

Práticas de Desenvolvimento de Software


#

Aula 09

Desenvolvimento Web

PROJETO FINAL

AULA 2/3



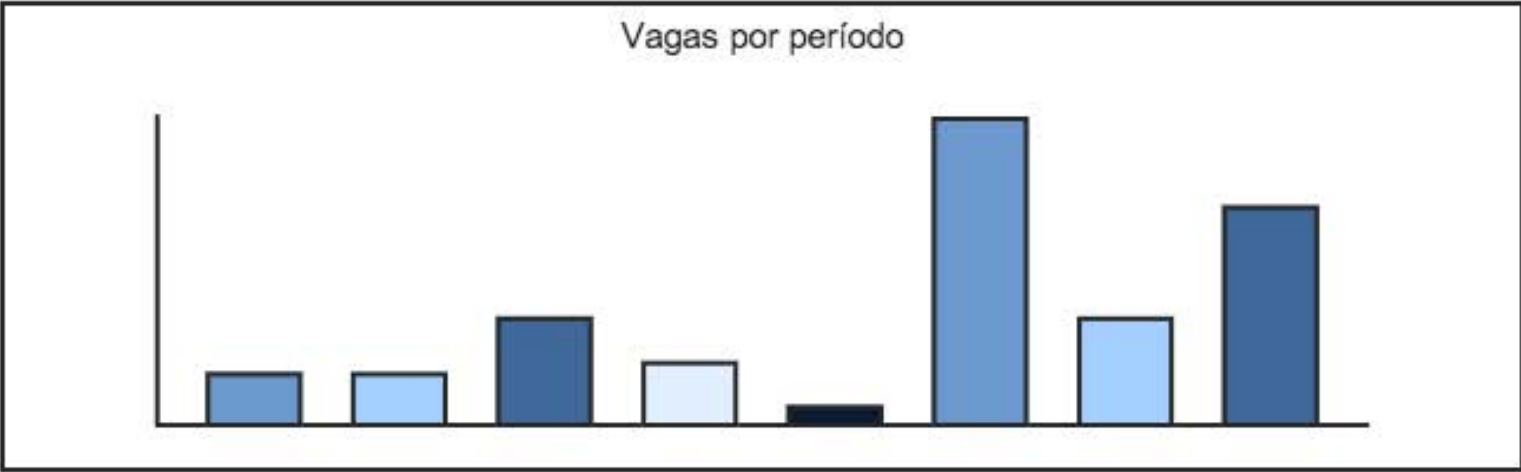
Armazenar e
manipular dados

Extrair dados da
internet


Continuação do
projeto final

Nome do aluno		Práticas de Desenvolvimento de Software
Resumo	<div>Busca</div> <div><input type="text" value="Digite o termo de busca"/> <input type="button" value="Pesquisar vagas"/></div> <div>Campos considerados:</div> <div><input type="checkbox"/> Habilitação <input type="checkbox"/> Empresa <input type="checkbox"/> Descrição <input type="checkbox"/> Benefícios <input type="checkbox"/> Número de vagas <input checked="" type="checkbox"/> Título <input type="checkbox"/> Área de atuação <input type="checkbox"/> Requisitos <input type="checkbox"/> Contatos</div> <div>Resultados</div> <div>Tempo de busca: 5s Vagas encontradas: 50</div> <div>Muitos resultados encontrados. Faça uma nova pesquisa para refinar os resultados.</div>	
Vagas por período		
Requisitos: Palavras-chave		
Descrição: Palavras-chave		
Busca		
Projeto desenvolvido para o curso Práticas de Desenvolvimento de Software		

Nome do aluno		Práticas de Desenvolvimento de Software
Resumo	<div>Resultados</div> <div>Tempo de busca: 5s Vagas encontradas: 50</div> <div><div>Vaga #1234 Habilitação: Cursando o 4o. ano do Cooperativo Título: Desenvolvimento Web Empresa: Foo -- Soluções para Web Área de atuação: Desenvolvimento Web (client-side e server-side) Descrição: Requisitos: Benefícios: Ótimo ambiente de trabalho Contatos: Maria da Silva Data do anúncio: 29/04/2014 Válido até: 29/04/2014 Número de vagas: 1</div><div>Vaga #1234 Habilitação: Cursando o 4o. ano do Cooperativo Título: Desenvolvimento Web Empresa: Foo -- Soluções para Web Área de atuação: Desenvolvimento Web (client-side e server-side) Descrição:</div></div>	
Vagas por período		
Requisitos: Palavras-chave		
Descrição: Palavras-chave		
Busca		
Projeto desenvolvido para o curso Práticas de Desenvolvimento de Software		

