Web Scraping

Prof. Dr. Jorge Zavaleta

Departamento de Ciências Ambientais (DCA)

Universidade Federal Rural do Rio de Janeiro (UFRRJ)

Pesquisador de Pós-doutorado (PDJ/CNPq)

De onde vêm os dados?

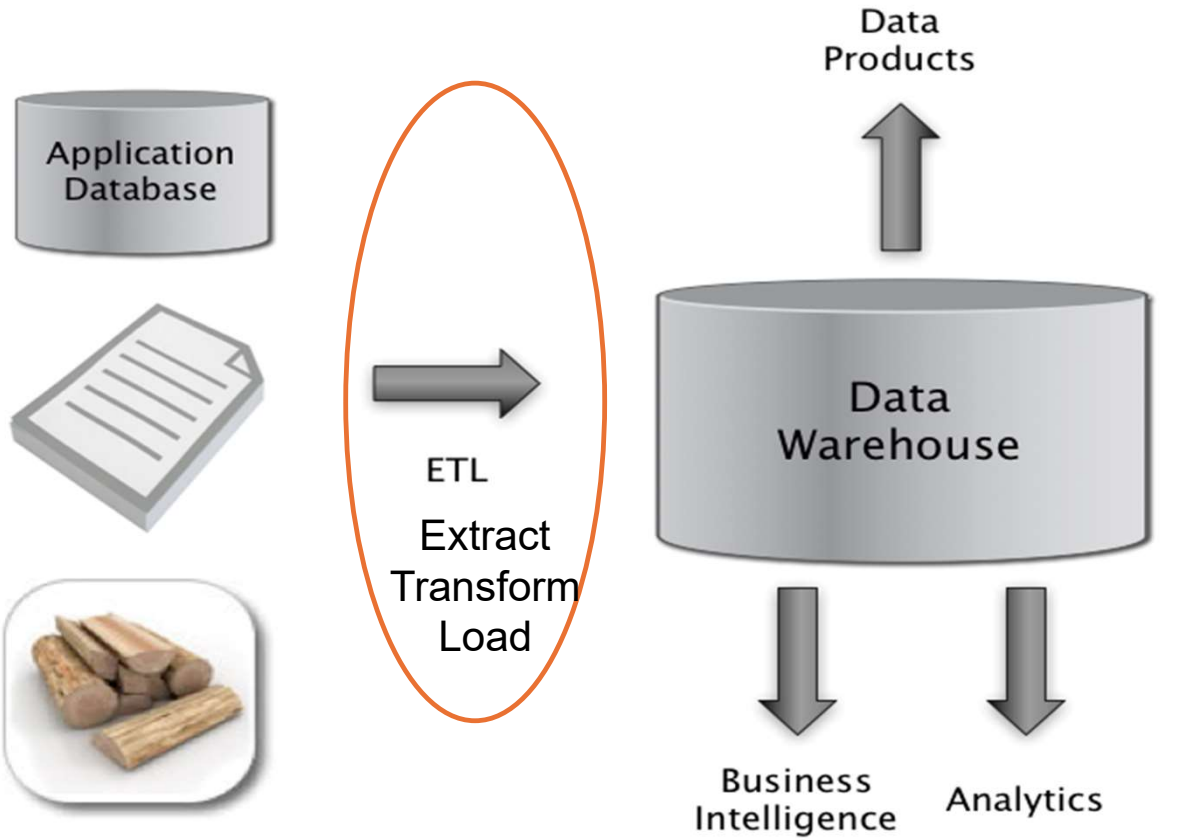
- **Fontes internas**: já coletadas ou fazem parte da coleta geral de dados da sua organização.
  - Por exemplo: dados centrados nos negócios disponíveis no banco de dados da organização para registrar as operações do dia a dia; dados científicos ou experimentais são obtidos de um ensaio.
- **Fontes externas existentes**: disponível em formato pronto para leitura de uma fonte externa gratuitamente ou por uma taxa.
  - Por exemplo: bancos de dados públicos do governo, dados do mercado de ações, esportes, COVID-19.
- **Fontes externas que exigem esforços de coleta**: disponível em fonte externa, mas a aquisição requer processamento especial.
  - Por exemplo: dados que aparecem apenas em formato impresso ou dados em sites.

Maneiras de coletar dados online

Como obter dados gerados, publicados ou hospedados online?

