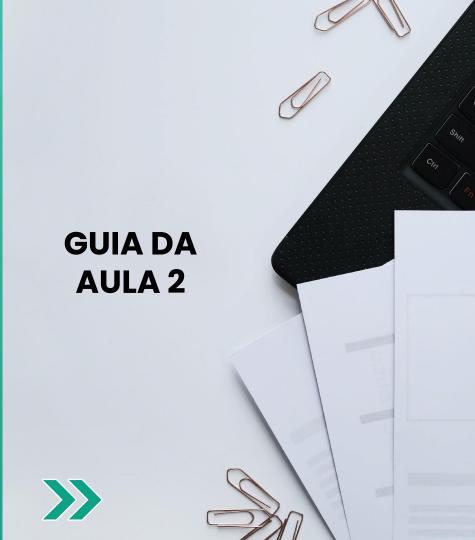


## Profissão: Analista de dados





## **COLETA DE DADOS II**







# Extraia dados da web (scrapping)



- Pacote beautifulsoup4
- Web Scrape



Acompanhe aqui os temas que serão tratados na videoaula





### **Web Scraping**

#### **Formato HTML**

Um arquivo texto semi-estruturado, organizado por tags,

**Exemplo:** Arquivo lotr.html

```
In [ ]:
         %%html
         <html>
          <head>
            <!-- metadata -->
          </head>
          <body>
           <h3>Senhor dos Anéis</h3>
           Filmes:
           <111>
                     <b>2001:</b> 0 Senhor dos Anéis: A Sociedade do Anel
                     <b>2002:</b> 0 Senhor dos Anéis: As Duas Torres
                     <b>2003:</b> 0 Senhor dos Anéis: 0 Retorno do Rei
                \langle\langle\langle \mid a \rangle\rangle\rangle
            </body>
        </html>
```



```
In [ ]:
```

```
%%writefile lotr.html
<html>
 <head>
  <!-- metadata -->
 </head>
 <body>
  <h3>Senhor dos Anéis</h3>
  Filmes:
  <01>
       <b>2001:</b> 0 Senhor dos Anéis: A Sociedade do Anel
       <b>2002:</b> 0 Senhor dos Anéis: As Duas Torres
       <b>2003:</b> 0 Senhor dos Anéis: 0 Retorno do Rei
  </body>
</html>
```



#### Pacote beautifulsoup4

Pacote Python para extrair informações de arquivos HTML.



```
Exemplo: Extrair os filmes e anos do arquivo lotr.html em um dicionário.
```

```
In []:
    from bs4 import BeautifulSoup
        pagina = BeautifulSoup(open('lotr.html', mode='r'), 'html.parser')

In []:
    filmes_li = pagina.find_all('li')
        print(filmes_li)

In []:
    print(list(set(map(lambda filme_li: type(filme_li), filmes_li))))
```







```
filmes = []

for filme_li in filmes_li: filme =
    filme_li.get_text()
    ano = int(filme.split(sep=':')[0].strip())
    titulo = ':'.join(filme.split(sep=':')[1:]).strip()
    filmes.append({'ano': ano, 'titulo': titulo})

for filme in filmes: print(filme)
```

#### **Web Scrape**

Aplicação que extrai conteúdo de páginas web de forma automatizada, em geral é aplicado após o processo de web crawling.



## **Exemplo:** Extrair todo o texto da página da Wikipédia sobre web crawlers e contar a ocorrência da palavra crawler



```
In [ ]:
         URL = 'https://en.wikipedia.org/wiki/Web crawler'
          conteudo = crawl website(url=URL)
          with open (file='wiki.html', mode='w', encoding='utf8') as arquivo:
           arquivo.write(conteudo)
In [ ]:
         from bs4 import BeautifulSoup
         pagina = BeautifulSoup(open('wiki.html', mode='r'), 'html.parser')
In [ ]:
         texto = pagina.get text()
          print(texto)
In [ ]:
         import re
         ocorrencias = len(re.findall('crawler', texto, re.IGNORECASE))
         print(ocorrencias)
```