Nome do aluno		Práticas de Desenvolvimento de Software
Resumo	Distribuição de vagas por data de publicação	
	Data inicial: DD/MM/YYYY Data final: DD/MM/YYYY	
Vagas por período		
Requisitos: Palavras-chave		
Descrição: Palavras-chave		
Busca	<p>(Seleção é opcional. Agrupamento padrão: quadrimestre)</p> <p><input type="radio"/> Mês <input checked="" type="radio"/> Quadrimestre <input type="radio"/> Ano</p>	
Projeto desenvolvido para o curso Práticas de Desenvolvimento de Software		

Formato de dados



Armazenar e
manipular dados

Extrair dados da
internet

Continuação do
projeto final

Como armazenar e manipular os dados?

 **JSON**

Como armazenar e manipular os dados?



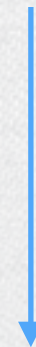
- JavaScript Object Notation
- Formato de dados leve para troca de informação estruturada
- Fácil de gerar e parsear
- Independente de linguagem

Objeto JSON

```
{  
  "key1": value1,  
  "key2": value2,  
  "key3": value3  
}
```


Tipos básicos de datos

{ "key": value }



String

Number

Boolean

Object

Array

null

String

```
{  
  "firstName": "João",  
  "lastName": "da Silva"  
}
```


Number

```
{  
  "age": 25,  
  "height": 1.82  
}
```


Boolean

```
{  
  "required": true,  
  "finished": false  
}
```


Object

```
{  
  "address": {  
    "street": "Avenida Paulista",  
    "number": 807,  
    "city": "São Paulo",  
    "state": "SP"  
  }  
}
```


Array

```
{  
  "phoneNumbers": [  
    { "type": "home", "number": "1234-5678" },  
    { "type": "work", "number": "9876-5432" }  
  ]  
}
```


Objeto JSON

```
{  
  "firstName": "João",  
  "lastName": "da Silva",  
  "age": 25,  
  "address": {  
    "street": "Avenida Paulista",  
    "number": 807,  
    "city": "São Paulo",  
    "state": "SP"  
  },  
  "phoneNumbers": [  
    { "type": "home", "number": "1234-5678" },  
    { "type": "work", "number": "9876-5432" }  
  ]  
}
```

The diagram illustrates the data types of the JSON object. Blue arrows point from specific values to boxes on the right:

- An arrow from the string value `"João"` points to a box labeled **String**.
- An arrow from the object value `{ "street": "Avenida Paulista", "number": 807, "city": "São Paulo", "state": "SP" }` points to a box labeled **Object**.
- An arrow from the number value `807` points to a box labeled **Number**.
- An arrow from the array value `[{ "type": "home", "number": "1234-5678" }, { "type": "work", "number": "9876-5432" }]` points to a box labeled **Array**.

Extração de dados

Armazenar e
manipular dados

Extrair dados da
internet

Continuação do
projeto final

O que é “Data Scraping”?

“Data Scraper é um software capaz de extrair dados da saída de outro programa. O tipo mais popular atualmente é o Web Scraper, cuja função é coletar dados de um website na internet.”

