

Estudo de Viabilidade: Extração de Dados do Site de Editais da UFU

Aluno: Waldemar Patrique Flores Silva


Matrícula: 31921BSI001

Objetivo

Este estudo visa avaliar a viabilidade da extração de dados do site de Editais da Universidade Federal de Uberlândia (UFU). O objetivo é coletar informações de editais publicados para o curso de Sistemas de Informação, incluindo os seguintes dados: número do edital, órgão responsável, título, tipo, data de publicação e link para o edital completo.

Ferramentas e Tecnologias Utilizadas

- **Linguagem de Programação:** Python
 - python.org
- **Bibliotecas:**
 - **Requests:** pypi.org/project/requests
 - **BeautifulSoup:** pypi.org/project/beautifulsoup4
 - **Pandas:** pandas.pydata.org
 - **Openpyxl:** openpyxl.readthedocs.io
- **Comandos:**
 - `pip install requests beautifulsoup4 pandas openpyxl`
- **URL do Site:** <http://www.editais.ufu.br/discente> [Figura 1.]



Portal de Editais e Concursos

UFU

Universidade Federal de Uberlândia

Aprovação

Edital PET/Procurador

Edital e Cultura

Conc. Púb. Docente

Concurso Téc. Adm.

Outros Editais/Oportunidades

Vestibular UFU

Fale conosco

Filtros

Número do Edital

Órgão Responsável

Título do Edital

Tipo

<Ampl>

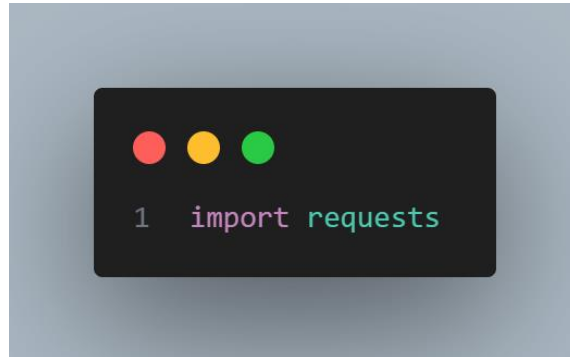
Result.

Nº	Órgão Resp.	Título	Tipo	Entra	Result. do Edital	Data de Publicação
50204	PPDCS	Edital de abertura das inscrições e do processo de seleção de alunos TURMA 2024/1 para ingresso no Programa de Pós-Graduação em Ciências da Saúde - Profissional (PPDCS)	Metod. Especializ.	Não	Não Publ.	14/09/2024 (04/09/2024)
60204	PET Física	EDITAL DE SELEÇÃO DE ALUNOS PARA INGRESSO NO PET FÍSICA	Programa PET	Sim	Não Publ.	14/09/2024 (04/09/2024)
60204	DIRBI	PROCESSO SELETIVO PARA ESTÁGIO INTERNO NÃO OBRIGATÓRIO NA DIRETORIA DO SISTEMA DE BIBLIOTECAS	Estágio UFU	Sim	Publicado	14/09/2024 (04/09/2024)
30204	CLAPET	EDITAL CLAPET Nº 30204	Programa PET	Não	Publicado	14/09/2024 (04/09/2024)
30204	DIRCO	PROCESSO SELETIVO PARA CONTRATAÇÃO DE ESTAGIÁRIO - EDITAL DIRCO Nº 30204	Estágio UFU	Não	Publicado	14/09/2024 (04/09/2024)
20204	HEADDEPT/PROF	SELEÇÃO DE ESTUDANTES NEGROS E NEGROS DOS PROGRAMAS DE PÓS-GRADUAÇÃO DA UNIVERSIDADE FEDERAL DE UBERLÂNDIA PARA INTERCÂMBIO NA UNIVERSIDADE PEDAGÓGICA DE BAPATO (BR/AMBIQUE) - PROGRAMA DE DESENVOLVIMENTO ACADÊMICO ADOÇÃO E NASCIMENTO - CAPES	Metod. Especializ.	Sim	Publicado	14/09/2024 (04/09/2024)
610204	PET Sistema de Informação - Monte Carmelo	EDITAL DE SELEÇÃO DE ALUNOS PARA INGRESSO NO PET SISTEMA DE INFORMAÇÃO, CAMPUS MONTE CARMELO	Programa PET	Sim	Publicado	14/09/2024 (04/09/2024)
620204	PET Estatística	Edital de seleção de alunos para ingresso no PET Estatística	Programa PET	Sim	Publicado	14/09/2024 (04/09/2024)
10204	DIRPO	PROCESSO SELETIVO PARA CONTRATAÇÃO DE ESTAGIÁRIOS DO CURSO DE ARQUITETURA E URBANISMO	Estágio UFU	Não	Publicado	14/09/2024 (04/09/2024)
910204	PET Ciências Contábeis	EDITAL DE SELEÇÃO DE ESTUDANTES PARA INGRESSO NO PET CIÊNCIAS CONTÁBEIS	Programa PET	Sim	Publicado	14/09/2024 (04/09/2024)
910204	PET Administração	EDITAL DE SELEÇÃO DE DISCENTES PARA INGRESSO NO PET ADM	Programa PET	Sim	Não Publ.	14/09/2024 (04/09/2024)
20204	DACIN	EDITAL DE SELEÇÃO DE MONITORES - MONITORIA DACIN DE APOIO E INCLUSÃO - UBERLÂNDIA-MG	Monitoria	Não	Não Publ.	14/09/2024 (04/09/2024)
10204	FAMED	PROCESSO SELETIVO PARA ESTÁGIO INTERNO NÃO OBRIGATÓRIO - EDITAL DINFAMED Nº 10204	Estágio UFU	Não	Não Publ.	14/09/2024 (04/09/2024)
10204	PPGOSBIO/UFU	EDITAL PPGOSBIO Nº 10204 - Edital de abertura das inscrições e do processo de seleção para ingresso no Programa de Pós-Graduação em Genética e Biomolécula	Metod. Especializ.	Sim	Não Publ.	14/09/2024 (04/09/2024)
10204	DIRSAR	EDITAL DE SELEÇÃO PARA CONTRATAÇÃO DE ESTAGIÁRIO - EDITAL DIRSAR Nº 10204	Estágio UFU	Não	Publicado	14/09/2024 (04/09/2024)

