**Map Position and justification**

The area I’ve chosen is [Johannesburg](http://s3.amazonaws.com/metro-extracts.mapzen.com/johannesburg_south-africa.osm.bz2) in South Africa, small enough to be manageable (126MB uncompressed - compared to 3GB for Amsterdam) and big enough to be interesting. (selected the standard option from Map Zen)

Reasoning was that Africa is a completely new region to me but all the names still make sense as ‘Afrikaans’ is similar enough to Dutch for me to understand and assess for value. Making it interesting on multiple levels.

**Dataset processing**

Problems are as one would expect with manually entered data - a lot of different spellings for similar things, roads, townships, municipalities etc. Though someone already noted a lot of these issues with ‘FIXME’ - which was nice. Most of the these referred to approximate locations or being uncertain about a position, using Mapquest (Google Maps refused to give text based output) I could correct them for my own dataset.

Additionally, there are different ways of calling certain locations or infrastructure in South Africa, hence there is a need to update the expected values in the previously mentioned code of the chapters. For instance, “commons”, “close” and “crescent” are perfectly acceptable ways of naming a road. As consequence there were typo’s by users for these inputs as well, such as “cresent” for “crescent”. In other cases it might be better to simply refer to “road” or “street” as South African “ways” are too small to incorporate on their own.

A separate problem is that there are very few inputs that actually have a address, the majority of them are just coordinates inputs, with an name and amenity but no more information than that.

**Data Analysis**

The top user was responsible for 10% (10.09) of the total changes. 80% of all changes were made by the top 4.25% of users. In total there were about 640K changes by 753 people of which 143 only made a single change.

{'bounds': 1,

'member': 19385,

'nd': 693696,

'node': 547910,

'osm': 1,

'relation': 1695,

'tag': 273234,

'way': 90852}

Playing with it:

> db.streetdata.find().count()

638762

> db.streetdata.find({"type":"node"}).count()

547910

> db.streetdata.find({"type":"way"}).count()

90852

> db.streetdata.find({"address.street":{$exists:true}}).count()

713

The majority of the amenities turned out to be parking spots, which considering the lack of public transport in the region makes perfect sense. In a similar fashion the next most common amenity were fuel stations only then to be followed by more general amenities such as restaurants and schools.

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

{ "\_id" : "parking", "count" : 1253 }

{ "\_id" : "fuel", "count" : 503 }

{ "\_id" : "restaurant", "count" : 495 }

{ "\_id" : "school", "count" : 419 }

{ "\_id" : "fast\_food", "count" : 366 }

{ "\_id" : "place\_of\_worship", "count" : 205 }

{ "\_id" : "atm", "count" : 160 }

{ "\_id" : "bank", "count" : 154 }

{ "\_id" : "cafe", "count" : 122 }

{ "\_id" : "pharmacy", "count" : 111 }

{ "\_id" : "swimming\_pool", "count" : 88 }

{ "\_id" : "post\_office", "count" : 82 }

{ "\_id" : "pub", "count" : 72 }

{ "\_id" : "hospital", "count" : 71 }

{ "\_id" : "police", "count" : 39 }

{ "\_id" : "kindergarten", "count" : 37 }

{ "\_id" : "toilets", "count" : 32 }

{ "\_id" : "car\_wash", "count" : 31 }

{ "\_id" : "bar", "count" : 29 }

{ "\_id" : "cinema", "count" : 28 }

The top user was responsible for 10% (10.09) of the total changes and 80% of all changes were made by the top 4.25% of users. In total there were about 640K changes by 753 people of which only 141 made a single entry.

db.streetdata.aggregate([{$match:{"created.user":{$exists:1}}},{$group:{\_id:"$created.user","count":{$sum:1}}},{$sort:{"count":-1}}])

{ "\_id" : "Firefishy", "count" : 64596 }

{ "\_id" : "thomasF", "count" : 44966 }

{ "\_id" : "Adrian Frith", "count" : 41434 }

{ "\_id" : "NicRoets", "count" : 33488 }

{ "\_id" : "Markus59", "count" : 30313 }

{ "\_id" : "Tinshack", "count" : 30237 }

{ "\_id" : "Gerhardus Geldenhuis", "count" : 20540 }

{ "\_id" : "titanbeos", "count" : 20198 }

{ "\_id" : "bahnpirat", "count" : 18402 }

{ "\_id" : "MatzeM", "count" : 17113 }

{ "\_id" : "mapryan", "count" : 16863 }

{ "\_id" : "CharlizeFerreira", "count" : 16786 }

{ "\_id" : "Lara (EWC)", "count" : 16387 }

{ "\_id" : "user\_634020", "count" : 15489 }

{ "\_id" : "SuneMomsen", "count" : 14136 }

{ "\_id" : "chdr", "count" : 13058 }

{ "\_id" : "Rezane van Vuuren (EWCop)", "count" : 9086 }

{ "\_id" : "Jan de Jager", "count" : 8738 }

{ "\_id" : "Teddy73", "count" : 8137 }

{ "\_id" : "mcdmx", "count" : 7822 }

In terms of the various cuisines offered by the restaurants, burgers were on top followed by pizza and chicken. Which is surprising since the region is primarily famous for steakhouses and grill restaurants as indicated by Anthony Bourdain in his television show. This would indicate there is still a lot of data missing from this set. db.streetdata.aggregate([{$match:{"cuisine":{$exists:1}}},{$group:{\_id:"$cuisine","count":{$sum:1}}},{$sort:{"count":-1}}])

{ "\_id" : "burger", "count" : 61 }

{ "\_id" : "pizza", "count" : 52 }

{ "\_id" : "chicken", "count" : 28 }

{ "\_id" : "indian", "count" : 18 }

{ "\_id" : "regional", "count" : 17 }

{ "\_id" : "chinese", "count" : 15 }

{ "\_id" : "italian", "count" : 13 }

{ "\_id" : "coffee\_shop", "count" : 8 }

{ "\_id" : "fish", "count" : 8 }

{ "\_id" : "fish\_and\_chips", "count" : 5 }

{ "\_id" : "thai", "count" : 4 }

{ "\_id" : "steak\_house", "count" : 4 }

{ "\_id" : "Italian", "count" : 3 }

{ "\_id" : "international", "count" : 3 }

{ "\_id" : "portuguese", "count" : 3 }

{ "\_id" : "sushi", "count" : 3 }

{ "\_id" : "salad", "count" : 2 }

{ "\_id" : "survey", "count" : 2 }

{ "\_id" : "japanese", "count" : 2 }

{ "\_id" : "seafood", "count" : 2 }

**Future Data Ideas**

The relatively small amount of data was quite surprising, especially when compared to other major cities. Out of interest I checked the size of Amsterdam, Netherlands (219 km² in area and has a 3.13 GB uncompressed .osm file) vs Johannesburg, South Africa (1,645 km² in area and has a 127MB uncompressed .osm file). This clearly highlights why the data from my dataset was not necessarily as complete as I was expecting.

Getting enough data maybe a bit difficult and perhaps incorperating data from a source like <http://what3words.com/> would perhaps help to increase the overall size & accuracy of the information. Additionally, we could look at including information from other geolocation based services like Facebook, Foursquare & Google My Business to enhance the usefulness of the dataset.