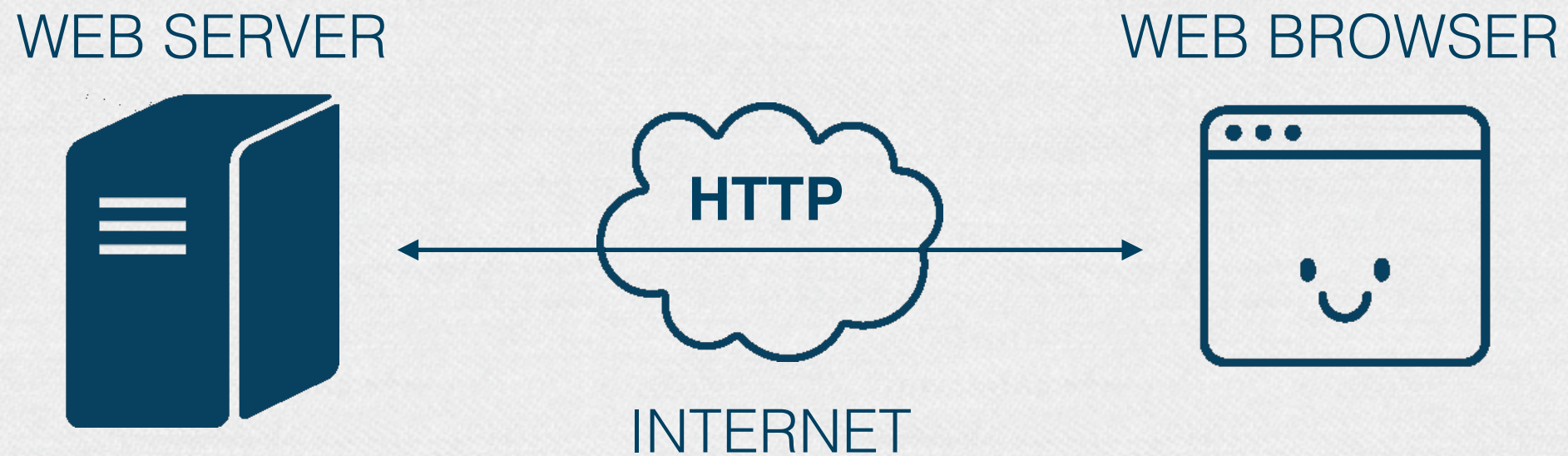
Não confunda com...

- **Crawlers:** navegação do tipo “força bruta”, sem filtros para identificar os links que acessa. Scrapers têm a característica de escolher os links que acessa e a navegação que executa.
- **Parsers:** processam dados estruturados, de forma a extrair informações mecanicamente. Um scraper contém um componente de inteligência para localizar informações não estruturadas.

Ferramentas

- Web browser
 - Google Chrome + Developer Tools
 - Mozilla Firefox + Developer Tools
- Ruby
 - Mechanize
 - Nokogiri
- Fiddler
 - Para Windows

Download de páginas



GET

POST

HEAD

PUT

DELETE

Extração de dados

The screenshot shows the Infosimples website in a browser. The main heading is "Soluções de tecnologia". The browser's developer tools are open to the Network tab, showing a list of requests. The first request, to "infosimples.com", is highlighted with a red box. Below the table, summary statistics are provided.

infosimples Serviços Read in English

A Infosimples atende empresas com excelência em **tecnologia**

Soluções de tecnologia

Elements Network Sources Timeline Profiles Resources Audits Console

☐ Preserve log

Name Path	Method	Status Text	Type	Initiator	Size Content	Time Latency	Timeline
infosimples.com	GET	200 OK	text/html	Other	4.9 KB 11.1 KB	164 ms 161 ms	
css?family=Open+Sans:300italic,400...	GET	200 OK	text/css	infosimples.com/:16	947 B 2.7 KB	154 ms 153 ms	

21 requests | 697 KB transferred | 823 ms (load: 864 ms, DOMContentLoaded: 848 ms)

Extração de dados

Portal de Estágios do Depa x

estagios.pcs.usp.br/semLogin/login.aspx

Escola Politécnica da USP
PCS - Departamento de Engenharia de Computação e Sistemas Digitais
PORTAL DE ESTÁGIOS

[Página inicial](#) [Empresa](#) [Aluno](#) [Coordenadoria](#) [Fale Conosco](#) [Procedimentos](#) [Login](#)

Login

Login

Senha

☐ Lembre-me

Logar

Clique [aqui](#) para lembrar a senha.

Elements Network Sources Timeline Profiles Resources Audits Console

Preserve log

Name Path	Method	Status Text	Type	Initiator	Size Content	Time Latency	Timeline	100 ms	150 ms
login.aspx /semLogin	POST	302 Found	text/html	Other	255 B 0 B	23 ms 23 ms			
login.aspx /semLogin	GET	200 OK	text/html	http://estagios.pcs... Redirect	5.0 KB 11.6 KB	21 ms 18 ms			
Theme1.css /App_Themes/Theme1	GET	200 OK	text/css	login.aspx:6 Parser	(from cac...)	0 ms 0 ms			
jquery-ui.css /Content/themes/base	GET	200 OK	text/css	login.aspx:17 Parser	(from cac...)	0 ms 0 ms			
WebResource.axd?d=qAoRo9r41vtRP...	GET	304 Not Modified	applicatio...	login.aspx:60 Parser	225 B 21.8 KB	18 ms 16 ms			

11 requests | 6.5 KB transferred | 175 ms (load: 176 ms, DOMContentLoaded: 164 ms)

Developer Tools

Conjunto de ferramentas de edição de depuração integradas ao navegador web.

