OpenStreetMap Sample Project

Data Wrangling with MongoDB

Talita Barcelos

Area Mapeada: São Paulo, SP, Brasil

1. Problemas encontrados no mapeamento

- Abreviações de Nomes de Ruas

- Foram identificados que algumas abreviações de nomes de ruas não estavam padronizadas como por exemplo Av. Agenor. Todas essas abreviações foram padronizadas, seguindo o exemplo citado anteriormente a padronização alterou o nome da avenida para Avenida Agenor. Alguns endereço foram encontrados sem especificação do tipo de endereço. Devido a isso algumas alterações para esse tipo de correção podem ter forçado algum endereço de forma errada.

- Devido aos endereços com caracteres alpha numéricos como o "ç" foi necessário fazer um encode para o Utf-8 durante as conversões de endereço.

- Alguns CEP´s também não estavam padronizados, existiam CEP´s com o "." no lugar do "-", alguns estavam sem o traço e outros possuíam o "." e o traço. Todos eles foram alterados para terem 5 digitos + "-" + 3 digitos

- Ao tentar padronizar os CEP foram identificados alguns CEP´s não pertencentes a cidade de São Paulo. Ao se listar todas as cidades que estavam vindo do arquivo do metro extract (para identificar se existia algum erro) foi identificado que as cidades da região metropolitana de São Paulo pertenciam a base e isso justificou os CEP`s.

- Não foi possível executar toda a base de dados de São Paulo devido a performance, o arquivo original possui 701.625KB, dessa forma foi gerado um arquivo de amostras com 354.739KB.

2. Visão Geral dos dados

Tamanho dos arquivos

sao-paulo_brazil.osm ..... 354 MB

sao-paulo_brazil.osm.json .... 388 MB

#Numero de Documentos

```
db.OPenStreet.find().count()
```

Existem 1815337 documentos na coleção

#Numero de Nodes

```
db.OPenStreet.find({"type":"node"}).count()
```

Existem 1595298 nodes

#Numero de usuários únicos

```
db.OPenStreet.distinct("created.user").length
```

Existem 1702 usuários que contribuem

#Usuário que mais contribuiu

```
db.OPenStreet.aggregate(
                    [   {"$group":{"_id":"$created.user", "contagem":{"$sum":1}}},
                       {"$sort":{"contagem":-1}},
                       {"$limit":10}   ]   )
```

```
_id = Bonix-Mapper
count = 771473,0
```

O usuário que mais contribuiu foi o Bonix-Mapper

 # Número de usuários com apenas 1 post

```
db.OPenStreet.aggregate(
                    [ {"$group":{"_id":"$created.user", "contagem":{"$sum":1}}},
                     {"$match":{"contagem":{"$eq":1}}},
                     {"$group":{"_id":"contagem", "num_usuario":{"$sum":1}}},
                     {"$sort":{"contagem":-1}},
                     {"$limit":1}   ]  )
```

#Número de Usuários com apenas 1 post: 359

Existem 668 usuários que fizeram apenas uma contribuição.

3. Additional Ideas

#Número registros com rua

Existem 9098 registros que possuem um nome de rua.

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

#Número registros com rua e sem CEP

Desses registros 3333 (36,64%) não possuem CEP

```
db.OPenStreet.aggregate(
        [   {"$match":{"address.street":{"$exists":1},"address.postcode":{"$exists":0} }},
        {"$group":{"_id":"$address",
                "contagem":{"$sum":1}}},
        {"$group":{"_id":"contagem", "num_ruas":{"$sum":1}}},
        {"$sort": {"count": -1}},
        {"$limit":1}    ]   )
```

#Número registros com rua e sem Bairro

E 7851 (77,49%) não possuem Bairro.

```
db.OPenStreet.aggregate(
        [ {"$match":{"address.street":{"$exists":1},"address.suburb":{"$exists":0} }},
        {"$group":{"_id":"$address",
                "contagem":{"$sum":1}}},
```

{"$group":{"_id":"contagem", "num_ruas":{"$sum":1}}},

{"$sort": {"count": -1}},

{"$limit":1}])

#Número registros com rua e sem cidade

No entanto 1422 (15%) não possuem a cidade

```
db.OPenStreet.aggregate(
            [  {"$match":{"address.street":{"$exists":1},"address.city":{"$exists":0} }},
            {"$group":{"_id":"$address",
                    "contagem":{"$sum":1}}},
             {"$group":{"_id":"contagem", "num_ruas":{"$sum":1}}},
             {"$sort": {"contagem": -1}},
             {"$limit":1}  ] )
```

#quantidade por tipo de rua

O tipo de endereço que mais existe é rua com 5993 registros, seguido de Avenida que possui 3134

```
    db.OPenStreet.aggregate(
                [  {"$group":{"_id":"$address.streettype", "contagem":{"$sum":1}}},
                  {"$sort":{"count":-1}},
                 {"$limit":10}  ] )
```

Isso representa 94,6% de todos os tipos de ruas existentes:

```
  db.OPenStreet.find({"address.streettype":{"$exists":1}}).count()
```

#verificando % de contribuições dos usuários

Dos 1702 usuários que contribuem, 39,2% contribuiram apenas uma vez.

O usuário que mais contribuiu (Bonix-Mapper) representa 42,5% do total de contribuições.

```
   db.OPenStreet.find({"created.user":{"$exists":1}}).count()
```

 # Top 10 amenidades

A amenidade que mais aparece é Parking com 1469 seguida de posto de abastecimento com 867. Existem 568 escolas diferentes, 467 bancos e 156 hospitais

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

#Top 10 emergencias

Existem poucos registros de postos de emergencia, por exemplo estão registrados apenas 6 postos de ambulancia. É muito pouco quando comparado ao tamanho da cidade.

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

#Top 10 shops

O maior tipo de comércio registrado é supermercado, seguido de lojas de roupas com 277. No entanto também são números muito baixos para o número da cidade.

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

 # Top 10 cozinhas

O tipo de cozinha com mais registros é a cozinha regional com 108 seguido de hamburquerias com 74.
O tipo de cozinha com menor número de restaurantes é a arábica com 8.

```
db.OPenStreet.aggregate(
            [  {"$match":{"cuisine":{"$exists":1}}},
            {"$group":{"_id":"$cuisine",
                    "contagem":{"$sum":1}}},
             {"$sort": {"contagem": -1}},
             {"$limit":10}    ]
```

Conclusion

Após as análises dos dados é possível verificar que várias informações como cozinha, emergência, comércio etc. estão faltando. Também existem vários endereços sem CEP, Bairro ou até mesmo cidade cadastrado.

Poucos usuários contribuem muito para o projeto, foi possível visualizar que apenas 1 usuário representa quase 50% dos cadastros, mas são poucos usuários que contribuíram com apenas um registro quando comparado ao total de usuários que contribuem.

Também a impressionante a quantidade de informação que existe quando comparado aos usuários, é uma taxa de 1066 contribuições por usuário.

Uma sugestão para melhoria desses dados seria a integração do open street maps com aplicativos como waze e foursquare. Esses aplicativos possuem uma vasta gama de informações de lugares e endereços e poderia contribuir para melhoria e complementação das informações já existentes. A implementação dessa melhoria seria um desafio principalmente em informações que existem nos dois lugares, seria difícil dizer qual informação manter. Outro problema seriam as padronizações diferentes em cada um desses aplicativos já que cada um deles tem uma forma de armazenar essas informações e isso poderia causar outro problema em formato de dados.