LabRedes de Conhecimento

Chatbots: da teoria ao deploy, com IBM Watson

João Paulo de Melo
jpmdik@gmail.com
Tecnólogo em Sistemas para Internet

Aula 06: Web scraping

Web scraping

Web scraping é uma técnica de extração de dados utilizada para coletar dados de sites. Por meio de processos automatizados, implementados usando um rastreador bot, esse tipo de “raspagem” de informações é uma forma de realizar cópias de dados em que informações específicas são coletadas e copiadas da web, tipicamente em um banco de dados ou planilha local central, para posterior recuperação ou análise.

Web scraping

No entanto, o web scraping tem se tornado uma prática maliciosa utilizada por criminosos para roubar conteúdos protegidos e cometer fraudes, repassando informações de produtos e serviços de uma empresa para a concorrência, o que pode causar grandes prejuízos aos negócios.

Web scraping - como funciona?

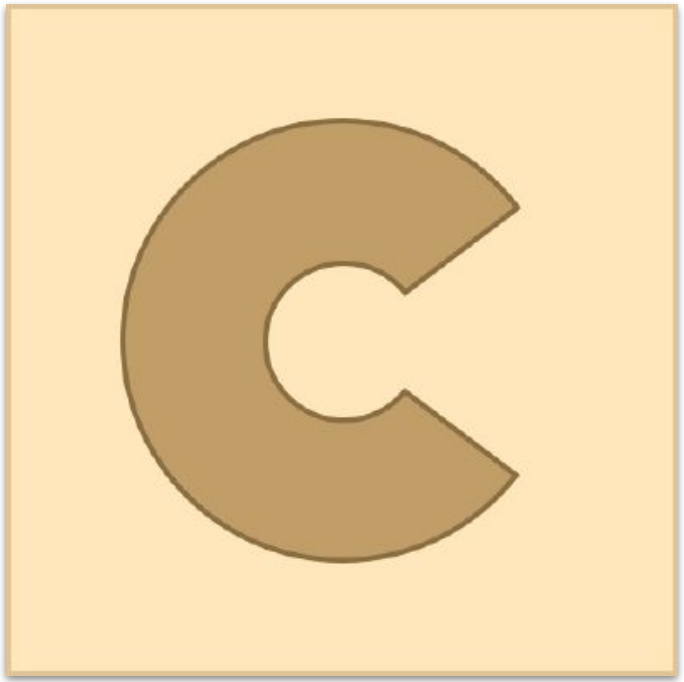
Web scraping - python

Link: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Web scraping - ruby on rails

Link: https://nokogiri.org/

Web scraping - nodejs



Link: https://github.com/cheeriojs/cheerio

Web scraping - genérico

Biblioteca que faça requisição de uma página + expressões regulares

+

Expressões regulares por aí...

- Python
  - `import re`
  - `regex = r'(\d\d)/(\d\d)/(\d\d\d\d)'`
  - `match = re.match(regex, '13/04/2011')`
  - `match.group(0)` → `'13/04/2011'`
  - `match.group(1)` → `'13'`
  - `match.group(2)` → `'04'`
  - `match.group(3)` → `'2011'`

Iniciando a aplicação

Iniciando a aplicação

Instalando as dependências:

Comando: pip install beautifulsoup4

Exercício Sala

Aula 06

Extrair os dados dos filmes para um arquivo json do site http://www.henancius.com/henancius/top100.html

Exercício Casa

Aula 06

Extrair Informações úteis de um site, de forma que seja necessário varrer um ou mais links dentro da página.

Referências

- Blog Brasil Westcon. **O QUE É WEB SCRAPING?**. 2019. Disponível em: <https://blogbrasil.westcon.com/o-que-e-web-scraping>. Acesso em: 20 de maio de 2019.