

Módulo | Análise de Dados: Coleta de Dados

Caderno de Exercícios

Professor André Perez

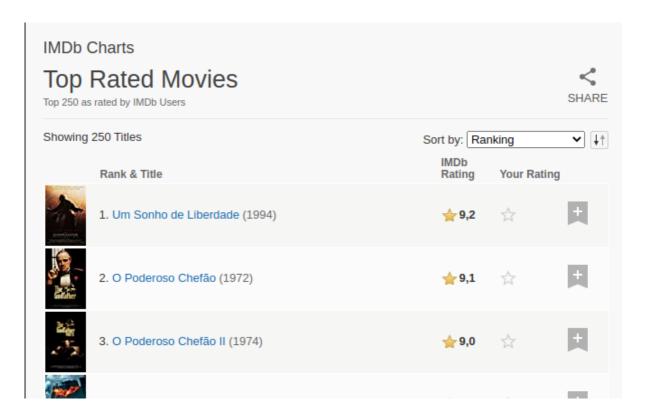
Tópicos

- 1. Web Crawling;
- 2. Web Scraping;
- 3. Web API.

Exercícios

1. Filmes populares do IMDB

O IMDB é um famoso site de reviews de filmes e seriados. Uma das páginas mais acessadas do website é o ranking de filmes mais bem votados. Neste exercício, vamos extrair informações deste website:



1.1. Arquivo Robots.txt

Utilize o pacote Python requests para fazer o download do conteúdo do arquivo robots.txt do site do IMDB e salve numa variável chamada robots, este é o link:

```
https://www.imdb.com/robots.txt
```

Com o conteúdo na variável robots , verifique se a palavra top ou charts está presente no conteúdo do texto. Se sim, imprima True , senão imprima False .

```
In []: # solução do exercício 1.1
```

Dica: Você pode colar o endereço do arquivo robots.txt no seu navegador para visualizar o conteúdo do arquivo.

1.2. Crawling & Scraping

Utilize os pacotes Python requests e beautifulsoup4 para extrair os 10 filmes mais populares do IMDB (titulo, ano e nota), este é o link:

```
https://www.imdb.com/chart/top/
```

Escreva os dados extraídos no arquivo csv imdb.csv separado por ; no seguinte formato:

```
ranking;titulo;ano;nota
1;The Shawshank Redemption;1994;9.2
2;The Godfather;1972;9.1
3;The Godfather: Part II;1974;9.0
```

```
In []: # a) Utilize o pacote requests para fazer o download da página na variável
# conteudo
import requests
```

```
from requests.exceptions import HTTPError
         conteudo = None
         URL = 'https://www.imdb.com/chart/top/'
         ... # continue o codigo aqui
In [ ]:
         # b) Utilize o pacote beautifulsoup4 para carregar o HTML da variavel
         # conteudo na variavel pagina
         from bs4 import BeautifulSoup
         pagina = ... # continue o codigo aqui
In [ ]:
         # c) Utilize o código abaixo para iterar nas linhas e colunas da tabela e
         # preencher a variavel conteudo extraido
         conteudo extraido = []
         tabela = pagina.find('table', {'class': 'chart'})
         for linha in tabela.find all('tr'):
           textos_coluna = list()
           for coluna in linha.find all('td'):
             texto coluna = coluna.get text().strip().split('\n')
             textos coluna += texto coluna
           print(textos coluna)
           ... # continue o codigo aqui
```

Dica: O código na letra c já extrai o conteúdo das linhas na lista textos_coluna, basta que você extraia o conteúdo de interesse dela. Como exemplo:

```
['', '1.', ' The Shawshank Redemption', '(1994)', '9.2', '12345678910 ', '', '', 'NOT YET RELEASED', ' ', '', 'Seen', ''] ['', '2.', ' The Godfather', '(1972)', '9.1', '12345678910 ', '', '', '', 'NOT YET RELEASED', ' ', '', 'Seen', ''] ['', '3.', ' The Godfather: Part II', '(1974)', '9.0', '12345678910 ', '', '', 'NOT YET RELEASED', ' ', '', 'Seen', ''] ['', '4.', ' The Dark Knight', '(2008)', '9.0', '12345678910 ', '', '', '', '', 'NOT YET RELEASED', ' ', '', 'Seen', ''] ['', '5.', ' 12 Angry Men', '(1957)', '8.9', '12345678910 ', '', '', '', '', 'NOT YET RELEASED', ' ', '', 'Seen', '']

In []: # d) Escreva o arguivo imdb.csv com o conteudo da variavel conteudo extraido
```

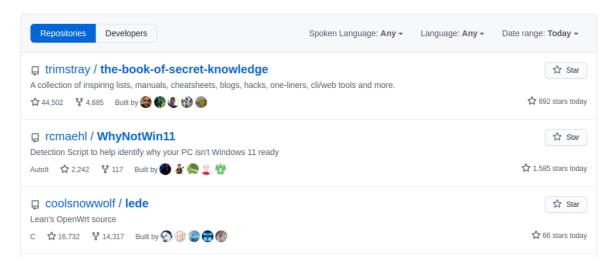
2. Bônus: Projeto em destaque do GitHub

Nota: Este exercício não é obrigatório.

... # continue o codigo aqui

O GitHub é o maior repositória de código aberto na internet. Nele, você pode encontrar o código fonte de diversos projetos, alguns inclusive utilizamos em nossas aulas, como o

Pandas. O GitHub apresenta uma página de projetos em destaque, que são os projetos que estão recebendo muita atenção da comunidade:



Utilize os pacotes Python requests e beautifulsoup4 para extrair os 10 projetos mais populares do GitHub, este é o link:

https://github.com/trending

Escreva os dados extraídos no arquivo csv github.csv separado por ; no seguinte formato:

```
ranking;project;language;stars;stars_today;forks
1;the-book-of-secret-knowledge;;44502;692;4685
2;whynotwin11;autoit;2242;1585;117
3;lede;c;16732;66;14317
```

Nota: Confira o arquivo robots.txt do website.

```
In [ ]: # solução do exercício 2
```