

Web Scraper

INE 5454

Bianca Mazzuco Verzola (22202621)
Mariana Amaral Steffen (22200511)
Vitor Praxedes Calegari (22200379)

Sumário

- 1. Introdução**
- 2. Objetivos**
- 3. Entradas e Saídas**
- 4. Desenvolvimento do Crawler**
- 5. Aplicação Exemplo**
- 6. Análises**
- 7. Conclusão**

Domínio

- Ecossistema de plataformas de avaliação de filmes.
- Ambiente distribuído entre diferentes fontes, cada uma com seus padrões, estruturas e políticas de disponibilização de dados distintas.



Entidades

- Sites de classificação e avaliações sobre cinema;
- Filmes, avaliações, e dados relacionados a ambos.



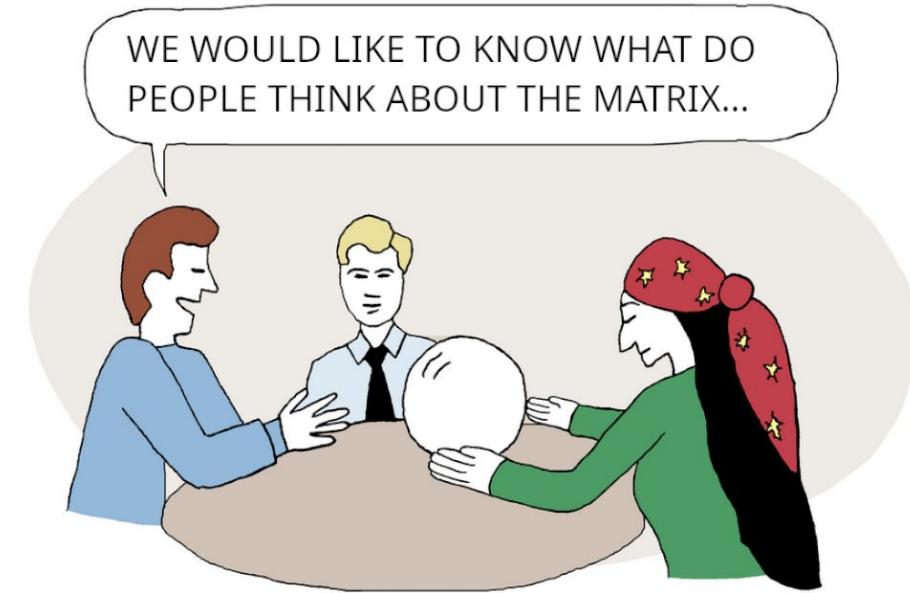
Perfis

- Entusiastas de cinema;
- Usuários que buscam recomendações mais informadas;
- Pessoas que desejam obter informações cinematográficas de forma mais simples.



Motivação

- Cada site apresenta dados diferentes;
- Integrar os dados das diferentes plataformas em um só local.
- Gosto pessoal dos integrantes do grupo;



Motivação

- Reunir tudo isso em um único local permite análises mais amplas:
 - Detectar discrepâncias entre crítica e público
 - Verificar onde um filme pode ser assistido;
 - Agregar estatísticas e alimentar aplicações.



Objetivo

- Criar um Web Crawler que, a partir de URLs iniciais de cada site, obtenha informações relevantes sobre filmes, suas avaliações e dados relacionados;

IMDb

Base de dados global sobre filmes, séries, elenco e equipe técnica, focada em informação detalhada da indústria audiovisual.

IMDb All

All topics

The Matrix

1999 · R · 2h 16m



Play trailer 2:26

499 299

19 VIDEOS 99+ PHOTOS

Action Epic Artificial Intelligence Cyberpunk Dystopian Sci-Fi Gun Fu Martial Arts

When a beautiful stranger leads computer hacker Neo to a forbidding underworld, he discovers the shocking truth--the life he knows is the elaborate deception of an evil cyber-intelligence.

8.7/10 2.2M Rate 178 ▾ 8

IMDB

Search HTML

```
<!DOCTYPE html>
<html class=" scriptsOn" lang="en-US" xmlns:og="http://opengraphprotocol.org/schema/" xmlns:fb="http://www.facebook.com/2008/fbml" style="--ipt-focus-outline-on-base: none; --ipt-focus-outline-on-baseAlt: none;"> [event]
  > <head>[...]</head>
  > <body class=""> [event]
    > <div>[...]</div>
    > <div id="__next"> [event]
      > <script>[...]</script>
      > <nav id="imdbHeader" class="sc-d620d5d7-0 bExpVV imdb-header celwidget" data-csa-c-id="99g6at-cs4emq-t7a2be-4bhj0l" data-cel-widget="imdbHeader">[...]</nav> [flex]
      > <script>[...]</script>
      > <script>[...]</script>
      > <main class="ipc-page-wrapper ipc-page-wrapper--base" role="main">
        > <div class="ipc-page-content-container ipc-page-content-container--full sc-b8154748-0 fAXDUN" role="presentation">
          > <div>[...]</div>
          > <script async="">[...]</script>
          > <section class="ipc-page-background ipc-page-background--base sc-358297d7-0 CHcbB">
            > <section class="ipc-page-background ipc-page-background--baseAlt sc-2fb918b7-0 kbBJjv atf-background-theme-dark" data-testid="atf-wrapper-bg">
              > <div id="ipc-wrap-background-id" style="position:absolute;width:100%;height:100%"></div>
              > <section class="ipc-page-background ipc-page-background--baseAlt inline20-page-background">[...]</section>
              > <div class="ipc-page-content-container ipc-page-content-container--center sc-2fb918b7-1 dT0xG" role="presentation"></div>
              > <script>[...]</script>
              > <div class="ipc-page-content-container ipc-page-content-container--center" role="presentation">
                > <section class="ipc-page-background ipc-page-background--baseAlt sc-14a487d5-0 gLkolc">
                  > <div class="sc-14a487d5-1 dyzPFD"></div>
                  > <section class="ipc-page-section ipc-page-section--baseAlt ipc-page-section--tp-none ipc-page-section--bp-xs sc-14a487d5-2 kmEeUD" data-testid="hero-parent">
                    > ::before
                    > <div class="sc-9194d746-0 qhHiIf">[...]</div>
```

pc-page-wrapper.ipc-page-wrapper-> div.ipc-page-content-container.ipc-page-> section.ipc-page-background.ipc-page-bac... > section.ipc-page-background.ipc-page-bac... > div.ipc-pag >

Letterboxd

Rede social de
cinéfilos voltada a
registrar, avaliar e
compartilhar
experiências pessoais
com filmes.

The screenshot shows a movie page for "The Matrix" on the Letterboxd website. At the top, there's a navigation bar with "SIGN IN", "CREATE ACCOUNT", "FILMS", "LISTS", "MEMBERS", "JOURNAL", and a search icon. Below the navigation is a large image of Keanu Reeves as Neo. Underneath the main image is a smaller thumbnail of the movie poster featuring Neo, Trinity, Agent Smith, and Persephone. To the right of the poster, the movie title "The Matrix" is displayed along with its release year "1999", director "Lana Wachowski, Lilly Wachowski", and a "Sign in to log, rate or review" button. Below the title is a "Share" button. A "RATINGS" section shows a bar chart with a rating of 4.2 and a row of names for the cast and crew. At the bottom of the page, there are buttons for "WHERE TO WATCH" (Amazon US, RENT, BUY, DISC) and "Trailer".

Letterboxd

Search HTML

```
<!DOCTYPE html>
<html id="html" class="no-mobile js context-client-not-app cssanimations backdropfi...xlegacy objectfit object-fit svg no-touchevents has-no-touch" lang="en" data-useragent="Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:145.0) Gecko/20100101 Firefox/145.0" data-platform="Linux x86_64">
  <head>::</head>
  <body class="film backdropped logged-out backdrop-loaded" data-type="film" data-tmdb-type="movie" data-tmdb-id="603"> (event)
    <div class="backdrop-container">
      <div id="backdrop" class="backdrop-wrapper backdrop-fade-in -loaded" data-backdrop="https://a.ltrbx.com/resized/sm/upload/o3/er/ey/ie/matrix-1200-1200-675-675-crop-000000.jpg?v=eea86a2195" data-support-custom="true" data-production-uid="film:51518" data-custom-backdrop-template-url="/ajax/backdrop/film/the-matrix/:backdropId/?v=2" data-backdrop2x="https://a.ltrbx.com/resized/sm/upload/o3/er/ey/ie/matrix-1920-1920-1080-1080-crop-000000.jpg?v=eea86a2195" data-backdrop-mobile="https://a.ltrbx.com/resized/sm/upload/o3/er/ey/ie/matrix-960-960-540-540-crop-000000.jpg?v=eea86a2195" data-offset="80">
        <div class="backdropimage js-backdrop-image" style="background-position: center -15px; background-image: url("ht.../ie/matrix-1200-1200-675-675-crop-000000.jpg?v=eea86a2195");"></div>
      <div class="backdropmask js-backdrop-fade">
        ::before
      </div>
    </div>
    <script type="module" crossorigin="" src="https://s.ltrbx.com/static/js/es/js/screens/production-C7QygMgu.js"></script>
    <link rel="modulepreload" crossorigin="" href="https://s.ltrbx.com/static/js/es/chunks/B50t9EMX.js">
    <link rel="modulepreload" crossorigin="" href="https://s.ltrbx.com/static/js/es/chunks/Duz0x7wZ.js">
    <link rel="modulepreload" crossorigin="" href="https://s.ltrbx.com/static/js/es/chunks/DHOGRMit.js">
    <link rel="modulepreload" crossorigin="" href="https://s.ltrbx.com/static/js/es/chunks/CmVCXTtj.js">
    <link rel="modulepreload" crossorigin="" href="https://s.ltrbx.com/static/js/es/chunks/bKpgursT.js">
    <link rel="modulepreload" crossorigin="" href="https://s.ltrbx.com/static/js/es/chunks/Di0w5U6i.js">
    <link rel="modulepreload" crossorigin="" href="https://s.ltrbx.com/static/js/es/chunks/C3e4t58V.js">
    <link rel="modulepreload" crossorigin="" href="https://s.ltrbx.com/static/js/es/chunks/Dxzbedgu.js">
    <link rel="modulepreload" crossorigin="" href="https://s.ltrbx.com/static/js/es/chunks/BOYOeiUl.js">
    <link rel="modulepreload" crossorigin="" href="https://s.ltrbx.com/static/js/es/chunks/DjoGib27.js">
    <link rel="modulepreload" crossorigin="" href="https://s.ltrbx.com/static/js/es/chunks/03U_i1H8.js">
  
