Data Wrangling with MongoDB and OpenStreetMap data

André Marais

Map Area: Cape Town and surrounding area, South Africa

Map data gathered from https://mapzen.com/data/metro-extracts

1. Concerns with the Data

The data from Cape town and the surrounding area is surprisingly clean. I did pick up on two issue which required further investigation.

- Very high concentration of certain postcodes and city names.
- I'm slightly curious as to *why* the data is so clean

High Concentration of certain postcodes and city names

Regarding the postcodes, postcode 7405 appears more than 62% of the time, which is reason for further investigation. If you look at the user frequency, you'll see there's one person in particular who is very active (Adrian Firth). More than 75% of the address updates were done by him. Doing a group by city and user, you'll see he's the most active in Pinelands (which is also the most frequent city). So it is not a surprise that the Pinelands data in particular is so abundant and clean. See details below for city frequency (users per city)

Why the data is so clean

This is mainly thanks to Adrian. All of the post codes are 4 digits (the South African standard), and within the right range (between 7100 and 8099 for Cape Town and the surrounding areas). Apart from replacing abbreviate street types, there was almost no other cleaning left to do.

2. Data Overview

Cape town has a relative small data set, compared to major capital cities in the world. The entire data set is but only 21.5 MB. Cleaned and converted to JSON it's sitting at 31.6 MB.

Document count

```
[{"$group" : {"_id" : "$type", "count" : {"$sum" : 1}}},
 {"$sort" : {"count" : -1}}]
```

shows that there are 96,972 nodes and 17,242 ways, totalling 114,221 documents. There are also 2 'palms' and 2 'routes' we can conveniently ignore.

User statistics

As stated earlier, Adrian Firth is the top contributor. Below is a list of the top 5 contributors:

```
[{"$match":{"address.city":{"$exists":1}}},
 {"$group":{"_id":"$created.user", "count":{"$sum":1}}},
 {"$sort":{"count":-1}},
 {"$limit" : 5}]
```

shows:
[{u'_id': u'Adrian Frith', u'count': 404},
 {u'_id': u'Stefanoodle', u'count': 50},
 {u'_id': u'adjuva', u'count': 35},
 {u'_id': u'Kelerei', u'count': 16},
 {u'_id': u'olivierk', u'count': 2}]

User per city

```
[{"$match":{"address.city":{"$exists":1}}},
 {"$group": {"_id": {"user" : "$created.user",
                                "city" : "$address.city"},
                     "count":{"$sum":1}}},
 {"$sort":{"count":-1}},
 {"$limit" : 10}]
```

You can see Adrian is very active in Pineland specifically:
[{u'_id': {u'city': u'Pinelands', u'user': u'Adrian Frith'}, u'count': **325**},
 {u'_id': {u'city': u'Rosebank', u'user': u'Adrian Frith'}, u'count': 39},
 {u'_id': {u'city': u'Mowbray', u'user': u'Adrian Frith'}, u'count': 36},
 {u'_id': {u'city': u'Cape Town', u'user': u'adjuva'}, u'count': 35},
 {u'_id': {u'city': u'Loevenstein', u'user': u'Stefanoodle'}, u'count': 35}]

Other Ideas

Some street name information

This is just interesting, the street names in Pinelands are named quite aptly :)

```
[{"$match":{"address.city":'Pinelands',}},
 {"$group":{"_id": {"street name" : "$address.street"},
"count":{"$sum":1}}},
 {"$sort":{"count":-1}},
 {"$limit" : 10}]
```

[{u'_id': {u'street name': **u'Forest** Drive'}, u'count': 15},
 {u'_id': {u'street name': **u'Ringwood** Drive'}, u'count': 15},
 {u'_id': {u'street name': **u'Sunny** Way'}, u'count': 7},
 {u'_id': {u'street name': u'Victory Avenue'}, u'count': 6},
 {u'_id': {u'street name': **u'Woodside** Drive'}, u'count': 6},
 {u'_id': {u'street name': u'Peak Drive'}, u'count': 6},
 {u'_id': {u'street name': u'Rhone'}, u'count': 6},
 {u'_id': {u'street name': u'Achilles Way'}, u'count': 5},
 {u'_id': {u'street name': **u'Field** Close'}, u'count': 5},
 {u'_id': {u'street name': u'Manatoka Avenue'}, u'count': 5}]

Restaurants and Fast food cuisine types

```
[{"$match":{"amenity":"fast_food"}},
 {"$group":{"_id":"$name", "count":{"$sum":1}}},
  {"$sort":{"count":-1}}]
```

and

```
[{"$match":{"amenity":"restaurant"}},
  {"$group":{"_id":"$cuisine", "count":{"$sum":1}}},
   {"$sort":{"count":-1}}]
```

Unfortunately shows very little information. Most of the restaurants haven't been classified, but at least all but one fast food joint have been named.

Conclusion

It seems like it was a bad decision to use data from Cape Town. The data set is very small and there's only a small pool of contributors for this data on OpenStreetMap. Adrian did quite an amazing job at cleaning the data he had, but there's still a lot of work left to do for Cape Town alone, never mind the other major cities in South Africa. We can apply the process that was set up by Lesson 6 and this assessment to do the same for Johannesburg as well, but I imagine the data would be *this* sparse again at the very best.