Data Analyst Nanodegree - Project #3

Data Wrangling with MongoDB

Adam Wright

Map Area: Santa Cruz, CA, United States

https://mapzen.com/data/metro-extracts/#santa-cruz-california

1. Problems Encountered in the map

Postal Codes

There were two primary problems with the postal codes:

1. Some 5 digit Zips were preceded by the state code ('CA')

2. Some 5 digit Zips were followed by a hyphen and the Zip Four (e.g. '95065-1711')

In order to strip out any alpha characters I used the Regular Expression module:

```
re.sub('r\D', '',  subtag.get(v))
```

That accomplished, all that has to be done to exclude the Zip 4 is to only pass through the first 5 characters in the resulting string:

```
re.sub('r\D', '',  subtag.get(v))[:5]
```

City Names

There were a number of cities other than Santa Cruz stored in the address[cityname] dictionaries: Scott Valley, Captiola, Felton, Aptos, and Soquel. This turned out to not be a real issue though, as Google Maps proved that these municipalities are all immediately contingent to Santa Cruz and can be reasonably included in a Santa Cruz data extract.

'tiger' and 'gnis' attributes

Each element included a large number of 'tiger' and 'gnis' attributes. The value of these attributes were completely redundant to data that I was already capturing. In the interest of simplifying the resulting .json documents, I simply excluded these attributes (using continue) from my shape element function.

2. Data Overview

File Sizes

santa_cruz.osm - 52 MB

santa_cruz.json - 58 MB

Number of Documents

```
> db.santa_cruz.find().count()
```

271,884

Number of Nodes

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

250,739

Number of Ways

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

21,086

Number of Unique Users

```
> db.santa_cruz.distinct('created.user').length
```

434

Top User

```
> db.santa_cruz.aggregate([{'$group':'$created.user',
        'count':{'$sum':1}}},
        {'$sort':{'count':-1}},
        {'$limit':1}])
```

{'_id': 'stevea', 'count': 154498}

3. Additional Ideas

Is the University overrepresented?

While running queries to try and uncover any abnormalities in that data, two things really stuck out:

1. There was a Zip Code - 95064 - that had nearly 5 times as many tags as the next closest:

```
db.char.aggregate([{"$match":{"address.postcode":{"$exists":1}}},
        {"$group":{"_id":"$address.postcode", "count":{"$sum": 1}}},
         {"$sort":{"count":1}}])
```

{'_id': '95064', 'count': 402}
{'_id': '95064', 'count': 84}

2. The most tagged street was 'Porter-Kresge Road':

```
db.char.aggregate([{"$match":{"address.street":{"$exists":1}}},
        {"$group":{"_id":"$address.street", "count":{"$sum": 1}}},
         {"$sort":{"count":1}}])
```

{'_id': 'Porter-Kresge Road', 'count': 35}

It turns out that Zip Code 95064 is exclusive to the University of Calfornia, Santa Cruz and that Porter-Kresge Road is a main thoroughfare on the university grounds!

4. Additional Queries

Top 10 amenities

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

{'_id':'parking', 'count': 950}
{'_id':'bicycle_parking', 'count': 263}
{'_id':'restaurant', 'count': 249}
{'_id':'toilets', 'count': 221}
{'_id':'bench', 'count': 201}
{'_id':'place_of_worship', 'count': 141}
{'_id':'school', 'count': 105}
{'_id':'recycling', 'count': 96}
{'_id':'cafe', 'count': 89}
{'_id':'fast_food', 'count': 68}

It would appear that people and Santa Cruz both drive and bike a lot as parking for cars and bikes are the top two tagged amenities. You also have to love a place with more tagged cafes than fast food restaurants!

Most popular cuisines

```
db.santa_cruz.aggregate([{"$match":{"amenity:{"$exists":1}, "amenity":"restaurant"}},
    {"$group":{"_id":"$cuisine", "count":{"$sum":1}}},
    {"$sort":{"count":-1}},
    {"$limit":5}])
```

{'_id':'mexican', 'count': 32}
{'_id':'chinese', 'count': 22}
{'_id':'pizza', 'count': 21}
{'_id':'italian', 'count': 14}
{'_id':'japanese', 'count': 13}

Mexican was the most popular cuisine followed by the now ubiquitous American takeout experience: Chinese. The top 5 was rounded out by pizza, Italian, and Japanese. Mexican being the top cuisine and Japanese being in the top 5 cuisines strikes me as being quintessentially Californian.

5. Conclusion

By far the most interesting feature of this dataset is the degree to which the University is "over tagged" compared to the rest of Santa Cruz. While the characteristics of a University population - younger, better educated, more technologically sophisticated - are indicative of a group more comfortable with internet age technologies like open source mapping the degree of the difference - nearly 5 times as many Zip tags - is remarkable. While this isn't a problem per se, encouraging students/faculty to branch out and cover the surrounding area to the degree the campus is covered certainly represents an opportunity. As an incentive, Santa Cruz has some of the best surfing in the country and you have to pass through the city to get to the beach from the campus! On a more serious note, if the Computer Science department at UCSC is in a civic mood, creating an automated process to tag the rest of the city to the level of the campus could be an interesting project for a group of undergraduates.