OpenStreetMap Project - Data Wrangling with MongoDB

Xiaomin Xu

Sep 20th, 2015
Data used : MapZen Weekly OpenStreetMaps Metro Extracts

Map Areas: Santa Monica, CA, U.S.A.

1. Project Steps

The city I choose to do this data wrangling project with OpenStreetMaps data is Santa Monica due to the respectively small file size.

Take the course materials as reference, I do this project by the following steps:

- read the data
- find out mistakes in state, city names, street names and postcode
- fix the problem and update the data with the code written in the practice
- upload data to MongoDB database
- explore the data and do basic analysis

2. Problems Encountered

Theere are four main problems with the row data extracted:

- Unconsistent State name (i.e. CA or California)
- Multiple cities(i.e. Santa Monica or Los Angels) and expired city names like 'Venice'.
- Street names with different types of abbreviations. (i.e. 'Ave' or 'ave' or 'Pico')
- Inconsistent postal codes (i.e. 'CA 90291' or '90401-2405')

Unconsistent State name

There are 3 types of state names showing up in the dataset: "CA", "CA," and "California". To keep consistent, "CA" is selected as the standard version as it has the highest count and the other twos were updated to "CA".

Multiple Cities and Expired City Names

This city we aimed to investigate is Santa Monica. In this dataset, despite of Santa Monica, we also found 4 other city names:

- Los Angeles

- Pacific Palisades
- Marina Del Rey
- Venice

These 4 cities are neighbors of Santa Monica. In addition, Venice is no longer an independent city since 1926 when it became part of Los Angeles.
City transformation results are shown as below:

- santa Monica => Santa Monica
- Pacific Palisades => Pacific Palisades
- Marina Del Rey CA => Marina Del Rey
- Los Angeles-Venice => Los Angeles
- Marina Del Rey => Marina Del Rey
- Santa Monica => Santa Monica
- Venice => Los Angeles
- West Los Angeles => Los Angeles
- Los Angeles => Los Angeles
- Marina del Rey => Marina Del Rey
- Marina del Ray => Marina Del Rey
- Venice CA => Los Angeles

Problem with Abbriviations to Street Names

This showed some new abbreviations which needed to be accounted for, such as directions (Ave,Dr,ave,etc) and location specific (Ste,Lp,etc.). Below is only part of the update to the street names.

- Walnut Ln => Walnut Lane
- Main St. => Main Street
- Olive ave => Olive Avenue
- Santa Monica Bvd => Santa Monica Boulevard
- Entrada Dr => Entrada Drive
- Ocean Bd. => Ocean Boulevard.
- Ohio Ave => Ohio Avenue
- 15th Street Ste. 1101 => 15th Street
- Wilshire Blvd => Wilshire Boulevard

Inconsistent Postcode

Initially, some cleanup was needed for the zip codes. This included

- Removing the 4 digit postcode suffix.
- Removing state letters from postcode
- Converting to int (not strictly required)

After adding a zip code cleaning function, all zip codes adhered to a 5 digit code.

Below shows how the zip codes were updated.

- 90025-9998 => 90025
- CA 90291 => 90291
- 90401-2405 => 90401 ...

3. Data Overview

(1) Number of documents

">db.cities.find().count()"

72368

(2) Number of Nodes

">db.cities.find({"type":"node"}).count()"

62954

(3) Number of ways

">db.cities.find({"type":"way"}).count()"

9406

(4)Number of Contributors

">len(db.cities.distinct("created.user"))"

278

(5) Top 5 Contributors

contributor =

db.cities.aggregate([{"$match":{"created.user":{"$exists":1}}},{"$group": {"_id":
{"City":"$city_name","User":"$created.user"}, "count": {"$sum": 1}}},

{"$project": {'_id':0, "City":"$_id.City","User":"$_id.User","Count":"$count"}},

{"$sort": {"Count": -1}},

{"$limit" : 5 }])

pprint.pprint(list(contributor))

[{u'Count': 13242, u'User': u'techlady'},

{u'Count': 8784, u'User': u'Rovastar'},

{u'Count': 6808, u'User': u'StellanL'},

{u'Count': 5206, u'User': u'bdiscoe'},

{u'Count': 2876, u'User': u'mdapol'}]

4.Additional data exploration using MongoDB queries

(1) Top 5 appearing amenities

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

pprint.pprint(list(amenity))
```

[{u'_id': u'fountain', u'count': 2},

{u'_id': u'bus_station', u'count': 2},

{u'_id': u'community_centre', u'count': 2},

{u'_id': u'child care', u'count': 2},

{u'_id': u'cemetery', u'count': 2}]

Since Santa Monica is a beach city, it makes sense that "fountain" is on top of the amenities.

(2) Top Religion in this area

```
religion =

db.cities.aggregate([{"$match":{"amenity":{"$exists":1},"amenity":"place_of_worship"}},

{"$group":{"_id": {"City":"$city_name","Religion":"$religion"},"count":{"$sum":1}}},

{"$project":{"_id":0,"Religion":"$_id.Religion","Count":"$count"}},

{"$sort":{"Count":-1}}])

pprint.pprint(list(religion))
```

[{u'Count': 92, u'Religion': u'christian'},

{u'Count': 4, u'Religion': u'buddhist'},

{u'Count': 2, u'Religion': u'jewish'}]

(3)Most popular food in this area

food =

db.cities.aggregate([{"$match":{"amenity":{"$exists":1}, "amenity":"restaurant"}},

{"$group":{"_id":{"City":"$city_name","Food":"$cuisine"},"count":{"$sum":1}}},

{"$project":{"_id":0, "Food":"$_id.Food", "Count":"$count"}},

{"$sort":{"Count":-1}},

{"$limit":6}])

pprint.pprint(list(food))

[{u'Count': 16, u'Food': u'american'},

{u'Count': 10, u'Food': u'italian'},

{u'Count': 10, u'Food': u'mexican'},

{u'Count': 8, u'Food': u'burger'},

{u'Count': 4, u'Food': u'pizza'}]

American food is most popular in Santa Monica, followed by Italian food and Mexican food.

5.Conclusion

When checking the output of the data as it was being cleaned, I found that data for one area would be easily messy with its neighbors, expecially for cities like Santa Monica and Los Angeles, which two are bounded by each other and share same first two digits in the zipcodes.

As a famous beach city attracting many visitors, it suprised me that data for Santa Monica is not that complete, for example, when cheking most popular amenities, even the top twos "fountain" and "bus stations" only appeared twice. Maybe combined with Los Aneles into checking would provide us

more infomation as it is very possible that Santa Monica was considered as part of Los Angeles instead of an independent city durig data entry.