Google Chrome

<https://developers.google.com/chrome-developer-tools/>

Mozilla Firefox

<https://developer.mozilla.org/en-US/docs/Tools>

Mechanize

Biblioteca utilizada para automatizar interação com websites.

<http://mechanize.rubyforge.org/>

Mechanize

```
require 'mechanize'

agent = Mechanize.new

# Example of HTTP GET request
agent.get 'http://www.google.com'

form = agent.page.forms[0]
form['q'] = 'poli usp'

# Example of form submission
form.submit
```


Parsing

Processar um documento em busca de conteúdos específicos.

Ferramentas úteis para parsing de HTML: seletores CSS, XPath e Regular Expressions.

Extração de dados

poli usp - Google Search

https://www.google.com/search?q=poli+usp&oq=poli+usp&aqs=chrome..69i57j69i61.1119j0j1&sourceid=chrome&es_sm=91&ie=UTF-8

Google poli usp

Web Images Maps Shopping News More Search tools

About 884,000 results (0.33 seconds)

div#resultStats 245px x 43px

Escola Politécnica da USP

www... Translate this page Polytechnic School of the University of São Paulo



Escola Politécnica da USP - Formando engenheiros e líderes. ... Carro da Poli/USP é o primeiro em velocidade máxima na competição Baja SAE Brasil

4.3 ★★★★★ 22 Google reviews · Write a review

Avenida Prof Luciano Gualberto, 380, travessa 3 - Butantã, São Paulo - SP, 05508-010, Brazil
+55 11 3091-5295

Ensino
A excelência do ensino na Poli é reconhecida no Brasil e no ...

Concursos e processo seleti...
... Comunicação · Início · Acesso
Rápido Concursos e processos ...



Polytechnic School of the University of São Paulo

Elements Network Sources Timeline Profiles Resources Audits Console

```
<script>...</script>
<div class="mw">...</div>
<div data-jibp="h" data-jiis="uc" id="bst" style="display:none">...</div>
<div data-jibp="h" data-jiis="uc" id="top_nav">...</div>
<div data-jibp="h" data-jiis="uc" id="appbar">
  <div id="extabar">
    <div id="topabar" style="position:relative">
      <div class="ab_tnav_wrp" id="slim_appbar">
        <div id="sbfrm_l">
          <div id="resultStats">
            "About 884,000 results"
            <nobr> (0.33 seconds)&nbsp;</nobr>
          </div>
        </div>
      </div>
    </div>
    <div id="botabar" style="display:none"></div>
  </div>
</div>
```

html body#gpr.srp.tbo.vasq.vsh div#main div#cnt.mdm div#appbar div#extabar div#topabar div#slim_appbar.ab_tnav_wrp div#sbfrm_l div#resultStats

Styles Computed Event Listeners

element.style {

body.vasq search?q=poli+u...91&ie=UTF-8:10

#resultStats {

line-height: 43px;

#resultStats { search?q=poli+u...1&ie=UTF-8:215

color: #808080;

#resultStats { search?q=poli+u...91&ie=UTF-8:10

padding-left: 16px;

padding-top: 0;

padding-bottom: 0;

padding-right: 8px;

#resultStats { search?q=poli+u...91&ie=UTF-8:10

Parsing: Nokogiri

```
require 'mechanize'

agent = Mechanize.new

# Example of HTTP GET request
agent.get 'http://www.google.com'

form = agent.page.forms[0]
form['q'] = 'poli usp'

# Example of form submission
form.submit

doc = agent.page.parser
selector = '#resultStats'
puts doc.css(selector).text
```


Parsing: XPath

```
require 'mechanize'

agent = Mechanize.new

# Example of HTTP GET request
agent.get 'http://www.google.com'

form = agent.page.forms[0]
form['q'] = 'poli usp'

# Example of form submission
form.submit

doc = agent.page.parser
selector = '//div[@id="resultStats"]'
puts doc.xpath(selector).text
```


Parsing: Regular Expression

```
require 'mechanize'

agent = Mechanize.new

# Example of HTTP GET request
agent.get 'http://www.google.com'

form = agent.page.forms[0]
form['q'] = 'poli usp'

# Example of form submission
form.submit

regex = /(Aproximadamente [0-9.]* resultados)/
match = agent.page.body.match(regex)
puts match[1]
```