```

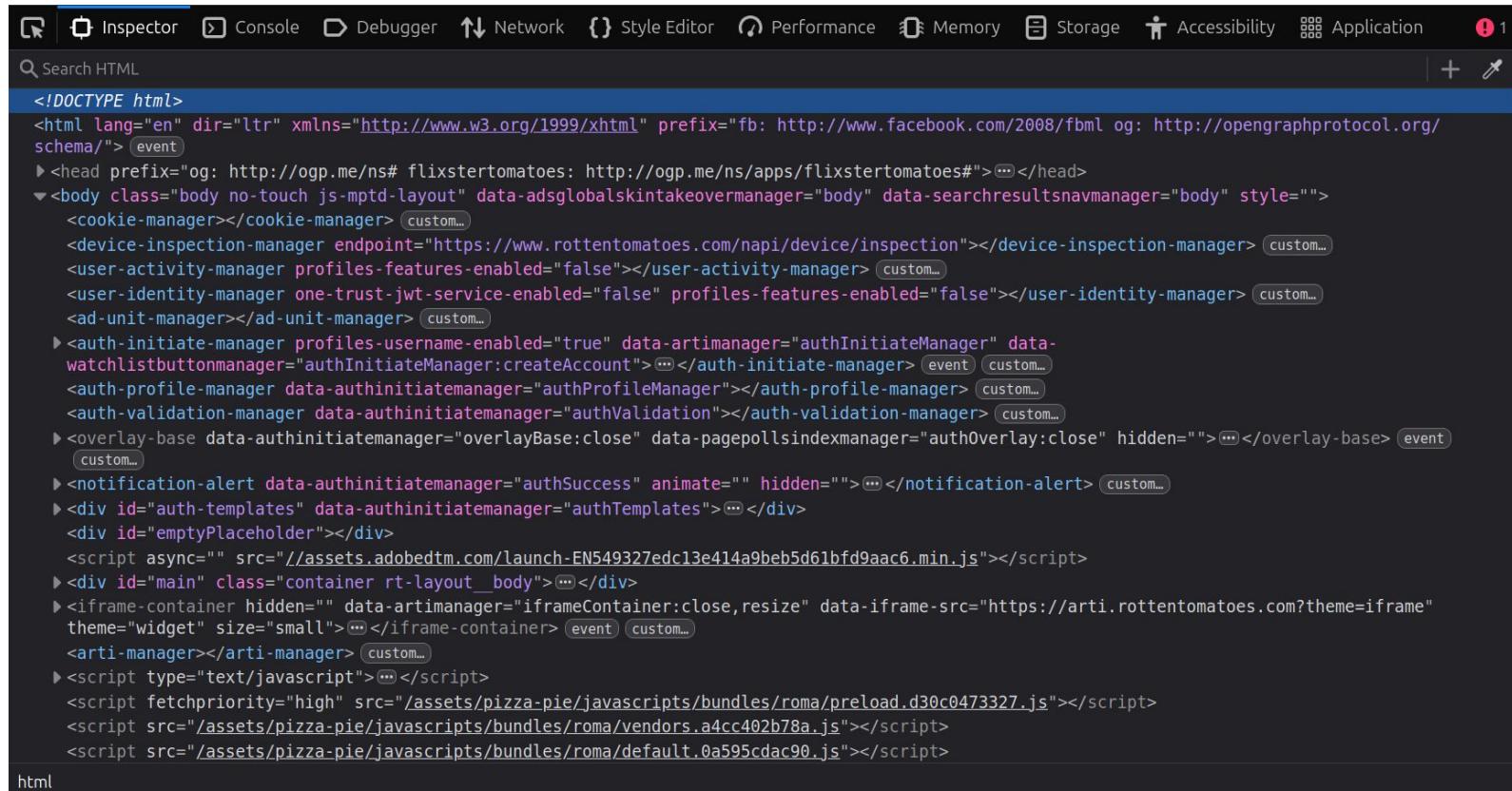
html#html.no-mobile.js.context-client-no... > body.film.backdropped.logged-out.backdro...

Rotten Tomatoes

Plataforma que agrupa críticas profissionais e avaliações do público para calcular índices de aprovação de filmes e séries.

The screenshot shows the Rotten Tomatoes homepage with a search bar at the top. Below it, a banner for 'TRENDING ON RT' includes links for 'Holiday Streaming', 'Renewed and Cancelled TV', 'Watch for Free', and 'The Rotten Tomatoes App'. A large image of Trinity from 'The Matrix' is on the right. The main content area features 'The Matrix' (1999) with a 2h 16m runtime, described as a Sci-Fi/Action/Mystery & Thriller. A 'WATCHLIST' button is present. Below the movie info, there's a smaller image of the movie poster and two rating boxes: 'Tomatometer' at 83% (209 reviews) and 'Popcornmeter' at 85% (250,000+ ratings). A quote from Neo is shown: 'Neo believes that Morpheus, an elusive figure considered to be the most dangerous man alive, can answer his question – What is the Matrix? Neo...'. Buttons for 'Watch on Fandango at Home' and 'STREAM NOW' are below the quote. At the bottom, tabs for 'Where to Watch', 'What to Know', 'Reviews', 'Cast & Crew', 'More Like This', and 'Related' are visible. A 'Where To Watch' section lists platforms like Amazon, Apple TV, Google Play, HBO Max, and Telecine. On the right sidebar, sections for 'What to Watch' (In Theaters and At Home), 'Most popular' (with icons and titles like 'Five Nights at Freddy's', 'Train Dreams', etc.), and 'Oh. What. Fun.' are displayed.

Rotten Tomatoes



The screenshot shows the Chrome DevTools interface with the "Elements" tab selected. The top navigation bar includes tabs for Inspector, Console, Debugger, Network, Style Editor, Performance, Memory, Storage, Accessibility, Application, and a refresh icon. A search bar labeled "Search HTML" is present. The main area displays the HTML code of the Rotten Tomatoes homepage, with various components and scripts highlighted in purple. The code includes DOCTYPE, head, body, script, and style elements, along with numerous custom and event handlers.

```
<!DOCTYPE html>
<html lang="en" dir="ltr" xmlns="http://www.w3.org/1999/xhtml" prefix="fb: http://www.facebook.com/2008/fbml og: http://opengraphprotocol.org/schema/"> [event]
  > <head prefix="og: http://ogp.me/ns# flixstertomatoes: http://ogp.me/ns/apps/flixstertomatoes#">[...]
```

```
    <body class="body no-touch js-mptd-layout" data-adsglobalskintakeovermanager="body" data-searchresultsnavmanager="body" style="">
      <cookie-manager></cookie-manager> [custom...]
      <device-inspection-manager endpoint="https://wwwrottentomatoes.com/napi/device/inspection"></device-inspection-manager> [custom...]
      <user-activity-manager profiles-features-enabled="false"></user-activity-manager> [custom...]
      <user-identity-manager one-trust-jwt-service-enabled="false" profiles-features-enabled="false"></user-identity-manager> [custom...]
      <ad-unit-manager></ad-unit-manager> [custom...]
    > <auth-initiate-manager profiles-username-enabled="true" data-artimanager="authInitiateManager" data-watchlistbuttonmanager="authInitiateManager:createAccount">[...]
```

```
      </auth-initiate-manager> [event] [custom...]
      <auth-profile-manager data-authinitiatemanager="authProfileManager"></auth-profile-manager> [custom...]
      <auth-validation-manager data-authinitiatemanager="authValidation"></auth-validation-manager> [custom...]
    > <overlay-base data-authinitiatemanager="overlayBase:close" data-pagepollsindexmanager="authOverlay:close" hidden="">[...]
```

```
      </overlay-base> [event] [custom...]
    > <notification-alert data-authinitiatemanager="authSuccess" animate="" hidden="">[...]
```

```
      </notification-alert> [custom...]
    > <div id="auth-templates" data-authinitiatemanager="authTemplates">[...]
```

```
      </div>
      <div id="emptyPlaceholder"></div>
      <script async="" src="//assets.adobedtm.com/launch-EN549327edc13e414a9beb5d61bfd9aac6.min.js"></script>
    > <div id="main" class="container rt-layout__body">[...]
```

```
      </div>
      <iframe-container hidden="" data-artimanager="iframeContainer:close,resize" data-iframe-src="https://arti.rottentomatoes.com?theme=iframe" theme="widget" size="small">[...]
```

```
        </iframe-container> [event] [custom...]
      <artimanager></artimanager> [custom...]
    > <script type="text/javascript">[...]
```

```
      </script>
      <script fetchpriority="high" src="/assets/pizza-pie/javascripts/bundles/roma/preload.d30c0473327.js"></script>
      <script src="/assets/pizza-pie/javascripts/bundles/roma/vendors.a4cc402b78a.js"></script>
      <script src="/assets/pizza-pie/javascripts/bundles/roma/default.0a595cdac90.js"></script>
```

html

JSON Gerado: movies.json

- **url** – Endereço web da página oficial do filme usada como fonte dos dados. qualitativo
- **título** – Nome oficial pelo qual o filme é conhecido. qualitativo
- **gêneros** – Categorias que caracterizam o estilo ou tema do filme. qualitativo
- **data de lançamento** – Quando o filme estreou oficialmente. quantitativo
- **classificação indicativa** – Faixa etária recomendada pelos órgãos reguladores. qualitativo

JSON Gerado: movies.json

- **sinopse** – Resumo breve da história do filme. qualitativo
- **duração** – Tempo total do filme em minutos. quantitativo
- **diretor** – Pessoa (ou pessoas) responsável pela direção do filme. qualitativo
- **elenco** – Lista dos principais atores e atrizes envolvidos. qualitativo
- **onde assistir** – Plataformas em que o filme está disponível, com links diretos. qualitativo

JSON Gerado: movies.json

- **link do poster** – URL da imagem oficial de divulgação do filme. qualitativo
- **nota média dos críticos e dos usuários** quantitativo
- **taxa de recomendação dos críticos e usuários** quantitativo
- **quantidade de reviews de críticos e usuário** quantitativo
- **reviews de críticos e usuário** – Lista das avaliações de críticos e usuários, incluindo nota, texto e data.
 - quantitativo
 - qualitativo
 - quantitativo

JSON Gerado: movies.json

- Estrutura do arquivo JSON gerado a partir do crawling:

```
movie_dict = {
    "url": list[str],
    "titulo": str,
    "generos": list[str],
    "data de lançamento": str | None,
    "classificacao indicativa": str | None,
    "sinopse": str | None,
    "duracao": str | None,
    "diretor": list[str],
    "elenco": list[str],
    "onde assistir": list[
        {
            "plataforma": str,
            "link": str
        }
    ],
    "link do poster": str | None,
    "nota média dos críticos": float | None,
    "taxa de recomendação dos críticos": float | None,
    "quantidade de reviews de críticos": int | None,
    "reviews de críticos": list[
        {
            "avaliação (nota até 10)": float | None,
            "texto": str | None,
            "data": str | None,
        }
    ],
    "nota média dos usuários": float | None,
    "taxa de recomendação dos usuários": float | None,
    "quantidade de reviews de usuários": int | None,
    "reviews de usuários": list[
        {
            "avaliação (nota até 10)": float | None,
            "texto": str | None,
            "data": str | None
        }
    ]
}
```

JSON Gerado: movies.json

Entradas e Saídas

```
"onde assistir": [
  {
    "plataforma": "HBO Max",
    "link": "https://play.max.com/video/watch/9fb854be-3d3c-44d8-a7c0-d3122b989be8"
  },
  {
    "plataforma": "Prime Video",
    "link": "https://www.primevideo.com/detail/amzn1.dv.gti.dea9f6b7-e0a0-38c9-7467-d6079c6fb4f0"
  }
],
"link do poster": "https://m.media-amazon.com/images/M/MV5BN2NmN2VhMTQtMDNiOS00NDlhLTlMjgtODE2ZTY0ODQyNDRhXkE",
"nota média dos críticos": 7.3,
"taxa de recomendação dos críticos": null,
"quantidade de reviews de críticos": 36,
"reviews de críticos": [
  {
    "avaliação (nota até 10)": 9.1,
    "texto": "The Matrix slams you back in your chair, pops open your eyes and leaves your jaw hanging slackly.",
    "data": null
  },
  {
    "avaliação (nota até 10)": 9.0,
    "texto": "The Wachowskis do it so playfully well, keeping The Matrix's potentially confusing plot interestingly mysterious.",
    "data": null
  }
],
"nota média dos usuários": 8.7,
"taxa de recomendação dos usuários": null,
"quantidade de reviews de usuários": 2208724,
"reviews de usuários": [
  {
    "avaliação (nota até 10)": 10.0,
    "texto": "When this came out, I was living with a roommate. He went out and saw it, came home and said it was the best movie he had ever seen. I have never seen a movie that has affected me like that since.",
    "data": "2019-10-31"
  },
  {
    "avaliação (nota até 10)": 10.0,
    "texto": "The Matrix - 1999This was a real change in filmmaking. Like watching it again in 2020, i.e. it still holds up well and is as good as ever.",
    "data": "2020-09-30"
  }
],
```

JSON Gerado: movies_united.json

- Como os dados vêm de vários sites, o mesmo filme pode aparecer mais de uma vez.
- Detecção de filmes iguais via **similaridade de título (fuzz.ratio)**.
- Filmes considerados iguais quando similaridade > 80%.
- Resultado: um único registro por filme.

