1. Problems encountered in the map

The Toronto dataset was remarkably well maintained. The main issue was with abbreviated street names (eg. Blvd., St. Ave.). I decided that it was best to remove these in order to maintain consistency. The abbreviations which appear at the end of the string were easy enough to deal with. The challenge was those that appear in the substring (eg. Jerry Blvd N). Luckily, small update to the code from problem set 6 solved my issues.

The most challenging part of this was going through the huge variety of road names like “Treeway”, “Harbour” and “Gateway”. There were a lot more than we saw in the problem sets. Definitely was interesting going through the dataset!

2. Data Overview

This section contains basic statistics about the dataset and the MongoDB queries used to gather them.

File sizes
Toronto_canada.osm 1.24gb
Sample_toronto.osm.json 1.76gb

However in the interest of saving my macbook from exploding, I used a smaller subset of the data for this assignment (about 700mb)

```
# Number of documents
> db.cities_features.find().count()
2273700

# Number of Nodes
> db.cities_features.find({“type”: “node”}).count()
1747114

# Number of ways
> db.cities_features.find({“type”: “way”}).count()
526586

# Number of unique users
> db.cities_features.distinct(“created.user”).length
1500

# Top 1 contributing user
> db.cities_features.aggregate([{"$group": {"_id": "$created.user", "count": {"$sum": 1}}}, {"$sort": {"count": -1}}, {"$limit": 1}])
{ "_id" : "andrewpmk", "count" : 1403806 }
```

Number of users appearing only once (having 1 post)
> db.cities_features.aggregate([{"$group":{"_id":"$created.user", "count":{"$sum":1}}}, {"$group":{"_id":"$count", "num_users":{"$sum":1}}}, {"$sort":{"_id":1}}, {"$limit":1}])
500

3. Additional Ideas

Top 10 appearing amenities

db.cities_features.aggregate([{"$match":{"amenity":{"$exists":1}}}, {"$group":{"_id":"$amenity", "count":{"$sum": 1}}}, {"$sort": {"count": -1}}, {"$limit": 10}])

{ "_id" : "parking", "count" : 2550 }
{ "_id" : "fast_food", "count" : 298 }
{ "_id" : "restaurant", "count" : 269 }
{ "_id" : "school", "count" : 259 }
{ "_id" : "bench", "count" : 206 }
{ "_id" : "place_of_worship", "count" : 184 }
{ "_id" : "post_box", "count" : 175 }
{ "_id" : "cafe", "count" : 142 }
{ "_id" : "waste_basket", "count" : 114 }
{ "_id" : "fuel", "count" : 96 }

Biggest religion (no surprise here)

db.cities_features.aggregate([{"$match":{"amenity":{"$exists":1}, "amenity":"place_of_worship"}},{"$group":{"_id": "$religion", "count": {"$sum": 1}}}, {"$sort": {"count": -1}}, {"$limit": 1}])

{ "_id" : "christian", "count" : 149 }

Conclusion

The structure and quality of the data has been well maintained by contributors. Toronto's data sees to be surprisingly complete compared to other cities, which I think speaks highly of the open source community we have here.