- **API** (Application Programming Interface): Usando um conjunto pré-integrado de funções desenvolvidas por uma empresa para acessar seus serviços. Muitas vezes pagas para usar.
  - Por exemplo: API do Google Maps, API do Facebook, API do Twitter
- **RSS** (Rich Site Summary): resume o conteúdo online atualizado com frequência em formato padrão. Livre para ler se o site tiver um.
  - Por exemplo: sites relacionados a notícias, blogs
- **Web scraping** (rastreamento): usando software, scripts ou extraindo manualmente dados do que é exibido em uma página ou do que está contido no arquivo HTML (geralmente em tabelas).

Bons velhos tempos…



O que é Web Scraping?

- A construção de um agente para baixar, analisar e organizar dados da web de uma forma automatizada.
- Um usuário final humano clica em um navegador da web e copiar e colar partes interessantes em uma planilha. A automatização desta tarefa usando um computador é chamada de “Web Scraping”
- A coleta automatizada de dados da Internet é provavelmente tão antiga quanto a própria Internet.

Web Scraping para ciência de dados?

- A web contém muitas fontes de dados interessantes que fornecem um tesouro para todos os tipos de coisas interessantes.
- Infelizmente, a atual natureza não estruturada da web nem sempre facilita a coleta ou exportação desses dados de maneira fácil.
- Os navegadores da Web são muito bons em mostrar imagens, exibir animações e organizar sites de uma forma que seja visualmente atraente para os humanos, mas não apresentam uma maneira simples de exportar seus dados, pelo menos não na maioria dos casos.

Web scraping

- **Por que fazer isso?**
- Sites de notícias mais antigos do governo ou menores podem não ter APIs para acessar dados
- Publique feeds RSS ou tenha bancos de dados para download.
- Não se tem $$ para pagar para usar a API ou a API fornecida é limitada por taxa.
- Monitorar um site de notícias em busca de novas histórias sobre um tópico de interesse específico
- **Fazer análises de redes sociais** usando dados de perfil encontrados em um fórum da web.

Web scraping

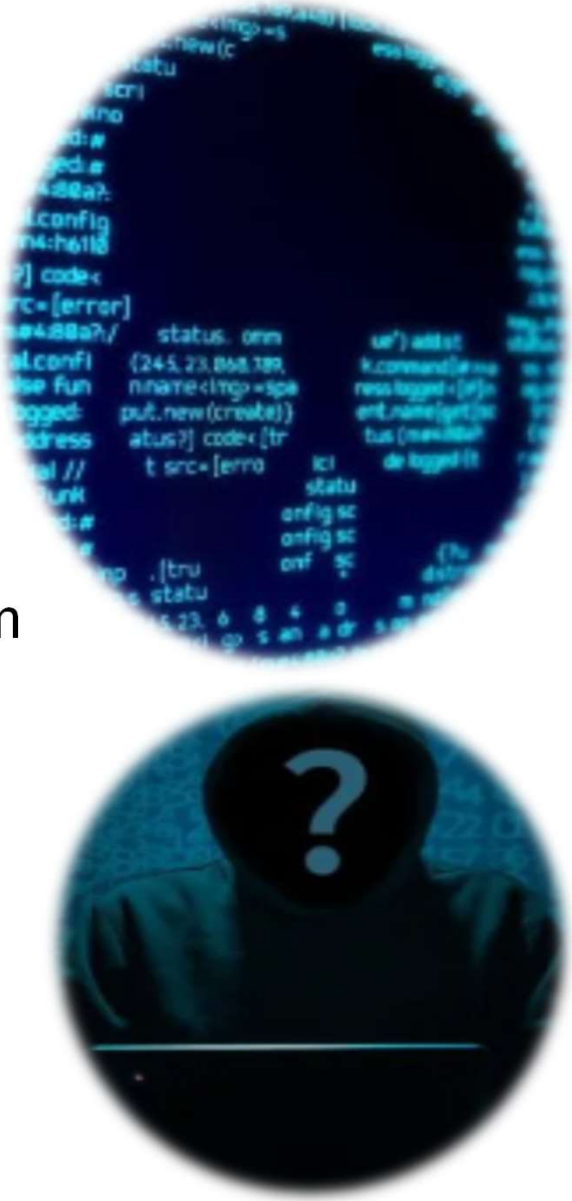
- **Quem deve fazer isso?**
- Só deseja explorar:
  - Está violando os termos de serviço deles?
  - Preocupações com a privacidade do site e de seus clientes?
- Deseja publicar uma análise ou produto:
  - Eles têm uma API ou taxa que você está ignorando?
  - Eles estão dispostos a compartilhar esses dados?
  - Está violando os termos de serviço deles?
  - Existem preocupações com a privacidade?
- **Como se faz isso?**

