Map Area :Vancouver, Canada

Mapzen Link : https://s3.amazonaws.com/metro-extracts.mapzen.com/vancouver_canada.osm.bz2

Openstreet Link : https://www.openstreetmap.org/relation/1852574

Introduction

I firstly want to choose a city from India, but the data I found from India is too much small or too much heavy. So, finally I decided to choose Vancouver, Canada.

Vancouver is one of the major city in Canada with a population of 603,000. City is ethnically and linguistically diverse, with 52% do not speak English as first language. And almost 30% inhabitants are Chinese.

Problem Encountered

Small and Capital Letter Problem

In some places the whole street name is in Capital Letters and in some small.

```
changes_required = { 'WEST BROADWAY': 'West Broadway'}
```

```
changes_required = { 'east keith Road': 'East Keith Road',}
```

Spelling Mistake Problem

The problem with the data there were lots of problem with usage of abbreviation Road, like in many cases it was only Rd, in many there was spelling mistake like RAOD.

Street Abbreviations

There are lots of abbreviations are used in the dataset like St.,St, Street, Av, Dr etc. I tried to standardized these abbreviations, some of them is given below

```
{       'Ave': 'Avenue',
        'Ave.': 'Avenue',
        'Blvd': 'Boulevard',
        'Dr': 'Drive',
        'Dr.': 'Drive',
        'Hwy': 'Highway',
        'Hwy.': 'Highway',
        'RD': 'Road',
        'Rd': 'Road',
        'Rd.': 'Road',
        'S.': 'South',
        'St': 'Street',
        'St.': 'Street',
        'Street3': 'Street',
        'av': 'Avenue',
        'road': 'Road',
        'st': 'Street',
        'street': 'Street'}
```

Using id in place of Object Id

I checked the data and found there is a id existed in data, which is unique. So, I decided to use this id in place of Object Id, and used the id as _id.

Field Names With Colon (:)

The lower_colon regex match with most of the problems, but it failed in a few fields which had uppercase as well. For dealing this situation, I just removed the first part of the field from the colon.

DATA OVERVIEW

DATA SIZE

Size of XML Data (agra_india.osm) : 151 MB

Size of JSON DATA(agra_india.osm.json) : 162.3 MB

MongoDB Queries

No. of Documents

```
count = db.vancouver.find().count()
```

```
783789
```

No. of Nodes

```
nodes = db.vancouver.find({"type" : "node"}).count()
```

```
678705
```

No. of Ways

```
way = db.vancouver.find({"type" : "way"}).count()
```

```
105084
```

No. of Unique Contributors

```
unique_users = len(db.vancouver.distinct("created_by"))
```

```
10
```

No. of Unique Sources

```
unique_sources = len(db.vancouver.distinct("source"))
```

97

Top Contributor

```
top_contributor = db.vancouver.aggregate([{"$match" : {"created_by" : {"$exists" : 1}}},
                                          {"$group" : {"_id" : "$created_by", "count" : {"$sum" : 1}}},
                                          {"$sort" : {"count" : -1}},{"$limit" : 1}])
```

```
{u'_id': u'JOSM', u'count': 529}
```

Top Sources of data

```
top_sources = db.vancouver.aggregate([{"$match" : {"created_by" : {"$exists" : 1}}},
                                      {"$group" : {"_id" : "$created_by", "count" : {"$sum" : 1}}},
                                      {"$sort" : {"count" : -1}}])
```

```
{u'_id': u'JOSM', u'count': 529}
{u'_id': u'JOSM', u'count': 529}
{u'_id': u'Potlatch 0.10f', u'count': 172}
{u'_id': u'Potlatch 0.8a', u'count': 130}
{u'_id': u'Potlatch 0.7b', u'count': 11}
{u'_id': u'Potlatch 0.8b', u'count': 6}
{u'_id': u'Potlatch 0.10b', u'count': 5}
{u'_id': u'Potlatch 0.9a', u'count': 4}
{u'_id': u'Potlatch 0.5b', u'count': 3}
{u'_id': u'Potlatch 0.9', u'count': 3}
{u'_id': u'Potlatch 0.9c', u'count': 1}
```

Amenity Types

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

```
{u'_id': u'parking', u'count': 856}
{u'_id': u'bench', u'count': 502}
{u'_id': u'restaurant', u'count': 501}
{u'_id': u'cafe', u'count': 285}
{u'_id': u'fast_food', u'count': 207}
{u'_id': u'bicycle_parking', u'count': 180}
{u'_id': u'post_box', u'count': 170}
{u'_id': u'bank', u'count': 141}
{u'_id': u'school', u'count': 118}
{u'_id': u'toilets', u'count': 93}
```

Top 10 Restaurants

```
top_restaurant = db.vancouver.aggregate([{"$match" : {"amenity" : "restaurant"}},
                                         {"$group" : {"_id" : "$name", "count" : {"$sum" : 1}}},
                                         {"$sort" : {"count" : -1}},{"$limit" : 10}])
```

```
{u'_id': u'White Spot', u'count': 8}
{u'_id': u'Cactus Club Cafe', u'count': 4}
{u'_id': u'Boston Pizza', u'count': 4}
{u'_id': u'Rogue Kitchen & Wetbar', u'count': 3}
{u'_id': u"Denny's", u'count': 3}
{u'_id': u'Subway', u'count': 3}
{u'_id': u'Earls', u'count': 2}
{u'_id': u"Joe's Grill", u'count': 2}
{u'_id': u'De Dutch', u'count': 2}
{u'_id': u"Earl's", u'count': 2}
```

OTHER IDEAS ABOUT THE DATASET

The number of documents in dataset is healthy, but most of the data is only filled with id, type and location info, and not with any other attribute. This can may give problems in developing sophisticated prediction models.

NODES

The node type exist in 86.59% data which look like this.

{u'_id': u'254482874',

u'changeset': u'12448654',

u'id': u'254482874',

u'pos': [49.3146867, -123.065267],

u'timestamp': u'2012-07-23T11:51:31Z',

u'type': u'node',

u'uid': u'32360',

u'user': u'pdunn',

u'version': u'5'}

WAY

The way type exist in 105084 documents, that means 13.41 % of total documents.

I ran this query to check how many documents are with address.street

street = len(db.vancouver.distinct("address.street"))

And I found only 387 streets are described in this dataset, which is a small number for a major city like Vancouver.

CONCLUSION

The data provide clear picture of the Vancouver, but that will be naïve to draw any general inference and conclusion from this dataset. The dataset is definitely large but do not provide any insight about the city Vancouver.

As the data is 86.5% made of node type, the interesting thing can be done if way type is removed from the dataset (if not required) then it drastically reduce the side of the data. The other thing need to note, the data is populated from one source that OpenStreetMaps, while the data is from many sources, but still data is pretty outdated and may be possible some entries are wrong. It will be a interesting to see, if the some data is pulled from the Google API to crosscheck the current data, by nodes latitude and longitude.

This is also possible to expand this dataset by adding the user reviews about like the good or bad areas, housing price data in a certain society, hospital and school reviews. This can help users in choosing a good neighborhood.