Limpando Dados do OpenStreetMap com mongoDB

Márcio Ozório de Jesus

Introdução

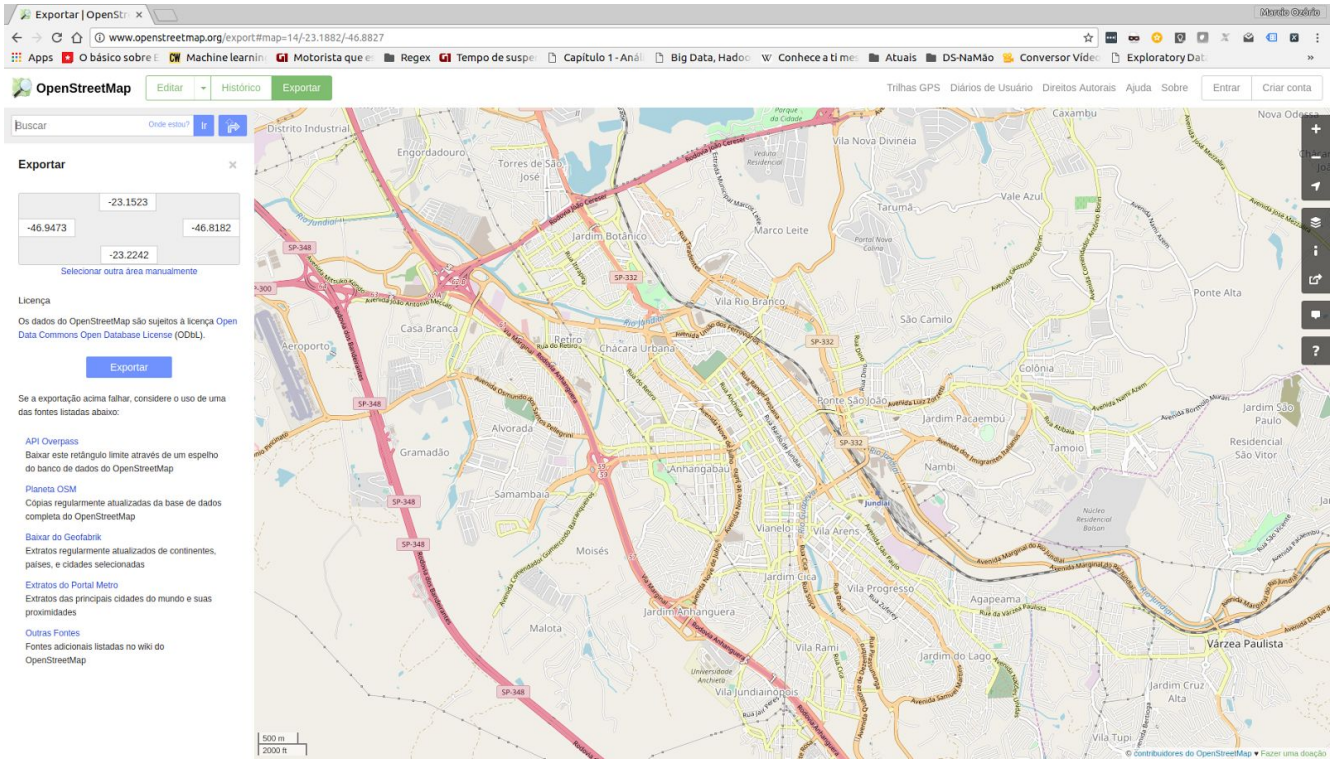
O OpenStreetMap (OSM) é um projeto de mapeamento colaborativo para criar um mapa livre e editável do mundo. Traduzindo para português o nome significa Mapa Aberto de Ruas.
Diversas pessoas do mundo adicionam novas mapas geográficos ou realizam correções e melhoria das informações já existente. O formato padrão para envio e recebimento de informações é o XML. O site oficial do OpenStreetMap é o http://www.openstreetmap.org. Para saber mais informações sobre o conteúdo, padrões, etc, acesse o site http://wiki.openstreetmap.org/. O endereço http://wiki.openstreetmap.org/wiki/Pt:Main_Page tem parte do conteudo traduzido, guia para iniciantes e informações sobre a comunidade do OpenStreetMap no Brasil.

Área do mapa: Cidade de Jundiaí e Região, São Paulo, Brasil.

Escolhi esta localização, pois, nasci e moro atualmente em Jundiaí e achei interessante saber como estão os dados no OpenStreetMap, além da maior facilidade em analisar uma vez que já conheço detalhes da região. Caso existam muitos problemas com os dados, posso até contribuir com a melhoria.

