Data Analyst Nanodegree: P2_Data Wrangle OpenStreetMaps Data

Zhihui Xie
xiewisdom@gmail.com

Project Summary

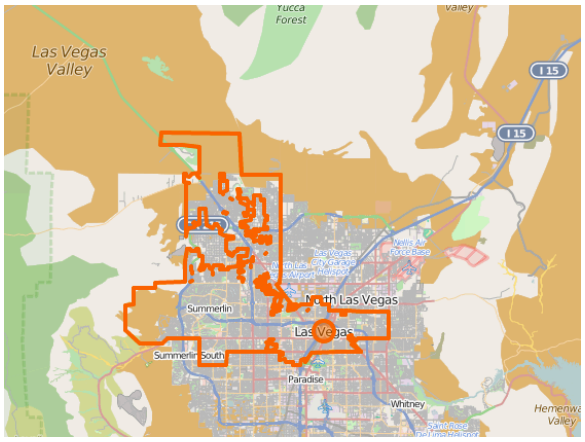
OpenStreetMap, an open project to create a free map around the world, is a powerful tool for viewing maps and humanitarian aid. The initial data of the map were collected using a handheld GPS and a notebook, digital camera, or a voice recorder. These data heavily rely on the human input, which may cause inconsistent input, misspellings or error (e.g. inconsistent input of Street, street, st., St.). This project is going to focus on map of Las Vegas, Nevada, USA and wrangling the data: 1) overview data (code: mapparser.py); 2) check the “k” value for each “<tag>” (code: tags.py); 3) find out number of unique users contributed to the map (code: users.py); 4) fix unexpected street types (e.g. street, st., St. to be Street) (code: audit.py); 5) transform the shape of data and insert data into mongodb (code: data.py and mongodb.py); 6) use MongoDB queries to find number and names of hotels, and shopping malls (code: query.by). The codes mentioned above are included in a separated file and tested by sample dataset (a part of the whole dataset) before run for the whole dataset.

Data source

I chose Las Vegas for data wrangling because it is one of the most popular tourist destination and my dream place for vacation. I hope I can have a vacation there this holiday and would like to know it well. I hope this map to be improved in quality!

I explored the data from the following link:

URL: https://www.openstreetmap.org/relation/170117#map=11/36.2501/-115.2531



Is there a list of Web sites, books, forums, blog posts, github repositories etc that you referred to or used in this submission (Add N/A if you did not use such resources)? The above images were cited from OpenStreetMap and wiki:
https://www.openstreetmap.org/relation/170117#map=11/36.2551/-115.2387
http://en.wikipedia.org/wiki/Las_Vegas#mediaviewer/File:LasVegasMontage6.jpg

Problems encountered in the map

In this dataset, there are 7 types of problem for street types (table 1). Some of them are inconsistent input, for example, “AVE”, “Ave” and “Ave.”. Some are abbreviation: “Ln.”, “Pkwy”, “St” and “Rd”. Using attached auditing code (audit.py), these inconsistent input or abbreviation were fixed (table 1).

In addition to street type, the keys were also validated (using code: tags.py). In this dataset, there are 240883 keys including 'lower' regular expression ('^([a-z]|_)*$'): e.g. “landuse” and “name”, 263369 keys including 'lower_colon' regular expression ('^([a-z]|_)*:([a-z]|_)*$'): e.g. “addr:state” and “gnis:feature_id”, 2 keys including 'problemchars' expression ('[=\+/&<>;\'"\?%#$@\,\. \t\r\n]'): “(new tag)” and “Loading Docks 2F-2H”. The keys with 'problemchars' expression will be ignored during data shape transformation (code: data.py).

Table 1 Fixing problem of street types

Types of problem	I			II		III		IV	V	VI	VII
Before auditing	AVE	Ave	Ave.	Blvd	Blvd.	Dr	Dr.	Ln.	Pkwy	St	Rd
After auditing	Avenue			Boulevard		Drive		Lane	Parkway	Street	Road

Overview of the data

To overview the data, the following items were recorded:

- size of the file (las-vegas_nevada.osm): 166.9 MB;
- number of unique users (code: users.py): contributed by 618 unique users;
- number of nodes and ways (code: mappraser.py): nodes: 723370, ways: 78394;
- number of top 10 amenities with most amount of units (attached code: query.py, function: get_query (data, db)):
  [('parking', 746), ('school', 535), ('place_of_worship', 367), ('fountain', 269), ('restaurant', 200), ('fast_food', 158), ('fuel', 139), ('hotel', 104), ('fire_station', 70), ('hospital', 66)]

