Data Wrangling with MongoDB Final Project

Evan Harley

Map Area: Victoria, British Columbia, Canada

Link to Area

1 Problems Encountered:

After downloading the data set for Victoria, and parsing it for upload to MongoDB, I found 3 problems with the data in the data set. I will discuss them in order of those I found simplest to most complex to solve.

1. Inconsistent Postal Code format
2. Data from Washington State in the data set
3. Erroneous City values

Inconsistent Postal Code Format:

After importing the data into MongoDB, I knew that the first likely problem would be encountered with Postal Codes. So, I ran the following query:

```
>db.victoria.aggregate([{'$match': {'address.postcode': {'$exists': 1}}},
                    {'$group': {'_id': '$address.postcode', 'count': {'$sum': 1}}},
                    {'$sort': {'count': -1}}])
{ "_id" : "98221", "count" : 63 }
{ "_id" : "V8K 1V5", "count" : 26 }
{ "_id" : "V9B2S3", "count" : 12 }
{ "_id" : "V8W 1H2 ?", "count" : 2 }
```

The above text shows the 4 types of postal code values found in my data set. Valid British Columbia postal codes with a space, valid British Columbia postal codes without a space, valid British Columbia postal codes with a unicode character appended to the end, and lastly Washington State postal codes. For the sake of uniformity I formatted all of t he BC postal codes as Post Canada does with a space.

Washington State Data:

As above I found postal codes from Washington State included in my data set. The following query confirmed it for me.

```
> db.victoria.aggregate([{'$match': {'address.city': {'$exists': 1}}},
                                   {'$group': {'_id': '$address.city', 'count': {'$sum': 1}}}, {'$sort': {'_id': 1}}])

{ "_id" : "Anacortes", "count" : 42 }

{ "_id" : "Becher Bay 1", "count" : 30 }

{ "_id" : "Bellingham", "count" : 1 }
```

Of the three city names above only one is from the Victoria Area. I decided that as I was working with the Victoria, Canada data set, I would remove the Washington State Values.

Erroneous City Values:

I found 3 types of erroneous city values in my data set they are in order of least problematic to most.

1. City or Township names that do not match Post Canada's listing for the area
2. City or Township names that have a number appended to the end of them
3. 792 entries with the completely erroneous value “Capital H (Part 1)”

Demonstrated by the following results of the previous query.

```
{ "_id" : "Becher Bay 1", "count" : 30 }

{ "_id" : "Cole Bay 3", "count" : 27 }

{ "_id" : "Capital H (Part 1)", "count" : 792 }
```

The first of these was easy to fix. Simply changing all of the city values to match their Post Canada listing. The second too was very simple to fix. I updated the address.city value to the appropriate name without the trailing character. The third one however took a lot of time to fix, as the area covered by the erroneous values was significant, and covered 4 different cities. The code I used to fix it is found in my additional code document.

2 Data Overview:

The following section contains general information regarding the data set and the queries used to get the information.

File Size:

victoria_canada.osm	247 MB
victoria_canada.osm.json	274 MB

Total Number of Documents:

```
> db.victoria.find().count()
1259504
```

Total Number of Nodes:

```
> db.victoria.find({'type': 'node'}).count()
1161590
```

Total Number of Ways:

```
> db.victoria.find({'type': 'way'}).count()

97804
```

Total Number of Users:

```
> db.victoria.aggregate([{'$match': {'created.user': {'$exists': 1}}},
                 {'$group': {'_id': '$created.user', 'count': {'$sum': 1}}},
                 {'$group': {'_id': 'Number unique Users', 'count': {'$sum': 1}}}])
{ "_id" : "Number unique Users", "count" : 642 }
```

Top 3 Amenities:

```
> db.victoria.aggregate([{'$match': {'amenity': {'$exists': 1}}},
                        {'$group': {'_id': '$amenity', 'count': {'$sum': 1}}},
                        {'$sort': {'count': -1}}])
{ "_id" : "parking", "count" : 1626 }
{ "_id" : "bench", "count" : 1019 }
{ "_id" : "waste_basket", "count" : 326 }
```

3 Other Ideas:

Involvement of the University Community:

After running my amenity query, I was surprised to find that only a very small percentage of the records were in the City of Victoria proper. A mere 2.01% of the total number of records with a city value. While Saanich, known for parks and hiking trails, has 41.59% of the total. Similarly, benches account for 19.88% of the total number of records with an amenity value. While restaurants, and bars account for 3.61% and, 0.15% of the total respectively. I find this particularly surprising given the size of the local University. I would have thought that, given the size of their computer science program, that something like Open Street Map data would be a tool used by Professors to introduce students to scripting tasks. I think that by engaging Universities' student populations could drive improvement both in the quality of the data set, and in the variety of information supplied. Particularly if the gamification idea found in the sample project were used.

Other Queries:

Top 3 Schools:

```
> db.victoria.aggregate([{'$match': {'amenity': {'$exists': 1}, 'amenity': 'school'}},

                         {'$group': {'_id': '$name', 'count': {'$sum': 1}}},

                         {'$sort': {'count': -1}}])

{ "_id" : null, "count" : 41 }

{ "_id" : "Gulf Islands Secondary School", "count" : 4 }

{ "_id" : "Sequim Community School", "count" : 2 }
```

Top 3 Cuisines:

```
> db.victoria.aggregate([{'$match': {'amenity': {'$exists': 1}, 'amenity': 
'restaurant'}},
                     {'$group': {'_id': '$cuisine', 'count': {'$sum': 1}}},
                     {'$sort': {'count': -1}}])
{ "_id" : null, "count" : 119 }
{ "_id" : "pizza", "count" : 8 }
{ "_id" : "thai", "count" : 8 }
```

I find it interesting how few of the amenity listings have information attached to them.

4 Conclusion:

After processing this data set, and cleaning it as well as I can, I have found that the data set is very incomplete. Most of the records with amenities seem to be tied to Parks and Hiking Trails. Which does make sense given the population of Victoria. However, I think with some involvement of students that could be broadened, to include many more interesting locations. It is also my opinion that by involving computer science students the data could also be cleaned regularly resulting in a more robust and useful data set.