Web Scraping

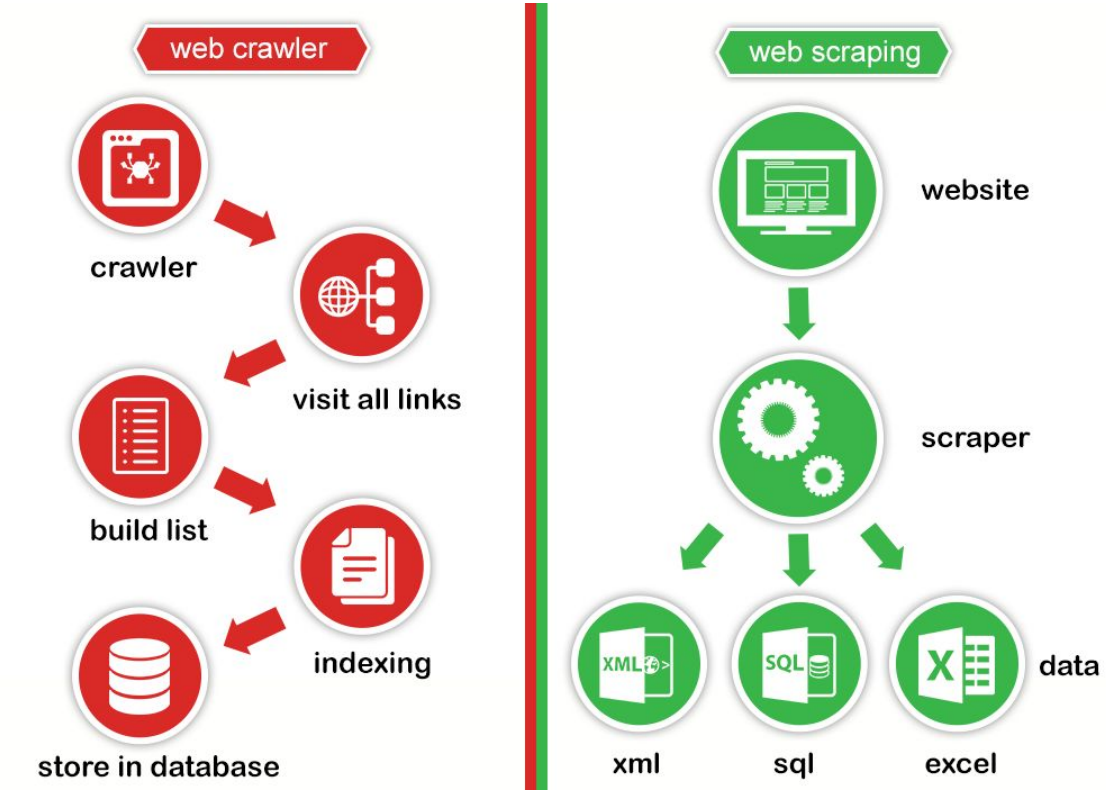
Por Tales Luna

Minerando dados na Web

Que "troço" é esse?

É a técnica de navegar através de páginas web colhendo informações (misturadas em HTML) que não estão disponíveis em API's públicas transformando-as em um formato de estrutura de dados melhor legível por você ou por outros softwares, tais como JSON, CSV ou XML.

Web SCRAPING não é Web Crawler!



Quem usa? e por quê?

Muitas empresas usam **Web Scraping** pois muitas vezes precisam de informações que são públicas mas não estão legíveis para seus sistemas e seria muito custoso manter departamentos inteiros apenas para coletar essas informações, além de demorado.

- ➔ Empresas de redes sociais (Facebook, Twitter)
- ➔ Search Engines (Google, Bing)
- ➔ Empresas de Marketing Digital
- ➔ Empresas de precificação (Trivago, Buscapé, etc..)
- ➔ Startups

Startups?

Sim, startups de variados ramos amam **Web Scraping** pois essa é a melhor maneira dessas empresas conseguirem dados a relativos ao negócio por um baixo custo e sem a necessidade de de firmar parcerias para isso.

Exemplo Prático

Exemplo musiquinhas

Imagine que o site *www.musiquinhaslegais.com* publica posts diarios sobre músicas e por algum motivo você gostaria dessas informações para seu software ou negócio mas o site não tem ou não libera uma API ou Banco de Dados para você consumir, é aí que entra o **Web Scraping**, você pode utilizar essa técnica para navegar pelo site e recolher a informação que precisa no formato que deseja.

Exemplo musiquinhas

Quando acessamos *www.musiquinhaslegais.com* temos a página:

Músicas mais populares

Artista	Nome	Album	Arquivo
Pink Floyd	Comfortably Numb	The Wall	Download
Radiohead	Exit Music (for a film)	Ok Computer	Download

HTML

Se olharmos o código fonte HTML da página então teremos o conteúdo da imagem ao lado

Como poderíamos transformar isso em JSON ?

