Udacity
Data Analyst nanodegree program
Project 2- Wrangling Open Street Map Data

Jamey McCabe – jm6819@att.com

Project completed: July 17, 2015

The project goal was to download a large data set of Geographic data from the opensource OpenStreetMap (OSM) facility and use techniques to understand, find issues and cleanup the data found. The purpose of doing this is;

- to practice new techniques and tools, notably the Xml ElementTree python library concentrating automated discovery, cleanup and mongodb based analysis
- to contribute to a 'large data' project (gain experience as well as example work for a portfolio)
- to demonstrate skills with data analysis in order to give the instructors information about my skill level

The activities conducted during this project were the following:

1. Browsed and learned about the openStreet maps project largely using the OpenstreetMaps Wiki.
2. Researched, downloaded and used 2 Android apps (OSM Tracker and keypadMapper) who's purpose running on your mobile device is to capture Ways and Nodes as you move about .
3. Played with various techniques of downloading the currently available data in OSM for the Santa Fe NM metropolitan area.
4. Coded routines to investigate, clean and transform data to Mongodb and summarize all Ways in the Santa fe metropolitan area, concentrating on handling a convention known as street-pretype
5. Submitted fixed data back into OSM
6. Using Mongodb queries, Analyzed the Santa Fe data in comparison to other cities:
    - Los Alamos, NM a nearby bu relatively recent city
    - Oak Park, IL a midwestern, non Spanish history city but with similar population
    - Los Angeles, CA – A significantly large dataset and also Spanish history city

Table of Contents

Problems encountered in the map....3
1. street-pretypes are found in Santa Fe....3
2. Lack of Node Data with Addresses so switch to Way Names....5
3. Way Data Issues....6
Overview of the Data....8
Other ideas about the datasets....10

Problems encountered in the map

1. street-pretypes are found in Santa Fe

The first issue encountered and which had to be coded for was not actual dirty data but serve to illustate that new data will usually come with significant hurdles even before you can find dirty data and begin to clean it up.

Santa Fe has a significant amount of Spanish language content and Spanish naming convention for streets. From the class provided routine to summarize node address data

by street type, a large portion of street names in Santa Fe did not fall into the normal American Street Type categories. In Exhibit 1 below, see the highlighted data which is not being usefully summarized:

Exhibit 1

```
{'Anasazi': set(['Camino Anasazi']),
 'Ancha': set(['750 Canada Ancha']),
 'Avenue': set(['Don Gaspar Avenue',
                'Lincoln Avenue',
                'Park Avenue',
                'West Palace Avenue']),
 'Chaco': set(['Placita Chaco']),
 'Chelly': set(['Camino de Chelly']),
 'Court': set(['Office Court']),
 'Hualapai': set(['Camino Hualapai']),
 'Lejo': set(['Camino Lejo']),
 'Madre': set(['Acequia Madre']),
 'Marcos': set(['Vuelta San Marcos']),
 'Oraibi': set(['Camino Oraibi']),
 'Palace': set(['Cliff Palace']),
 'Peralta': set(['Paseo de Peralta']),
 'Road': set(['Canyon Road',
              'Cerrillos Road',
              'Country Club Road',
              'East Barcelona Road',
              'West Barcelona Road',
              'West Cordova Road']),
 'Sol': set(['Paseo del Sol']),
 'Street': set(['Agua Fria Street',
                'Alto Street',
                'Baca Street',
                'East Booth Street',
                'East San Francisco Street',
                'East de Vargas Street',
                'Galisteo Street',
                'Hickox Street',
                'Market Street',
                'North Guadalupe Street',
                'Sandoval Street',
                'West Alameda Street',
                'West San Francisco Street',
                'West Water Street']),
 'Trail': set(['Old Santa Fe Trail'])}
```

