DATA WRANGLING WITH MONGO DB

Analyzing OpenStreetMaps data

By Nikita Barsukov:

nikita@barsukov.net

barsukov.n@gmail.com

Map Area: Kharkiv, Ukraine

Table of Contents

Introduction

This text is a grading paper for Udacity's course "Data Wrangling with MongoDB". For this paper it was required to download an OpenStreetMaps dataset representing any part of the world, and perform a brief analysis of this dataset, according to given specifications. Minimum required size of an uncompressed dataset is 50 Mb.

I chose an area around Kharkiv, a large the city on the East of Ukraine, where I grew up. The dump is available on http://1drv.ms/1SnZXM3

I visualized some aspects of this dataset, made an overview of the dataset, tried to identify and, if possible, fix inconsistencies, and suggested ways to improve the quality of dataset. When I identified problems and directions for improvement, I also dataset for Kharkiv with two other data dumps for Stockholm and Copenhagen.

I crested quite a few amount of code to generate plots and figures in this paper. I used MongoDB for data wrangling, as required in this course. All of the queries in MongoDB were wrapped in Python scripts, using pymongo library. When required I included Mongo queries wrapped in Python. Plots were created using R programming language.

All the code is available on my GitHub repository, https://github.com/nikita-barsukov/data_wrangling

1. Data Overview

File size of original data dump is 179,5 Mb. It contains 901.917 elements

```
> client = MongoClient("mongodb://localhost:27017")
```

```
> coll = client['osm']['kharkiv']
```

```
> coll.count()
```

550 users added data to the dataset.

Top ten users are given in table 1 below.

Table 1 - Top ten contributors to OSM data of Kharkiv area

Username	Number of contributions
_sev	123.736
Rereader	77.624
Vort	59.236
Svargref	36.323
dimoster	32.980
headlong22	32.176
dima_ua_import	28.700
baleyko	18.041
bakasana	17.346

Breakdown by type of data looks like this:

Table 2 - node types in Open Street Maps data of Kharkiv area

Type	Count
node	766.864
way	134.983
route	24
heat	18
broad_leaved	15

coniferous	4
water	2
deciduous	2
mixed	1
public	1
gas	1
private	1
swamp	1

Types “node” and “way” are by far the most popular in the dataset. Other types are less frequent by several orders of magnitude.

There are 3432 amenities in the dataset. Top 10 amenities for Kharkiv area are:

Table 3 – Top 10 amenities in OSM data of Kharkiv area

Amenity	Count
'parking'	563
'fuel'	247
'school'	237
'pharmacy'	196
'bank'	189
'cafe'	182
'kindergarten'	145
'atm'	120
'hospital'	111
'restaurant'	98

Total number of distinct amenities in the dataset is 76.

Let’s see how the dataset was filled over the time.

Plot on Figure 1 below shows number of created elements in this dataset by month.

