

# Movie Web Scraper

Lucas Broering dos Santos





# Domínio do Problema

O domínio do problema escolhido é o de review de filmes. Envolve a coleta e apresentação de dados relacionados a filmes de dois sites famosos, Metacritic e The Movie Database. Existem algumas particularidades nesse domínio, a diversidade de dados e como os mesmos são apresentados, cada site tem uma metodologia diferente para avaliar os filmes. Existe também a questão que novos filmes são lançados todos os dias, então há atualização constante desses dados.

Entidades do mundo real envolvidas:

- Filmes
- Avaliações
- Usuários
- Plataformas de Avaliação



# Domínio do Problema

## Público Interessado

- Cinéfilos
- Críticos de Cinema
- Indústria do Entretenimento
- Desenvolvedores

Ao final do trabalho se espera criar um scraping eficiente que navegue por diversas páginas relacionadas a filmes e extraia informações relevantes. Fazendo a limpeza dos dados e armazenando eles em um formato acessível (json).

O código faz um trabalho de crawler para percorrer as páginas e extractor para extrair as informações da estrutura html.



# Objetivo

Receber páginas html como entrada e gerar um json com as informações limpas e organizadas.

# Exemplo dos dados de entrada

Foram utilizadas duas url diferentes no trabalho, do metacritic e The Movie Database (TMDB)



The screenshot shows the TMDB page for the movie 'Inside Out 2 (2024)'. At the top, there are navigation tabs: Overview (selected), Media, Fandom, and Share. On the left is the movie poster. The main content area includes the title 'Inside Out 2 (2024)', release date '06/20/2024 (BR)', genres 'Animation, Family, Drama, Adventure, Comedy', and runtime '1h 37m'. Below this is a 'User Score' of 79% with a 'What's your Vibe?' button. There are icons for list, heart, bookmark, and a 'Play Trailer' button. The 'Overview' section features the tagline 'Make room for new emotions.' and a synopsis: 'Teenager Riley's mind headquarters is undergoing a sudden demolition to make room for something entirely unexpected: new Emotions! Joy, Sadness, Anger, Fear and Disgust, who've long been running a successful operation by all accounts, aren't sure how to feel when Anxiety shows up. And it looks like she's not alone.' At the bottom, credits are listed for Kelsey Mann (Director, Story), Meg LeFauve (Screenplay, Story), and Dave Holstein (Screenplay).

Overview ▾ Media ▾ Fandom ▾ Share ▾

**Inside Out 2 (2024)**  
06/20/2024 (BR) • Animation, Family, Drama, Adventure, Comedy • 1h 37m

79% User Score 🥰 😬 What's your Vibe? 0

📋 ❤️ 📌 ▶ Play Trailer

*Make room for new emotions.*

**Overview**

Teenager Riley's mind headquarters is undergoing a sudden demolition to make room for something entirely unexpected: new Emotions! Joy, Sadness, Anger, Fear and Disgust, who've long been running a successful operation by all accounts, aren't sure how to feel when Anxiety shows up. And it looks like she's not alone.

**Kelsey Mann**  
Director, Story

**Meg LeFauve**  
Screenplay, Story

**Dave Holstein**  
Screenplay

Exemplo de página do The Movie Database (TMDB).

# Exemplo dos dados de entrada

```
▼ <div class="header large border first">
  ▼ <div class="keyboard_s custom_bg"> flex
    ▼ <div class="single_column">
      ▼ <section id="original_header" class="images inner"> flex
        ▶ <div class="poster_wrapper false"> ... </div>
        ▼ <div class="header_poster_wrapper false"> flex
          ...
          ▼ <section class="header_poster"> flex == $0
            ▶ <div class="title ott_false" dir="auto"> ... </div> flex
            ▶ <div class="flex flex-col"> ... </div> flex
            ▶ <div class="header_info"> ... </div>
            </section>
          </div>
        </section>
```

Exemplo da estrutura html da página do The Movie Database (TMDB).