Santa Fe is an older city and many of its streets were put in place while it was still a Spanish colony (1608-1850). Because of this, most streets have names in Spanish and many have a Spanish style street-pretype. A an exmple of how this appears in Exhibit 2 below, the OSM XML data for a well known building in Santa Fe on ***Paseo De*** Peralta:

```
2121842-       <node id="357614810" lat="35.6916495" lon="-105.9364852" version
="2" timestamp="2015-04-25T20:22:14Z" changeset="30483946" uid="360392" user="ma
xerickson">
2121999-               <tag k="addr:city" v="Santa Fe"/>
2122035-               <tag k="addr:housenumber" v="463"/>
2122073-               <tag k="addr:postcode" v="87501"/>
2122110:               <tag k="addr:street" v="Paseo de Peralta"/>
2122156-               <tag k="amenity" v="social_centre"/>
2122195-               <tag k="gnis:feature_id" v="935960"/>
2122235-               <tag k="name" v="Scottish Rite Temple"/>
2122278-               <tag k="phone" v="505-982 4414"/>
2122314-               <tag k="website" v="http://www.nmscottishrite.org/"/>
2122370-               <tag k="wikipedia" v="en:Scottish Rite Temple (Santa Fe,
 New Mexico)"/>
2122444-       </node>
```

Exhibit 2

To address this in the code the custom alorithm in Exhibit 3 was developed to find and handle street types wether they were pre or post type.

```
def determineStreetType(street_name):
    # function takes a street name and first splits it into words so it can test
    # the first word and then the first and second words befre finally finding
    # the last word using a regular expresion all done to  determines it's
    # street type and then populates a group variable "street types"
    # which groups all the street names by street type
    SpanishInd = False
    nameWords = street_name.split()
    #print allNameWords
    if nameWords[0] in ["E","East","N","North","S","South","W","West"]:
    # strip direction
       del nameWords[0]
    if len(nameWords)> 2 and nameWords[1] in ["de","del","a"]:
    # if 2nd word is equivalent of “of” first word is pretype
        street_type = nameWords[0]
        SpanishInd = True
    elif len(nameWords)> 0 and nameWords[0] in reverseTypes:
    # lookup first work in reversetypes table which was built from discovery of SantaFe.osm
        street_type = nameWords[0]
        SpanishInd = True
    elif len(nameWords)> 2 and nameWords[0] + ' ' +nameWords[1] in twoWordTypes:
    # lookup other discovered 2 word pre-types
        street_type = nameWords[0] + ' ' + nameWords[1]
    else:
    # pull out the post street type
        m = street_type_re.search(street_name)
        if m:
            street_type = m.group()
        else:
            street_type = ""
        #if you just need a list of unexpected streets uncomment the following
        #if street_type not in expected:
    return(better_name, SpanishInd)
```

Exhibit 3

2. Lack of Node Data with Addresses so switch to Way Names

A second issue is also apparent from the above. There is not much node data with addresses in the Santa Fe area, only 38 streets.

Running the class provided routine to count the number of tags in this data set gives the following:

```
{'bounds': 1,
 'member': 428,
 'nd': 56348,
 'node': 47892,
 'osm': 1,
 'relation': 30,
 'tag': 42284,
 'way': 6029}
```

Exhibit 4

The dataset is 11.8 Mb and has 47,892 node tags which would seem to indicate there would be significantly more addresses. In Santa Fe however, all but 38 of the Node tags are documenting the shape of the ways not the actual addresses on the street. A somewhat non-technology oriented city like Santa Fe has not yet had significant Address-Nodes captured in the OpenStreetMap community.

3. Way Data Issues

In the ~ 6000 Ways in the Santa fe dataset there are approximately ~2400 unique way names. In that data via the technique of summarizing streets by their street type the following issues were discovered. Beyond implenting the code to clean them for the project, all 57 were submitted as changes to OpenStreetMaps data, giving the author (Jamey McCabe) their first contribution to the OSM community:

