Implementando a Técnica de Web Scraping

Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas Professor: Paulo Battistella, Dr.



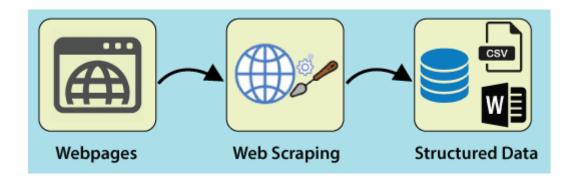
Objetivos

- Introduzir os conceitos básicos de web scraping;
- Instalar bibliotecas BeautifulSoup; Requests e XlsxWriter;
- Demonstrar como extrair dados de páginas web;
- Demonstrar como salvar os dados extraídos em um arquivo excel.

Introdução

O que é Web Scraping?

- Termo inglês "scraping" significa raspagem;
- Extração de dados de páginas da web.









Introdução

- Ética e legalidade:
 - Muitas vezes é usado para fins comercias, sem levar em conta os direitos autorais e termos de uso;
 - Pode criar uma sobrecarga no site pelo excesso de requisições;
 - Não é especificado com os dados serão usados.



Introdução

- Ética e legalidade:
 - O ato de extrair dados não é ilegal. Mas existem direitos autorais e termos de uso que devem ser respeitado;
 - Quando extraímos dados de outras pessoas, sem permissão ou autorização, podemos ter problemas legais.



Aplicações de Web Scraping

- Monitoramento de preços online;
- Encontrar oportunidade de investimento;
- Analisar dados de mídia social;
- Aplicar técnicas para aprendizagem de máquina.



Aplicações de Web Scraping

- Coletar dados automaticamente da web;
- Extrair informações de contatos;
- Monitorar notícias.



Pré-requisitos

- Conhecimento básico de Python e HTML;
- IDE: Pycharm, Spyder, VSCode;
- Python3.X;
- Bibliotecas:
 - Requests, BeautifulSoup e XlsxWriter.



Atividade 1

Criar um projeto em Python e instalar as Libs:

\$ python -m pip install **requests**

\$ python -m pip install **beautifulsoup4**

\$ python —m pip install **XIsxWriter**



Atividade 2

 Crie um script em Python que extraia as novidades da linguagem Python e salve os dados em um arquivo Excel.



Atividade 3

 Faça scraping de um site de Ações, Previsão do Tempo e Notícias. Ao final salve em uma planilha do Excel.



Obrigado!

paulo. battistella@faculdadeces usc. edu. br









cesusc.edu.br f/faculdadecesusc 48 3239.2600