

Relatório de Web Scraping - Raspagem de Dados de Jogos

1. Introdução

Este relatório documenta o desenvolvimento de um projeto de **web scraping** aplicado ao site [Torrent Jogos](#). O objetivo do projeto é coletar informações detalhadas sobre jogos disponíveis para download via torrent, incluindo seus **magnet links**, a fim de construir uma base de dados consolidada para consulta rápida.

2. Bibliotecas Utilizadas

Para a implementação da raspagem de dados, foram utilizadas as seguintes bibliotecas da linguagem Python:

- **pandas**: Manipulação e estruturação dos dados coletados.
- **BeautifulSoup**: Extração e parse de dados estruturados em HTML.
- **requests**: Envio de requisições HTTP para obtenção do conteúdo das páginas.
- **Selenium**: Interação dinâmica com elementos da página que exigem carregamento dinâmico.

3. Metodologia

A coleta de dados foi realizada em três etapas principais:

1. **Coleta da lista de jogos:**
 - O script **scrap-cards.ipynb** percorre as páginas do site listando os jogos disponíveis.
 - Exemplo de URL: <https://www.torrentjogos.com.br?page=4>
2. **Coleta dos detalhes de cada jogo:**
 - O script **scrap-cards-details.ipynb** acessa a página específica de cada jogo e extrai informações adicionais.
 - Exemplo de URL: <https://www.torrentjogos.com.br/Split-Fiction>
3. **Processamento e limpeza dos dados:**
 - Os dados obtidos são armazenados em três arquivos Excel:
 - [dados-jogos.xlsx](#): Lista de jogos.
 - [dados-jogos-details.xlsx](#): Detalhes completos de cada jogo.

- **dados-jogos-completo.xlsx**: Arquivo consolidado contendo todas as informações.
- Os scripts **juntar-excels** e **limpar-excel-completo** realizam a fusão e limpeza dos dados, garantindo sua qualidade e organização.

4. Instruções para Execução

Para executar corretamente o projeto, siga os passos abaixo:

1. **Acesse a pasta do projeto** anexada junto com este relatório.
2. **Instale as bibliotecas necessárias**
3. **Execute os scripts na seguinte ordem:**
 - **scrap-cards.ipynb**: Raspagem das páginas de lista de jogos.
 - **scrap-cards-details.ipynb**: Raspagem dos detalhes individuais dos jogos (esta etapa pode demorar um pouco).
 - **juntar-excels.ipynb**: Consolidação dos dados coletados em um único arquivo.
 - **limpar-excel-completo.ipynb**: Processamento final para remoção de dados duplicados e formatação.
4. **OBS:** No código, configurei para apenas fazer o scraping da primeira página do site, para ter uma execução mais rápida, mas pode ser alterado na variável **pages** dentro de **scrap-cards.ipynb**

```
pages = 1
cards = []
cards_details = []
for page in range(1,pages+1):
    cards = cards + retornar_cards(page)
```

✓ 0.4s

<https://www.torrentjogos.com.br/games?page=1>

Após a execução dos scripts, os arquivos resultantes estarão disponíveis em formato **Excel**:

- **dados-jogos.xlsx**
- **dados-jogos-details.xlsx**
- **dados-jogos-completo.xlsx**

5. Informações Coletadas

Os dados extraídos incluem:

- **Título do jogo**
- **Link para a página do jogo**
- **Data da última atualização**
- **Tamanho do jogo**
- **URL da imagem do banner**
- **Link da loja para compra (Steam, Epic, etc.)**
- **URL para imagens de screenshots do jogo**
- **Descrição do jogo**
- **Categorias e gêneros**
- **Magnet link para download via torrent**
- **URL do trailer**
- **Requisitos mínimos e recomendados**