Exhibit 5

General Type of issue	Fix	Count
Acronyms need to be spelled out		
	Change Ave to Avenue	1
	Change Blvd to Boulevard	1
	Change Cam to Camino	2
	Change Cii to Calle	1
	Change Cli to Calle	1
	Change Cll to Calle	1
	Change Ct to Court	9
	Change DR to Drive	1
	Change Dr to Drive	7
	Change Ln to Lane	1
	Change Rd to Road	18
	change St to Street	1
	Change vis to Vista	2
Mispelled		
	CAmino to Camino	1
	Caltamira to Calle Altimira Court	1
	Paso to Paseo	1
	Change Vis to Via	1
	Entrade to Entrada	1
Street is not accurrately named		
	change Cereza to Plaza Rojo	1
	change Ristra to Ristra Plaza	1
	Drop Circle from Via Janna Circle	1
	remove Curcle from Cinco Pintores Curcle	1
Other issues		
	Gwendloyn does not exist – not a	1
	Fix nodes that make up this way	1
	Total Result	**57**

To address this in the code, the following dictionary and function was added mostly using class provide algorithm:

Exhibit 6

```
 # list of street names to fix
mapping = { "Ave":"Avenue",
            "Blvd":"Boulevard",
            "Caltamira":"Calle Altimira Court",
            "Cam":"Camino",
            "CAmino ":"Camino",
            "Cereza":"Plaza Rojo",
            "Cii":"Calle",
            "Cli":"Calle",
            "Cll":"Calle",
            "DR ":"Drive",
            "Dr":"Drive",
            "Entrade":"Entrada",
            "Ln":"Lane",
            "Paso":"Paseo",
            "Rd":"Road",
            "Ristra":"Ristra Plaza",
            "St": "Street",
            "Vis":"Via",
            "vis":"Vista",
            }
fixed = 0

def update_name(name, mapping):
    # This function looks up a passed in street type and replaces it with one from the array
called mapping
    global fixed
    if name in mapping:
        fixed = fixed + 1
        print ("{} - found : {}, replaced with: {}").format(fixed,name,mapping[name])
        name = name.replace(name, mapping[name])
    return name
```

Overview of the Data

Size of the data for the Open Street Maps was over 1Gb for the Los Angeles Area. Note that all 4 were converted to JSON and loaded to mongodb for the following queries;

```
jamey@jamey-CM6850:~/uIDS/P2/ud032/OSM$ ls -s  -S *.osm|head
1,100,048 los-angeles_california.osm
  11,344 santafe.osm
   7,884 LosAlamos.osm
   4,184 OakPark.osm
```

1972 Unique Users

```
> db.ways.aggregate(
...     { $group: { _id: "$created.user"}  },
...     { $group: { _id: 1, count: { $sum: 1 } } }
... );
{ "_id" : 1, "count" : 1972 }
```

4,964041 Nodes and 524,894 Ways

```
> db.ways.aggregate({"$group":{_id:"$type","jcount":{"$sum":1}}}   )
{ "_id" : "node", "jcount" : 4964041 }
{ "_id" : "way", "jcount" : 524894 }
```

Top 15 Street Types – note that 13 and 15 are Spanish Style Street Pretypes.

```
> db.ways.aggregate({"$group":{_id:"$address.streetType","jcount":
{"$sum":1}}},    {"$sort":{"jcount":-1}                    }
... )
{ "_id" : null, "jcount" : 283497 }
{ "_id" : "Street", "jcount" : 41419 }
{ "_id" : "Avenue", "jcount" : 38308 }
{ "_id" : "Drive", "jcount" : 28715 }
{ "_id" : "Road", "jcount" : 20167 }
{ "_id" : "Lane", "jcount" : 14589 }
{ "_id" : "Court", "jcount" : 14540 }
{ "_id" : "Place", "jcount" : 8961 }
{ "_id" : "Way", "jcount" : 8709 }
{ "_id" : "Circle", "jcount" : 6839 }
{ "_id" : "Boulevard", "jcount" : 3935 }
{ "_id" : "Freeway", "jcount" : 3852 }
{ "_id" : "Via", "jcount" : 3490 }
```

```
{ "_id" : "Trail", "jcount" : 3021 }
{ "_id" : "Calle", "jcount" : 1904 }
```

Santa Fe has a clearly significant amount of spanish style Pre-typed Way names over those found in the other 3 cities:

	SpanishInd		Spanish Cultur
County	FALSE	TRUE	
OakPark.osm	641		**0%**
LosAlamos.os	620	50	**7%**
Los-angeles...	224,529	12,232	**5%**
Santa Fe, NM	2,166	1,159	**35%**
Total Result	**227,956**	**11,118**	**5%**

```
> db.ways.aggregate(        {"$group":
...             {
...                 "_id":
...
{"files":"$file","spanishInd":"$address.spanishInd"},
...                 "hwCount":{"$sum":1}
...             }
...         },
...         { "$sort":
...             { "hwCount": -1 }
...         },
...         { "$group":
...             {
...                 "_id": "$_id.files", "spanishInds":
...                     { "$push":
...                         { "aSpanishInd": "$_id.spanishInd",
"count": "$hwCount" },
...                     }
...             }
...         },
...         { "$sort":
...             { "count": 1 }
...         })
{ "_id" : "OakPark.osm", "spanishInds" : [ { "count" : 1646 },
{ "aSpanishInd" : false, "count" : 641 } ] }
{ "_id" : "LosAlamos.osm", "spanishInds" : [ { "count" : 2091 },
{ "aSpanishInd" : false, "count" : 620 }, { "aSpanishInd" : true,
"count" : 50 } ] }
{ "_id" : "santafe.osm", "spanishInds" : [ { "count" : 2704 },
{ "aSpanishInd" : false, "count" : 2166 }, { "aSpanishInd" : true,
"count" : 1159 } ] }
```

```
{ "_id" : "los-angeles_california.osm", "spanishInds" : [ { "count" :
277056 }, { "aSpanishInd" : false, "count" : 224529 }, { "aSpanishInd"
: true, "count" : 12232 } ] }
```


Other ideas about the datasets

In the exploration of the Santa Fe (and Los Alamos and Oak Park) data the following were observed but not futher studied:

1. High Use of the key Bicycle in Santa Fe when this is not a common tag. In researching where the key Bicycle would be used in OSM it was not documented other than as a sub key value for a limited scenario “Road (UK) or path (USA, Canada) dedicated to cyclists on separate right of way.” See: http://wiki.openstreetmap.org/wiki/Bicycle . Instead of Bicycle= the tag used ot indicate Bicycle paths is Highway=Cycleway. In my knowledge of Santa Fe there are no or at least very few separated Bicycle paths. Further investigation of these and perhaps recoding if they are coded incorrectly is warranted.
2. Why would Los Alamos with the smallest population have lots of buildings in the OSM data and Oak Park has none? Los Alamos had 812 keys for Building and Oak Park had none. Investigation as to Los Alamos and what buildings are captured is warranted. It also seems to indicate room for a valuable contribution to Oak park for their numerous and famous Frank Lloyd Wright buildings.
3. Why does Los Alamos have so many Node's compared to ways? Los Alamos has 13 nodes per way as compared to Santa fe which has 8 nodes per way and Oak Park which has 7 nodes/way. This seems to call for some further investigation to see if this is an issue for Los Alamos or some technique which could be used in Santa Fe and Oak Park.
4. Explore cultural diversity of an area based on Street Types or other attributes of street names beyond street pretype as an indicator of Spanish culture.. From the street-pretype analysis it is clear that street names can be a metric of sorts to reflect age and culture of a city. Applying this across cities and some sort of US heat map would be interesting. Developing the criteria for cultural indicators could be challenging as street pretype was fairly simple but other cultural indicators (e.g. Chinese) would likely be less clear (e.g. perhaps use Chinese names such as Wei or Han) and more difficult to computationally derive (e.g. being sure that Wei was a the chinese name and not a mispelling).
5. Low “other” character count in Los Alamos values. From the comparison of Character quality between Santa Fe NM, Los Alamos NM and Oak Park, IL it was discovered that Los Alamos had a very low count of “other” characters in it's population of key-value pair – values. “other” is when there is a character found which is not “a-z” or in the problem character set “[=\ +/&<>;\'"\?%#$@”. It is usually dificult to answer such a question “why aren't there any?'. The hope though is by categorizing what the “other' characters are in Santa Fe and Oak Park data sets we might discover a practice that had been employed in Los Alamos data that makes it cleaner.