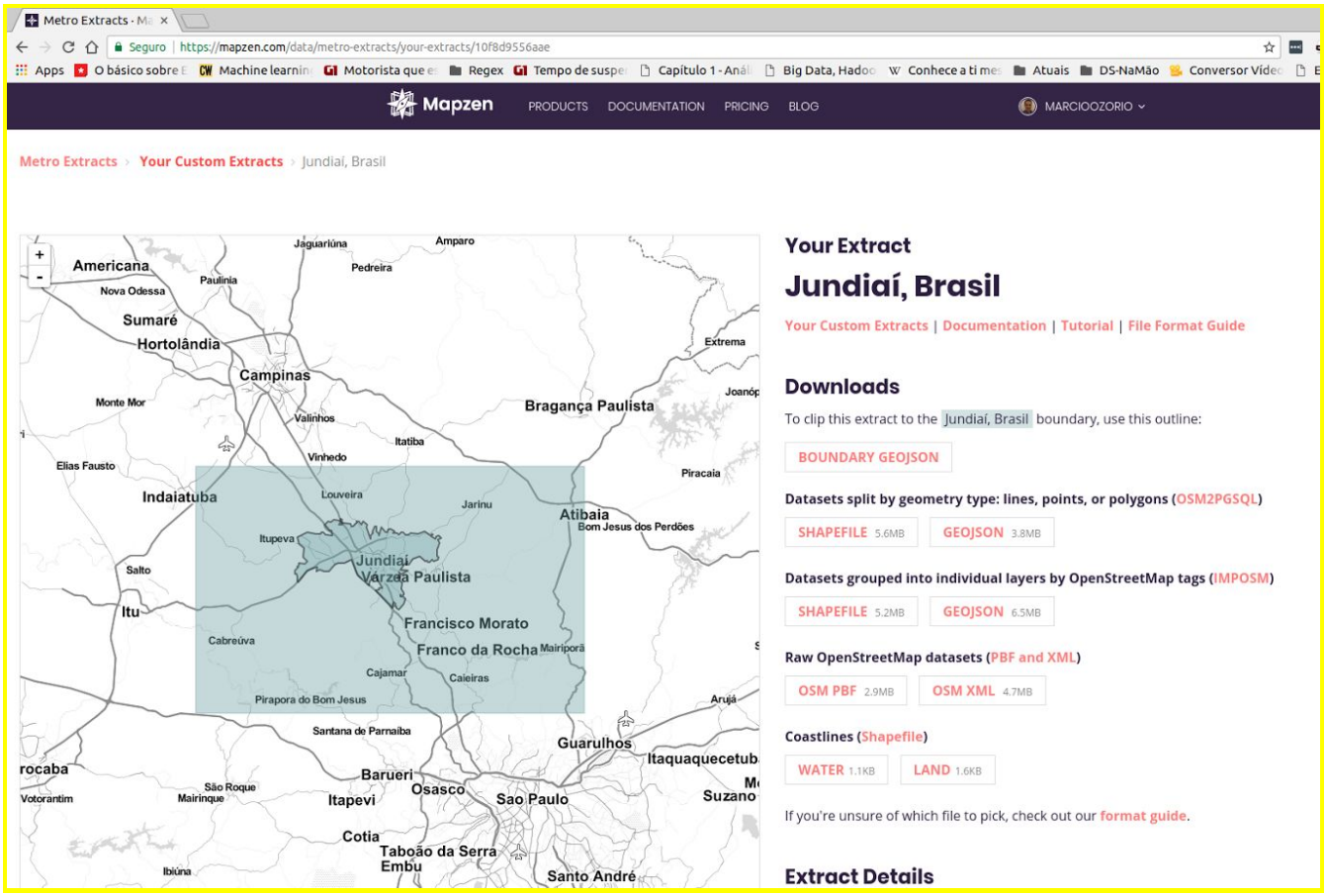
Para conhecimento da base de dados, baixei uma área reduzida da Cidade de Jundiaí diretamente do OpenStreetMap:

http://www.openstreetmap.org/export#map=14/-23.1882/-46.8827



Mapa completo extraído do MapZen utilizando um recorte considerando Jundiaí e região:

https://mapzen.com/data/metro-extracts/your-extracts/10f8d9556aae



1. Problemas encontrados no Mapa
    a. Nomes de cidades incorretos
    b. Telefones fora do formato padrão
    c. Códigos postais (CEP) fora do formato padrão
    d. Tipos de ruas fora do padrão e com caracteres especiais
2. Limpando os dados através de programação
3. Preparação do arquivo e importação no mongoDB
4. Visão geral dos dados (Data Overview)
5. Ideias adicionais
6. Conclusão

1. Problemas encontrados no Mapa

a. Nomes de cidades incorretos

Segue as cidades incorretas destacadas abaixo juntamente com as quantidades de ocorrências, identificadas pelo programa: 1.a. - audit_cidades.ipynb

```
Araçariguama: 3
Atibaia: 167
Botujuru: 1
Caieiras: 4
Cajamar: 13
Campo Limpo Paulista: 11
Campo0 Limpo Paulista: 1
Francisco Morato: 9
Franco da Rocha: 86
Indaiatuba: 1
Itupeva: 13
Jarinu: 1
Jundiai: 11
JundiaÌ: 2
Jundiaí: 350
JUndiaí: 1
Jundiái: 1
Louveira: 1
Mairipora: 1
Mairiporã: 382
Mariporã: 2
Pirapora do Bom Jesus: 2
São Bernardo do Campo: 2
São Paulo: 11
Vinhedo: 4
Várzea Paulista: 18
Várzea Paulista, SP: 1
```

b. Telefones fora do formato padrão

Na documentação do OpenStreetMap (que pode ser acessada clicando aqui), podemos encontrar o formato padrão para telefones.

