Extração de Dados em Python

Índice

CSV

Arquivos CSV (Comma-Separated Values) são usados para armazenar dados tabulares. Exemplo de leitura de um arquivo CSV usando a biblioteca `pandas`:

```
import pandas as pd

# Lendo o arquivo CSV
df = pd.read_csv('dados.csv')

# Exibindo as primeiras linhas do DataFrame
print(df.head())
```

Excel

Arquivos Excel (.xlsx, .xls) são amplamente utilizados para armazenar dados em formato de planilha. Exemplo de leitura de um arquivo Excel usando a biblioteca `pandas`:

```
import pandas as pd

# Lendo o arquivo Excel
df = pd.read_excel('dados.xlsx')

# Exibindo as primeiras linhas do DataFrame
print(df.head())
```

JSON

Arquivos JSON (JavaScript Object Notation) são usados para armazenar dados estruturados. Exemplo de leitura de um arquivo JSON:

```
import json

# Lendo o arquivo JSON
with open('dados.json') as file:
    data = json.load(file)

# Exibindo os dados
print(data)
```

XML

Arquivos XML (eXtensible Markup Language) são usados para armazenar dados hierárquicos. Exemplo de leitura de um arquivo XML usando `ElementTree`:

```
import xml.etree.ElementTree as ET

# Lendo o arquivo XML
tree = ET.parse('dados.xml')
root = tree.getroot()

# Exibindo os dados
for elem in root:
    print(elem.tag, elem.attrib)
```

TXT

Arquivos TXT são usados para armazenar dados em texto plano. Exemplo de leitura de um arquivo TXT:

```
# Lendo o arquivo TXT
with open('dados.txt') as file:
    lines = file.readlines()

# Exibindo as linhas do arquivo
for line in lines:
    print(line.strip())
```

HTML

Arquivos HTML são usados para armazenar dados estruturados em páginas web. Exemplo de extração de dados de um arquivo HTML usando `BeautifulSoup`:

```
from bs4 import BeautifulSoup

# Lendo o arquivo HTML
with open('dados.html') as file:
    soup = BeautifulSoup(file, 'html.parser')

# Exibindo todos os links da página
for link in soup.find_all('a'):
    print(link.get('href'))
```

SQL

Arquivos SQL podem ser usados para armazenar comandos de banco de dados. Exemplo de execução de uma consulta SQL usando `sqlite3`:

```
import sqlite3

# Conectando ao banco de dados
conn = sqlite3.connect('banco_de_dados.db')
cursor = conn.cursor()

# Executando uma consulta SQL
cursor.execute("SELECT * FROM tabela")

# Exibindo os resultados
for row in cursor.fetchall():
    print(row)

# Fechando a conexão
conn.close()
```

PDF

Arquivos PDF são frequentemente utilizados para relatórios. Exemplo de extração de texto de um PDF usando `PyPDF2`:

```
import PyPDF2

# Lendo o arquivo PDF
with open('documento.pdf', 'rb') as file:
    reader = PyPDF2.PdfReader(file)
    page = reader.pages[0]
    text = page.extract_text()

# Exibindo o texto extraído
print(text)
```

Parquet

Arquivos Parquet são usados para armazenar dados em formato colunar. Exemplo de leitura de um arquivo Parquet usando `pandas`:

```
import pandas as pd

# Lendo o arquivo Parquet
df = pd.read_parquet('dados.parquet')

# Exibindo as primeiras linhas do DataFrame
print(df.head())
```

HDF5

Arquivos HDF5 (Hierarchical Data Format) são usados para armazenar grandes volumes de dados numéricos. Exemplo de leitura de um arquivo HDF5 usando `h5py`:

```
 import h5py

# Lendo o arquivo HDF5
with h5py.File('dados.h5', 'r') as file:
    data = file['dataset_name'][:]

# Exibindo os dados
print(data)
```