JSON Gerado: movies_united.json

- **Mesmos atributos do movies.json, mas com unificação:**
 - **url** – Todos os endereços web do filme.
 - **título** – Um dos títulos.
 - **gêneros** – União de todos, sem duplicatas.
 - **data de lançamento** – Data mais antiga.
 - **classificação indicativa** – Uma das faixas etárias.
 - **sinopse** – Versão mais longa.
 - **duração** – Maior duração registrada.

JSON Gerado: movies_united.json

- **Mesmos atributos do movies.json, mas com unificação:**
 - **diretor** – União de todos, sem duplicatas.
 - **elenco** – União de todos, sem duplicatas.
 - **onde assistir** – União das plataformas, sem duplicatas.
 - **link do poster** – Um dos links.
 - **nota média dos críticos e dos usuários** – Média das médias dos sites.
 - **taxa de recomendação dos críticos e usuários** – Média das taxas dos sites.

JSON Gerado: movies_united.json

- **Mesmos atributos do movies.json, mas com unificação:**
 - **quantidade de reviews de críticos e usuário** – Soma das quantidades dos sites
 - **reviews de críticos e usuário** – União todas as avaliações, incluindo nota, texto, data e url do site de onde veio a review.

JSON Gerado: movies_united.json

- Novos atributos:
 - **notas dos críticos e dos usuários** – Notas médias de cada site, sem combinar.
 - **taxas de recomendação dos críticos e dos usuários** – Taxas de recomendação de cada site, sem combinar.
 - **quantidades de reviews dos críticos e dos usuários** – Quantidade registrada em cada site, sem combinar.

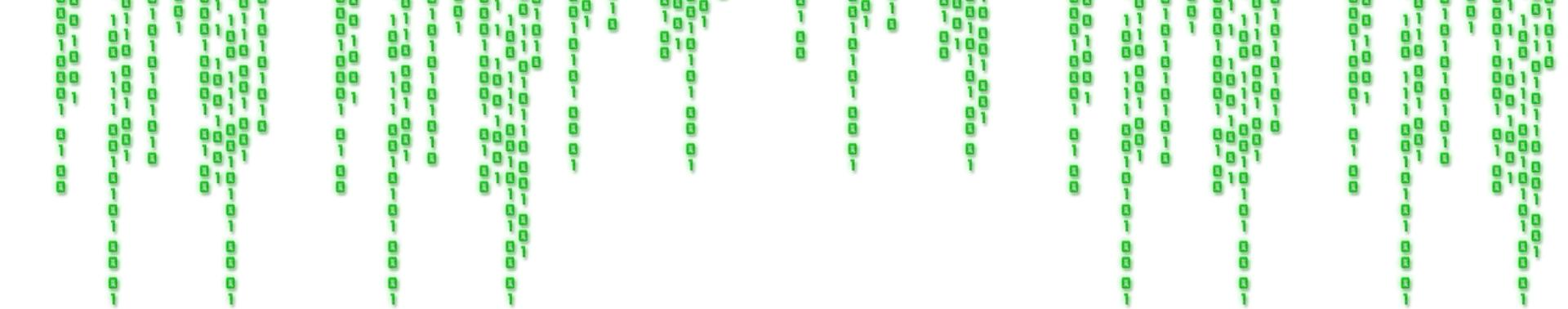
JSON Gerado: movies_united.json

- Estrutura do arquivo JSON gerado a partir da união dos filmes:

```
"nota média dos críticos": float | None,  
"notas dos críticos": list[  
    {  
        "link": str,  
        "nota": float  
    }],  
"taxa de recomendação dos críticos": float | None,  
"taxas de recomendação dos críticos": list[  
    {  
        "link": str,  
        "taxa": float  
    }],  
"quantidade de reviews de críticos": int | None,  
"quantidades de reviews dos críticos": list[  
    {  
        "link": str,  
        "quantidade": int  
    }],  
"reviews de críticos": list[  
    {  
        "avaliação (nota até 10)": float | None,  
        "texto": str | None,  
        "data": str | None,  
        "link": str  
    }]  
}  
}  
"  
"nota média dos usuários": float | None,  
"notas dos usuários": list[  
    {  
        "link": str,  
        "nota": float  
    }],  
"taxa de recomendação dos usuários": float | None,  
"taxas de recomendação dos usuários": list[  
    {  
        "link": str,  
        "taxa": float  
    }],  
"quantidade de reviews de usuários": int | None,  
"quantidades de reviews dos usuários": list[  
    {  
        "link": str,  
        "quantidade": int  
    }],  
"reviews de usuários": list[  
    {  
        "avaliação (nota até 10)": float | None,  
        "texto": str | None,  
        "data": str | None,  
        "link": str  
    }]  
}  
}
```

JSON Gerado: movies_united.json

Entradas e Saídas

A background image consisting of a grid of binary digits (0s and 1s) in green, arranged in a matrix pattern.