Segue o resultado dos telefones inconsistentes identificados através do programa: 1.b. - audit_telefones.ipynb

```
node tag {'k': 'phone', 'v': '11 4038-2655'}
node tag {'k': 'phone', 'v': '+55 11 45272373'}
node tag {'k': 'phone', 'v': '(11) 4522-3149'}
node tag {'k': 'phone', 'v': ': 11 3917-0751'}
way tag {'k': 'phone', 'v': '+55 11 4531 0082;+55 11 4531 0083'}
way tag {'k': 'phone', 'v': '+55 11 44466767'}
way tag {'k': 'phone', 'v': '+55 11 4526-1246;+55 11 4588-9446'}
way tag {'k': 'phone', 'v': '+55 11 45841402'}
way tag {'k': 'phone', 'v': '+ 55 11 4525 5000'}
way tag {'k': 'phone', 'v': '(11) 4496-2754'}
```

Os casos com mais de um telefone na mesma linha, serão verificados e se necessários corrigidos no item **2. Limpando os dados através de programação**

c. Códigos postais (CEP) fora do formato padrão

Conforme a documentação do projeto (acesse aqui) o formato do código postal depende do local do mapa. Como estamos no Brasil, utilizei como referência a definição realizada Correios, que pode ser consultada clicando aqui.

Utilizando o programa 1.c - audit_codigo_postal.ipynb, foram identificadas as inconsistências abaixo, juntamente com as quantidades:

```
07600000: 129
12940700: 1
12942540: 1
12942655: 10
12943000: 2
12943310: 7
12943320: 13
12943330: 31
12943340: 38
12943350: 16
12943370: 13
12943380: 3
12943500: 12
12947452: 8
13203280: 1
13221550: 1
13240000: 6
3221-390: 1
```

d. Tipos de ruas fora do padrão e com caracteres especiais

Os nomes de rua abaixo foram identificados como fora do padrão. No entanto, não significam que todos estão errados. Na fase de limpeza eles serão verificados, e corrigidos se necessário. O programa utilizado para a auditoria foi o 1.d. - audit_nome_das_ruas.ipynb.

```
{'k': 'addr:street', 'v': 'estrada Tahira Eki'}
{'k': 'name', 'v': 'Rua " 30 "'}
{'k': 'name', 'v': 'Rua " 26 "'}
{'k': 'name', 'v': 'Rua "1"'}
{'k': 'name', 'v': 'Rua "2"'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Rua "04"'}
{'k': 'name', 'v': 'Avenida dos Alpes; Avenida Pauliceia'}
{'k': 'name', 'v': 'Avenidas das Aves Marinhas'}
{'k': 'name', 'v': 'Rau Diogo Alveres correia'}
{'k': 'name', 'v': 'rua Raul Breesane Malta'}
{'k': 'name', 'v': 'Miguel De Barros'}
{'k': 'name', 'v': 'José Pereira da Silva'}
{'k': 'name', 'v': 'Rod. Manoel Silvério Pinto'}
{'k': 'name', 'v': 'Rua Vitoria,'}
{'k': 'name', 'v': 'Rua "R-5"'}
```

```
{'k': 'name', 'v': 'José Vitórino Ferreira Filho'}
{'k': 'name', 'v': 'Rua Manoel Pinto Rodrigues,'}
{'k': 'name', 'v': 'Rua " 16 "'}
{'k': 'name', 'v': 'Pedro Barsanelli'}
{'k': 'name', 'v': 'Antiga Estrada da Mata Fria'}
{'k': 'name', 'v': 'Ponte Velha'}
{'k': 'name', 'v': 'Doutor Osvaldo Urioste'}
{'k': 'name', 'v': 'Ru Fernão Dias'}
{'k': 'name', 'v': 'Complexo Viário dos Emancipadores'}
{'k': 'name', 'v': 'Jean Anastace Kovelis'}
{'k': 'name', 'v': 'Rua Atílio Simoneti;Rua José Lopes'}
{'k': 'name', 'v': 'estrada da bucolica'}
{'k': 'name', 'v': 'Siqueira de Moraes'}
{'k': 'name', 'v': 'Professora Clélia de Barros Leite'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Av.enida João Casarotto'}
{'k': 'name', 'v': 'Parque da Uva'}
{'k': 'name', 'v': 'Professor Antônio Garrido'}
{'k': 'name', 'v': 'R. Jurandir Rodrigues de Castro'}
{'k': 'name', 'v': 'Avenida Eliza Bárbaro Carraro,'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Rua "4"'}
{'k': 'name', 'v': 'Rua "4"'}
{'k': 'name', 'v': 'Rua "5"'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Complexo Viário Osmar Antonio Müller'}
{'k': 'name', 'v': 'Complexo Viário Benedito Pedro Fagundes'}
{'k': 'name', 'v': 'Complexo Viário Benedito Pedro Fagundes'}
{'k': 'name', 'v': 'Complexo Viário Osmar Antonio Müller'}
{'k': 'name', 'v': 'Complexo Viário Benedito Pedro Fagundes'}
{'k': 'name', 'v': 'Complexo Viário Osmar Antonio Müller'}
{'k': 'name', 'v': 'Complexo Viário Osmar Antonio Müller'}
{'k': 'name', 'v': 'Complexo Viário Osmar Antonio Müller'}
{'k': 'name', 'v': 'Complexo Viário Tobias Muzaiel'}
{'k': 'name', 'v': 'Complexo Viário Benedito Pedro Fagundes'}
{'k': 'name', 'v': 'Complexo Viário Osmar Antonio Müller'}
{'k': 'name', 'v': 'Complexo Viário Osmar Antonio Müller'}
{'k': 'name', 'v': 'Complexo Viário Benedito Pedro Fagundes'}
{'k': 'name', 'v': 'Complexo Viário Benedito Pedro Fagundes'}
{'k': 'name', 'v': 'Rua Kromberg & Schubert'}
{'k': 'name', 'v': 'Complexo Viário Benedito Pedro Fagundes'}
{'k': 'name', 'v': 'Complexo Viário Benedito Pedro Fagundes'}
{'k': 'name', 'v': 'Complexo Viário Osmar Antonio Müller'}
{'k': 'name', 'v': 'Complexo Viário Osmar Antonio Müller'}
{'k': 'name', 'v': 'Complexo Viário Benedito Pedro Fagundes'}
{'k': 'name', 'v': 'Complexo Viário Benedito Pedro Fagundes'}
```

```
{'k': 'name', 'v': 'Complexo Viário Benedito Pedro Fagundes'}
{'k': 'name', 'v': 'CDP de Jundiaí'}
{'k': 'name', 'v': 'CCR - AUTOBAN'}
{'k': 'name', 'v': 'Uva Paulistinha'}
{'k': 'name', 'v': 'Uva Patricia'}
{'k': 'name', 'v': 'Rua " 12 "'}
{'k': 'name', 'v': 'Complexo Viário Benedito Pedro Fagundes'}
{'k': 'name', 'v': 'Rua " D "'}
{'k': 'name', 'v': 'Rua " 19 "'}
{'k': 'name', 'v': 'Rua " 20 "'}
{'k': 'name', 'v': 'Rua " 15 "'}
{'k': 'name', 'v': 'Complexo Viário Osmar Antonio Müller'}
{'k': 'name', 'v': 'Rua " 16 "'}
{'k': 'name', 'v': 'Professora Clélia de Barros Leite'}
{'k': 'name', 'v': 'Professora Clélia de Barros Leite'}
{'k': 'name', 'v': 'Professora Clélia de Barros Leite'}
{'k': 'name', 'v': 'Rua " 02 "'}
{'k': 'name', 'v': 'Rua " 02 "'}
{'k': 'name', 'v': 'Rua "05"'}
{'k': 'name', 'v': 'Rua "05"'}
{'k': 'name', 'v': 'Rua "04"'}
{'k': 'name', 'v': 'Sol Maior'}
{'k': 'name', 'v': 'rua Pedro alvares Cabral'}
{'k': 'name', 'v': 'rua Deputado Emilio Carlos'}
{'k': 'name', 'v': 'rua joão Rais'}
{'k': 'name', 'v': 'Bromelias'}
{'k': 'name', 'v': 'Autódromo Fazenda Capuava'}
{'k': 'name', 'v': 'Complexo Viário Benedito Pedro Fagundes'}
{'k': 'name', 'v': 'Saida Terminal'}
{'k': 'name', 'v': 'Rod.Manoel Silverio Pinto'}
{'k': 'name', 'v': 'Rod.Manoel Silverio Pinto'}
{'k': 'name', 'v': 'R. Joana Forest Storani'}
{'k': 'name', 'v': 'Rua " 11 "'}
{'k': 'name', 'v': 'Bento Flores'}
{'k': 'name', 'v': 'Doutor Osvaldo Urioste'}
{'k': 'name', 'v': 'Nazareno Rossi'}
{'k': 'name', 'v': 'AvenidaAssembléia de Deus Ministério de Belém'}
{'k': 'name', 'v': 'Av Jeronimo de Camargo'}
{'k': 'name', 'v': 'Kartódromo de Atibaia'}
{'k': 'name', 'v': 'RuaJuscelino Kubitschek de Oliveira'}
```