Figura 1 (Portal de Editais e Concursos - <http://www.editais.ufu.br/discente>)

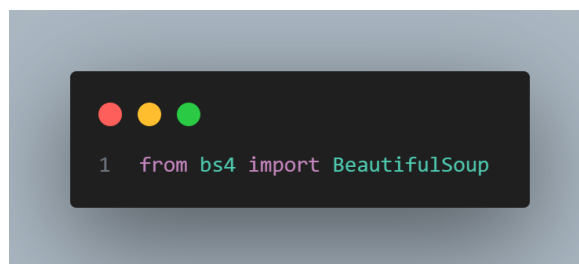
Requisição HTTP

Foi utilizada a biblioteca requests para realizar uma requisição HTTP à página de editais, recuperando o conteúdo para posterior processamento.



Parse do HTML

A biblioteca BeautifulSoup foi utilizada para fazer o parse do HTML, identificando e extraíndo os blocos de dados relevantes, como a tabela de editais.



Iteração e Extração

Iteramos sobre as linhas da tabela de editais (tags <tr>) e extraímos os campos necessários (tags <td>), obtendo as informações relevantes para cada edital.

```
1 for edital in editais_blocks:
2     numero_edital = edital.find('td', class_='views-field-field-nro-value')
3     numero_edital = numero_edital.get_text(strip=True) if numero_edital else 'N/A'
4     orgao_responsavel = edital.find('td', class_='views-field-field-setor-responsavel-value')
5     orgao_responsavel = orgao_responsavel.get_text(strip=True) if orgao_responsavel else 'N/A'
6     titulo = edital.find('td', class_='views-field-field-titulo')
7     titulo_text = titulo.get_text(strip=True) if titulo else 'N/A'
8
9     link_edital = None
10    if titulo:
11        a_tag = titulo.find('a', href=True)
12        if a_tag:
13            link_edital = urljoin(url, a_tag['href'])
14
15    tipo = edital.find('td', class_='views-field-field-tipo-value')
16    tipo = tipo.get_text(strip=True) if tipo else 'N/A'
17    data_publicacao = edital.find('td', class_='views-field-field-data-publicacao-value')
18    data_publicacao = data_publicacao.get_text(strip=True) if data_publicacao else 'N/A'
19    ano_edital = '{}'.format(data_publicacao[4:9].strip())
20
21    if numero_edital != 'N/A' or orgao_responsavel != 'N/A' or titulo_text != 'N/A' or tipo != 'N/A' or data_publicacao != 'N/A' or link_edital:
22        edital_info = {
23            'numero_edital': numero_edital,
24            'orgao_responsavel': orgao_responsavel,
25            'titulo': titulo_text,
26            'tipo': tipo,
27            'data_publicacao': data_publicacao,
28            'link': link_edital
29        }
30
31    if (orgao_responsavel in filtros_orgs) and (tipo in filtros_tipos) and ano_edital == ano_atual:
32        editais_data.append(editais_info)
```

