Data Wrangling Open Street Map Data

Paul Stephens

Map Area: Kitchener-Waterloo-Cambridge, Ontario, Canada

https://www.openstreetmap.org/export#map=12/43.4279/-80.4539

Problems Encountered in you map

- Inconsistent Street Type Names
- Streets not found in 'node' data
- Fake Data

Inconsistent Street Type Names

Data cleaning was done directly using Python prior to loading data into MongoDB. After applying some of the code from exercises in lesson six of the Data Wrangling course it became clear that within the geographical area there are many types of streets, and many ways to spell (misspell) and abbreviate these. After pulling out some of the most common types ("Street", "Avenue", "Boulevard", "Drive", "Court", "Place", "Square", "Lane", "Road") I reviewed the remaining types for additional acceptable street types. I ultimately had a list of 58 acceptable street name types including many numbers which reference highways, and rural roads.

I completed a mapping to update misspelled and abbreviated types.

I took an iterative approach to parsing my source xml file and writing each line to a new file. Where needed I would complete a mapping to update a name to the proper version.

Streets Not found in 'node' data

The quizzes from lesson six only dealt with 'node' data where street names all started with 'addr:'. I found in my dataset that many road, primarily residential streets that did not have a particular address of interest on them were only found as 'ways'. The 'ways' did not contain 'addr:' fields. I still needed to clean many street types for these entries so I coded a second check for the ways to look for a 'highway' tag, and the audit the 'name' tag.

Fake Data

After filtering out acceptable street types, and mapping the mistakes and abbreviations there were many entries left that were clearly not legitimate streets. I did a further iteration where if I found an unacceptable street type I did not write the element to the new file.

Overview of the Data

Size of Original OSM File	71.8mb
After First Cleaning Pass (mapping)	73mb
Second Pass (removing bad data)	72.6mb
Third Pass(leaving only nodes and ways)	71.9mb
JSON file	80mb

Initial MongoDB Code

```
In [1]: from pymongo import MongoClient

In [2]: import pprint

In [3]: client = MongoClient("mongodb://localhost:27017")

In [4]: db = client.NewDB
```

Number of Records

```
db.StreetData.count()                                   ->      368960
```

Number of Unique Users

```
len(db.StreetData.distinct("created.user"))             ->      342
```

Number of Nodes and Ways

```
db.StreetData.find({"type":"node"}).count()             ->      328176

db.StreetData.find({"type":"way"}).count()              ->      40742
```

Total Amenities

```
db.StreetData.find({"amenity":{"$exists" : 1}}).count() ->      3182
```

Most Frequent Restaurant, Fast_Food, or Café

```
db.StreetData.aggregate([

        { "$match" : { "amenity" : {"$in":["restaurant","fastfood",”café”]}}}

        { "$group" : { "_id" : "$name", "count" : { "$sum" : 1}}},

        { "$sort" : { "count" : -1}}])

                                                        ->      ‘Tim Hortons’ with 46
```

Most Frequent Bank

```
db.StreetData.aggregate([

        { "$match" : { "amenity" : "bank"}},

        { "$group" : { "_id" : "$name", "count" : { "$sum" : 1}}},

        { "$sort" : { "count" : -1}}])

                                                        ->      ‘TD Canada Trust’ with 22
```

Other Ideas about the Datasets

- Completeness of postal codes. The dataset is not complete. The only way to address elements that are missing entirely is to have someone add them, but we can make the existing data more complete. There are only 321 Postal Codes found in the dataset. There are over 3000 amenities alone. Each element as a latitude and longitude. These value could be used to lookup a postal code and fill in this data. Also, but not relying on users to fill in this data we avoid issues with different formatting of this field.
- Inconsistencies in naming. There is only one node listed as a ‘college’ but two additional ‘schools’ with the word college right in the name. Another example is there being 30 ‘pubs’ and 13 ‘bars’. There really doesn’t need to be a distinction between these.