Raspagem de Dados Web

(Data Scraping)

Por: Marcos Valadão e Nicolas Ferranti

Departamento de Ciência da Computação

www.ufjf.br/deptocomputacao

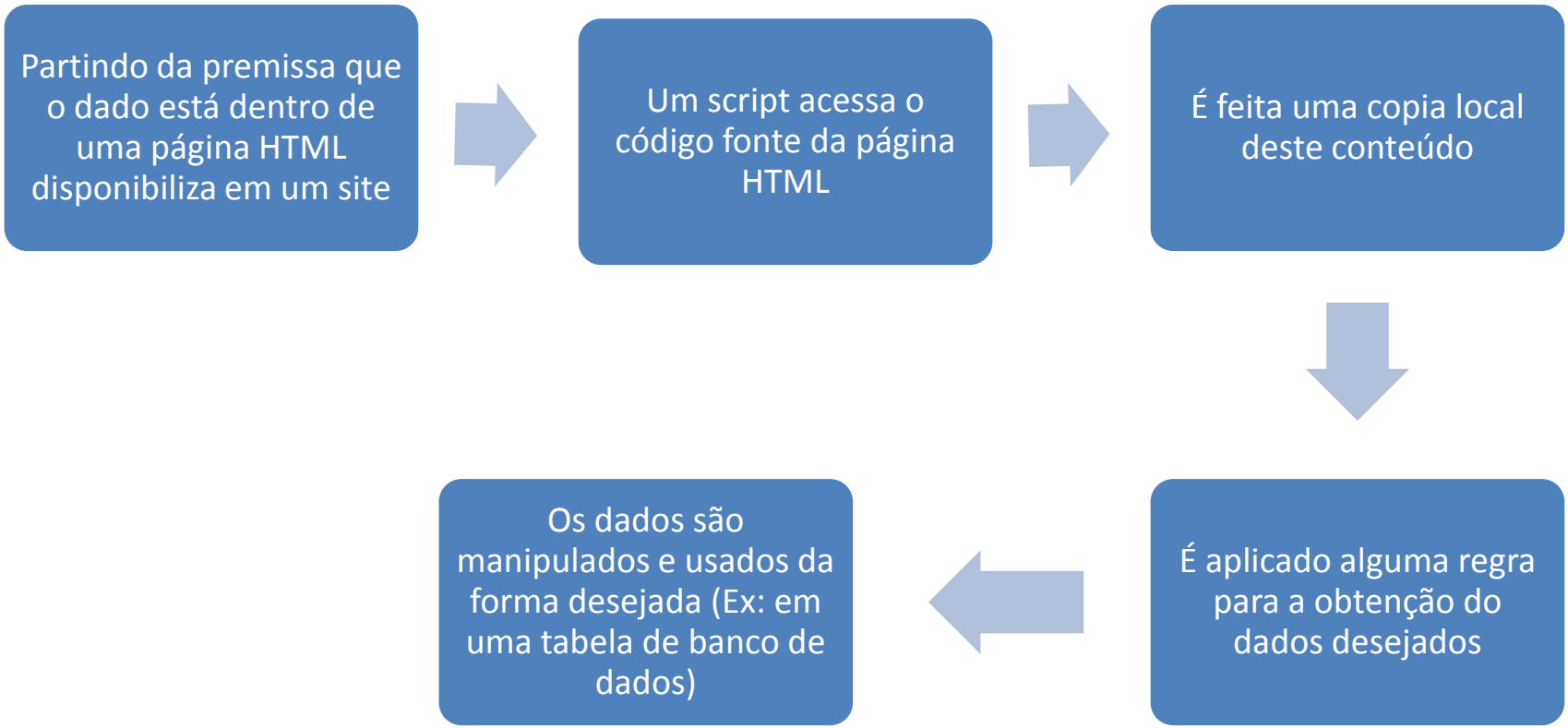
O QUE É?