MATRIX

Movie Analysis, TRacking & Information eXtractor

Desenvolvimento do Crawler

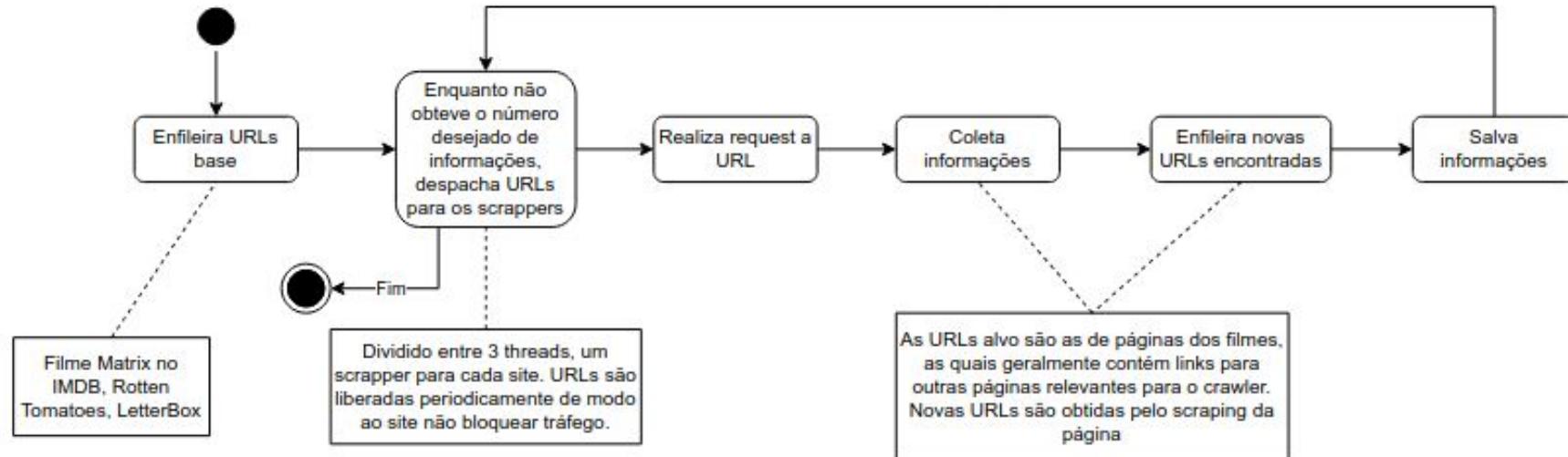
Infraestrutura Tecnológica

- Python
- Requests
- BeautifulSoup
- Numpy
- Selenium
- Playwright
- Pandas

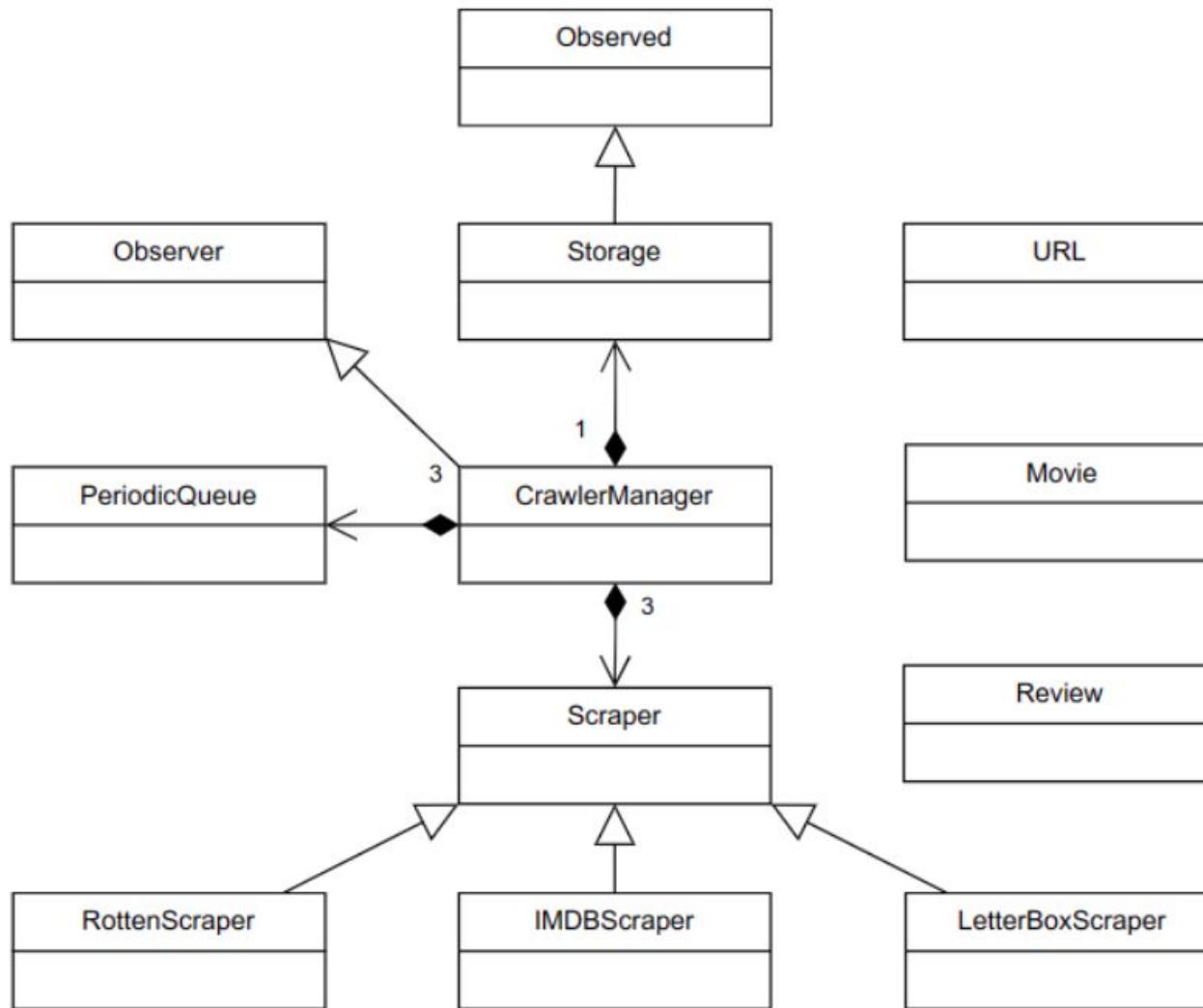


Fluxo de Desenvolvimento

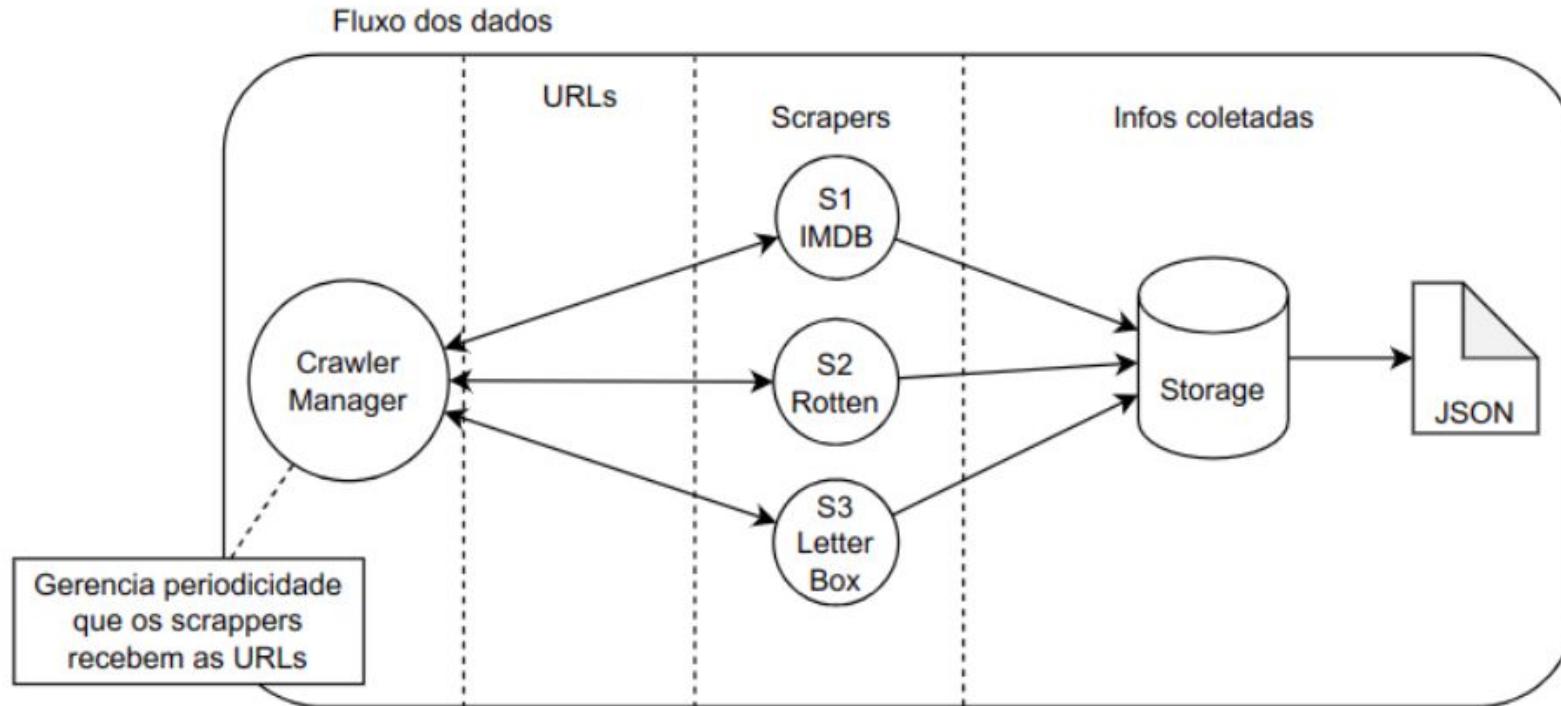
1. Definição da arquitetura inicial



o Crawler

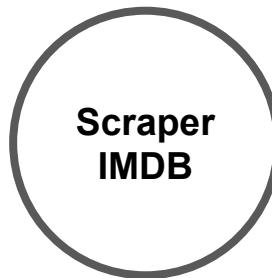


Fluxo de Desenvolvimento

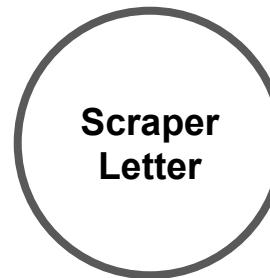


Fluxo de Desenvolvimento

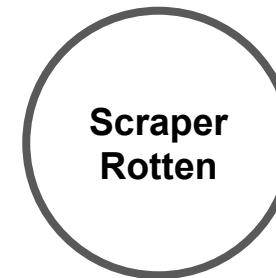
2. Divisão de tarefas



Vitor



Mariana



Bianca

Cada membro do grupo ficou responsável por um
scraper

Fluxo de Desenvolvimento

3. Coleta dos dados e análise de estatísticas

- Executar o crawler para gerar o JSON
- Realizar análises sobre os dados obtidos



UHM, YES, I'D LIKE TO ORDER SOME
DATA ANALYTICS...

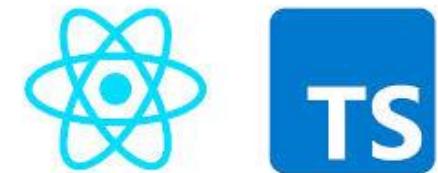
Coleta

- Data da coleta: 07/12/2025
- Tempo de coleta: 5505 s ≈ 1h 31min
- Idioma: inglês
- Local: Não especificamos, portanto o conteúdo obtido na resposta é servido pelos servidores mais próximos de nós.

Fluxo de Desenvolvimento

4. Desenvolvimento da Aplicação Exemplo

- Criado com Lovable, gerando um site em React + TypeScript com Vite
- Site consome o `movies_united.json`
- Exibe todos os filmes do JSON em uma interface
- Funciona como um agregador de informações de filmes



Aplicação Exemplo

 CINEBASE

The complete view of each movie

Quickly find ratings from top sites — IMDb, Rotten Tomatoes, and Letterboxd — all in one place

Buscar filmes...

1365 filmes encontrados



The Matrix
1999



Fight Club
1999



Inception
2010



The Dark Knight
2008



Forrest Gump
1994



The Lord of the Ri...
2002



Pulp Fiction
1994



Star Wars: Episode II - Attack of the Clones
2002



Gladiator
2000



The Godfather: Part II
1974

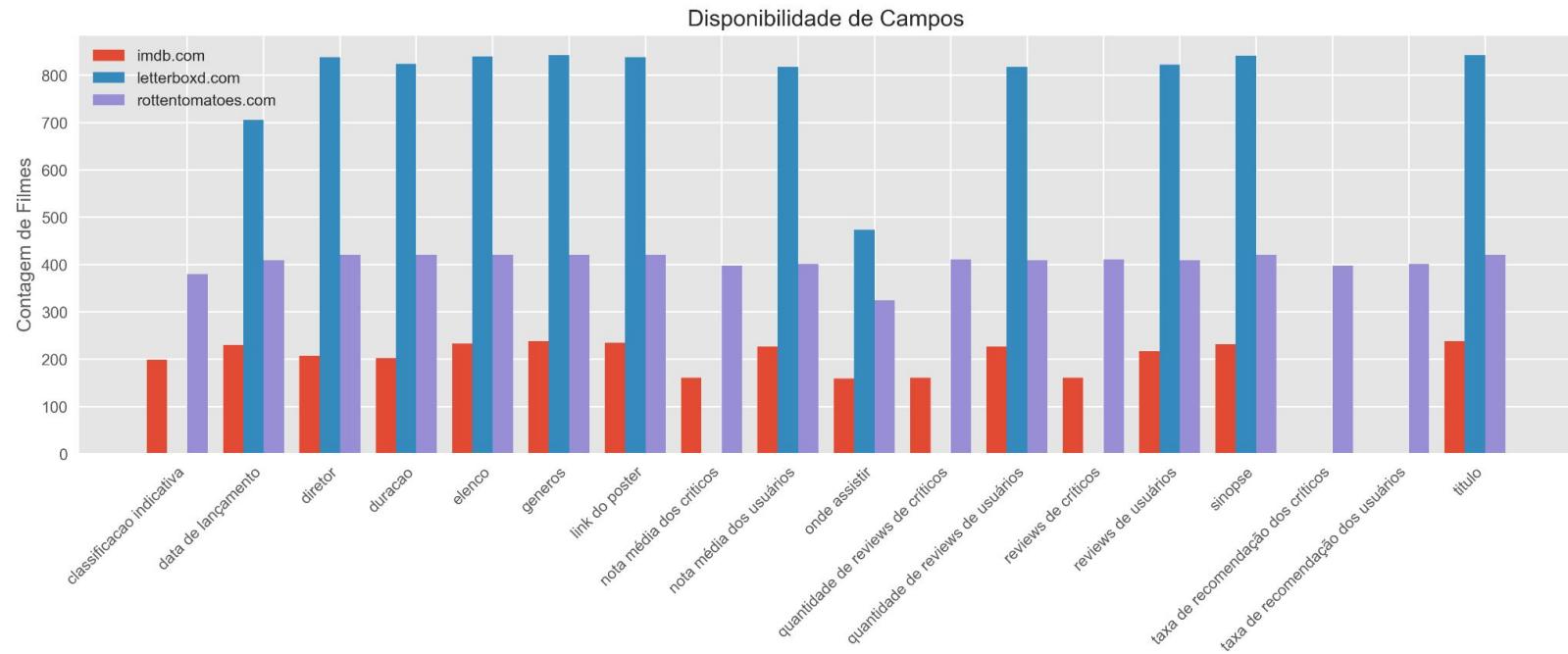


The Shawshank Redemption
1994

Análise dos dados

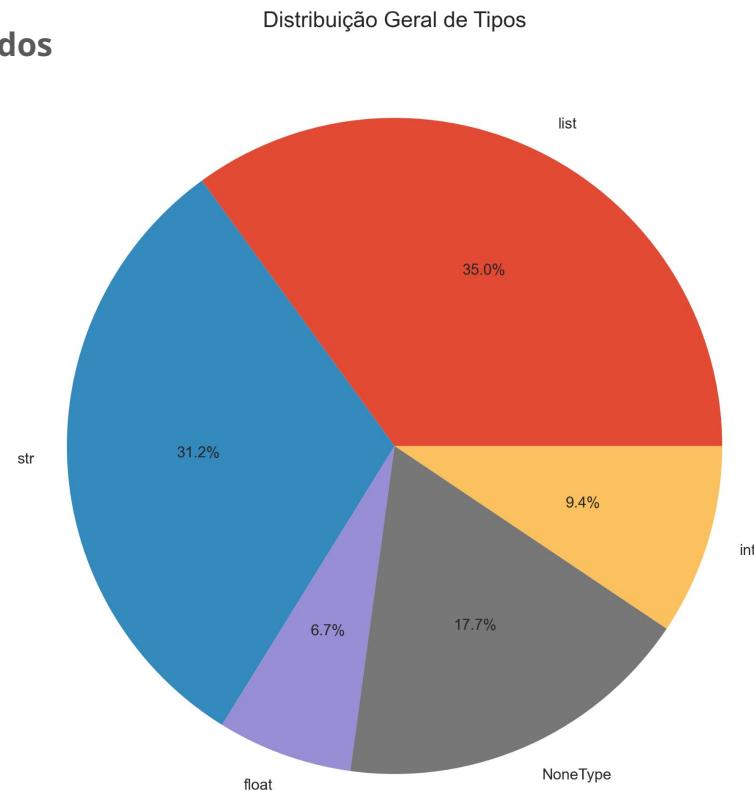
1. Análise estrutural

Gráfico de disponibilidade dos campos por fonte.



