

GABRIELE ROMANATO

Estrarre testo da un documento PDF con Python

L'estrazione del testo da un documento PDF può essere un'operazione molto utile in una vasta gamma di applicazioni, come il data mining, l'analisi dei documenti e l'automazione dei processi aziendali. Python, grazie alla sua vasta gamma di librerie, offre numerosi strumenti per eseguire questa operazione in modo semplice ed efficace. In questo articolo esploreremo come estrarre testo da un documento PDF utilizzando PyPDF2.

Prima di tutto, è necessario installare la libreria PyPDF2. Si può fare tramite il comando:

```
pip install PyPDF2
```

PyPDF2 è una delle librerie più popolari per lavorare con file PDF in Python. Offre funzionalità di base come l'unione di PDF, la rotazione delle pagine e, ovviamente, l'estrazione di testo.

Ecco un semplice esempio di come estrarre il testo da un PDF utilizzando PyPDF2:

```
import PyPDF2

# Apriamo il file PDF
with open('document.pdf', 'rb') as file:
    # Creiamo un lettore PDF
    pdf_reader = PyPDF2.PdfReader(file)

    # Inizializziamo una stringa vuota per il testo estratto
    text = ""

    # Iteriamo attraverso tutte le pagine del PDF
```

```
for page in range(len(pdf_reader.pages)):
    # Estraiamo il testo di ogni pagina
    text += pdf_reader.pages[page].extract_text()

print(text)
```

Spiegazione del codice:

- Apriamo il file PDF in modalità di lettura binaria ('rb').
- Creiamo un oggetto PdfReader che legge il contenuto del PDF.
- Iteriamo attraverso tutte le pagine del documento e utilizziamo il metodo `extract_text()` per ottenere il testo di ogni pagina.
- Il testo estratto viene poi memorizzato in una variabile e può essere stampato o elaborato ulteriormente.

Conclusione

L'estrazione del testo è solo una delle tante operazioni che si possono effettuare sui PDF con Python. Tra le altre possibilità ricordiamo l'estrazione di immagini, la manipolazione di pagine e molto altro ancora, aprendo la strada a una vasta gamma di applicazioni automatizzate per la gestione dei documenti.