Udacity Data Analyst Nanodegree P3:

Wrangle OpenStreetMap Data

Umang Agarwal

Map Area: Kolkata, West Bengal, India

<https://www.openstreetmap.org/node/245707150>

<https://mapzen.com/data/metro-extracts>

<https://en.wikipedia.org/wiki/Kolkata>

Problems encountered:

1. Inconsistent abbreviation of street names. For example, “Hana Para Rd” (Here, “Rd” stands for road)
2. There are about 1005 contributions (node and way tags) created by a particular user named “FuckOSMFandODbL” As the name suggests, and as reconfirmed by manual inspection, the contributions made by the user are incorrect and spurious.

Fake user:

> db.osm.remove({"created.user":"FuckOSMFandODbL"})

WriteResult({ "nRemoved" : 1005 })

There are only 30 zip codes in the dataset, all of which are valid. Used zipcode\_checker.py.

The size of the uncompressed “kolkata\_india.osm” file is 651.6 MB and the “kolkata\_india.osm.json” file is 767 MB.

**Overview of the data (after removal of contributions by one user):**

Number of documents:

> db.osm.find().count()

3581294

Number of nodes:

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

2960601

Number of ways:

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

620692

Number of unique users (after removal of one user as stated earlier):

> db.osm.distinct("created.uid").length

346

Number of cafes:

> db.osm.find({"amenity":"cafe"}).count()

15

Number of universities:

> db.osm.find({"amenity":"university"}).count()

29