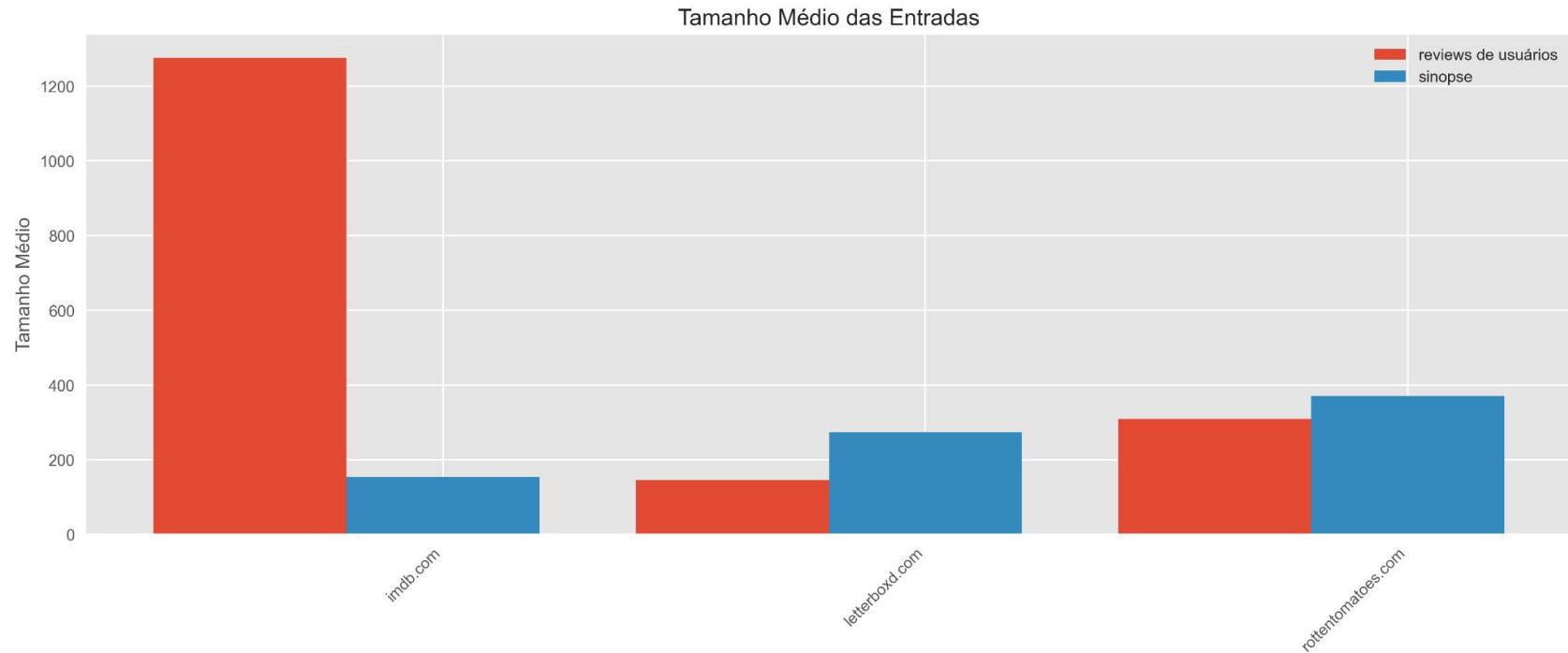
1. Análise estrutural

Distribuição de tipos de dados



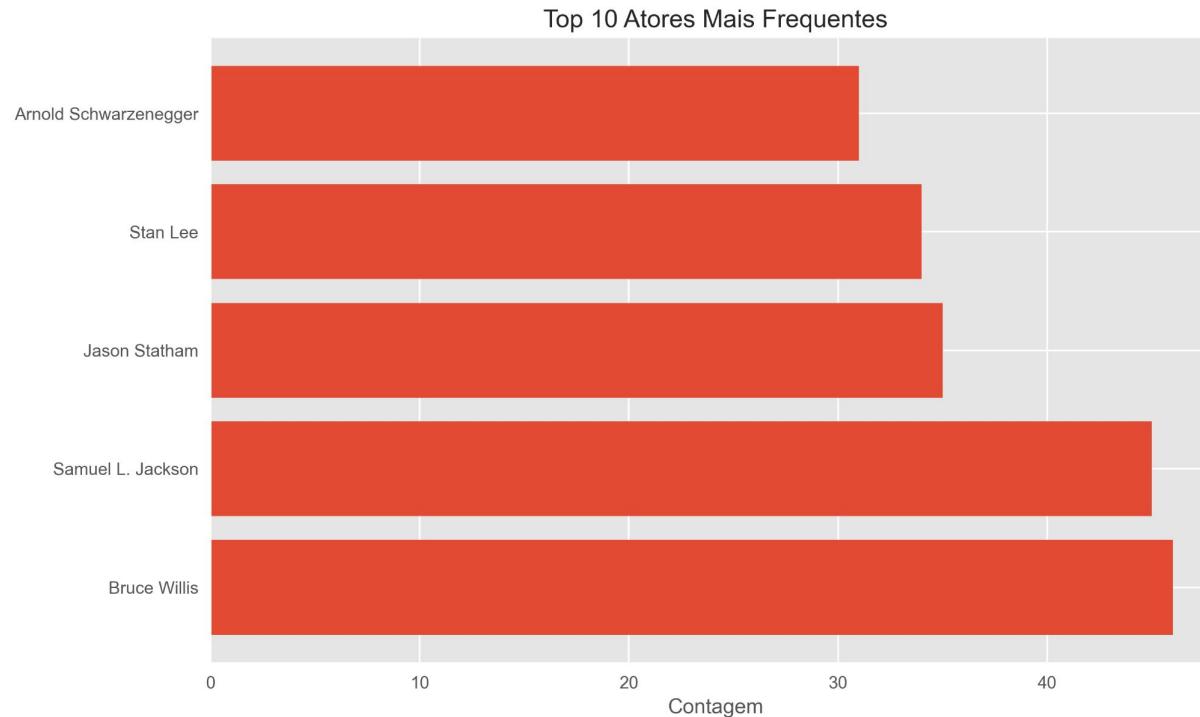
1. Análise estrutural

Tamanho médio das entradas por fonte, aplicada aos campos de texto.



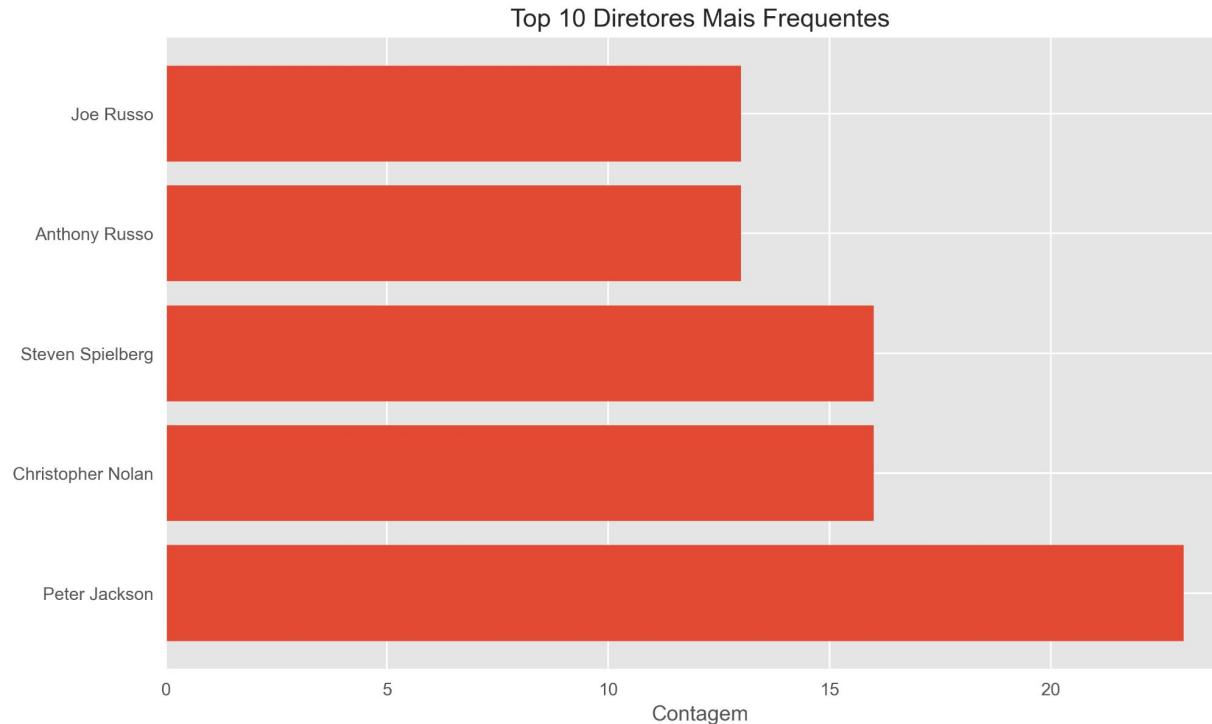
2. Análise estatística descritiva

Frequência de valores únicos: top atores.