2. Limpando os dados

Para a limpeza dos dados, foi criado um programa para correção de todos os problemas identificados na auditoria, seja Cidade, CEP, Telefone ou Rua.

Após a execução do programa 2. - limpando os dados.ipynb, é exibido um log com todas as correções realizadas. Segue o log da execução realizada:

```
=======> Inicio cidade......:  2017-12-28 18:57:22.964610

Mairipora  =>  Mairiporã
Jundiai  =>  Jundiaí
Jundiai  =>  Jundiaí
Jundiai  =>  Jundiaí
```

```
Jundiai  =>  Jundiaí
Várzea Paulista, SP  =>  Várzea Paulista
Jundiai  =>  Jundiaí
JUndiaí  =>  Jundiaí
JundiaÌ  =>  Jundiaí
JundiaÌ  =>  Jundiaí
Jundiái  =>  Jundiaí
Jundiai  =>  Jundiaí
Jundiai  =>  Jundiaí
Jundiai  =>  Jundiaí
Jundiai  =>  Jundiaí
Campo0 Limpo Paulista  =>  Campo Limpo Paulista
Mariporã  =>  Mairiporã
Mariporã  =>  Mairiporã
Jundiai  =>  Jundiaí
Jundiai  =>  Jundiaí


=======> Inicio telefone...:  2017-12-28 18:57:36.453419

11 4038-2655  =>  +55 11 4038-2655
+55 11 45272373  =>  +55 11 4527 2373
(11) 4522-3149  =>  +55 11 4522-3149
: 11 3917-0751  =>  +55 11 3917-0751
+55 11 4531 0082;+55 11 4531 0083  =>  +55 11 4531 0082;+55 11 4531 0083
+55 11 44466767  =>  +55 11 4446 6767
+55 11 4526-1246;+55 11 4588-9446  =>  +55 11 4526-1246;+55 11 4588-9446
+55 11 45841402  =>  +55 11 4584 1402
+ 55 11 4525 5000  =>  +55 11 4525 5000
(11) 4496-2754  =>  +55 11 4496-2754


=======> Inicio cód.postal.:  2017-12-28 18:57:49.066478
```