Armazenar e
manipular dados

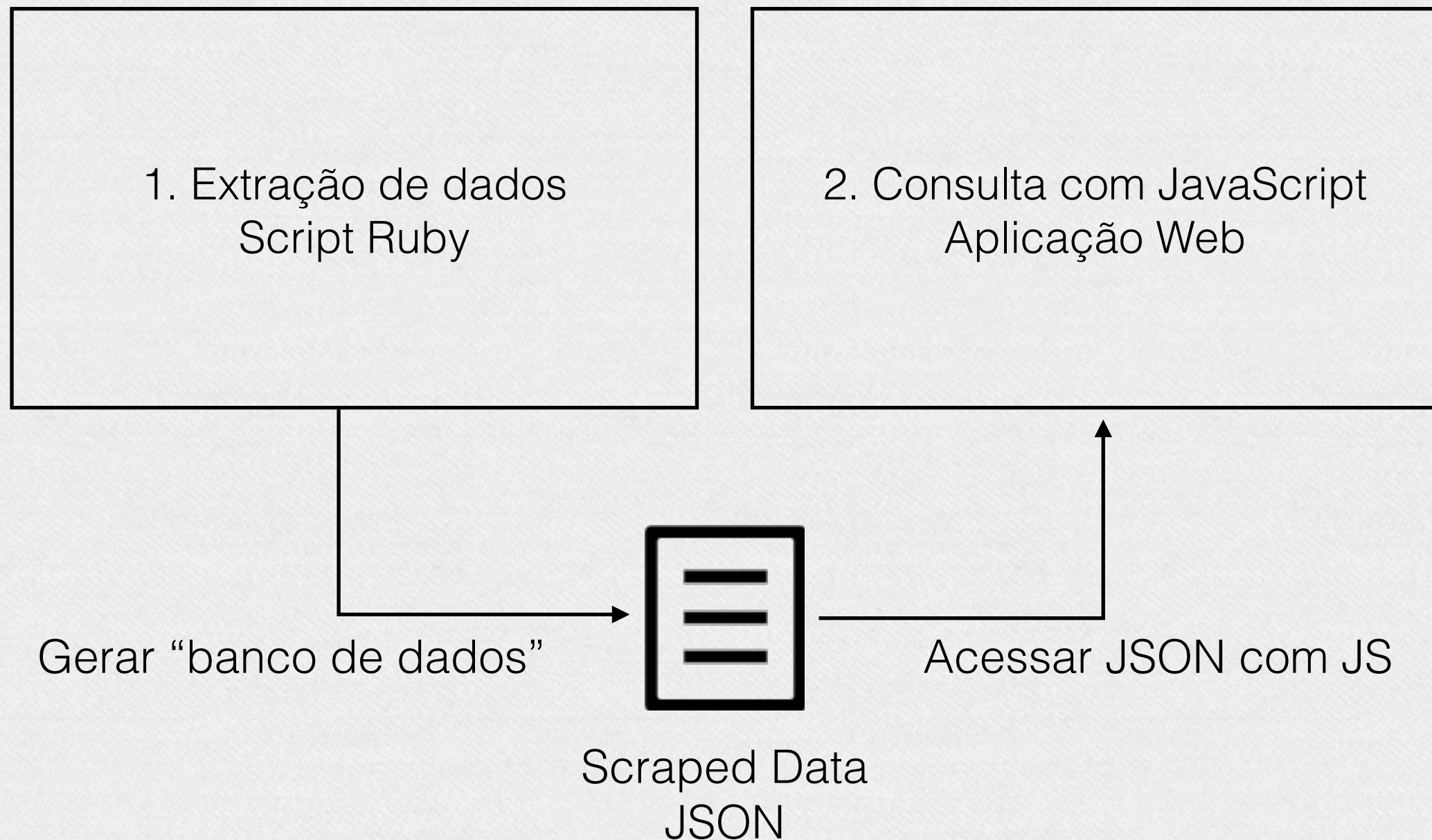
Extrair dados da
internet

Continuação do
projeto final

PROJETO FINAL

<https://gomockingbird.com/mockingbird/#vneou5d>

Dividir para conquistar



Dividir para conquistar

1. Extração de dados
Script Ruby

2. Consulta com JavaScript
Aplicação Web

Fazer um script Ruby capaz de extrair as informações do portal de estágios do PCS e gerar um arquivo no formato JSON com as informações estruturadas.

Dividir para conquistar

1. Extração de dados
Script Ruby

2. Consulta com JavaScript
Aplicação Web

Inserir os dados formatados em JSON no JavaScript da aplicação e criar funções para manipulá-los.