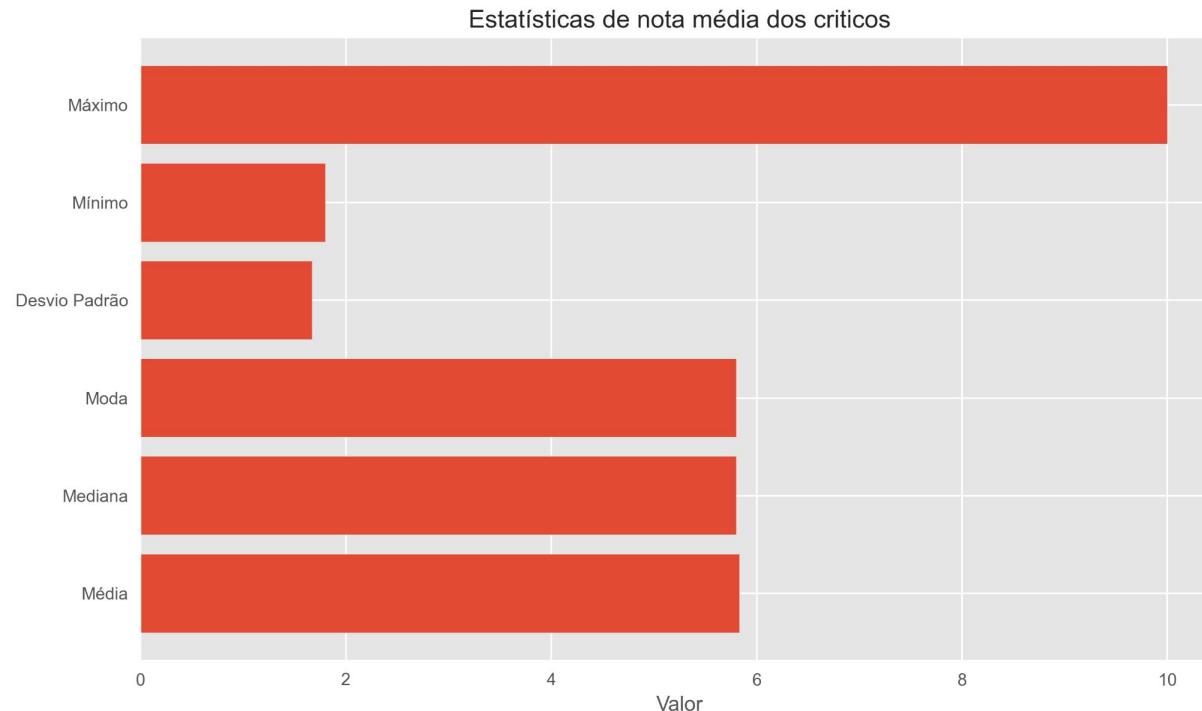
2. Análise estatística descritiva

Frequência de valores únicos: top diretores



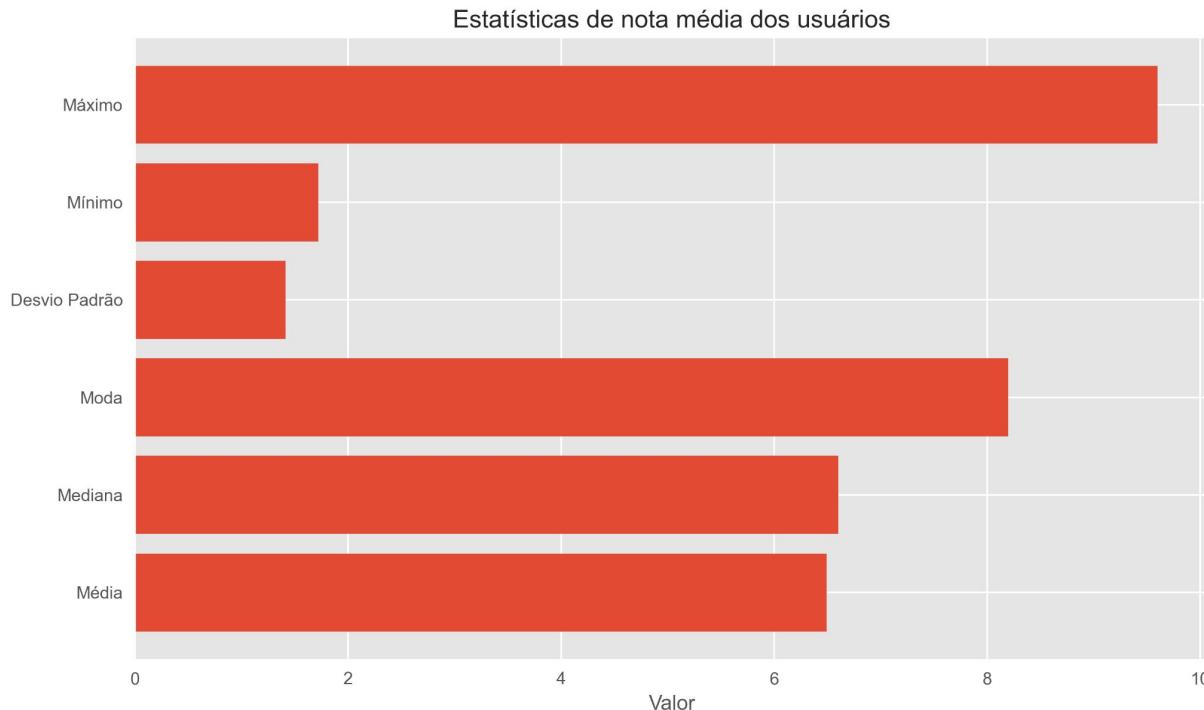
2. Análise estatística descritiva

Média, moda, mediana, desvio padrão: nota média dos críticos



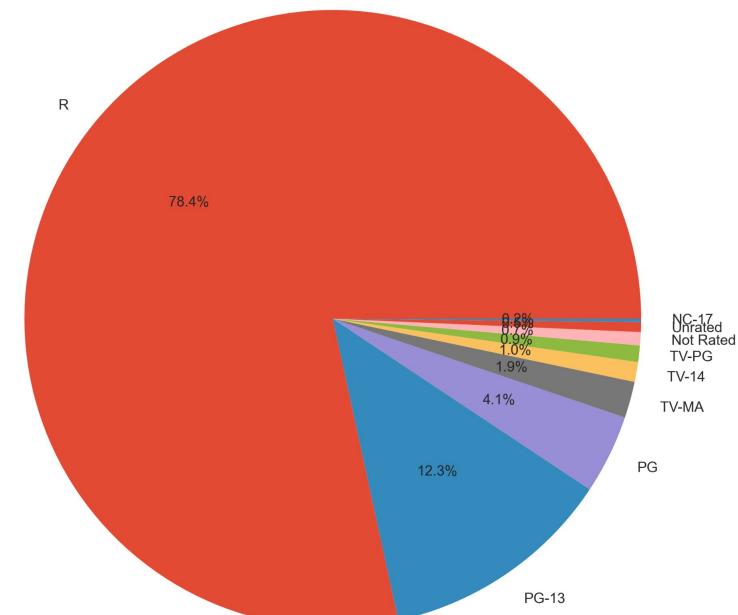
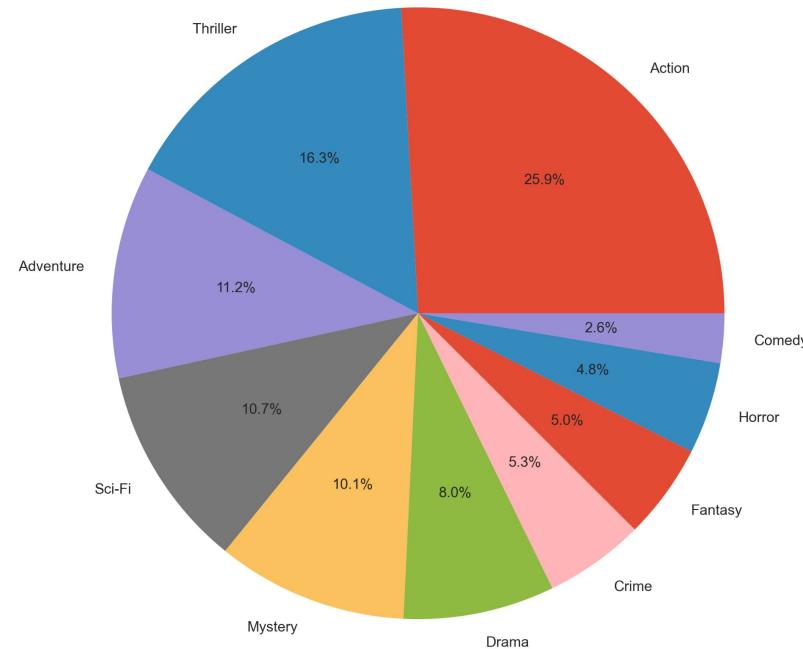
2. Análise estatística descritiva

