



escola
britânica de
artes criativas
& tecnologia

Módulo | Análise de Dados: Coleta de Dados II

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:

IMDb Charts

Top Rated Movies

Top 250 as rated by IMDb Users

Showing 250 Titles

Sort by: Ranking

Rank & Title	IMDb Rating	Your Rating
1. Um Sonho de Liberdade (1994)	★ 9,2	☆ +
2. O Poderoso Chefão (1972)	★ 9,1	☆ +
3. O Poderoso Chefão II (1974)	★ 9,0	☆ +

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 (título, 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 código aqui

```

```

In [ ]: # b) Utilize o pacote BeautifulSoup4 para carregar o HTML da variável
# conteudo na variável pagina

from bs4 import BeautifulSoup

pagina = ... # continue o código aqui

```

```

In [ ]: # c) Utilize o código abaixo para iterar nas linhas e colunas da tabela e
# preencher a variável 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 código 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 arquivo imdb.csv com o conteúdo da variável conteudo_extraido

... # continue o código aqui

```

2. Bônus: Projeto em destaque do GitHub

Nota: Este exercício não é obrigatório.

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:

The screenshot shows the GitHub trending page with filters for 'Repositories', 'Spoken Language: Any', 'Language: Any', and 'Date range: Today'. It lists three trending repositories:

Repository	Description	Stars	Forks	Builds	Stars Today
trimstray / the-book-of-secret-knowledge	A collection of inspiring lists, manuals, cheatsheets, blogs, hacks, one-liners, cli/web tools and more.	44,502	4,685	5	692 stars today
rcmaehl / WhyNotWin11	Detection Script to help identify why your PC isn't Windows 11 ready	2,242	117	5	1,585 stars today
coolsnowwolf / lede	Lean's OpenWrt source	16,732	14,317	5	66 stars today

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
```