Web scraping

- Usando programas Python (ou R) para obter dados online
- Muitas vezes muito mais rápido do que copiar dados manualmente!
- Transfira os dados para um formato compatível com seu código
- Questões legais e morais

Aviso: Web scraping

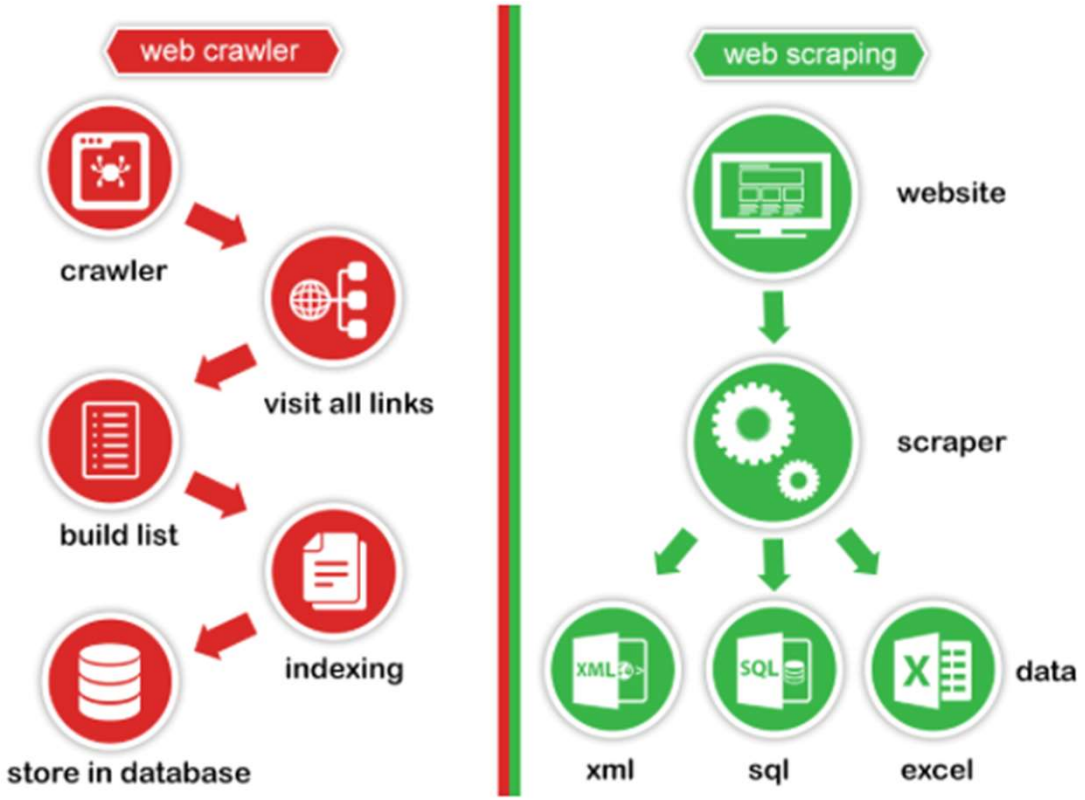
- **Dicas**:
  - Vasta fonte de informação; pode ser combinado com vários conjuntos de dados
  - Automatizar tarefas repetitivas
  - Acompanhar os sites / dados em tempo real
  - Seja cuidadoso e educado (não visite um servidor com muita frequência)
  - Seja robusto e imune a armadilhas e outros comportamentos maliciosos de servidores web
  - Dê o devido crédito!
  - Preocupar-se com a lei de mídia / obedecer licenças / privacidade
  - **Não se esqueça da proveniência dos dados!**



