Relatório WebScraping

KAUAN TORRES

IPOG - DATA SCIENCE
BUSINESS INTELLIGENCE

PROCESSO

URL E O INÍCIO

O processo de raspagem foi realizado no site OLX, uma plataforma de compra, venda e aluguel de produtos. O objetivo específico foi obter informações sobre casas disponíveis para aluguel. Para realizar a raspagem, utilizamos o serviço ZenRows, que permite a navegação por sites que bloqueiam requisições automatizadas. A URL base utilizada para a raspagem de dados foi:

- URL Base: https://www.olx.com.br/imoveis/aluguel/casas

Cada página da OLX contém uma listagem de imóveis que são exibidos por meio de blocos de anúncios. Para acessar diferentes páginas, a URL foi ajustada dinamicamente, incrementando o valor do parâmetro de página (o={page}), o que nos permitiu paginar pelos resultados e obter mais dados.

ESTRUTURA DA PÁGINA

A estrutura da página da OLX consiste em uma série de blocos de anúncios, cada um com as principais informações sobre o imóvel, como valor, número de quartos, banheiros, vagas de garagem e área total. Durante a raspagem, os principais elementos HTML utilizados foram:

- div com a classe "sc-9d8733cc-0 UnonC", que continha a listagem de anúncios.
- Dentro de cada anúncio (section), os dados estavam distribuídos em tags específicas:
  - Local: armazenado em uma tag h3, contendo a localização do imóvel.
  - Valor: presente em uma tag p com a classe "olx-text olx-text--caption olx-text--block olx-text--regular".
  - Atributos adicionais: como quartos, vagas de garagem e banheiros, armazenados em uma lista de li dentro de uma tag ul.

Esses elementos foram identificados através da análise manual do HTML da página e selecionados usando o BeautifulSoup para extração.

Desenvolvimento da Raspagem

REQUISIÇÃO DA PÁGINA

Utilizamos a função get_page_content() para realizar a requisição do conteúdo HTML da página via a API do ZenRows, que requer o uso de uma chave de API (fornecida como 'apikey' no código). Isso permitiu contornar as proteções da OLX contra raspagem direta, obtendo o conteúdo desejado para análise.

EXTRAÇÃO DE DADOS

Os dados foram extraídos iterativamente de cada página de listagem. Utilizamos a função get_table_items() para localizar o bloco principal que continha os anúncios (div com a classe sc-9d8733cc-0 UnonC). Dentro deste bloco, percorremos cada section correspondente a um anúncio, de onde extraímos:

- Link do anúncio: a URL para o anúncio completo.
- Número de quartos, vagas de garagem e banheiros: extraídos da lista de atributos, filtrando pela palavra-chave (ex: "quartos", "garagem", "banheiros").
- Área: metros quadrados do imóvel.
- Valor: o preço do aluguel.
- Local: cidade ou bairro onde o imóvel está localizado.

Esses dados foram armazenados em um dicionário (casas) com chaves específicas para cada atributo, permitindo fácil conversão para um DataFrame do Pandas no final.

PAGINAÇÃO

Para cobrir o maior número possível de anúncios, o código foi desenvolvido para percorrer várias páginas de resultados, aumentando o valor do parâmetro page na URL. O loop de raspagem continuava até que o HTML retornado não contivesse mais os elementos esperados, sinalizando o fim das páginas disponíveis.

DIFICULDADES ENFRENTADAS

Durante o desenvolvimento, algumas dificuldades surgiram, como:

- Bloqueio de Raspagem: A OLX implementa mecanismos para bloquear requisições automatizadas, o que inicialmente causou falhas ao tentar acessar o conteúdo da página. Isso foi solucionado através da utilização de um proxy via ZenRows, permitindo que o conteúdo fosse carregado corretamente.
- Estrutura Inconsistente de HTML: Alguns anúncios não continham todas as informações (por exemplo, algumas listagens não tinham o número de quartos ou banheiros). Isso foi resolvido com checagens condicionais para evitar erros de extração quando algum dado não estivesse presente.

GERAÇÃO DO DATAFRAME

Após a coleta dos dados de todas as páginas, o dicionário casas foi convertido em um DataFrame do Pandas, que permitiu:

- A visualização estruturada dos dados.
- A remoção de duplicatas, utilizando o método drop_duplicates().
- Salvamento dos dados coletados em um arquivo Excel ou CSV, possibilitando análises futuras.

CONCLUSÃO

A raspagem de dados do OLX foi bem-sucedida após a superação de desafios técnicos relacionados à proteção contra bots e à estrutura inconsistente dos dados nos anúncios. A utilização do serviço ZenRows foi fundamental para acessar o conteúdo da página de forma contínua e eficiente. Com os dados extraídos, agora é possível realizar análises sobre os imóveis para aluguel na plataforma OLX, fornecendo insights valiosos sobre preços, características e localização.