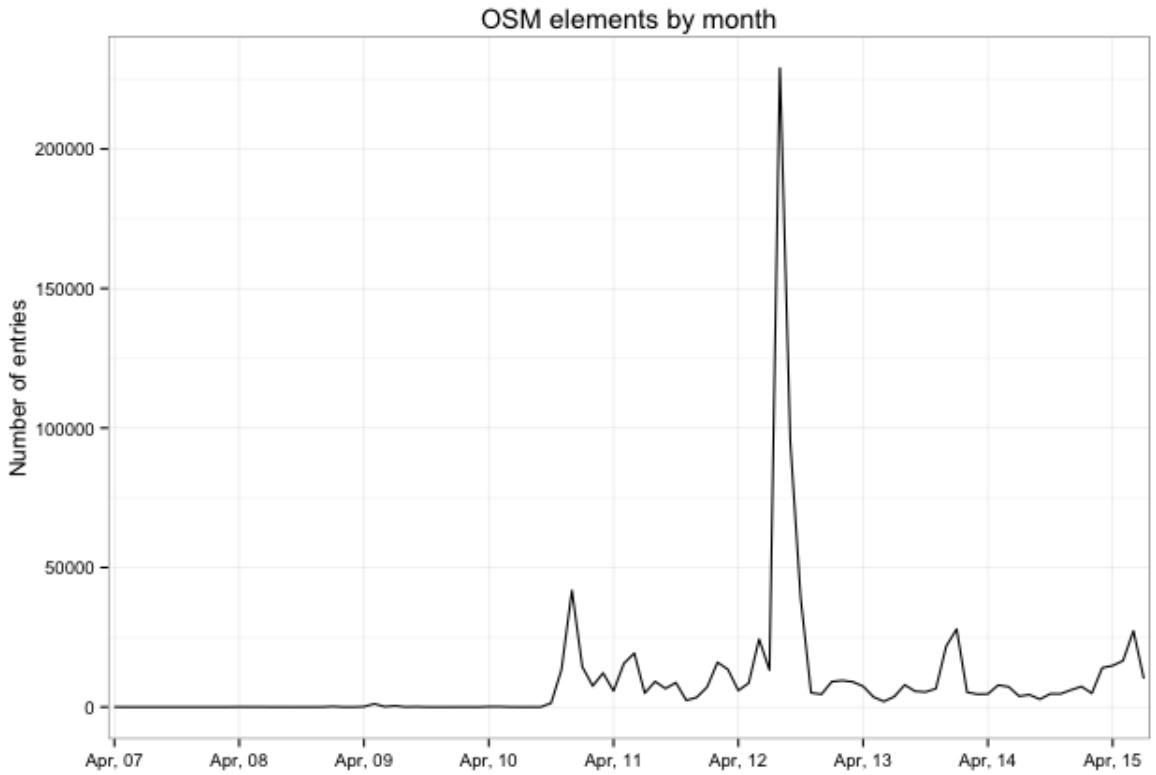

Figure 1

We can see that rate of adding elements to map of Kharkiv are is uneven. It was almost negligible during first three years, but it quickly started expanding after autumn 2010. We can also see a sharp spike in middle of 2012 (August 2012 to be precise). A bit less than 230.000 elements were created during this month, more than a quarter of all the elements in the dataset.

When we look at creation times at a different angle and break it down just by month and day of week, we will also see an interesting picture (see figure 2 below).

The breakdown by month shows us that most of entries were made in August, which corroborates with the previous line chart. However, breakdown by weekday is more or less even, there is no spike on that chart.

