OpenStreetMap Data Case Study: Seattle

Map area:

http://www.openstreetmap.org/export#map=11/47.6057/-122.3616

I chose Seattle, Washington as my map area. Seattle is one of my favorite cities and I am curious to see what the queries output regarding frequency of different amenities and hopefully be able to contribute my work the OSM.

Auditing

In order to find out what was wrong with the data I had to run scripts to audit it. I started with the street name and ran a script that parsed through the file and counted the instances of each street name. Using the same method, I then moved on to postal codes and found no issues with the data. That is when I decided to move to check shops, amenities and so on.

Problems Encountered

The data was in too bad of shape but there were some issues that I found and needed to address.

1. Abbreviated street names: There were occurrences of abbreviations such as ‘St’ for ‘Street’, ‘Pl’ for ‘Place’, and ‘S’ for ‘South’, to name a few examples.
2. Amenity repetition: There were instances of repetitive amenity names. I decided to change ‘fast_food’ occurrences to ‘restaurant’ and ‘atm’ to ‘bank’.

Abbreviated Street Names

I had to audit the OSM data first, which revealed the issues I need to fix. I mapped the need changes and created a loop that would iterate through each word of the given string to fix not only words on the end of the string but also occurrences of the issue in the middle of a string. I made a noted to exclude change to words that could be abbreviations of the word ‘Suite’, which otherwise would be mistaken and changed to ‘Street’. Below is Python code used to fix the street abbreviations.

```
for w in range(len(words)):
    if words[w] in mapping:
        if words[w].lower() not in ['suite', 'ste.', 'ste']:
            words[w] = mapping[words[w]]
            name = " ".join(words)
```

Below is the output:

```
S Bradner Pl => South Bradner Place
15th Ave S => 15th Avenue South
Westlake Ave => Westlake Avenue
```

Amenity Repetition

Next, I changed all of the occurrences of ‘fast_food’ and ‘atm’ and changed them to ‘restaurant’ and ‘bank’, respectively. Below is the output:

```
fast_food => restaurant
atm => bank
```

Data Overview

Size of file:

```
seattle_WA.osm...52.4 MB
```

Number of nodes:

```
sqlite> SELECT COUNT(*) FROM nodes;
41633
```

Number of ways:

```
sqlite> SELECT COUNT(*) FROM ways;
2529
```

Number of restaurants:

```
sqlite> SELECT COUNT(*)
   ...> FROM nodes_tags
   ...> WHERE value='restaurant';
48
```

Top 5 most popular amenities:

```
sqlite> SELECT value, count(*) as total
   ...> FROM nodes_tags
   ...> WHERE key='amenity'
   ...> GROUP BY value
   ...> ORDER BY total DESC
   ...> LIMIT 5;
```

```
bicycle_parking,122
restaurant,48
cafe,33
waste_basket,26
bench,18
```

Top 5 most popular cuisines:

```
sqlite> SELECT nodes_tags.value, COUNT(*) as num
   ...> FROM nodes_tags
   ...>     JOIN (SELECT DISTINCT(id) FROM nodes_tags WHERE value='restaurant') i
   ...>     ON nodes_tags.id=i.id
   ...> WHERE nodes_tags.key='cuisine'
   ...> GROUP BY nodes_tags.value
   ...> ORDER BY num DESC
   ...> LIMIT 5;
```

```
american,8
italian,8
mexican,8
pizza,6
seafood,6
```

Additional Ideas

Improvements

I think the main issue is the that usually one user contributes a huge proportion to the data in a given area. For my area, user ‘Glassman’ contributed to a quarter of the information. This may bring some consistency in the way the data is cleaned but with such few people contributing, it will be difficult for a lot OSM to ever be cleaned. I was not aware of the fact that one could actually contribute their information until this project, and I think that if OSM promoted it a little more there wouldn’t be so much dirty data.

Benefit:

- This will result in more clean data in OpenStreetMap.
- Exposure for OpenStreetMap.

Anticipated Issues:

- A possible issue of a lot more people cleaning the data is inconsistency.
- Also, just because more people are cleaning it, doesn't mean then really know how which could just make the problem worse.

Analysis

An additional idea for analysis is using to research for business ventures. When a company or person is looking to open up shop somewhere, this data analysis would be useful in researching postal codes and the occurrences of certain businesses in the area. For example, if I want to open up a Mexican restaurant in a certain location, I can use the data to find how many Mexican restaurants are already in the area. If I find that there are already 5, then I may choose to open shop somewhere else.

Conclusion

While the data was not in terrible shape, there were some improvements to be made. I addressed the issues with the street names and repetitive amenity categories. It was interesting to find the coffee and cafes were Seattle's most popular cuisine and amenity. Although not surprising, it was fun to see that Seattle's reputation for being a coffee hub actually holds water.