum":1}}},{"$sort":{"count":-1}},{"$limit":5})
```

```
{ "_id" : null, "count" : 12 }
{ "_id" : "Pigalle", "count" : 4 }
{ "_id" : "Calveira", "count" : 3 }
{ "_id" : "Trouville", "count" : 3 }
{ "_id" : "Farmashop", "count" : 2 }
```

Here I notice that a lot of data is missing, I know that Farmashop has at least 20 stores around the city, yet only 2 appear.

Perhaps if OpenStreetMap data were more visible, these big chain stores might like their data to be correct in order to direct customers to the nearest pharmacy, maybe through a phone app.

Restaurants and cuisine

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

{ "_id" : null, "count" : 66 }
{ "_id" : "steak_house", "count" : 16 }
{ "_id" : "pizza", "count" : 10 }
{ "_id" : "regional", "count" : 4 }
{ "_id" : "pizza;regional", "count" : 2 }
```

Again, of the 114 restaurants in the data, more than 50% do not have a cuisine attribute. The rest have repeating tags, although it is no surprise to me that steak houses and pizza places are on top as they are the favourite kind of food in the city.

Conclusion

Although the data is quite clean, there is a ton of missing information. Most of the information was provided by one or two users as GPS devices are not mainstream.

I proposed an idea of paid people working for the city council or students at universities adding to the data. Also, some big chains might like their stores to be more easily accessible through phone apps and might want to add to the data.

I also explained why it would be beneficial, trying to work around potential lack of interest from authorities.