Capacitação em PyPDF2

**Data da reunião:**

**Horário**:

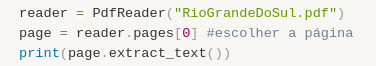
**Responsáveis:**

* Bianca Patrocinio
* Matheus de Mello

**Ata da capacitação**

Este código Python utiliza a biblioteca **PyPDF2** para extrair o texto de um arquivo PDF chamado "RioGrandeDoSul.pdf". O código realiza duas tarefas principais:

1. **Extraindo e Imprimindo a Página:** Neste trecho de código comentado:



* Importa a classe **PdfReader** da biblioteca **PyPDF2**.
* Abre o arquivo PDF "RioGrandeDoSul.pdf" e cria um objeto **PdfReader** para lê-lo.
* Seleciona a primeira página do PDF usando **reader.pages[0]**.
* Extrai o texto da página selecionada usando **page.extract\_text()** e imprime na saída padrão.

1. **Ignorando o Cabeçalho e Salvando em um Arquivo:** Neste trecho de código:



* Novamente, cria um objeto **PdfReader** para abrir o arquivo PDF "RioGrandeDoSul.pdf".
* Seleciona a terceira página do PDF usando **reader.pages[2]**.
* Define uma lista vazia chamada **parts** que será usada para armazenar partes do texto.
* Define uma função **visitor\_body** que será usada como um callback pelo método **page.extract\_text()**. Essa função é chamada para cada bloco de texto na página.
* A função **visitor\_body** verifica a posição vertical (**y**) de cada bloco de texto na página. Se o valor de **y** estiver entre 50 e 720 (o que provavelmente corresponde à área do corpo do texto, ignorando o cabeçalho), o texto é adicionado à lista **parts**.
* Após a extração de texto da página usando o método **page.extract\_text()**, os textos selecionados são concatenados em uma única string chamada **text\_body**.
* O código cria um arquivo de texto chamado "arq02.txt" no modo de escrita (**'w'**), escreve o conteúdo de **text\_body** no arquivo e fecha o arquivo.

No geral, o código extrai o texto da terceira página do PDF "RioGrandeDoSul.pdf", ignorando o cabeçalho com base na posição vertical do texto, e salva o texto resultante em um arquivo de texto chamado "arq02.txt".