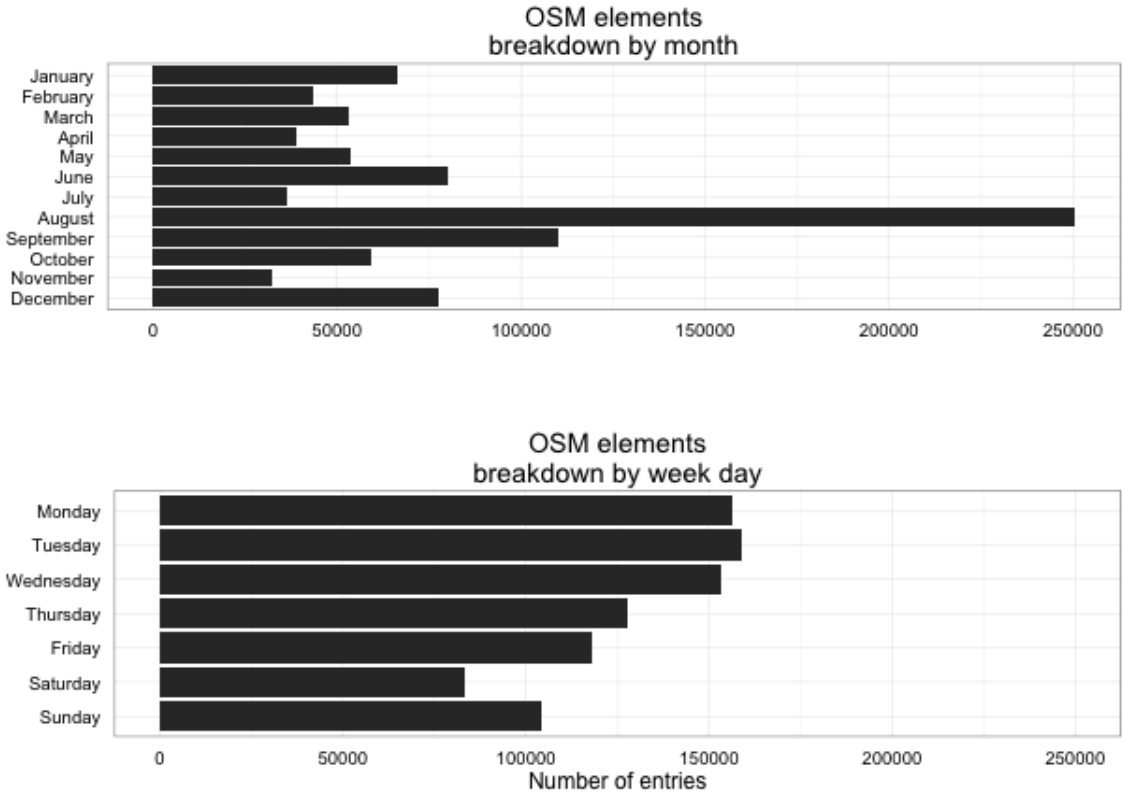

Figure 2

2. Problems with the dataset

One of the first problems that I saw with the dataset is addresses there. A lot of given addresses contained just street name or a postal code. Some of it was valid: small towns around Kharkiv have just a postcode, like this:

```
  <node id="337533668" lat="50.1178118" lon="36.3922213" version="4"
timestamp="2014-07-07T12:49:16Z" changeset="24003351" uid="204049"
user="olehz">

    <tag k="addr:postcode" v="62440"/>

  </node>
```

However, many nodes representing houses have simply a house number without street name, postal code or other address data. Counts of nodes with address by number of address fields looks like this:

Table 4 - Address fields in OSM dataset for Kharkiv area

Number of address fields	Number of documents
1	663

2	145
3	95
4	36
5	2

As we can see most of the nodes with given address do not have a full address. Compare this with open street maps data for Stockholm:

Table 5- Address fields in OSM dataset for Stockholm area

Number of address fields	Number of documents
1	48
2	32
3	478
4	290
5	415
6	19

Another problematic area of dataset of Kharkiv area is the fact that many proper names are given in several languages, typically English, Ukrainian and Russian. It creates some inconsistency, with plenty of named elements having only name in Ukrainian or Russian without English translation in it.

Table 6 - Number of translated names in OSM dataset of Kharkiv area

Total elements with name	**10.127**
Elements without English name	5.310
Elements without Russian name	3.806

Finally, there is another problem area. Some data points have website field, and as it happens so often, some of them are no longer active. Moreover, the site URLS are not normalized, many of them do not have http:// prefix, trailing slash is present in some cases, and absent in another.

Out of 221 documents with website only 8 had broken links. This problem is in fact easy to fix simply by removing broken websites from document.

3. Additional ideas for improving dataset.

The dataset can be improved in several different ways. An obvious thing would be to add more English variants of street names, amenities and other objects in the OSM database of the area.

Another area of improvement is to add better addresses, and I pointed out in previous section.

Also it looks like there are too phew cafes and restaurants in the OSM dataset for Kharkiv. In the dataset there are only 182 cafes and 98 restaurants, and note that Kharkiv's population is around 1,5 million. Compare this to Stockholm: 1371 restaurants and 747 cafes, or Copenhagen: 626 restaurants and 358 cafes.

This can indicate that plenty of objects within Kharkiv are still not added to OSM.

Sources

Since both Python and MongoDB are new technologies for me, I used on-line documentation and code snippets extensively.

1. Python Software Foundation. Python Language Reference, version 3.4.2. Available at http://www.python.org
2. R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
3. H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009.
4. http://docs.mongodb.org/
5. http://www.cookbook-r.com/Graphs/
6. https://docs.python.org/
7. http://stackoverflow.com/questions/18337407/saving-utf-8-texts-in-json-dumps-as-utf8-not-as-u-escape-sequence
8. http://stackoverflow.com/questions/3996904/generate-random-integers-between-0-and-9
9. http://stackoverflow.com/questions/2824157/random-record-from-mongodb
10. http://stackoverflow.com/questions/15092884/how-can-i-return-an-array-of-mongodb-objects-in-pymongo-without-a-cursor-can
11. http://stackoverflow.com/questions/3086973/how-do-i-convert-this-list-of-dictionaries-to-a-csv-file-python
12. http://stackoverflow.com/questions/13281733/is-it-possible-to-flatten-mongodb-result-query
13. http://stackoverflow.com/questions/16406329/python-dictionary-count-of-unique-values
14. http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_%28ggplot2%29/
15. https://stat.ethz.ch/pipermail/r-help/2007-September/140509.html
16. https://jira.mongodb.org/browse/SERVER-6074