* sim, o “inspecionar elemento” tem muita utilidade!!

```
<html lang="pt-BR">
  <title>Musiquinhas Legais</title>
  <meta charset="utf-8" />
  <body>
    <div class="posts">
      <h4 class="posts-title">Músicas mais populares</h4>
      <table border="1">
        <thead>
          <tr>
            <th>Artista</th>
            <th>Nome</th>
            <th>Album</th>
            <th>Arquivo</th>
          </tr>
        </thead>
        <tbody>
          <tr>
            <td>Pink Floyd</td>
            <td>Comfortably Numb</td>
            <td>The Wall</td>
            <td><a href="http://mp3.musiquinhaslegais.com/files/1.mp3">Download</a></td>
          </tr>
          <tr>
            <td>Radiohead</td>
            <td>Exit Music (for a film)</td>
            <td>Ok Computer</td>
            <td><a href="http://mp3.musiquinhaslegais.com/files/2.mp3">Download</a></td>
          </tr>
        </tbody>
      </table>
    </div>
  </body>
</html>
```

Repetição

Parser, água em vinho

Para transformar o HTML em JSON precisaremos analisar o HTML e descobrir suas particularidades, tais como:

- ➔ Estruturas de repetição [looooooooooping]
- ➔ Classes [.eu-sou-muitos]
- ➔ IDs [#eu-sou-alone]
- ➔ Tags [<eu>]
- ➔ Demais atributos

Estruturas de repetição

É natural que para formar o HTML com informações vindas de um banco de dados será feito uma estrutura de repetição (for, foreach, etc) que deve ser nosso foco inicial, principalmente se os dados que queremos extrair são uma coleção.

Exemplo:

Para gerar o HTML o site *www.musiquinhaslegais.com* faz foreach de um array gerando um bloco dentro de uma tag com class “.post”, logo se pegarmos todas as divs com class “.post” temos uma coleção dessas informações

CLASSES

Como já dito anteriormente classes são essenciais para o web scraping e devem ser utilizadas para adquirir coleções de informações que repetem na página web, tanto no primeiro nível quanto nos demais momentos da extração das informações

TAGS

Assim como classes e ID’s, podermos utilizar as tags diretamente para filtrar e extrair conteúdo, mas cuidado, as TAGS podem repetir de forma indiscriminada, por isso não as use como elemento “pai” para iniciar uma coleção de dados.

ENGENHARIA (COM MÃO NA MASSA)

De posse da análise da página de onde se quer extrair as informações e de que como se comporta o HTML é hora de criar o que chamamos de “engenharia”, em palavras diretas é onde vamos criar uma maneira (que será implementada em código) de extrair informações.

- ➔ **Tenho o HTML**
- ➔ **Sei os dados que quero**
- ➔ **Sei onde começa a estrutura dos dados**
- ➔ **Sei as tags e classes que os compõe**
- ➔ **É hora de criar o script!**

Código

Agora vamos criar nosso software, mas primeiro vamos esclarecer algumas coisas:

1. Esqueça padrões de projetos e código “limpo”
2. Aceite que o site pode mudar e seu código também
3. Programação “clássica” lhe espera a frente *(muitos for e arrays)*
4. Cada código é específico a uma página ou site e dificilmente poderá ser reutilizado.
5. Você não conseguirá extrair dados que foram inseridos no DOM via JS, apenas dados estáticos, aplicações SPA’s e partes populadas por JQuery podem ser problema.

O que usar?

São boas linguagens para Web Scraping:

- ➔ Python [Mechanize, Scrapy]
- ➔ Perl [Mechanize]
- ➔ JavaScript[Cheerio]

Motivos:

São linguagens intuitivas, de script, com muita documentação sobre scraping além de possuir bibliotecas e frameworks já consolidados para essa finalidade.

COMO SERIA EM JS?

Terminal:

$ git clone https://github.com/talesluna/web-scraping-cheeriojs-example
$ npm install
$ node scraping.js

JSON (obtido de www.musiquinhaslegais.com.br)

```
[
  {
    "description": "Músicas mais populares",
    "playlist": [
      {
        "artist": "Pink Floyd",
        "name": "Comfortably Numb",
        "album": "The Wall",
        "download": "http://mp3.musiquinhaslegais.com/files/1.mp3"
      },
      {
        "artist": "Radiohead",
        "name": "Exit Music (for a film)",
        "album": "Ok Computer",
        "download": "http://mp3.musiquinhaslegais.com/files/2.mp3"
      }
    ]
  }
]
```

```
▼ array [1]
  ▼ 0 {2}
      description : Músicas mais populares
    ▼ playlist [2]
      ▼ 0 {4}
          artist : Pink Floyd
          name : Comfortably Numb
          album : The Wall
          download : http://mp3.musiquinhaslegais.com/files/1.mp3
      ▼ 1 {4}
          artist : Radiohead
          name : Exit Music (for a film)
          album : Ok Computer
          download : http://mp3.musiquinhaslegais.com/files/2.mp3
```

Ética e Legalidade

Realizar **Web Scraping** não é ilegal, uma vez que você está colhendo informações públicas como visitante de um site, mas deve-se ficar atentos a algumas questões:

- ➔ Não fazer requisições excessivas ao site (DDoS?)
- ➔ Não utilizar os dados obtidos para fraudes ou usar o nome da empresa de onde se retirou os dados em hipótese alguma
- ➔ Não usar os dados para Spam (no caso de Marketing Digital)

Todavia é sempre bom consultar advogados.

Web Scraping é bom, é viável e eficaz