Courtney Harano

June 6, 2015

Project 2

Map Area: Atlanta, GA, United States

Sources:

http://www.openstreetmap.org/relation/119557

https://s3.amazonaws.com/metro-extracts.mapzen.com/atlanta_georgia.osm.bz2

Problems Encountered in the Map

After doing some provisional analysis on the downloaded data set before loading it into MongoDB (**see included IPython notebook for code, results, and annotations on step-by-step analysis**), I noticed the following problems:

- "TIGER" and "GNIS" project tags
- Different postal code standards
- Incorrect postal codes
- Variety of street abbreviations

"TIGER" project tags

When checking on the types of key values available in the data, I noticed that there were a large number of keys that started with "tiger" or "gnis" (i.e. "tiger:name_base_1", "gnis:state_id"). After some searching on the OpenStreetMap data, it appeared that a lot of this was a relic from the TIGER project that was originally used to populate the OpenStreetMap data (http://wiki.openstreetmap.org/wiki/TIGER_to_OSM_Attribute_Map) and the USGS GNIS bulk import in 2009 (http://wiki.openstreetmap.org/wiki/USGS_GNIS). Digging through the articles, it looked like these were still being used to populate some areas, but when I checked the Battle Grid (http://184.73.220.107/battlegrid/#12/33.7506/-84.3441) to see how much of it still needed to be checked in the Atlanta area, it looked like just referring to the OpenStreetMap user-contributed data would still be pretty accurate, so I decided to ignore any data where the key started with "tiger" or "gnis" in order to standardize the dataset.

Different postal code standards

When I did an audit on the postal code values, it appeared that while most of the OpenStreetMap data used the 5-digit standard, there were still a few that used a 5+4 digit code (i.e. "30058-8351"), while others were prefixed with "GA" (i.e. "GA 30342"). I cleaned up the 5+4 and the GA-prefixed codes so that they would conform to the 5-digit standard.

Incorrect postal codes

I also noticed that there were a number of postal codes that were less than for 4 digits (i.e. “0019”), between 5 and 9 digits (i.e. “300313”), and some were non-numeric characters (i.e. “Atlanta” or “GA”). I ignored these postal codes.

Variety of street abbreviations

There was also a wide variety of street abbreviations, some of which were the standard that we cleaned up in Lesson 6 (i.e. “St” -> “Street”), but Atlanta also has a lot of streets that are distinguished directionally (i.e. “Peachtree Street Northwest”), so I also had to clean up those directional abbreviations as well (i.e. “NW”, “SE”, etc.).

(See data_transformation.py for code or see included IPython notebook for code)

Data Overview

(See data_overview.py for code or see included IPython notebook for code + results)

File Sizes

atlanta_georgia.osm: 2.21 GB

atlanta_georgia.osm.json: 3.24 GB

Basic Statistics

Number of documents: 11,603,114

Number of nodes: 10,881,364

Number of ways: 721,615

Number of unique users: 1,528

Top Contributor

Username: Liber, # of entries: 5,623,416 (48% of the total entries!)

Number of Unique Zip Codes

Even with the cleanup on zipcodes that I did, I still ended up with 215 unique zipcodes in my dataset.

Number of Cities

I did not initially think to do any cleanup on the cities, so I ended up with 120 unique cities (including variants of Atlanta such as “Atlanta, GA”). However, the top 10 cities listed in the data set all made sense, as they included Atlanta and major suburbs of the greater Atlanta metropolitan area (such as Decatur and Union City).

Amenities

These were the top amenities in the greater Atlanta metropolitan area:

1. Place of Worship: 5,518
2. Parking: 2,768
3. School: 2,366
4. Graveyard: 2,206
5. Restaurant: 1,067
6. Parking Space: 837
7. Fuel: 615
8. Fast Food: 605

The results show how old and large of a city it is (with the number of places of worship, graveyards, and schools). It also shows how dependent the residents of Atlanta are on their cars, with parking areas and fueling stations making up a large number of the marked amenities.

Top Religions

Christianity (4,269 occurrences) was by far the highest represented religion in the places of worship in the dataset, although additional exploration may be needed to examine the ones where the religion was not listed (1,228 occurrences).

Top Cuisines

Interestingly, many of the restaurants did not have their cuisine designated (372 occurrences). The rest of the top cuisines was pretty standard for an American city:

1. American: 90
2. Pizza: 80
3. Mexican: 63
4. Burger: 33
5. Sandwich: 31

I was surprised by the overall low number of restaurants with ethnic cuisines (considering Atlanta is a pretty metropolitan city with many diverse restaurants), so I decided to investigate this phenomenon with 2 well-known suburbs – Marietta (I expected standard American fare and some Latin American restaurants) and Decatur (I expected more gastropubs and a more eclectic selection). It looked like Marietta turned out as expected, with 'pizza', 'american', and 'mexican' occurring frequently, but Decatur seemed to be sorely lacking (only had 4 restaurants associated with it). This indicates to me that the population of the OpenStreetMap data may be skewed to certain neighborhoods.

Schools/Colleges/Universities

This dataset also seems to need some work, as most of the nodes tagged as a "school" did not have a name associated with it. Even the colleges and universities datasets seemed to be missing Georgia State University and did not seem to be very standardized (for example, university buildings would show up without any association to the university – i.e. "S1 - Recreation and Wellness Center").

Contributions by Year

I wanted to double-check if OpenStreetMap was actually being maintained (was actually current), so I also just checked the number of entries created every year. While 2009 seems to be the highest volume year (63%), 2014 was the second highest contribution year and 2015 seems to be trending to also add many more entries, so I am confident that at least folks are still contributing to this effort.

Additional Ideas

As I started to aggregate and explore the data, I realized that it was very hard to be confident in the counts I was seeing because there seemed to be such a discrepancy in the amount of information available in each of the different suburbs of Atlanta. As discussed above, when I checked on the number of restaurants in two big suburbs of Atlanta, I saw a big difference in number of results. After living in Atlanta for the last year and a half, I have noticed that Yelp is a pretty good source of data, especially on restaurants in the area. If I was to build a comprehensive map of Atlanta, I would use OpenStreetMap as a base for the basic mapping capabilities and would enhance it with the Yelp API to add more restaurants, bars, small businesses, etc. that do not exist as much in the OpenStreetMap dataset. While this would greatly add to the amount of content I would have on the area, there would be a significant amount of cleanup that would have to be done in order to standardize the data from both sources. For example, Yelp returns their address information in a "location" subtag instead of an "addr" tag. In addition, the standards with which the two sources tag business/point of interest types are very different. For example, Yelp considers everything a "business" and just designates what type of business through its "category" tag, while OpenStreetMap actually uses different "amenity" tags depending on if it is a restaurant, dentist, post office, etc. If I wanted to merge the two data sets and be able to query across all of them, I would need to map Yelp's category and OpenStreetMap's amenity tags to create a common definition across both sources.