# Exemplo dos dados de entrada

Filters

RELEASE YEAR [RESET](#) 16,189 results

1910  2024

STREAMING SERVICES

☐ Netflix

☐ Max

☐ Starz

☐ Paramount+

☐ Hulu

+ Show More

RELEASE TYPE

☐ Coming Soon

☐ In Theaters

Games **Movies** TV Shows

Metascore

**2,329. A Love Song**  
JUL 29, 2022 • Rated PG  
Faye (Dale Dickey) is a lone traveler bidding her...

**2,330. Speed**  
JUN 10, 1994 • Rated TV-14  
Keanu Reeves stars as Jack Traven an LAPD...

**2,331. At the Ready**  
OCT 22, 2021  
Ten miles from the Mexican border, students...

**2,332. Detective Story**  
OCT 24, 1951 • Rated...  
On one day in the 21st Precinct squad room,...

**2,333. Non-Fiction**  
MAY 3, 2019 • Rated R  
Set amidst the bohemian intelligentsia of the...

**2,334. Jazz Fest: A...**  
MAY 13, 2022 • Rated PG-13  
The New Orleans Jazz & Heritage Festival...

**2,335. Eve's Bayou**  
NOV 7, 1997 • Rated R  
After a daughter witnesses her father having an affa...

**2,336. Miracle in Milan**  
DEC 17, 1951 • Rated Not...  
An open-hearted, unrelentingly energetic...

**2,337. Dr. No**  
MAY 8, 1963 • Rated TV-PG  
A resourceful British government agent seeks...

Exemplo de página do Metacritic.



# Exemplo dos dados de entrada

```
<main class="c-pageBrowse g-grid-container u-grid-columns"> grid
  <aside class="c-pageBrowse_filters lg:u-col-3"> ... </aside>
  <section class="c-pageBrowse_content lg:u-col-9">
    <div class="c-FinderTabbedHeader u-flexbox"> flex
      <div class="c-FinderTabbedHeader_tabs"> ... </div>
    </div>
    ...
    <div class="c-finderControls g-outer-spacing-bottom-medium"> == $0
      <div class="c-finderControls_buttons u-flexbox"> ... </div> flex
      <div class="c-finderControls_total u-flexbox g-inner-spacing-top-medium"> ... </div>
    </div>
    <div section="detailed|98" class="c-productListings"> ... </div>
    <div class="c-navigationPagination u-flexbox u-flexbox-alignCenter u-flexbox-justifyCenter g-outer-spacing-top-large g-outer-spacing-bottom-large u-text-center c-pageBrowse
      __smaller-bottom-spacing"> ... </div> flex
    <span class="c-pageBrowse_footnote">Titles with fewer than 7 critic reviews are excluded.</span>
```

Exemplo da estrutura html da página do Metacritic.



# Exemplo dos dados de saída

```
1  [
2  {
3    "ID": "1",
4    "TITLE": "Dekalog (1988)",
5    "LAUNCH YEAR": "Mar 22, 1996",
6    "RATED": "TV-MA",
7    "DESCRIPTION": "This masterwork by Krzysztof Kieślowski is one of the twentieth century's greatest achievements in visual storytelling. Orig
8    "SCORE": "100/100",
9    "URL": "https://www.metacritic.com/browse/movie/?releaseYearMin=1910&releaseYearMax=2024&page=1"
10 },
11 {
12   "ID": "2",
13   "TITLE": "The Leopard (re-release)",
14   "LAUNCH YEAR": "Aug 13, 2004",
15   "RATED": "PG",
16   "DESCRIPTION": "Set in Sicily in 1860, Luchino Visconti's spectacular 1963 adaptation of Giuseppe di Lampedusa's international bestseller is
17   "SCORE": "100/100",
18   "URL": "https://www.metacritic.com/browse/movie/?releaseYearMin=1910&releaseYearMax=2024&page=1"
19 },
20 {
21   "ID": "3",
22   "TITLE": "The Godfather",
23   "LAUNCH YEAR": "Mar 24, 1972",
24   "RATED": "R"
```

Exemplo de saída dos dados extraídos no metacritic.

# Exemplo dos dados de saída

```
[
  {
    "ID": "1",
    "TITLE": "Ariel",
    "CERTIFICATION": "K-12",
    "RELEASE": "21/10/1988 (FI)",
    "GENRES": "Drama,Comédia,Romance,Crime",
    "RUNTIME": "1h 13m",
    "DESCRIPTION": "Taisto Kasurinen trabalha numa mina de carvão que passa por graves problemas. Seu pai comete suicídio e ele acaba sendo preso.",
    "DIRECTOR": "Aki KaurismäkiDirector, Writer",
    "URL": "https://www.themoviedb.org/movie/2"
  },
  {
    "ID": "2",
    "TITLE": "Sombras no Paraíso",
    "CERTIFICATION": "S",
    "RELEASE": "17/10/1986 (FI)",
    "GENRES": "Drama,Comédia,Romance",
    "RUNTIME": "1h 14m",
    "DESCRIPTION": "Depois de perder seu amigo e colega de trabalho para um ataque cardíaco repentino, o solitário catador de lixo Nikander está",
    "DIRECTOR": "Aki KaurismäkiDirector, Writer",
    "URL": "https://www.themoviedb.org/movie/3"
  },
]
```

Exemplo de saída dos dados extraídos no The Movie Database (TMDB).