Other ideas about the dataset

1. Hotels and malls in the datasets

I am interested in finding all the hotels and shopping malls in Las Vegas. To do that, I try to find hotels and malls listed as amenity using two pipelines (pipeline1 and pipeline3 in attached code: query.py, function: get_pipeline()) to query database inserted into MongoDB. However, only one hotel without name and two

malls were found, which does not make sense. I went back to check the original dataset and noticed that “hotel” and “mall” were listed as “tourism” and “shop”, respectively. Thus, I edited the query pipelines (pipeline2 and pipeline4) and got the following results: Hotels as amenity: 1, Hotels as tourism: 103; Malls as amenity: 2, Malls as shop: 8. Finally, in order to improve the map, I modified the data.py code and added hotels and malls to amenity. Execute the query pipeline1 and pipeline3 again, then I got 10 malls and 104 hotels as amenity. These results suggest the hotels and malls were successfully and correctly added into new database.

Pipelines:

```
#hotels as amenity
pipeline1 = [{"$match": {"amenity": "hotel"}},
{"$project":{"_id": "$name", "cuisine": "$cuisine", "phone": "$phone"}}]
#hotels as tourism
pipeline2 = [{"$match": {"tourism": "hotel"}},
{"$project":{"_id": "$name", "address": "$address", "phone": "$phone"}}]
#mall as amenity
pipeline3 = [{"$match": {"amenity": "mall"}},
{"$project":{"_id": "$name", "cuisine": "$cuisine", "phone": "$phone"}}]
#mall as shop
pipeline4 = [{"$match": {"shop": "mall"}},
{"$project":{"_id": "$name", "address": "$address", "phone": "$phone"}}]
```

Outputs:

Hotels as amenity: 104

```
{u'ok': 1.0,
 u'result': [{u'_id': u'SLS Hotel and Casino', u'phone': u'(702) 737-2111'},
            {u'_id': u'Four Seasons'},
            {u'_id': u'Tuscany'},
            {u'_id': u'Embassy Suites'},
            {u'_id': u'Silver Sevens Hotel & Casino',
             u'phone': u'(702) 733-7000'},
            {u'_id': u'Suncoast Hotel and Casino'},
            {u'_id': u'Lake Mead Lodge'},
            {u'_id': u'Encore Casino at Wynn'},
            {},
            {u'_id': u'Hacienda Hotel and Casino',
             u'phone': u'(702) 293-5000'},
            {u'_id': u'Hilton Grand Vacations Suites on the Las Vegas Strip',
             u'phone': u'702-765-8300'},
            {u'_id': u'Hilton Grand Vacations Suites - Las Vegas (Convention Center)',
             u'phone': u'702-946-9210'},
            {u'_id': u'Embassy Suites Convention Center Las Vegas',
             u'phone': u'702-893-8000'},
            {u'_id': u'Renaissance Las Vegas Hotel',
             u'phone': u'702-784-5700'},
            {u'_id': u'Extended Stay America'},
            {u'_id': u'Boulder Station Hotel and Casino'},
            {u'_id': u'Element Las Vegas Summerline, 10555 Discovery Drive'},
            {u'_id': u'Main Street Station Casino and Hotel'},
            {},
            {u'_id': u'Americas Best Value Inn'},
            {u'_id': u'Summer Bay Resort'},
            {u'_id': u'Candlewood Suites Las Vegas'},
            {u'_id': u'Quality Inn'},
            {u'_id': u'Boulder Dam Hotel'},
```

