OpenStreetMap-Data-Case-Study

Project 1 : Udacity Data Analyst(Advanced) Nanodegree

Author : Yuan Zhou

Map Area

Stockholm, Sweden

Dataset of preselected metro area

Explore the Dataset

After taking every 400-th top level element of the original data stockholm_sweden.osm, the sample.osm was created to have an overview of the dataset.

```
stockholm_sweden.osm | 1.45 GB
sample.osm           | 3.50 MB
```

1. Check tags type

```
{'lower': 4170, 'lower_colon': 1734, 'other': 90, 'problemchars': 0}
```

I am a bit curious about the 90 other unknown type tags. and then I find that there are some tags only contain uppercase `'FIXME'` , some tags contain swedish special letter `'ref:raä'` and some tags contain two or more colon `'seamark:light:1:sector_end'`

2. Check unusual street names and postcode.

Unlike the dataset learnt in the course, the swedish street names combine the name with street in one word. For example, `Adolfsbergsvägen` , 'Adolfsbergs' is name and 'vägen' means the street. For this specific dataset, there is no need to audit street names which are uniform.

There are only two different postcode. 5 digits, or (3 digits + space + 2 digits), e.g., ('11415', '114 15')

3. Preparing dataset for desired "node" and "way" file.

   Parse the elements in the OSM file and transform them to tabular format. Thus resulting in "nodes.csv", "nodes_tags.csv", "ways.csv", "ways_node.csv" and "ways_tags.csv" which can be easily imported to a SQL database as tables for further analysis.

Problems Encountered in the Map

After exploring the sample dataset and get the desired CSV files, I noticed

CSV file written with Python has blank lines between each row

The solution is found on stackoverflow. The reason is that Python 2 CSV writer produces wrong line terminator on Windows.

Swedish special characters

Swedish special characters åäöÅÄÖ are messed up in the CSV files. For example, "Örbyhus" is encoded as "Ã–rbyhus".

To resolve this issue, manually change the exported CSV format from UTF-8 to ISO-8859-1 encoding.

```
def writerow(self, row):
  super(UnicodeDictWriter, self).writerow({
    k: (v.encode('ISO-8859-1','ignore') if isinstance(v, unicode) else v)
    for k, v in row.iteritems()
    })
```

Inconsistent postcode

The Swedish postcode system is based on a five-digit number combination, divided into two groups of three and two digits. The goal is to make sure all postcodes in the right format.

- "11619" to "116 19"

```
  def transfer_postcode(postcode):
  """Transform postcode to correct format.
  :param postcode:
  :return:

  >>> transfer_postcode("11619")
  '116 19'
  >>> transfer_postcode("116 19")
  '116 19'
  """
  match = re.search(r"(\d{3})\s*(\d{2})", postcode)
  if match:
      return match.group(1) + " " + match.group(2)
```

Data Overview

Now it is time to import .csv files as tables in Stockholm database and use the SQL queries to have an overview of the city.

File sizes

```
stockholm.db        735 MB
nodes.csv           529 MB
nodes_tags.csv      19.4 MB
ways.csv            43.7 MB
ways_tags.csv       55.2 MB
ways_nodes.csv      193 MB
```

USERS

Number of unique users

```
SELECT COUNT(DISTINCT(e.UID))
FROM (SELECT UID FROM NODES UNION ALL SELECT UID FROM WAYS) e;
```

```
2954
```

Top 10 contributing users

```
SELECT e.USER, COUNT(*) as NUM
FROM (SELECT USER FROM NODES UNION ALL SELECT USER FROM WAYS) e
GROUP BY e.USER
ORDER BY num DESC
LIMIT 10;
```

USER	NUM
MichaelCollinson	723392
Fringillus	700885
huven	461503
emj	424292
jordgubbe	252176
Tooga	239063
SA0BJW	207343
TheOddOne2	168680
Snusmumriken	167677
Zorac	159200

WAY and NODES

Number of nodes

```
SELECT COUNT(*) FROM NODES;
```

```
6839865
```

Top 10 nodes type

```
SELECT TYPE as node_type, count(*)
FROM NODE_TAGS
GROUP BY TYPE
ORDER BY count(*) DESC
LIMIT 10;
```

node_type	count
regular	307372
addr	251882
seamark	12934
light	5424
is_in	1184
recycling	943
name	884
contact	494
ref	476
payment	434

The result shows that most nodes do not have specific type which is not convenient for analysis.

