Open Street Map Case Study

Date: March 9th, 2018

Author: Yilin Qiu

Map Area

Toronto, Canada

I moved to Toronto after I graduated, so I want to investigate the user contributions to this area in the open street map and city amenities.

The link of the dataset:

https://www.openstreetmap.org/export#map=11/43.7183/-79.3767

Generate sample file

The complete dataset is around 476MB, so I need to obtain a sample which is less than 10MB. I used the get_element function to obtain a sample of the Toronto osm file. I changed the k to 50 to obtain a sample osm file which is around 10MB.

Data Audit

1. First, I iterated through the dataset using element tree.iterparse() method and count the number of tags in the dataset. The result is shown below:

```
defaultdict(<class 'int'>,
            {'member': 2038,
             'nd': 44057,
             'node': 38630,
             'osm': 1,
             'relation': 105,
             'tag': 44209,
             'way': 6639})
```

2. Then I checked the k value for each tag and separated them into three different groups: lowercase letters, lowercase letters with a colon, problematic characters and other characters. The result is shown below:

```
{'lower': 27209, 'lower_colon': 16292, 'other': 708, 'problemchars': 0}
```

From the result we can see that there is no problem character in the k value, which means that the data quality is quite good.

3. I then explored the number of unique users who have contributed to this map. I found there 639 unique users who have entered information.

Problems encountered in the Map

1. Abbreviated street name
2. Inconsistent province

Abbreviated street name:

After parsing the xml data, I found there are some abbreviated street names such Augusta Ave, Hazelton Ave. I used the regular expression to match the last word of the street name which is the street type. Then I returned the matched results which are not in expected street type list. I created a mapping to correct the wrong street types.

Inconsistent province:

I found the province name entries are not consistent. Some entries use the province code “ON”, while others use the full name “Ontario”. So, I replaced the full name “Ontario” with the province code “ON” to make the province name consistent.

Data overview

This section contains basic statistics information about the Toronto open street map dataset, the sql queries used to gather them and some additional ideas about the data.

File sizes

```
The toronto.osm file size is 499.577352 MB
The sample.osm file size is 10.047653 MB
The nodes.csv file size is 3.562357 MB
The nodes_tags.csv file size is 0.954338 MB
The ways.csv file size is 0.454131 MB
The ways_nodes.csv file size is 1.069027 MB
The ways_tags.csv file size is 1.074861 MB
The toronto.db file size is 8.450048 MB
```

Numer of Nodes

```
sqlite> SELECT COUNT(*)
   ...> FROM Nodes;
38630
```

Number of Ways

```
sqlite> SELECT COUNT(*)
   ...> FROM way;
6639
```

Number of unique users

```
sqlite> SELECT COUNT(user.uid)
   ...> FROM (SELECT uid FROM Nodes UNION SELECT uid FROM way) user;
630
```

Top 10 Contributing users

```
sqlite> SELECT user.user as User, COUNT(*) as User_Count
   ...> FROM (SELECT user FROM Nodes UNION ALL SELECT user FROM way)
user
   ...> GROUP BY User
   ...> ORDER BY User_Count DESC
   ...> LIMIT 10;
"b'andrewpmk'",27253
"b'Kevo'",3285
"b'Matthew Darwin'",2767
"b'Victor Bielawski'",2001
"b'Bootprint'",1167
"b'Mojgan Jadidi'",790
"b'andrewpmk_imports'",526
"b'MikeyCarter'",475
"b'TristanA'",463
"b'Nate_Wessel'",427
```

Number of users appearing only once

```
sqlite> SELECT COUNT(*)
   ...> FROM (SELECT user.user AS User, COUNT(*) AS User_Count FROM
   ...> (SELECT user FROM Nodes UNION ALL SELECT user FROM way) user GROUP
BY User
   ...> HAVING COUNT(*) = 1);
274
```

Additional ideas:

Top 10 popular amenities

```
sqlite> SELECT COUNT(*), value
   ...> FROM Node_tags
   ...> WHERE key = "b'amenity'"
   ...> GROUP BY value
   ...> ORDER BY COUNT(*) DESC
   ...> LIMIT 10;
135|b'fast_food'
111|b'bench'
111|b'restaurant'
72|b'post_box'
69|b'parking'
51|b'bank'
48|b'cafe'
39|b'pharmacy'
39|b'waste_basket'
39|b'waste_basket;recycling'
```

Top 10 popular cuisines

```
sqlite> SELECT Node_tags.value, COUNT(*)
   ...> FROM Node_tags
```

```
   ...> JOIN (SELECT DISTINCT(id) FROM Node_tags WHERE value =
"b'restaurant'") i
   ...> ON Node_tags.id = i.id
   ...> WHERE Node_tags.key = "b'cuisine'"
   ...> GROUP BY Node_tags.value
   ...> ORDER BY COUNT(*) DESC
   ...> LIMIT 10;
b'chinese'|9
b'indian'|6
b'pizza'|6
b'brazilian'|3
b'greek'|3
b'italian'|3
b'mexican'|3
b'sushi'|3
```

Number of Tim Hortons

```
sqlite> SELECT COUNT(*)
   ...> FROM Node_tags
   ...> WHERE value LIKE "%Tim Hortons%";
24
```

Number of Starbucks

```
sqlite> SELECT COUNT(*)
   ...> FROM Node_tags
   ...> WHERE value LIKE "%Starbucks%";
18
```

Top 10 popular bank

```
sqlite> SELECT Node_tags.value, COUNT(*)
   ...> FROM Node_tags
   ...> JOIN (SELECT DISTINCT (id) FROM Node_tags WHERE value = "b'bank'")
i
   ...> ON Node_tags.id = i.id
   ...> WHERE Node_tags.key = "b'name'"
   ...> GROUP BY Node_tags.value
   ...> ORDER BY COUNT(*) DESC
   ...> LIMIT 10;
b'TD Canada Trust'|15
b'RBC Royal Bank'|6
b'RBC'|6
b'Scotiabank'|6
b'CIBC'|3
b'Callian Capital'|3
b'CityCan Mortgage Services'|3
b'RBC Financial Group'|3
b'Shinhan Bank'|3
b'Tangerine'|3
```

Other ideas about the dataset

1. In my opinion, the Toronto open street map dataset is quite incomplete. Around half of the users appear only once. In addition, many nodes do not contain any useful information such as address, amenity and name etc. Therefore, user activity should be improved to have more users enter useful information. Measures such as rewards and missions can be used to improve the user activity.

   Benefits:
   This solution will result in a more detailed and complete open street map database. Moreover, the open street map can be updated regularly, and data analysis based on it will become more beneficial for users and companies.

   Issues:

Since this solution requires many rewards, it might be cost-effective. In addition, missions that are used to improve user activity need to be designed and planned, which requires lots of human resources, time and effort.

2. Since there are many users enter information, human input errors often exist. To improve data quality, we can use google map API and PokemonGo API to correct input errors and add more detailed information to the dataset.
   Benefits:
   This makes the open street map more accurate and more reliable. Data scientists can obtain more accurate information on the city's amenities and public transit systems.
   Issues:
   It might be hard to validate the data using other APIs since the data format might be inconsistent.

Conclusion

Overall, the Toronto open street map is a well-cleaned data with large amounts of user contributions. Although there are still some problems such as abbreviated street names and inconsistent province names, the data quality is quite high. After solving those two problems, the dataset can be analyzed using SQL queries. The results obtained from SQL queries are quite accurate. However, the open street map still need to obtain updated information from other sources such as google map API and PokemonGo API etc.