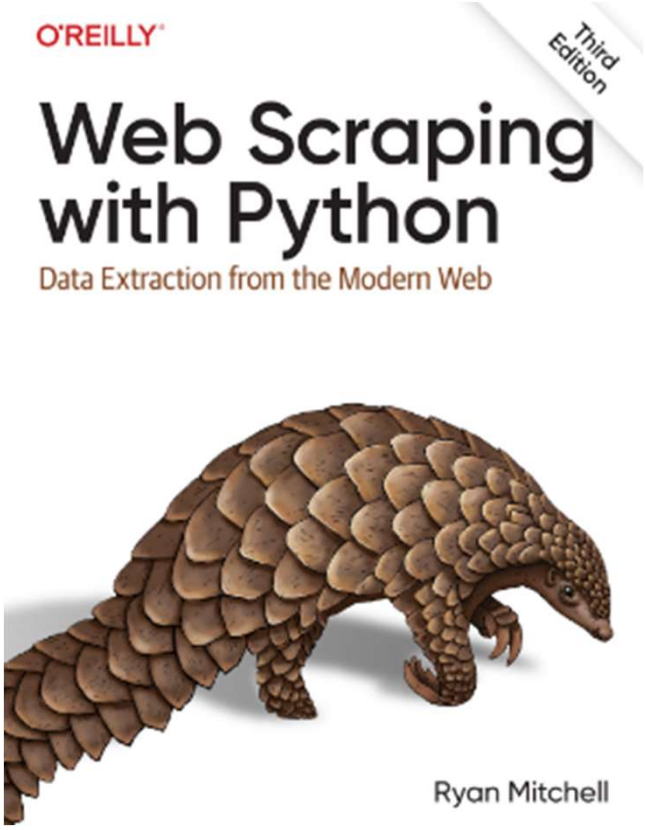
Obtendo dados: Web scraping

- **Robots.txt**
- Protocolo para dar aos spiders ("robôs") acesso limitado a um site, originalmente de 1994 www.robotstxt.org/wc/norobots.html
- Site anuncia sua solicitação sobre o que pode (não) ser rastreado
  - Especificado (restrições de acesso) pelo proprietário do site
  - Fornece instruções para robôs da web (por exemplo, seu código)
  - Localizado no diretório de nível superior do servidor Web
  - Por exemplo, http://google.com/robots.txt

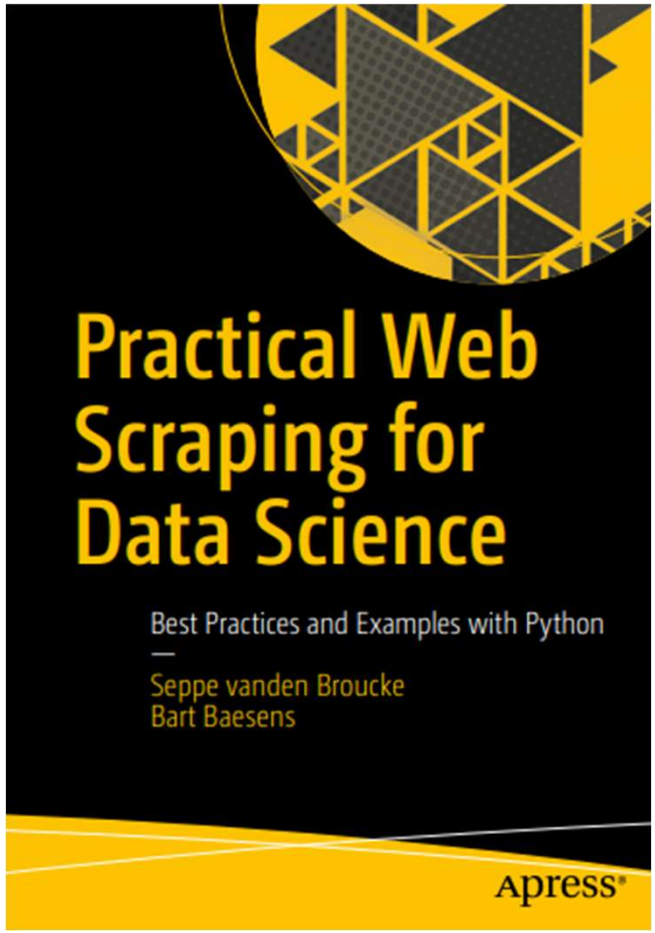
Web Crawler x web scraper



Referências



Todos os slides de esta aula foram adaptados das aulas de **Fundamentos de Ciencia de Dados**. PPGI, 2024. Sergio Serra e Jorge Zavaleta.



Beautiful Soup Documentation — Beautiful Soup 4.9.0 documentation (crummy.com)

Hands on...

NOTEBOOK:
BEAUTIFULSOAP