Number of ways

```
  SELECT COUNT(*) FROM WAYS;
```

```
776961
```

Each way has `6839865\776961=8.8` nodes in average.

Top 10 way types

```
SELECT TYPE as way_type, count(*)
FROM WAY_TAGS
GROUP BY TYPE
ORDER BY count(*) DESC
LIMIT 10;
```

way_type	count
regular	1300065
addr	375157
building	22884
roof	16345
railway	8732
lst	3409
mtb	2831
source	899
name	700
maxspeed	503

Top 10 ways with most nodes

```
SELECT ID as WAY_ID, COUNT(*) as NUM_OF_NODES
FROM WAY_NODES
GROUP BY ID
ORDER BY COUNT(*) DESC
LIMIT 10;
```

WAY_ID	NUM_OF_NODES
513726309	1906
57037180	1902
208588290	1809
307214937	1804
251485316	1800
244800188	1789
309608363	1763
241085854	1757
461615668	1757
271742624	1751

Districts

Stockholm Municipality is divided into 14 boroughs. The boroughs are subdivided into **districts**.

Top 10 districts with most nodes information

```
SELECT tags.VALUE as CITY, COUNT(*) as count
FROM  (SELECT * FROM NODE_TAGS ) tags
WHERE tags.KEY='city'
GROUP BY tags.VALUE
ORDER BY COUNT(*) DESC
LIMIT 10;
```

CITY	count
Stockholm	14859
Uppsala	3123
Älvsjö	2070
Nacka	1819
Bromma	1579
Årsta	1250
Johanneshov	1157
Upplands Väsby	1118
Hägersten	1084
Enskede	873

Data Exploration

After having an overview of the data. I would like to investigate something related with recycling since Scandinavia countries are famous for their environmentally friendly recycling system.

1. How many kinds of recycling?

```
SELECT count(*)
FROM (SELECT KEY as category
FROM NODE_TAGS
WHERE TYPE='recycling' AND VALUE='yes'
GROUP BY KEY) as recycle
```

   Amazing, there are `33` categories in total for recycling

2. What categories are for recycling?

```
SELECT KEY as category, count(*)
FROM NODE_TAGS
WHERE TYPE='recycling' AND VALUE='yes'
GROUP BY KEY
ORDER BY count(*) DESC;
```

category	count
glass	140
paper	136

cans	72
plastic	54
batteries	48
glass_bottles	43
scrap_metal	34
plastic_packaging	33
newspaper	32
cardboard	28
clothes	27
plastic_bottles	26
cartons	24
paper_packaging	22
magazines	18
waste	10
aluminium	9
metal	4
books	3
electrical_appliances	2
green_waste	2
low_energy_bulbs	2
small_appliances	2
wood	2
wrapping	2
bottles	1
bulbs	1
compost	1
engine_oil	1
metal_packaging	1

mobile_phones	1
organic	1
plastic_bags	1

3. How many nodes have recycling station?

```
SELECT count(ID)
FROM NODE_TAGS WHERE KEY='amenity' and VALUE='recycling';
```

```
716
```

4. How many ways have recycling nodes?

```
SELECT count(WAY_NODES.ID)
FROM WAY_NODES
JOIN (SELECT ID FROM NODE_TAGS WHERE KEY='amenity' and
VALUE='recycling') as NODE_ID
ON WAY_NODES.NODEID= NODE_ID.ID
```

```
6
```

Conclusion

After reviewing the data, I found that there are many interesting things can be explored. Since I am not so familiar with the facilities and traffic, I choose the recycling topic to discussed. As we can see that the recycling rules are particular, and there are many nodes have the recycling station in the Stockholm area, however, I am not sure the whether way data is complete to see how many ways have recycling service cause 6 is too less.

Reference

Charlotte

Toronto