BeautifulSoup

Marcos Castro e Thomas William

O que é BeautifulSoup ?

Biblioteca Python projetada para facilitar a extração de dados nos documentos html e xml.

Principais Características

- Fornece alguns métodos simples e expressões pythonicas para navegar, pesquisar e modificar uma árvore de análise.

- Converte automaticamente documentos de entrada para Unicode e documentos de saída para UTF-8.

- Utiliza-se bibliotecas html parse lxml, html5lib, permitindo-lhe experimentar estrategias de análise diferentes ou velocidade comercial para a flexibilidade.

- Bastante robusta para trabalhar com html/xml mal formatado.

PARSER

Parser	Modo de uso	Vantagem	Desvantagem	
html.parser	BeautifulSoup(markup, "html.parser")	• Incluso biblioteca • Velocidade moderada		
lxml	BeautifulSoup(markup, "lxml") BeautifulSoup(markup, "lxml-xml") BeautifulSoup(markup, "xml")	• Rápida • Suporta xml	• Dependência externa	C
html5lib	BeautifulSoup(markup, "html5lib")	• Analisa páginas da mesma forma que um browser • Cria HTML5 válido	• Lenta • Dependência externa	Python

OBRIGADO !