Filtragem dos Dados

Filtros foram implementados para incluir apenas editais relacionados ao curso de Sistemas de Informação. Esses filtros consideram o órgão responsável e o tipo de edital, sendo facilmente adaptáveis para expandir a novos cursos.

```
1 filtros_orgs = {"PET Sistemas de Informação", "PET Sistemas de Informação - Monte Carmelo", "PETSIMC"}
2 filtros_tipos = {"Programa PET", "Estágio UFU", "Mest./Dout./Especializ.", "Monitoria"}
```

```
1 if (orgao_responsavel in filtros_orgs) and (tipo in filtros_tipos) and ano_edital == ano_atual:
2     editais_data.append(editais_info)
```

Exportação dos Dados

Formato JSON

Os dados extraídos são convertidos para o formato JSON utilizando a biblioteca json. Esse formato facilita a integração futura com o front-end da aplicação.

```
1 editais_json = json.dumps(editais_data, ensure_ascii=False, indent=4)
2 json_filename = "editais.json"
3 with open(json_filename, 'w', encoding='utf-8') as json_file:
4     json_file.write(editais_json)
```

Formato XLSX

Os dados também foram exportados para um arquivo Excel (.xlsx) com o auxílio da biblioteca pandas, garantindo que as informações possam ser analisadas e compartilhadas em um formato acessível.

```
1 xlsx_filename = 'editais.xlsx'
2 df = pd.DataFrame(editais_data)
3 df.to_excel(xlsx_filename, index=False)
```

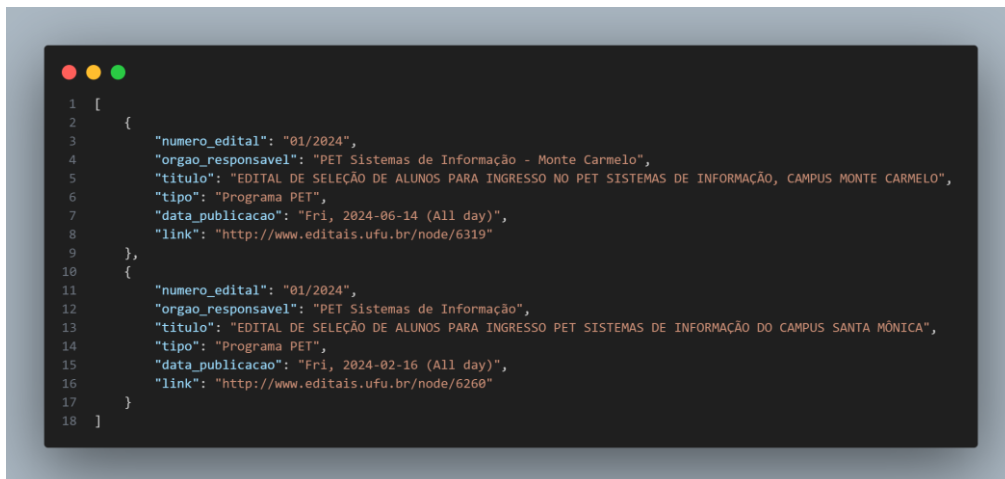
Conclusão

A extração de dados do site de Editais da UFU é viável e relativamente simples de ser implementada. A utilização das bibliotecas requests, BeautifulSoup e pandas em Python permite uma extração e exportação eficiente dos dados de interesse, que podem ser utilizados em diferentes formatos (JSON e XLSX) para futuras análises ou integração com outras aplicações.

Arquivos de saída:

editais.xlsx

	numero_edital	orgao_responsavel	titulo	tipo	data_publicacao	link
2	01/2024	PET Sistemas de Informação - Monte Carmelo	EDITAL DE SELEÇÃO DE ALUNOS PARA INGRESSO NO PET SISTEMAS DE INFORMAÇÃO, CAMPUS MONTE CARMELO	Programa PET	Fri, 2024-06-14 (All day)	http://www.editais.ufu.br/node/6319
3	01/2024	PET Sistemas de Informação	EDITAL DE SELEÇÃO DE ALUNOS PARA INGRESSO PET SISTEMAS DE INFORMAÇÃO DO CAMPUS SANTA MÔNICA	Programa PET	Fri, 2024-02-16 (All day)	http://www.editais.ufu.br/node/6260



