☐ README.md

OSM Data Cleanup Project Notes

Submission Details

1. This README is my project writup.
2. Project 6 solutions can be found in Udacity Data Wrangling Lesson 6.ipynb
3. See The Map Data writeup for more info on what OSM data I used and why I chose it.
4. A thinned version of the data I used is located in this repository, here.
5. My references can also be found in this repository, here.

Data Model

After a quick audit, I came up with an initial data model for the OSM data. The following ways are how I planned to clean the OSM data.

Problem Characters

Any key with any “problem characters”, defined by this regular expression `[=\+/&<>;\'"\?%#$@\,\. \t\r\n]` , will be removed.

address

Anything starting with "addr:” will be converted to an `address` object. Also, convert street abbreviations to full type (e.g., “Blvd” to “Boulevard”).

```
-Ex:
    "address": {
        "housenumber": 5158,
        "street": "North Lincoln Avenue”,
        ...
    }
```

is_in

The `is_in` property is generated in the same what the `address` object is.

phone

The phone number data comes in a variety of formats. To clean them up, phone numbers will be stripped of non-numeric characters and broken up into their three sections, the area code, the three digit part, and the four digit part, or, by their more appropriately technical names, npa, nxx, and xxx respectively (https://en.wikipedia.org/wiki/North_American_Numbering_Plan#Numbering_system).

```
-Ex.
    “phone”: {
        “npa”: “213”,
        “nxx”: “555”,
        “xxxx”: “5555”
    }
```

open_hours

The open hours data will be converted into a dict with keys being days of week and values being arrays of 2-element tuples, where the first element is an opening time and the second element is a closing time.

```
-Ex.
    “open_hours”: {
        “Monday”: [(“9:00”, “14:00”), (“17:00”, “22:00”)]
        “Tuesday”: [(“9:00”, “14:00”), (“17:00”, “22:00”)]
        “Wednesday”: [(“9:00”, “14:00”), (“17:00”, “22:00”)]
        “Thursday”: [(“9:00”, “14:00”), (“17:00”, “22:00”)]
        “Friday”: [(“9:00”, “14:00”), (“17:00”, “22:00”)]
        “Saturday”: [(“9:00”, “22:00”)]
        “Sunday”: [(“9:00”, “22:00”)]
    }
```

Difficulties with the Data

Street Name Type

In the data model above I decided to replace all street name abbreviations with their full names. This proved a little harder that I originally thought. My first (naïve) approach just took the last word of the street field. This resulted in many incorrect selections, since the street type didn't always come last. So I went for a more sophisticated approach, using the natural language address parsing module `usaddress` to suss out the “StreetNamePostType” from each street field, which worked much better.

Open Hours

Unfortunately, I ended up abandoning the above data model for `open_hours` and leaving the data unchanged. This model proved to be too difficult to encode due to the wide range of formats for this field (a real testament to the difficulty of using human entered data). I even tried using a natural language time parser (parsedatetime), but it wasn't able to consistently get the correct times. Here are some examples of the data:

```
-24/7
-07:00-22:00
-May 15-Nov 16
-Mo-Fr 7:00-21:00; Sa-Su 7:00-21:00
-Monday:  Closed Tues-Sat: 11:00AM-2:00PM;5:00-8:00PM Sunday :   11:00AM-2:00PM
```

Exploring the Data

Total Elements

```
Total Number of Nodes:      5247813
Total Number of Ways:       562383
Total Number of All Documents:  5810196
```

Unique Users

```
Unique Users in Nodes:      2628
Unique Users in Ways:       1975
Unique Users in All Documents:  2871
```

Top Controbutors

```
Top 10 Contributors in Nodes:
woodpeck_fixbot               546730    (10.42%)
The Temecula Mapper           452435    (8.62%)
AM909                         429283    (8.18%)
nmixter                       329645    (6.28%)
Brian@Brea                    206737    (3.94%)
Aaron Lidman                  154956    (2.95%)
SJFriedl                      137614    (2.62%)
Jon Schleuss                  132206    (2.52%)
jerjozwik                     130263    (2.48%)
mattmaxon                     116291    (2.22%)
```

```
Top 10 Contributors in Ways:
balrog-kun                  62022     (11.03%)
The Temecula Mapper         39083     (6.95%)
AM909                       33124     (5.89%)
Aaron Lidman                26817     (4.77%)
Brian@Brea                  26430     (4.70%)
SJFriedl                    17471     (3.11%)
NE2                         11960     (2.13%)
Jon Schleuss                10677     (1.90%)
DaveHansenTiger             10607     (1.89%)
jerjozwik                   10292     (1.83%)

Top 10 Contributors in All Documents:
woodpeck_fixbot             546732    (9.41%)
The Temecula Mapper         491518    (8.46%)
AM909                       462407    (7.96%)
nmixter                     330838    (5.69%)
Brian@Brea                  233167    (4.01%)
Aaron Lidman                181773    (3.13%)
SJFriedl                    155085    (2.67%)
Jon Schleuss                142883    (2.46%)
jerjozwik                   140555    (2.42%)
mattmaxon                   121211    (2.09%)
```

Top Amenities

```
Top 10 Amenity Types:
Place Of Worship            3788
School                      3123
Restaurant                  1725
Fast Food                   1254
Fuel                        745
Cafe                        576
Fountain                    493
Parking                     490
Toilets                     483
Drinking Water              458
```

Top Cuisines

```
Top 15 Cuisines:
American                    155
Mexican                     145
Pizza                       84
Italian                     67
Chinese                     57
Japanese                    47
Thai                        41
Burger                      40
Sushi                       39
Sandwich                    33
Seafood                     24
Steak House                 17
Indian                      16
Asian                       15
Regional                    15
```

Top Restaurants

```
Top 20 Restaurants:
Denny's                     23        (American, Diner)
Subway                      17        (Sandwich)
IHOP                        15        (Breakfast, American, Pancake)
Panda Express               10        (American, Chinese)
Chipotle                    9         (Mexican)
Pizza Hut                   9         (Pizza)
```

```
California Pizza Kitchen      9        (Pizza, Italian)
Sizzler                       7        (American, Steak)
Red Robin                     7        (Burger, American)
Ruby's Diner                  7        (American)
Cheesecake Factory            7        (American)
Rubio's                       6        (Mexican)
Chipotle Mexican Grill        5        (Mexican)
Islands                       5        (Burger, American)
Round Table Pizza             5        (Pizza)
Corner Bakery                 5        (Chinese)
Carrows                       5        (American)
Olive Garden                  5        (Italian)
Chili's                       4        ()
BJ's                          4        (American)
```

Concusion and Other Ideas

I would suggest that the OSM database start combining user generated data with automated data retrieval methods. There is a upcoming federal government effort to install sensors to make our cities smarter (https://www.whitehouse.gov/the-press-office/2015/09/14/fact-sheet-administration-announces-new-smart-cities-initiative-help). The data from these sensors could be augment the OSM data which is more structural to include more information about how people move through and use the cities' infrastructure and amenities.

Also, the user generated effort should be continued as well to fill out the dataset even more. This data set, which already huge, can't be complete. (I mean, there are only 2 "In N Out"s in the dataset, for crying out lout!) This point is clearer when you realize that Los Angeles County has more people in it than the 11 smallest U.S. states combined (10,116,705 versus 9.870,265; http://dadaviz.com/s/population-extremes). However, with that many people, if the word about the OSM database were spread to even a small portion of them who would participate, it wouldn't take long to get a much more complete picture.