Média, moda, mediana, desvio padrão: nota média dos usuários



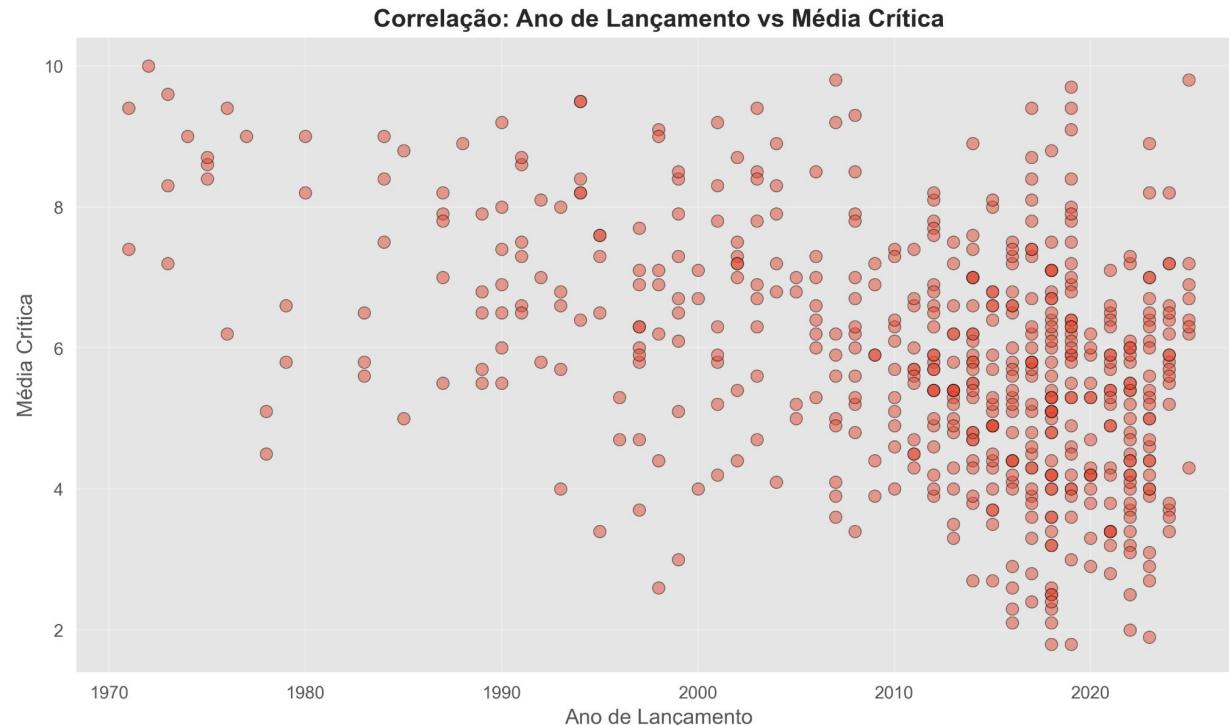
2. Análise estatística descritiva

Distribuição de classes: gêneros (top 10) e classificação indicativa



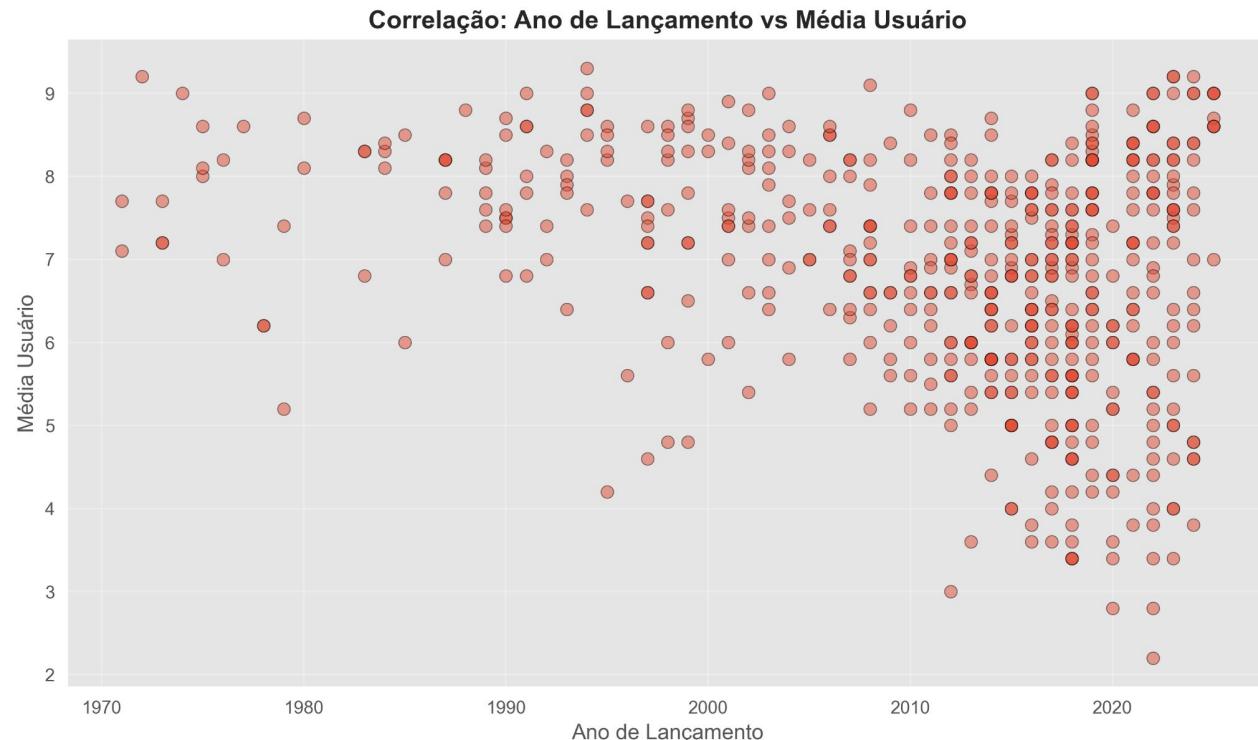
2. Análise estatística descritiva

Correlações relevantes: ano de lançamento e nota média dos críticos



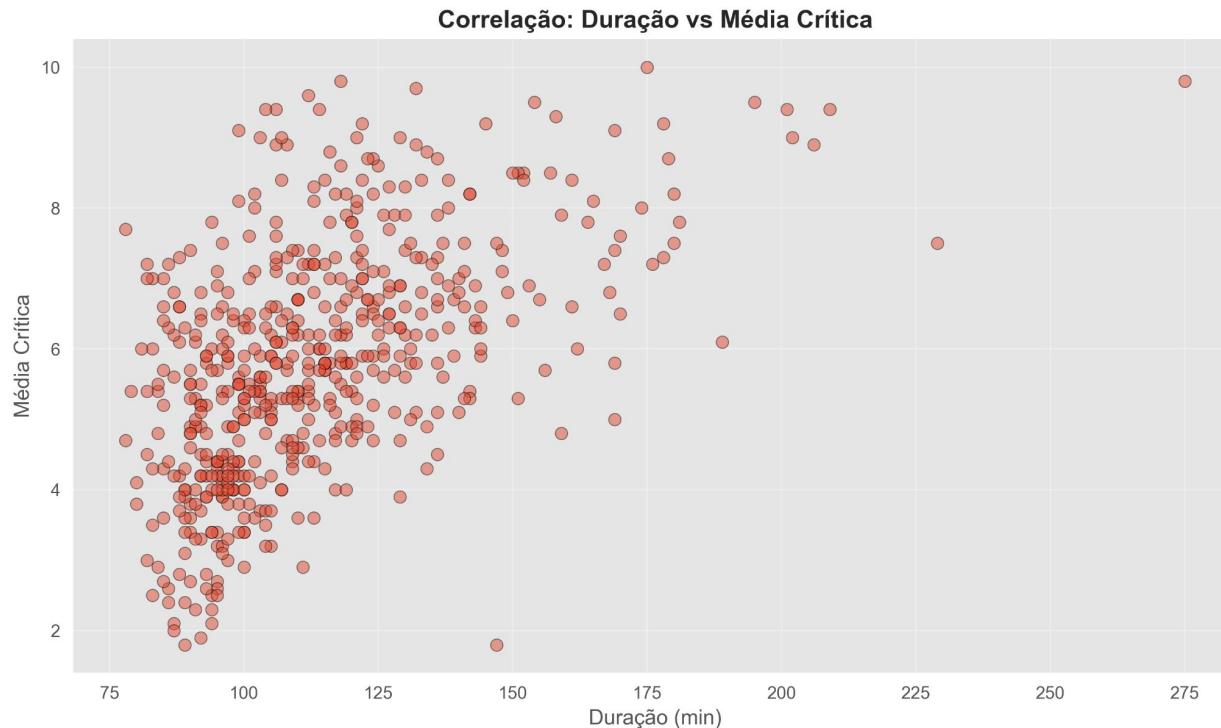
2. Análise estatística descritiva

Correlações relevantes: ano de lançamento e nota média dos usuários



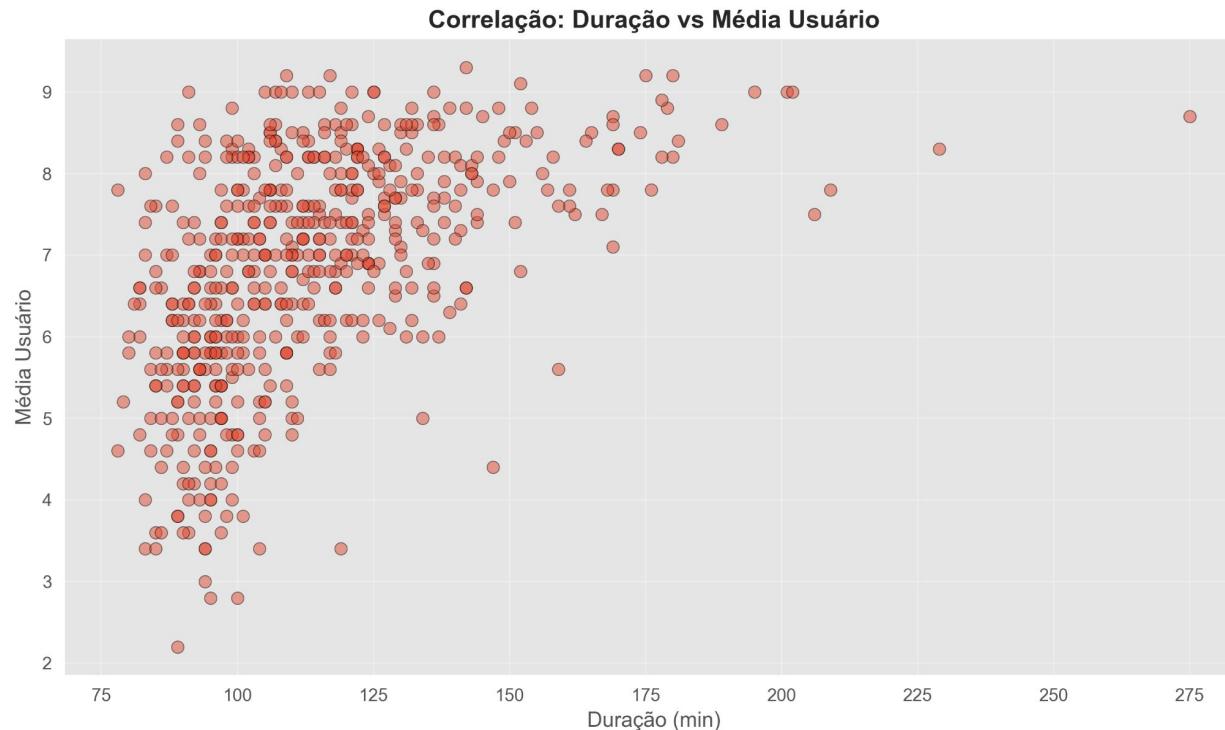
2. Análise estatística descritiva

Correlações relevantes: duração e nota média dos críticos



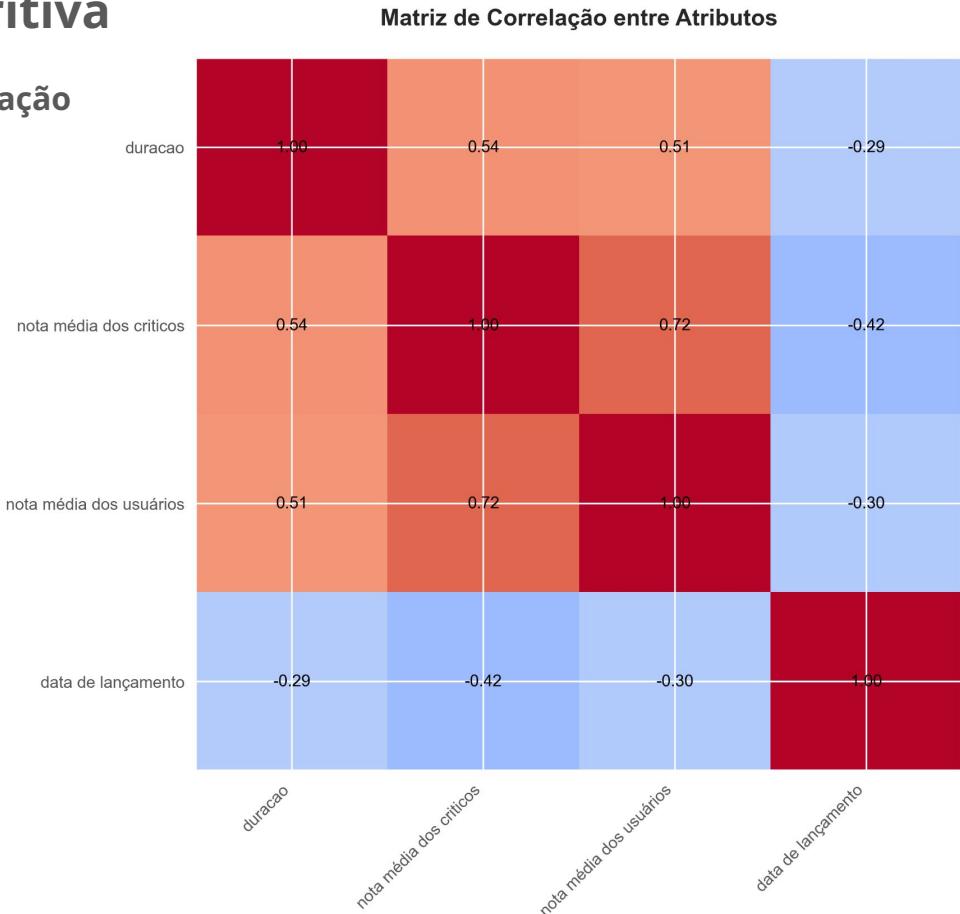
2. Análise estatística descritiva

Correlações relevantes: duração e nota média dos usuários



2. Análise estatística descritiva

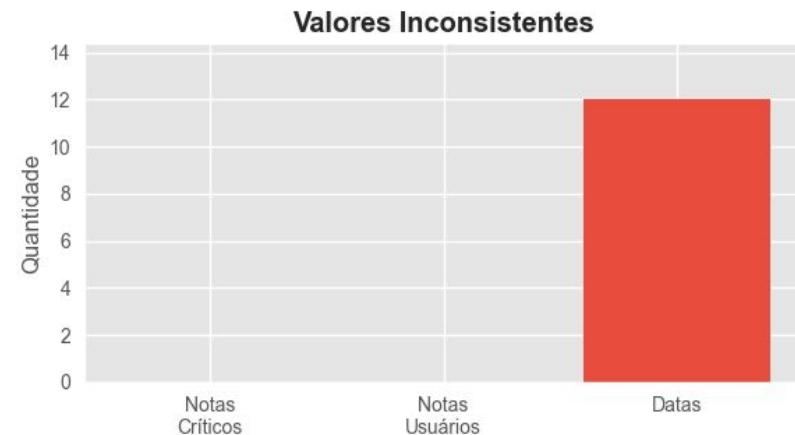
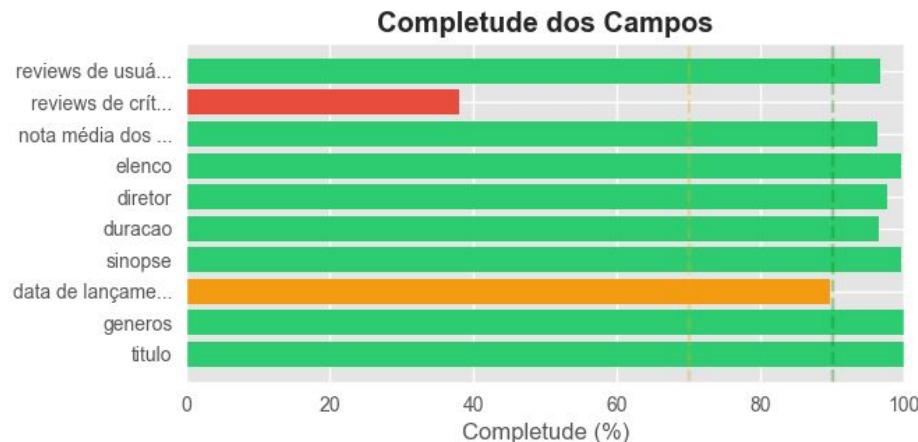
Correlações relevantes: matriz de correlação



3. Análise de qualidade

A análise de qualidade dos dados foi feita considerando a padronização prévia, realizada durante a montagem do json resultantes, o que explica a ausência de inconsistências.

Completude e valores inconsistentes



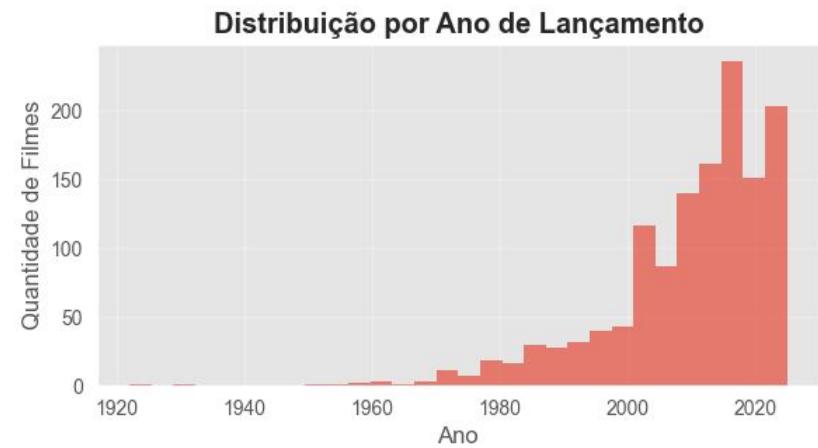
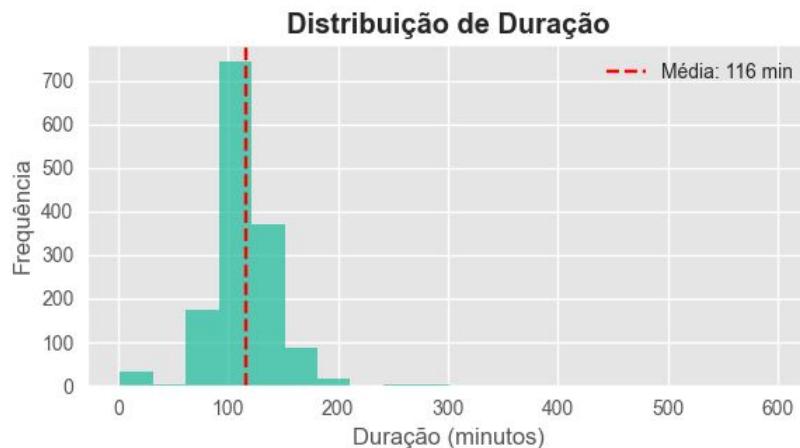
3. Análise de qualidade

Visualização de outliers: quantidades de reviews de críticos e usuários



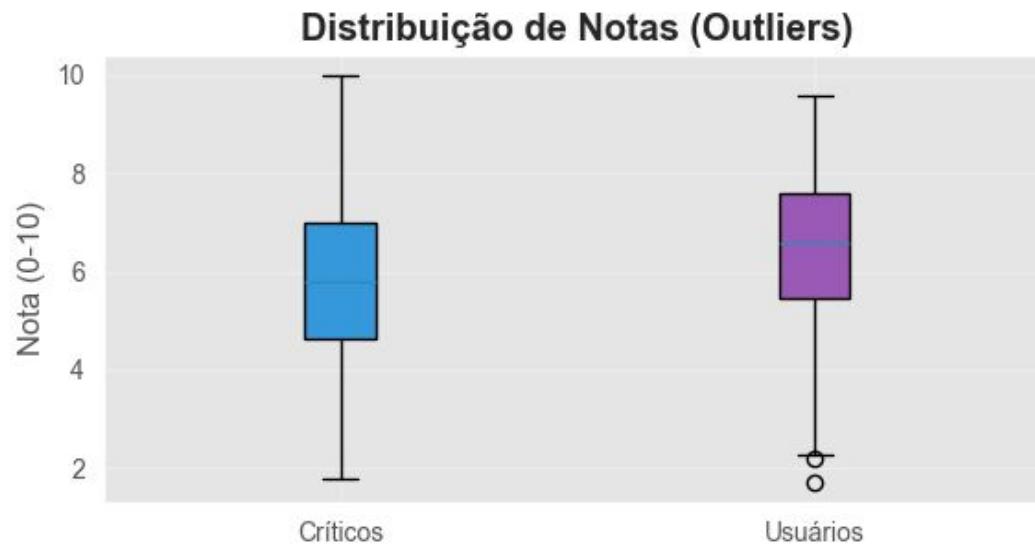
3. Análise de qualidade

Visualização de outliers: distribuição de data de lançamento e duração



3. Análise de qualidade

Visualização de outliers: boxplot nota média dos críticos e nota média dos usuários



3. Análise de qualidade