(Para facilitar a visualização, as linhas duplicadas foram removidas.
Total linhas original: 289)

```
07600000  =>  07600-000
13240000  =>  13240-000
3221-390  =>  13221-390
13221550  =>  13221-550
13203280  =>  13203-280
12943310  =>  12943-310
12943370  =>  12943-370
12943340  =>  12943-340
12940700  =>  12940-700
12943350  =>  12943-350
12943500  =>  12943-500
12947452  =>  12947-452
12943330  =>  12943-330
12942655  =>  12942-655
12943000  =>  12943-000
12942540  =>  12942-540
12943320  =>  12943-320
12943380  =>  12943-380
```

=======> Inicio ruas.......: 2017-12-28 18:58:03.434248

estrada Tahira Eki => Estrada Tahira Eki
Rua " 30 " => Rua 30
Rua " 26 " => Rua 26
Rua "1" => Rua 1
Rua "2" => Rua 2
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel
Rua "04" => Rua 04

Avenida dos Alpes; Avenida Pauliceia => Avenida dos Alpes; Avenida Pauliceia
Avenidas das Aves Marinhas => Avenida das Aves Marinhas
Rau Diogo Alveres correia => Rau Diogo Alveres correia
rua Raul Breesane Malta => Rua Raul Breesane Malta
Miguel De Barros => Miguel De Barros
José Pereira da Silva => José Pereira da Silva
Rod. Manoel Silvério Pinto => Rodovia Manoel Silvério Pinto
Rua Vitoria, => Rua Vitoria
Rua "R-5" => Rua R-5
José Vitórino Ferreira Filho => José Vitórino Ferreira Filho
Rua Manoel Pinto Rodrigues, => Rua Manoel Pinto Rodrigues
Rua " 16 " => Rua 16
Pedro Barsanelli => Pedro Barsanelli
Antiga Estrada da Mata Fria => Antiga Estrada da Mata Fria
Ponte Velha => Ponte Velha
Doutor Osvaldo Urioste => Doutor Osvaldo Urioste
Complexo Viário dos Emancipadores => Complexo Viário dos Emancipadores
Jean Anastace Kovelis => Jean Anastace Kovelis
Rua Atílio Simoneti;Rua José Lopes => Rua Atílio Simoneti;Rua José Lopes
estrada da bucolica => Estrada da bucolica
Siqueira de Moraes => Siqueira de Moraes
Professora Clélia de Barros Leite => Professora Clélia de Barros Leite
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel
Av.enida João Casarotto => Avenida João Casarotto
Av.enida João Casarotto => Avenida João Casarotto
Parque da Uva => Parque da Uva
Professor Antônio Garrido => Professor Antônio Garrido
R. Jurandir Rodrigues de Castro => Rua Jurandir Rodrigues de Castro
Avenida Eliza Bárbaro Carraro, => Avenida Eliza Bárbaro Carraro
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel
Rua "4" => Rua 4
Rua "4" => Rua 4
Rua "5" => Rua 5
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel
Complexo Viário Tobias Muzaiel => Complexo Viário Tobias Muzaiel

```
Complexo Viário Osmar Antonio Müller  =>  Complexo Viário Osmar Antonio Müller
Complexo Viário Benedito Pedro Fagundes  =>  Complexo Viário Benedito Pedro Fagundes
Complexo Viário Benedito Pedro Fagundes  =>  Complexo Viário Benedito Pedro Fagundes
Complexo Viário Osmar Antonio Müller  =>  Complexo Viário Osmar Antonio Müller
Complexo Viário Benedito Pedro Fagundes  =>  Complexo Viário Benedito Pedro Fagundes
Complexo Viário Osmar Antonio Müller  =>  Complexo Viário Osmar Antonio Müller
Complexo Viário Osmar Antonio Müller  =>  Complexo Viário Osmar Antonio Müller
Complexo Viário Osmar Antonio Müller  =>  Complexo Viário Osmar Antonio Müller
Complexo Viário Tobias Muzaiel  =>  Complexo Viário Tobias Muzaiel
Complexo Viário Benedito Pedro Fagundes  =>  Complexo Viário Benedito Pedro Fagundes
Complexo Viário Osmar Antonio Müller  =>  Complexo Viário Osmar Antonio Müller
Complexo Viário Osmar Antonio Müller  =>  Complexo Viário Osmar Antonio Müller
Complexo Viário Benedito Pedro Fagundes  =>  Complexo Viário Benedito Pedro Fagundes
Complexo Viário Benedito Pedro Fagundes  =>  Complexo Viário Benedito Pedro Fagundes
Rua Kromberg & Schubert  =>  Rua Kromberg & Schubert
Complexo Viário Benedito Pedro Fagundes  =>  Complexo Viário Benedito Pedro Fagundes
Complexo Viário Benedito Pedro Fagundes  =>  Complexo Viário Benedito Pedro Fagundes
Complexo Viário Osmar Antonio Müller  =>  Complexo Viário Osmar Antonio Müller
Complexo Viário Osmar Antonio Müller  =>  Complexo Viário Osmar Antonio Müller
Complexo Viário Benedito Pedro Fagundes  =>  Complexo Viário Benedito Pedro Fagundes
Complexo Viário Benedito Pedro Fagundes  =>  Complexo Viário Benedito Pedro Fagundes
Complexo Viário Benedito Pedro Fagundes  =>  Complexo Viário Benedito Pedro Fagundes
CDP de Jundiaí  =>  CDP de Jundiaí
CCR - AUTOBAN  =>  CCR - AUTOBAN
Uva Paulistinha  =>  Uva Paulistinha
Uva Patricia  =>  Uva Patricia
Rua " 12 "  =>  Rua 12
Complexo Viário Benedito Pedro Fagundes  =>  Complexo Viário Benedito Pedro Fagundes
Rua " D "  =>  Rua D
Rua " 19 "  =>  Rua 19
Rua " 20 "  =>  Rua 20
Rua " 15 "  =>  Rua 15
Complexo Viário Osmar Antonio Müller  =>  Complexo Viário Osmar Antonio Müller
Rua " 16 "  =>  Rua 16
Professora Clélia de Barros Leite  =>  Professora Clélia de Barros Leite
Professora Clélia de Barros Leite  =>  Professora Clélia de Barros Leite
Professora Clélia de Barros Leite  =>  Professora Clélia de Barros Leite
Rua " 02 "  =>  Rua 02
Rua "05"  =>  Rua 05
Rua "05"  =>  Rua 05
Rua "04"  =>  Rua 04
Sol Maior  =>  Sol Maior
rua Pedro alvares Cabral  =>  Rua Pedro alvares Cabral
rua Deputado Emilio Carlos  =>  Rua Deputado Emilio Carlos
rua joão Rais  =>  Rua joão Rais
Bromelias  =>  Bromelias
Autódromo Fazenda Capuava  =>  Autódromo Fazenda Capuava
Complexo Viário Benedito Pedro Fagundes  =>  Complexo Viário Benedito Pedro Fagundes
Saida Terminal  =>  Saida Terminal
Rod.Manoel Silverio Pinto  =>  Rodovia Manoel Silverio Pinto
Rod.Manoel Silverio Pinto  =>  Rodovia Manoel Silverio Pinto
R. Joana Forest Storani  =>  Rua Joana Forest Storani
Rua " 11 "  =>  Rua 11
Bento Flores  =>  Bento Flores
Doutor Osvaldo Urioste  =>  Doutor Osvaldo Urioste
Nazareno Rossi  =>  Nazareno Rossi
AvenidaAssembléia de Deus Ministério de Belém  =>  Avenida Assembléia de Deus Ministério de
```

```
Belém
Av Jeronimo de Camargo  =>  Avenida Jeronimo de Camargo
Kartódromo de Atibaia  =>  Kartódromo de Atibaia

=======> Fim..............:  2017-12-28 18:59:13.693601
```

