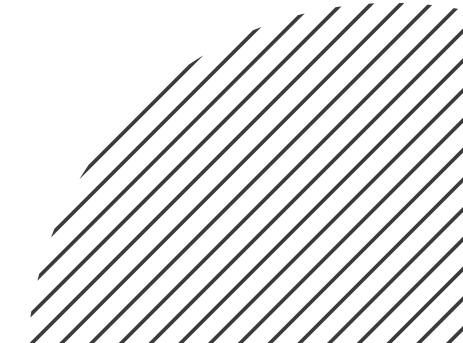
2018

Web Scraping com R

CURSO DE PROGRAMAÇÃO PARA CIÊNCIAS SOCIAIS

o que é web scraping?

extração de maneira automatizada de dados e conteúdos de websites



qual o plano?

1. O que será raspado: uma página? Várias páginas? São seguidas? Possuem numeração? Precisa construir um loop?
2. Capturar a página
3. Identificar a estrutura (parse) do texto lido pelo R
4. Eliminar sujeiras do documento
5. Procurar pelas tags e atributos que contenham as informações relevantes

OBS. Há outros métodos para lidar com APIs

conceitos importantes (1)

HTML

Páginas HTML servem para exibir informações. Frequentemente é como os navegadores mostram os websites. HTML é um tipo de XML.

XML

Linguagem com marcações conhecidas como 'tags' que possuem uma hierarquia e armazenam informações.

conceitos importantes (2)

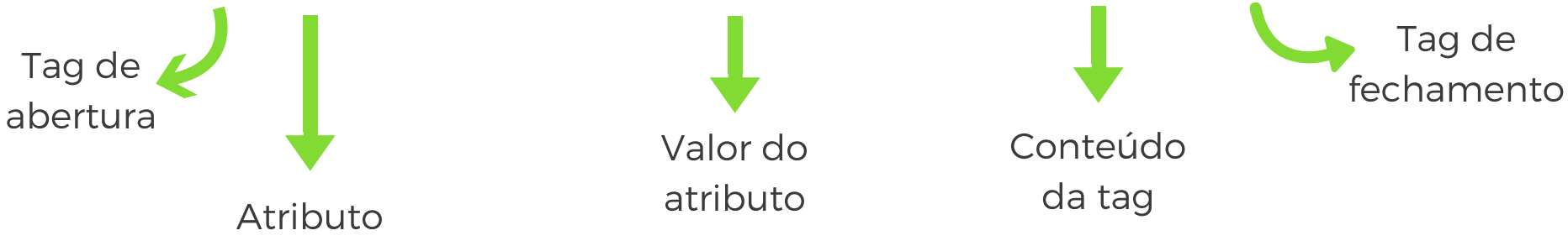
TAGS

Possuem tag de abertura e de fechamento e entre elas há um conteúdo, que podem ser outras tags.

<name> exemplo </name>

ATRIBUTOS

 Github



no mundo real...

http://dcp.fflch.usp.br/

conceitos importantes (3)

XPATH

Linguagem para expressar partes de um documento XML por meio de caminhos separados por barras ("/"). Possui elementos que permitem identificar o endereço de qualquer parte do documento.

CSS

Maneira de apontar um conteúdo da página baseado nas suas configurações de estilo. Todos os seletores em CSS podem ser traduzidos para um em XPath.

observações

LIMITAÇÕES

Muitas vezes, websites impõem um limite de o quanto é possível raspar antes de sofrer um bloqueio

JAVASCRIPT

Sites possuem camadas diferentes do que HTML e CSS precisam de outra solução para a extração de dados

o que faremos hoje?

Um tutorial dividido em três partes:

- Parte 1: revisão de loop e extração de tabelas
- Parte 2: HTML/XML, tags, atributos e XPath
- Parte 3: caso Datafolha