
Introdução a Web Scraping

Marcos Castro e Thomas William

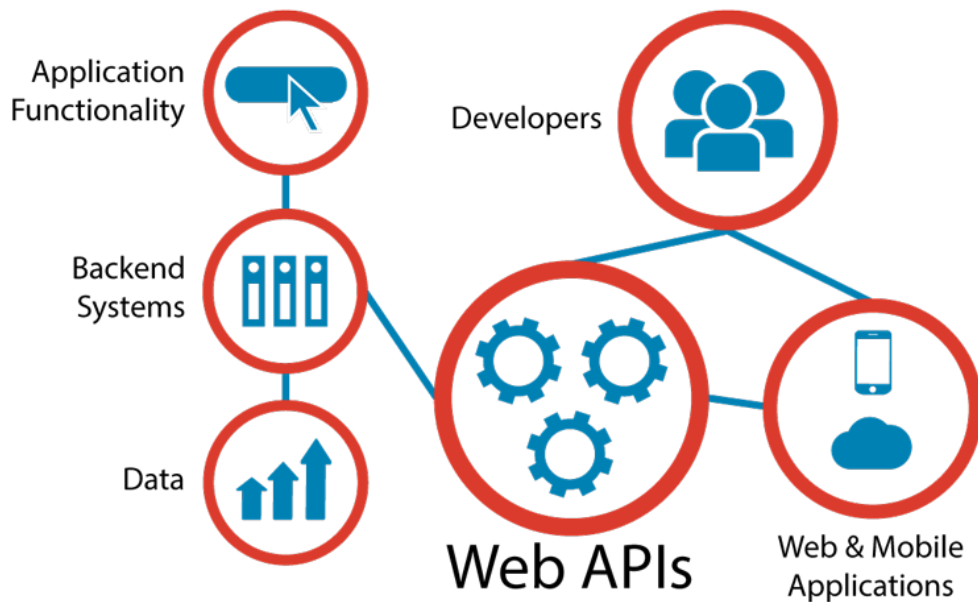
Web Scraping

Um *web scraping* é uma API para extrair dados de um web site



API

API é um conjunto de rotinas e padrões de programação para acesso a um aplicativo de software ou plataforma baseado na Web.

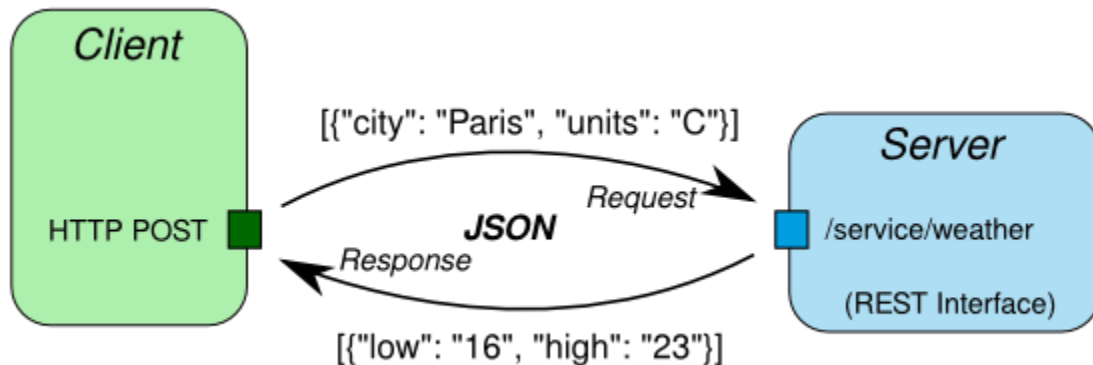


FORMATOS



JSON

Json é uma formatação de texto de troca de dados e completamente independente de linguagem.



JSON

```
3▼ {  
4   "title": "Example Schema",  
5   "type": "object",  
6▼  "properties": {  
7      "firstName": {  
8          "type": "string"  
9      },  
10     "lastName": {  
11         "type": "string"  
12     },  
13▼    "age": {  
14        "description": "Age in years",  
15        "type": "integer",  
16        "minimum": 0  
17    }  
18  },  
19  "required": ["firstName", "lastName"]  
20 }
```

XML

O XML é um formato para a criação de documentos com dados organizados de forma hierárquica.



XML

```
2 <!DOCTYPE glossary PUBLIC "-//OASIS//DTD DocBook V3.1//EN">
3 <glossary><title>example glossary</title>
4   <GlossDiv><title>S</title>
5     <GlossList>
6       <GlossEntry ID="SGML" SortAs="SGML">
7         <GlossTerm>Standard Generalized Markup Language</GlossTerm>
8         <Acronym>SGML</Acronym>
9         <Abbrev>ISO 8879:1986</Abbrev>
10        <GlossDef>
11          <para>A meta-markup language, used to create markup
12 languages such as DocBook.</para>
13          <GlossSeeAlso OtherTerm="GML">
14            <GlossSeeAlso OtherTerm="XML">
15              </GlossDef>
16            <GlossSee OtherTerm="markup">
17              </GlossEntry>
18            </GlossList>
19          </GlossDiv>
20        </glossary>
```


OBRIGADO !