RPA na PRÁTICA

Aula – PDFs

Explicação

Existem várias bibliotecas de leitura de PDF e aqui vão ser apresentadas algumas.
Vale destacar que PDFs não possuem um padrão assim como o Excel, ou seja, a leitura pode ser um pouco mais complexa de ser feita.

Biblioteca PDFminer

- Instalação da biblioteca

```
pip install pdfminer.six
```

- Importação da biblioteca

```
from pdfminer.high_level import extract_text
```

- Abertura e extração do texto

```
texto = extract_text('boleto.pdf')
print(texto)
# output:
#   Boleto de Pagamento
#   Referente à: Produto Exemplo 1
#   Emitido por: Seu nome ou o de sua empresa - CNPJ 22.517.527/0001-20
#   E-mail: contato@seuemail.com
#   ...
```

Biblioteca pdfplumber

- Instalação da biblioteca

```
pip install pdfplumber
```

- Importação da biblioteca

```
import pdfplumber
```

- Abertura do arquivo

```
pdf = pdfplumber.open('boleto.pdf')
```

- Seleção da página do documento à ser lida

```
page = pdf.pages[0]
```

- Extração do texto

```
texto = page.extract_text()
print(texto)
# output:
#   Boleto de Pagamento
#   Referente à: Produto Exemplo 1
#   Emitido por: Seu nome ou o de sua empresa - CNPJ 22.517.527/0001-20
#   E-mail: contato@seuemail.com
#   ...
```

Biblioteca PyMuPDF

- Instalação da biblioteca

```
pip install PyMuPDF
```

- Importação da biblioteca

```
import fitz
```

- Abertura do arquivo

```
pdf = fitz.open('boleto.pdf')
```

- Seleção da página do documento à ser lida

```
page = pdf.load_page(0)
```

- Extração do texto

```
texto = page.get_text()
print(texto)
# output:
#   Boleto de Pagamento
#   Referente à: Produto Exemplo 1
#   Emitido por: Seu nome ou o de sua empresa - CNPJ 22.517.527/0001-20
#   E-mail: contato@seuemail.com
#   ...
```

RPA na PRÁTICA