3. Preparação do arquivo e importação no mongoDB

a. Preparando o arquivo para importação

Para realizar a importação das informações no banco de dados mongoDB, iremos utilizar o comando **mongoimport**. Mas antes é necessário transformar o formato do arquivo de XML para o formato JSON.

Além disso, é importante realizar algumas modificações, permitindo que cada nó (node ou way) receba um ID, tornando-se assim assim um documento distinto dentro do mongoDB.

Para realizar essa transformação, foi criado o programa 3.a. - XML to JSON - OpenStreetMap.ipynb.

O programa irá ler o arquivo com estrutura XML chamado *output_jundiaieregiao.osm* e criará o arquivo *output_jundiaieregiao.osm.json*.

b. Importação no mongoDB

Conforme dito anteriormente, para realizar a importação do arquivo no banco de dados mongoDB, utilizaremos o comando **mongoimport** que deve ser executado diretamente na linha ou prompt de comando. Veja a execução do comando (ambiente linux) e log de execução:

```
$ mongoimport -d examples -c jundiaieregiao --file output_jundiaieregiao.osm.json
--type=json --drop
```

```
2017-12-28T19:12:18.259-0200 connected to: localhost
2017-12-28T19:12:18.268-0200 dropping: examples.jundiaieregiao
2017-12-28T19:12:20.722-0200 [........................] examples.jundiaieregiao 1.52MB/112MB (1.4%)
2017-12-28T19:12:23.722-0200 [##......................] examples.jundiaieregiao 11.0MB/112MB (9.8%)
2017-12-28T19:12:26.722-0200 [####....................] examples.jundiaieregiao 21.9MB/112MB (19.6%)
2017-12-28T19:12:29.722-0200 [######..................] examples.jundiaieregiao 28.6MB/112MB (25.5%)
2017-12-28T19:12:32.722-0200 [#######.................] examples.jundiaieregiao 37.3MB/112MB (33.3%)
2017-12-28T19:12:35.726-0200 [##########..............] examples.jundiaieregiao 47.4MB/112MB (42.3%)
2017-12-28T19:12:38.722-0200 [############............] examples.jundiaieregiao 56.0MB/112MB (50.0%)
2017-12-28T19:12:41.722-0200 [#############...........] examples.jundiaieregiao 64.6MB/112MB (57.7%)
2017-12-28T19:12:44.722-0200 [#################.......] examples.jundiaieregiao 76.0MB/112MB (67.8%)
2017-12-28T19:12:47.722-0200 [##################......] examples.jundiaieregiao 86.7MB/112MB (77.4%)
2017-12-28T19:12:50.722-0200 [####################....] examples.jundiaieregiao 94.8MB/112MB (84.7%)
2017-12-28T19:12:53.722-0200 [#####################...] examples.jundiaieregiao 101MB/112MB (89.8%)
2017-12-28T19:12:56.722-0200 [########################.] examples.jundiaieregiao 108MB/112MB (96.4%)
2017-12-28T19:12:57.791-0200 [########################] examples.jundiaieregiao 112MB/112MB (100.0%)
2017-12-28T19:12:57.791-0200 imported 344958 documents
```