{u'_id': u'Boulder Inn and Suites'},
{u'_id': u'Circus Circus, 2880 Las Vegas Boulevard, NV 89109 Las Vegas, Vereinigte Staaten von Amerika'},
{u'_id': u'Fortune Hotel & Suites', u'phone': u'702-830-4010'},
{u'_id': u'Super 8'},
{u'_id': u'Stratosphere Hotel and Casino'},
{u'_id': u'Residence Inn Las Vegas Convention Center'},
{u'_id': u'Courtyard Las Vegas Convention Center'},
{u'_id': u'Hilton Garden Inn Henderson'},
{u'_id': u'Palms Casino Resort'},
{u'_id': u'Gold Coast Hotel & Casino'},
{},
{u'_id': u'Tropicana Hotel and Casino'},
{u'_id': u'Luxor Hotel and Casino'},
{u'_id': u'Mandalay Bay Casino & Resort'},
{u'_id': u'Polo Towers'},
{u'_id': u'Encore'},
{u'_id': u'Mirage Hotel and Casino'},
{u'_id': u'Treasure Island Hotel and Casino'},
{u'_id': u'Best Western Mardi Gras Inn'},
{u'_id': u'Plaza Hotel', u'phone': u'(702) 386-2110'},
{u'_id': u'Paris Hotel and Casino'},
{u'_id': u'Fiesta Henderson'},
{u'_id': u'Hilton Garden Inn'},
{u'_id': u'Crestwood Suites'},
{u'_id': u'Manor Suites'},
{u'_id': u'Tahiti Village'},
{u'_id': u'Micdrotel Inn'},
{u'_id': u'Mandarin Oriental'},
{u'_id': u'Aria Resort & Casino'},
{u'_id': u'Vdara'},
{u'_id': u'The Cromwell'},
{u'_id': u"Marriott's Grand Chateau"},
{u'_id': u'The Carriage House', u'phone': u'702.798.1020'},
{u'_id': u'The California Hotel and Casino'},
{},
{u'_id': u'Trump Hotel'},
{u'_id': u'Bellagio Hotel and Casino'},
{u'_id': u'Excalibur Hotel and Casino'},
{u'_id': u'Monte Carlo Hotel and Casino'},
{u'_id': u'New York New York Hotel and Casino'},
{u'_id': u'Caesars Hotel and Casino'},
{u'_id': u'The Quad Resort & Casino',
u'phone': u'(800) 634-6441'},
{u'_id': u'Venetian Hotel and Casino'},
{u'_id': u'Mandalay Bay Hotel and Casino'},
{u'_id': u'Flamingo Hotel and Casino'},
{u'_id': u'MGM Grand Hotel and Casino'},
{u'_id': u"Bally's Hotel and Casino"},
{u'_id': u'Rio All Suite Hotel and Casino'},
{u'_id': u'The Hotel at Mandalay Bay'},
{u'_id': u'Hooters'},
{u'_id': u"Harrah's Hotel and Casino"},
{u'_id': u'the D Las Vegas Hotel & Casino',
u'phone': u'(702) 388-2400'},
{u'_id': u'La Quinta Inn'},
{u'_id': u'Hard Rock Hotel and Casino'},
{u'_id': u'Wyndham Grand Desert'},
{u'_id': u'LVH Las Vegas Hotel & Casino',
u'phone': u'702-732-5111'},
{u'_id': u'Planet Hollywood Resort and Casino',
u'phone': u'+1 866 919 7472'},
{u'_id': u'The Westin Casuarina'},
{u'_id': u'Circus Circus Las Vegas Hotel and Casino'},

{u'_id': u'Stratosphere Hotel and Casino'},
{u'_id': u'J. W. Marriot Resort'},
{u'_id': u'Hampton Inn and Suites'},
{u'_id': u'The Palazzo'},
{},
{u'_id': u'Gold Strike'},
{u'_id': u'Spring Hill Suites'},
{u'_id': u'Town Place Suites'},
{u'_id': u'Wynn Hotel & Casino'},
{u'_id': u'The Grand'},
{u'_id': u'Hyatt Place'},
{u'_id': u'Super 8'},
{u'_id': u'Golden Gate Hotel and Casino'},
{u'_id': u'Palace Station'},
{},
{u'_id': u'Blue Moon'},
{u'_id': u'Eastside Cannery Casino Hotel'},
{u'_id': u'Hampton Inn'},
{u'_id': u'La Quinta Inn & Suites'},
{u'_id': u'Best Western McCarran Inn'},
{u'_id': u'SpringHill Suites by Marriott'}]}

Malls as amenity: 10

{u'ok': 1.0,
u'result': [{u'_id': u'Fashion Show Mall'},
{},
{u'_id': u'Las Vegas Outlet Center Annex'},
{u'_id': u'Crystals'},
{u'_id': u'Miracle Mile Shops'},
{},
{},
{},
{u'_id': u'Fashion Show Mall'},
{u'_id': u'Downtown Container Park'}]}

2. Find top 5 contributors to the map

In addition, to find users who contribute to this map, I setup pipeline5 to query MongoDB and showed top 5 contributors as follow:

Pipelines:

```
#top 5 contributors to the map
pipeline5 = [{"$match": {"created.user": {"$exists": 1}}},
{"$group": {"_id": "$created.user", "count": {"$sum": 1}}},
{"$sort": {"count": -1}},
{"$limit": 5}]
```

Outputs:

Top 5 contributors to the map: 5

{u'ok': 1.0,
u'result': [{u'_id': u'alimamo', u'count': 254892},
{u'_id': u'woodpeck_fixbot', u'count': 79690},
{u'_id': u'nmixter', u'count': 70150},
{u'_id': u'gMitchellD', u'count': 48266},
{u'_id': u'robgeb', u'count': 43484}]}

Thank you for all contributors to this map instructors for this course and Udacity staffs whom help building and evaluating this project.

By Zhihui Xie
12/15/2014