# Descrição sobre o trabalho

```
def metacritic_scraper(pages):
    response = []

    while page_count < pages:

        url = f'https://www.metacritic.com/browse/movie/?releaseYearMin=1910&releaseYearMax=2024&page={page_count}'

        page = requests.get(url, headers=headers)

        soup = BeautifulSoup(page.text, 'html.parser')

        movies = soup.find_all("div", class_="c-finderProductCard_info u-flexbox-column")

        for movie in movies:
            title = movie.find("h3", class_="c-finderProductCard_titleHeading").get_text(strip=True)

            launch_year_span = movie.find("span", class_="u-text-uppercase")
            launch_year = launch_year_span.get_text(strip=True) if launch_year_span else "N/A"

            rated_span = movie.find("span", class_="u-text-capitalize")
            rated = rated_span.parent.get_text(strip=True) if rated_span else "N/A"

            description_div = movie.find("div", class_="c-finderProductCard_description")
            description = description_div.get_text(strip=True) if description_div else "N/A"

            score_div = movie.find("div", class_="c-siteReviewScore u-flexbox-column u-flexbox-alignCenter u-flexbox-justifyCenter g-text-bold c-siteR")
            score = score_div.get_text(strip=True) if score_div else "N/A"
            score_formatted = score + "/100"

            data = {'ID': str(movie_count), 'TITLE': title[2:], 'LAUNCH YEAR': launch_year, 'RATED': rated[5:], 'DESCRIPTION': description, 'SCORE': s

            response.append(data)

            movie_count += 1

        page_count += 1

    print(f'Scraping complete - {movie_count}')

    return response
```

Ao lado está a parte de percorrer as páginas e fazer a extração dos dados significantes da estrutura html.



# Descrição sobre o trabalho

Essa função pega a response e transforma em um arquivo json, já organizado.

```
def json_file(response):  
    with open('INE5454 - Tópicos Especiais em Gerência de Dados/metacritic_data.json', 'w', encoding='utf-8') as json_file:  
        json.dump(response, json_file, ensure_ascii=False, indent=4)
```



# Movie Web Scraper

Tecnologias:

Utilizei python e as bibliotecas, BeautifulSoup, requests e json, para extrair o conteúdo da páginas, acessar elas e transformar em arquivo json, respectivamente.

Abordagem, e métodos:

Os dados foram coletados de páginas html. Pessoas da indústria cinematográfica ou desenvolvedores, seriam os principais interessados nos dados extraídos. Dados extraídos em inglês/português em formato html.



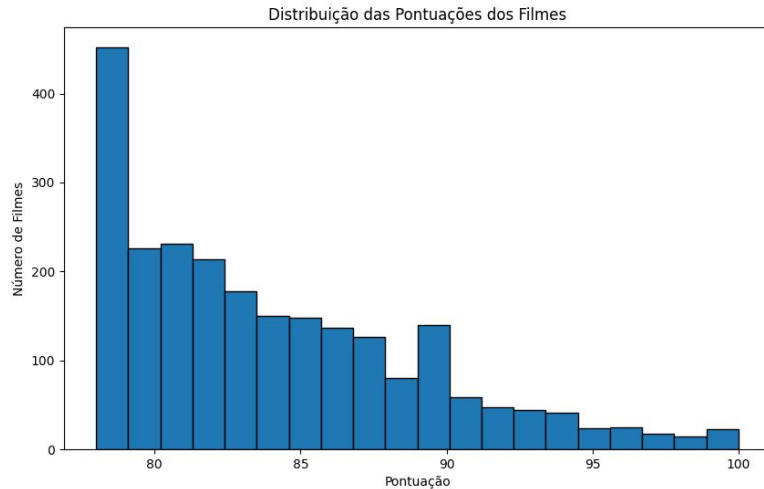
# Movie Web Scraper

Atributos coletados:

- TITLE - Qualitativo
- LAUNCH YEAR - Qualitativo
- RATED - Qualitativo
- DESCRIPTION - Qualitativo
- SCORE - Qualitativo
- URL - Qualitativo

# Análise dos resultados finais

Figure 1



Exemplo de gráfico que pode ser plotado utilizando os dados. Relaciona o número de filmes com a pontuação.



## Conclusão

Fazer o trabalho foi bem satisfatório para mim, nunca tinha mexido com web scraping e aprendi bastante. Acho que o mais difícil foi entender a estrutura html das url e fazer a extração corretamente, caminhando pela árvore html. Com certeza que se feito em maior escala, com mais url, seria um ótimo produto para o mercado cinematográfico, serviria fazer análises do que pode dar certo ou não no mundo do cinema, como analisar tendências de gêneros dos filmes.