Redefinindo linhas de bairros

Aqui na Loft usamos intensamente engenharia e ciência de dados para nossa tomada de decisão. Como nosso objetivo é revolucionar o mercado imobiliário, não nos prendemos a supostos "dogmas" existentes, e criamos o nosso entendimento de como a dinâmica de compra e venda de imóveis funciona usando os métodos mais robustos de machine learning.

Um dos conceitos que podemos pensar em redefinir é o de bairro. Bairros são, essencialmente, partições arbitrárias no espaço. Estamos acostumados a pensar que um bairro é homogêneo do ponto de vista imobiliário, mas isso não é necessariamente verdade. Por outro lado, bairros diferentes não necessariamente são entidades não-comparáveis. Acreditamos que é possível **redefinir as linhas de bairro** e construir outro particionamento no espaço que realmente agrupe apartamentos similares e seja regido pelas mesmas dinâmicas de mercado.

Desafio

Neste desafio, propomos que você **redefina** as linhas de bairros da cidade de São Paulo, utilizando dados do IPTU de 1.4 milhões de propriedades na cidade e tendo como objetivo criar novos bairros onde apartamentos são similares entre si e diferentes de apartamentos de outros bairros.. **Sugerimos o seguinte roteiro:**

1) **Proponha** uma métrica/forma de comparar propriedades ou regiões do espaço. Dica: métodos de seleção de variáveis e redução de dimensionalidade podem ajudar!
    - Seleção de variáveis no sklearn
    - Redução de dimensionalidade não-supervisionada no sklearn
    - Manifold learning no sklearn
    - Geohash para converter latlongs para regiões do espaço
    - Criando autoencoders no Keras

2) **Construa** um algoritmo que utilize esta métrica para particionar o espaço de cidade de São Paulo em "novos bairros". Os apartamentos de um mesmo bairro precisam ser mais similares quanto possível enquanto apartamentos de bairros diferentes precisam ser mais diferentes entre si quanto possível. Dica: métodos de clusterização podem ajudar!
    - Clusterização não-supervisionada no sklearn
    - "Clusterização supervisionada": árvores de decisão
    - Clusterização semi-supervisionada
    - Kepler: ferramenta de visualização de dados geográficos

3) **Crie** uma apresentação que responda às seguintes perguntas:
    - **Qual é o racional por trás da sua métrica de similaridade? Quais variáveis você utilizou?**
    - **Como seu algoritmo de "definição de bairros" funciona?**
    - **Quantos "novos bairros" você propõe? Por quê?**

Avaliação e prêmio

Envie sua apresentação e código até o dia 22/08 às 08h00 para guilherme.marmerola@loft.com.br. Sua solução será avaliada pelos nossos cientistas e engenheiros de dados. A melhor solução vai ganhar teclado mecânico da WASD modelo CODE V3 87-Key - Cherry MX Brown (link)!

Você poderá participar em grupos de até 3 pessoas.

Dados

Neste link você poderá baixar a base do IPTU de São Paulo do ano de 2018. Nesta base temos diversas variáveis, desde a área ocupada e o ano de construção até a latitude e longitude do imóvel. A descrição de cada variável pode ser encontrada aqui..

Rua Augusta, 2840. 17º andar Cerqueira César
São Paulo/SP - 01412-100