Extração de Dados: Web Scraping e suas Tecnologias mais recentes

Autor(res)

Sandro Teixeira Pinto
Arthur Henrique Inácio De Oliveira

Categoria do Trabalho

Trabalho Acadêmico

Instituição

CENTRO UNIVERSITÁRIO ANHANGUERA

Introdução

O Web Scraping ou "Raspagem de dados" é uma técnica de coleta de dados dinâmica podendo ser usada de diversas formas, seja extraindo dados através de plataformas online como redes sociais como em qualquer outro tipo de site.
Utilizando-se de inteligência artificial a raspagem de dados pode ser conhecida como "bots" que agem por contra própria através do script criado.
As tecnologias mais usadas em seu desenvolvimento são as linguagens: Python, Node JS, Ruby, C & C++ e PHP. Sendo a Python a mais eficiente entre elas é uma linguagem de programação orientada a objetos, de tipagem dinâmica e funcional podendo ser interpretada por scripts de alto nível, a extração dos dados passa ser mais rápida e de fácil entendimento.

Objetivo

Este artigo tem o objetivo de abordar as tecnologias mais recentes e eficientes de Web Scraping na atualidade com foco na linguagem Python e suas bibliotecas dinâmicas.

Material e Métodos

O procedimento para atingir o objetivo desse trabalho, foi fazer um levantamento bibliográfico em plataformas na internet sobre o tema. Para esta pesquisa foi utilizado palavras chave como "Web Crawling", "Web Scraping", "Raspagem de dados"e "Extração de dados", em sites como google acadêmico, Scielo e trabalhos da Capes.
Também compõe este trabalho diversos levantamentos de informações nas principais revistas de tecnologia e diretamente através da documentação de software.

Resultados e Discussão

Atualmente uma quantidade massiva de dados são geradas na internet, sendo de diferentes origens e formatos. Nesse contexto a necessidade de utilizar "bots" para coletar grandes volume dados de forma automatizada se torna cada vez mais necessário e viável.
Web Scraping (WS) pode ser definido como: "O processo de extração e combinação de conteúdos de interesse da Web de uma forma sistemática. Em tal processo, um agente de software, também conhecido como robô, imita a interação de navegação entre os servidores da Web e o humano." (LOURENÇO, 2013)

Já o BeautifulSoup é considerada uma biblioteca que faz uma identação de textos em HTML/XML extraídos das páginas, facilitando a leitura e aumentando a precisão da extração das informações.
Httpx é uma biblioteca na qual simula requisições padronizadas de forma rápida sendo a ferramenta mais leve em questão.
Playwright faz a simulação mais próxima do que seria um usuário acessando uma página.

Conclusão

Podemos concluir que as tecnologias mais utilizadas possuem vantagens e desvantagens, podendo ser aplicadas para cada tipo de situação. Outra coisa que podemos citar também é que essas bibliotecas podem ser usadas em conjunto, o que facilita a compatibilidade de sites que trabalham tanto com muitos elementos, ou com muitos payloads & requisições.

Referências

LOURENÇO, A. Web scraping technologies in an API world. Briefings in Bioinformatics, v. 15, n. 5, 04 2013, p. 788– 797. ISSN 1467-5463.
DOCUMENTACÃO Beautiful Soup. Disponível em https://www.crummy.com/software/ BeautifulSoup/bs4/doc.ptbr. Acesso em: 31 out 2022.
DOCUMENTACÃO Playwright. Disponível em: https://playwright.dev/python/. Acesso em: 31 out 2022.