“Data scraping (ou raspagem de dados) é uma técnica na qual um programa de computador extrai dados de saída legível para humanos, proveniente de um outro programa, e disponibiliza esses dados de modo que se tornem legíveis para outros programas de computador. ”

(Definição do Wikipédia)

No nosso caso, é uma técnica de retirada de dados web, utilizando métodos e ferramentas simples, para a manipulação e aproveitamento desejado.

Como é feita a Raspagem



1. Entenda a fonte de dados

Antes de obter o dado on-line é preciso conhecer muito bem a fonte de dados. Isso quer dizer que é preciso navegar pela página, verificar se as informações são paginadas, colocadas em uma tabela, frame, área dinâmica ou controle, se requerem um login, se são disponibilizadas em um arquivo, possuem acesso via API, seguem algum padrão de apresentação e organização e outros detalhes relevantes.

Não existe uma regra específica, mas geralmente alguma paginação é feita para evitar sobrecarga, que pode apresentar a
forma de links com o número da página, controles do tipo combo-box ou outros.

2. Planejar como será feita a obtenção bruta dos dados

Conhecendo os dados é hora de planejar como será feita a obtenção.

Isso quer dizer é preciso relacionar ferramentas, ambiente, plataforma, link de internet e outros recursos tecnológicos necessários para se obter os dados. Aqui também não há uma regra geral, pois cada situação pode requerer uma ferramenta específica.

Existe uma vasta extensão de ferramentas para a extração de dados, porém vamos utilizar o HTML DOM parser.

Esta página contém algumas das ferramentas mais utilizadas na raspagem
http://www.sharewareconnection.com/software.php?list=Pdf+Screen+Scraper

HTML DOM parser

- É uma ferramenta PHP5+ que permite manipular HTML de forma fácil.
- Exige PHP5 +.
- Encontra tags em uma página HTML com apenas seletores jQuery.
- Extraia o conteúdo de HTML em uma única linha .

Não é necessário conhecimento profundo de PHP, somente um parte bem básica.

A página http://simplehtmldom.sourceforge.net/manual.htm contém a documentação completa da ferramenta e um tutorial para utilizá-la.

É bem simples!

3. Programar a captura dos dados

Uma vez que o planejamento esteja pronto, basta programar o script que montava o loop e fazer a chamada ao arquivo que foi criado.

Obs: Não faça muitas requisições simultâneas à uma pagina, seu acesso pode ser bloqueado.

Aviso: fazer a raspagem de dados de um site pode violar termos de serviços. Você deve garantir que isso não vai acontecer antes de começar. Por exemplo, o Twitter proíbe completamente a raspagem de informação no site. Isso está nos Termos de Serviço:

"varrer o Serviço é permitido apenas em total acordo com as provisões do arquivo robots.txt, no entanto, ***raspar o Serviço sem consentimento prévio do Twitter é expressamente proibido****"* (tradução livre)

O Google também faz uma proibição similar para a raspagem de conteúdo em suas propriedades web:

Os Termos de Serviço do Google não permitem o envio automático de queries de qualquer tipo para o nosso sistema sem a permissão prévia do Google. (tradução livre)

Portanto, seja cauteloso

4. Verificar os dados capturados

Neste ponto da sua raspagem, você deve verificar os dados que foram gerados e analisar se estão corretos. Talvez algumas mudanças devam ser feitas no script para adequar ao que você está buscando.

5. Modelar os dados conforme necessidade

Neste ponto a raspagem dos dados já está concluída, basta usá-los da forma que quiser.

Você pode colocá-los em alguma tabela de banco de dados, utiliza-los para fazer uma listagem, encontrar elementos específicos e muitas outras aplicações...

Vamos começar a Raspagem?!

Faremos a listagem de produtos(de uma mesma origem) disponíveis no site www.submarino.com.br como exemplo de uma raspagem de dados web.

Vamos seguir o 5 passos:

1. Entenda a fonte de Dados.
2. Planejar como será feita a obtenção bruta dos dados.
3. Programar a captura dos dados.
4. Verificar os dados capturados.
5. Modelar os dados conforme necessidade.

ufjf