Código

```
import requests
from bs4 import BeautifulSoup
from urllib.parse import urljoin
import json
from datetime import date
import pandas as pd

# Os filtros são definidos nessas duas listas
filtros_orgs = {"PET Sistemas de Informação", "PET Sistemas de Informação - Monte Carmelo", "PETSIMC"}
filtros_tipos = {"Programa PET", "Estágio UFU", "Mest./Dout./Especializ.", "Monitoria"}

# Obtém o ano atual para comparar com a data dos editais e transforma em String
data_atual = date.today()
ano_atual = '{}'.format(data_atual.year)

# Inicializa variáveis de controle
id_pag = 0 # Página inicial
QTD_MAX = 20 # Última página
continuar = True # Variável que indica se devemos continuar a busca

# Lista que vai armazenar todos os dados de editais coletados
editais_data = []

# Laço que percorre as páginas de editais enquanto a quantidade de páginas for menor que QTD_MAX e devemos continuar
while (id_pag < QTD_MAX and continuar):
    # Formata a URL da página de editais com o número da página atual
    url = f"http://www.editais.ufu.br/discente?page={id_pag}"

    # Faz uma requisição HTTP para obter o conteúdo da página
    response = requests.get(url)

    # Verifica se a requisição foi bem-sucedida
    if response.status_code == 200:
        # Analisa o conteúdo da página HTML
```

```

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

# Encontra todos os blocos de editais (linhas da tabela de editais)
editais_blocks = soup.find_all('tr')

# Itera sobre cada edital encontrado na página extraíndo os dados
for edital in editais_blocks:
    numero_edital = edital.find('td', class_='views-field-field-nro-value')
    numero_edital = numero_edital.get_text(strip=True) if numero_edital else 'N/A'

    orgao_responsavel = edital.find('td', class_='views-field-field-setor-
responsavel-value')
    orgao_responsavel = orgao_responsavel.get_text(strip=True) if
orgao_responsavel else 'N/A'

    titulo = edital.find('td', class_='views-field-title')
    titulo_text = titulo.get_text(strip=True) if titulo else 'N/A'

    link_edital = None

    if titulo:
        a_tag = titulo.find('a', href=True) # Encontra a tag que contém o link
        if a_tag:
            # Cria a URL completa do edital usando urljoin
            link_edital = urljoin(url, a_tag['href'])

    tipo = edital.find('td', class_='views-field-field-tipo-value')
    tipo = tipo.get_text(strip=True) if tipo else 'N/A'

    data_publicacao = edital.find('td', class_='views-field-field-data-publicacao-
value')
    data_publicacao = data_publicacao.get_text(strip=True) if data_publicacao else
'N/A'

    # Extrai o ano da data de publicação
    ano_edital = '{}'.format(data_publicacao[4:9].strip())

    # Se qualquer uma dessas informações estiver disponível, cria um dicionário
com os dados do edital
    if numero_edital != 'N/A' or orgao_responsavel != 'N/A' or titulo_text !=
'N/A' or tipo != 'N/A' or data_publicacao != 'N/A' or link_edital:
        edital_info = {
            'numero_edital': numero_edital,
            'orgao_responsavel': orgao_responsavel,
            'titulo': titulo_text,
            'tipo': tipo,
            'data_publicacao': data_publicacao,
            'link': link_edital
        }

    # Adiciona o edital à lista se ele passar pelos filtros de órgão, tipo e ano.
    if (orgao_responsavel in filtros_orgs) and (tipo in filtros_tipos) and
ano_edital == ano_atual:
        editais_data.append(edital_info)

```

```

# Se o ano do edital não for o ano atual, interrompe a busca
if(ano_edital != ano_atual):
    continuar = False

else:
    # Exibe uma mensagem de erro se a página não for acessada com sucesso
    print(f"Erro ao acessar a página. Status code: {response.status_code}")
    break

# Incrementa o número da página para acessar a próxima
id_pag += 1

# Converte a lista de editais em um arquivo JSON e salva no disco
editais_json = json.dumps(editais_data, ensure_ascii=False, indent=4)
json_filename = "editais.json"
with open(json_filename, 'w', encoding='utf-8') as json_file:
    json_file.write(editais_json)

# Converte os dados dos editais em um DataFrame do pandas e salva em um arquivo Excel
xlsx_filename = 'editais.xlsx'
df = pd.DataFrame(editais_data)
df.to_excel(xlsx_filename, index=False)

print(f"Dados salvos em {xlsx_filename} e {json_filename}")

```