Como pode ver, foi criada a collection jundiaieregiao contendo **344.958** documentos.

4. Visão geral dos dados (Data Overview)

Detalhes dos arquivos utilizados:

Nome do arquivo	Tamanho	Comentário
jundiai-small.osm	6 MB	Arquivo inicial utilizado para estudo
jundiai_e_regiao_map_zen.osm	66 MB	Arquivo original extraido do MapZen
output_jundiaieregiao.osm	67 MB	Arquivo com realização da limpeza
output_street.osm.json	116 MB	Arquivo convertido para o formato JSON

Obs.: Os comandos abaixo foram executados diretamente no aplicativo do mongoDB. Também podem executados diretamente no Python através da API pymongo.

Número de documentos:

```
> db.jundiaieregiao.find().count()
344958
```

Número de nós (node e way):

```
> db.jundiaieregiao.find({"$or" : [{"type" : "node"},
                                   {"type" : "way"}]}).count()
344958
```

Número de usuários únicos:

```
> db.jundiaieregiao.distinct('created.user').length
328
```

Top 5 usuários que mais contribuíram:

```
> db.jundiaieregiao.aggregate([{"$group":{"_id" : "$created.user", "count" :
{"$sum" : 1}}}, {"$sort" : {"count" : -1}}, {"$limit" : 1}])

{ "_id" : "AjBelnuovo", "count" : 88960 }
{ "_id" : "natinacio", "count" : 43120 }
{ "_id" : "patodiez", "count" : 22154 }
{ "_id" : "Louis Goyard", "count" : 16178 }
{ "_id" : "Roberto Costa", "count" : 13975 }
```

Número de usuários que aparecem apenas uma vez:

```
> db.jundiaieregiao.aggregate([{"$group":{"_id" : "$created.user", "count" :
{"$sum":1}}}, {"$group":{"_id" : "$count", "num_users" : {"$sum":1}}},
{"$sort"  : {"_id":1}}, {"$limit":1}])

{ "_id" : 1, "num_users" : 62 }
```

Cidades e quantidade respectiva de documentos encontrados nesta base de dados:

```
> db.jdiregfinal.aggregate([{"$match" : {"address.city" : {"$exists" : 1}}},
{"$group":{"_id" : "$address.city", "count" : {"$sum" : 1}}}, {"$sort" :
{"count" : -1}}])

{ "_id" : "Mairiporã", "count" : 383 }
{ "_id" : "Jundiaí", "count" : 364 }
{ "_id" : "Atibaia", "count" : 167 }
{ "_id" : "Franco da Rocha", "count" : 85 }
{ "_id" : "Várzea Paulista", "count" : 19 }
{ "_id" : "Itupeva", "count" : 13 }
{ "_id" : "Cajamar", "count" : 13 }
{ "_id" : "Campo Limpo Paulista", "count" : 12 }
{ "_id" : "São Paulo", "count" : 11 }
{ "_id" : "Francisco Morato", "count" : 9 }
{ "_id" : "Caieiras", "count" : 4 }
{ "_id" : "Vinhedo", "count" : 4 }
{ "_id" : "Araçariguama", "count" : 3 }
{ "_id" : "São Bernardo do Campo", "count" : 2 }
{ "_id" : "Pirapora do Bom Jesus", "count" : 2 }
{ "_id" : "Mariporã", "count" : 2 }
{ "_id" : "Louveira", "count" : 1 }
{ "_id" : "Indaiatuba", "count" : 1 }
{ "_id" : "Jarinu", "count" : 1 }
{ "_id" : "Botujuru", "count" : 1 }
```

5. Explorações adicionais usando queries no mongoDB:

5 tipos de restaurantes mais populares:

```
> db.jundiaieregiao.aggregate([{"$match" : {"cuisine" : {"$exists" : 1},
"amenity" : "restaurant"}}, {"$group" : {"_id" : "$cuisine", "count" : {"$sum"
: 1}}}, {"$sort" : {"count" : -1}}, {"$limit" : 5}])

{ "_id" : "regional", "count" : 15 }
{ "_id" : "pizza", "count" : 6 }
{ "_id" : "italian", "count" : 6 }
{ "_id" : "japanese", "count" : 4 }
{ "_id" : "steak_house", "count" : 3 }
```