Duplicação de registros: os títulos apresentam duplicatas pois há uma verificação por fonte para evitar que o mesmo filme seja coletado duas vezes, mas nada impede que duas fontes coletem o mesmo filme. Esse comportamento é devido a natureza da aplicação proposta para os dados: um comparador de reviews entre plataformas.



Dificuldades

- Diferenças entre os sites, necessidade do uso de diferentes bibliotecas;
- Dificuldades específicas entre sites (uso de ipv6, ipv4, padronização dos dados, o que fazer com filmes repetidos entre sites e intra site?);
- Como unir dados coletados de diferentes sites?;
- Mudanças nos sites de repente (Rotten Tomatoes mudando forma de exibir reviews, de estática para dinâmica; Letterboxd saiu do ar temporariamente).

Casos de falha e O que teríamos feito diferente?

- Em alguns casos, Playwright e Selenium apresentaram timeout então não conseguiram localizar informações dinâmicas;
- Entrar em consenso sobre uma ferramenta de extração dinâmica (playwright ou selenium), para simplificar as dependências;
- Homogeneizar a quantidade de filmes coletada por cada fonte

Potencial para o mercado?

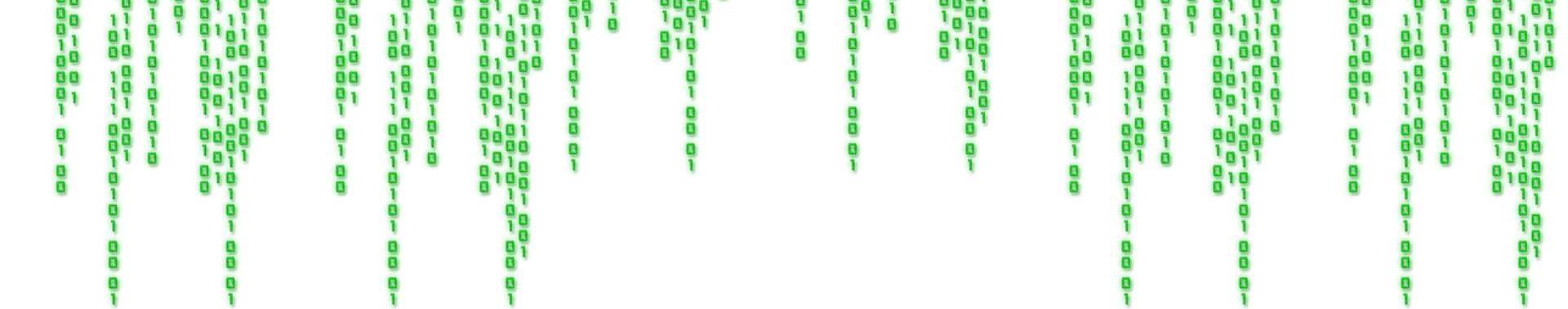
Oportunidades:

- **Público fragmentado:** Usuários precisam consultar 3-4 sites diferentes para tomar decisão de qual filme assistir
- **Público-alvo:** 50+ milhões de usuários ativos mensais somando IMDb, Letterboxd e RT
- **Dor real:** Perda de tempo navegando entre plataformas

Potencial para o mercado?

Vantagens

- **Conveniência:** Todas as notas e reviews semelhantes em um só lugar
- **Avaliação geral:** Recomendação de filmes a partir de avaliações de mais de uma plataforma



OBRIGADO!

Conclusão