Data Wrangling with MongoDB

Carl Centola

Project Summary

Map Area: Boston, MA, United States https://www.openstreetmap.org/relation/2315704

The goal of this project is to pick a place in the world that is important to you and to analyze the data available for that geographic location. To accomplish this task, we will use a series of Python scripts to audit and clean our data, and the open-source NoSQL database solution MongoDB to query and analyze our cleaned data.

I chose Boston, MA for my analysis not only because it is the part of the world in which I currently live, but also because I found amount of relevant data compressed into such a small geographic area to be intriguing in and of itself.

Section 1: Problems Encountered in the Map

The first task at hand is to identify any issues that occur throughout the course of our analysis and to deal with them as such. The main issues I came across in examining my Boston metro data are:

```
1. Over-abbreviated street names
2. Incorrect zip-codes
```

Over-abbreviated street names

Street names in our data set often have inconsistent street abbreviations. For example, "Street" alone had 6 variations: "St.", "St", "ST", "st", "St,", and "Street". Using a provisional audit.py script, we can audit our street names to ensure that we have consistent naming conventions for common street names.

Incorrect Zip-Codes

Zip-code is another element of out data set that seems to contain a few errors. Using a function built in to out audit.py file, we can narrow down our zip-code errors to include only those that do not begin with the Boston standard "02":

In [ ]:

```
01240 1
```

```
Mass Ave 1
MA 4
01250 1
01821 1
01238 1
MA 02118 1
MA 02116 3
01944 1
01125 1
01854 1
MA 02186 1
```

It is clear that we do have some errors in our data, most likely caused by simple user error when entering the data.

Section 2: Data Overview

Using MongoDB queries, we are able to generate some basic statistics about our map data.

File sizes:

boston_massachusetts.osm: 420 MB

boston_massachusetts.json: 485 MB

Number of documents:

In [ ]:

```
> db.boston_data.count()
2228884
```

Number of nodes:

In [ ]:

```
> db.boston_data.find({'type':'node'}).count()
1923138
```

Number of ways:

In [ ]:

```
> db.boston_data.find({'type':'way'}).count()
305746
```

Number of unique users:

In [ ]:

```
> db.boston_data.distinct("created.user").length
1095
```

Top 5 contributing users:

In [ ]:

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

{ "_id" : "crschmidt", "count" : 1206665 }
{ "_id" : "jremillard-massgis", "count" : 433507 }
{ "_id" : "OceanVortex", "count" : 92570 }
{ "_id" : "wambag", "count" : 81564 }
{ "_id" : "morganwahl", "count" : 69655 }
```

Number of users contributing only once:

In [ ]:

```
> db.boston_data.aggregate([
                              {'$group': {'_id': '$created.user', 'count':
 {'$sum': 1}}},
                              {'$group': {'_id': '$count', 'num_users': {'
$sum': 1}}},
                              {'$sort': {'_id': 1}},
                              {'$limit': 1}
                           ])

{ "_id" : 1, "num_users" : 273 }
```

Section 3: Additional Ideas

Boston is an historic, culturally-rich city, making it a popular tourist destination in the northeastern United States. Using MongoDB and our OpenStreetMap data, we can take a look at what visitors can expect when traveling to the area.

Top 5 amenities:

In [ ]:

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

{ "_id" : "parking", "count" : 1277 }
{ "_id" : "bench", "count" : 1021 }
{ "_id" : "school", "count" : 748 }
{ "_id" : "restaurant", "count" : 576 }
{ "_id" : "parking_space", "count" : 446 }
```

Of the top 5 amenities Boston has to offer, parking seems to be high up on the list. While the results of this query may come as a surprise to my fellow Bostonians, the number of parking spaces identified in our data set is relatively small in comparison the city's 2015 population estimate of 667,137 people, let alone the number of people who commute into there city to work each day.

Top 3 tourist attractions:

In [ ]:

```
> db.boston_data.aggregate([
                                {"$match":{"tourism":{"$exists": 1}}},
                                {"$group": {"_id": "$tourism", "count":{"$
sum": 1}}},
                                {"$sort": {"count": -1}}, {"$limit": 3}
                            ])

{ "_id" : "hotel", "count" : 72 }
{ "_id" : "museum", "count" : 53 }
{ "_id" : "artwork", "count" : 48 }
```

For such a small area (Boston has an area of 89.63 mi2), Boston not only has plenty of places to stay, but also plenty of attractions for museum and art-lovers.

Universities in Boston:

In [ ]:

```
> db.boston_data.aggregate([
                               {'$match': {'amenity': {'$exists': 1}, 'ame
nity': 'university', 'name': {'$exists': 1}}},
                               {'$group': {'_id': '$name', 'count': {'$sum
': 1}}},
                               {'$sort': {'count': -1}}
                           ])

{u'_id': u'Boston University', u'count': 41},
{u'_id': u'Massachusetts Institute of Technology', u'count': 10},
{u'_id': u'Suffolk University', u'count': 8},
{u'_id': u'Harvard University', u'count': 5},
{u'_id': u'University of Massachusetts Boston', u'count': 3},
{u'_id': u'Boston University Medical Campus', u'count': 3},
{u'_id': u'Boston College', u'count': 2},
{u'_id': u'Littauer Center', u'count': 2},
{u'_id': u'University Hall', u'count': 2},
{u'_id': u'Harvard Medical School', u'count': 2},
{u'_id': u'Northeastern University', u'count': 2},
...
```

Boston is commonly know for its many world-renowned universities. Based on our query, we can see that we have some familiar names among our list. While " university" is not considered to be among the top amenities in our data set, I would argue that the impact that a large congregation of students and universities is what makes Boston so unique in comparison to other major U.S. cities.

Other Ideas

As most people who work with data know (or quickly realize), While we were readily able to audit and clean our data set, it would have saved a considerable amount of time had we been able to trust that certain data was formatted in a predictable way. For example, OpenStreetMap could implement some data validation rules built into a form in order to verify the consistency of the data being entered. Using some of the concepts that we used in auditing process, identifying regular expressions for street and zip-code entries in particular, could really cut down on the cleaning process and there for allow data analyst more time to play around with the data to find meaningful insight.

One potential issue with this approach is that OpenStreetMap accepts data entry from geographic locations all over the globe. Different users may refer to the same element in different ways (i.e. referring to a street with 6 different abbreviations). While efforts could be made to validate data based on country, it would still prove a difficult task to maintain data validation rules for 196 different countries simply based on the variability of the sources of data entry.

References:

Data: https://www.openstreetmap.org/relation/2315704

Boston Square Milage: https://en.wikipedia.org/wiki/Boston

Boston Population Estimate http://www.census.gov/quickfacts/table/PST045215/2507000,00

MongoDB https://docs.mongodb.com/