Igrejas com maior quantidade (por denominação):

```
> db.jdiregfinal.aggregate([{"$match" : {"denomination" : {"$exists" : 1}}},
{"$group":{"_id" : "$denomination", "count" : {"$sum" : 1}}}, {"$sort" :
{"count" : -1}}, {"$limit" : 10}])

{ "_id" : "roman_catholic", "count" : 38 }
{ "_id" : "catholic", "count" : 15 }
{ "_id" : "pentecostal", "count" : 11 }
{ "_id" : "neo-pentecostal", "count" : 7 }
{ "_id" : "baptist", "count" : 5 }
{ "_id" : "mormon", "count" : 2 }
{ "_id" : "presbyterian", "count" : 2 }
{ "_id" : "adventist", "count" : 1 }
{ "_id" : "sunni", "count" : 1 }
{ "_id" : "Batista", "count" : 1 }
```

Anos que tiveram a maior quantidade de contribuição:

Podemos notar que os últimos anos foram os que mais tivemos contribuições, havendo um grande pico no ano de 2015.

```
> db.jundiaieregiao.aggregate([{"$project" : {"year" : {"$substr" :
["$timestamp", 0, 4]}}}, {"$group" : {"_id" : "$year", "count" : {"$sum" :
1}}}, {"$sort" : {"count": -1}}, {"$limit" : 5}])

{ "_id" : "2015", "count" : 109871 }
{ "_id" : "2017", "count" : 68979 }
{ "_id" : "2016", "count" : 61444 }
{ "_id" : "2013", "count" : 31429 }
{ "_id" : "2012", "count" : 30101 }
```

Locais comuns nas cidades com as respectivas quantidades:

```
> db.jundiaieregiao.aggregate([{"$match" : {"$or" : [{"amenity" : "hospital"},
{"amenity" : "police"}, {"amenity" : "prison"}, {"amenity" : "bank"},
{"amenity" : "cinema"}, {"amenity" : "university"}, {"amenity" : "teatre"},
{"amenity" : "marketplace"}]}}, {"$group" : {"_id" : "$amenity", "count" :
{"$sum" : 1}}}, {"$sort" : {"count": 1}}])

{ "_id" : "cinema", "count" : 2 }
{ "_id" : "university", "count" : 3 }
{ "_id" : "marketplace", "count" : 4 }
{ "_id" : "prison", "count" : 7 }
{ "_id" : "police", "count" : 16 }
{ "_id" : "hospital", "count" : 26 }
{ "_id" : "bank", "count" : 66 }
```

6. Conclusão

Através das pesquisas realizadas, foi contatado uma grande diversidade de informações referentes ao local do mapa que podem ser adicionadas e que existe uma variação significativa dependendo de região para região. Cada país realiza adequações nas informações para a utilização da estrutura.

É preciso se familiarizar com os termos utilizados para classificação das informações, pois estão todos na língua inglesa e pode não refletir a necessidade de todos os locais do planeta.

Para facilitar a padronização das informações, existem grupos de usuários do OpenStreetMap de acordo com a localidade, que trocam informações e se comunicam afim de criarem uma espécie de boas práticas para utilizarem um formato o mais padronizado possível.

No Brasil, temos uma comunidade com o foco ainda em algumas grandes capitais. Em contato com esta comunidade, pude conversar e tiver algumas dúvidas através de um grupo no telegram.

Hoje os membros da comunidade utilizam uma ferramenta chamada JOSM. Trata-se de um editor criado em java com o foco na manutenção das informações do OpenStreetMap.

Comentei da realização deste projeto com o objetivo da limpeza das informações e acharam interessante, no entanto, se mostraram preocupados com iniciativas deste tipo, pois é algo que precisa ser muito bem discutido, planejado, verificado e aceito. Disseram que é comum alguém realizar a correção e logo depois outro usuário realiza a sobreposição com dados incorretos novamente. E se houver atualizações em massa, fica mais difícil de identificar o que está causando o problema. Neste link segue o código de conduta utilizado pela comunidade.

Apesar dos erros de cadastros encontrados, a maior parte das informações da região extraída estão com uma boa qualidade. Ainda há muito o que melhorar, como por exemplo, na ultima consulta que fizemos “Locais comuns nas cidades”, identificamos somente 4 supermercados, no entanto, através das explorações realizadas na base de dados, pudemos identificar o crescente aumento de contribuições nos últimos anos. Isto significa que os dados do OpenStreetMap estão cada dia mais completos e ricos de informações.

Referências:

Mapas
http://www.openstreetmap.org
https://mapzen.com

OpenStreetMap - Wikipedia
https://pt.wikipedia.org/wiki/OpenStreetMap

OpenStreetMap - Wiki
http://wiki.openstreetmap.org/wiki/Main_Page
Comunidade OpenStreetMap no Telegram
https://web.telegram.org/#/im?p=@OSMBrasil_Suporte

Correios - Formato do código postal (CEP)
https://www.correios.com.br/para-voce/precisa-de-ajuda/o-que-e-cep-e-por-que-usa-lo/estrutura-do-cep

Documentação Pyton
https://docs.python.org/3/

MongoDB
https://docs.mongodb.com/

Expressões Regulares
https://docs.python.org/2/howto/regex.html
https://pythex.org/
https://tableless.com.br/o-basico-sobre-expressoes-regulares/

Converter XMl para Json
https://github.com/bestkao/data-wrangling-with-openstreetmap-and-mongodb/blob/master/data-wrangling-with-openstreetmaps.ipynb

Posts sobre XML, Python, Expressões regulares e MongoDB
https://stackoverflow.com