

Estrarre testo da PDF



Sommario

- Perché trasformare in testo un pdf
- Come possiamo farlo
- Cos'è PyPDF2
- Come funziona PyPDF2

Estrarre testo da PDF

Perché trasformare in testo un pdf

Spesso i dati che le aziende utilizzano sono racchiuse in documenti:

- Contratti
- Report
- Manuali tecnici e
- Documentazione
- etc...

I dati di questi documenti sono necessari alle aziende per effettuare analisi, previsioni e operatività.




Estrarre testo da PDF

Perché trasformare in testo un pdf

Per poter accedere a questi dati in maniera diretta è necessario estrarli e inserirli nei propri sistemi

- CRM
- ERP
- MAIL
- etc...

Svolgere questi lavori manualmente significa:

-  Impegnare una persona per parecchio tempo
-  Rischiare errori
-  Non avere i dati in RealTime .

Estrarre testo da PDF

Perché trasformare in testo un pdf

Recuperare i dati

Il modo migliore per recuperare i dati è quello di automatizzare questo processo.

Per farlo servono tre fasi distinte:

- 1. Recupero del testo dai documenti
- 2. Estrazione delle informazioni
- 3. Salvataggio delle informazioni nei sistemi aziendali

NOTA: In questa lezione ci occuperemo della prima fase.

Estrarre testo da PDF

Come possiamo fare

Per poter automatizzare questo processo, `python` fornisce diverse soluzioni, tra cui le due principali sono:

- `PyPDF2`
- `pdfminer`

Entrambe le librerie possono convertire file **PDF** in testo.

NOTA: In questa lezione ci concentreremo su `PyPDF2`

Estrarre testo da PDF

Cos'è PyPDF2

Cit. PyPDF2

PyPDF2 is a free and open-source pure-python PDF library capable of *splitting, merging, cropping, and transforming* the pages of PDF files.

It can also add custom data, viewing options, and passwords to PDF files. *PyPDF2 can retrieve text* and metadata from PDFs as well.

Estrarre testo da PDF

Cos'è PyPDF2

- PyPDF2 è una libreria Python per la manipolazione di file PDF
- Supporta operazioni come
 - *unire* testo
 - *dividere* testo
 - *cifrare* testo
 - *estrarre* testo
- Non supporta il riconoscimento ottico dei caratteri (OCR), lavora con testo selezionabile

Estrarre testo da PDF

Come funziona PyPDF2

PyPDF2 è un modulo python, dunque come ogni altro andrà prima di tutto installato:

```
pip install pypdf2
```

Dopo l'installazione sarà possibile utilizzare la libreria all'interno dei nostri programmi PYTHON.

NOTA: Attivare l'ambiente virtuale prima dell'installazione. Usare *venv* o *CONDA*

Estrarre testo da PDF

Come funziona PyPDF2

Metodi e Attributi di PdfReader

Di seguito i metodi e gli attributi principali di PdfReader:

Metodo/Attributo	Descrizione
<code>reader.pages</code>	Lista delle pagine del PDF
<code>reader.metadata</code>	Dizionario con i metadati del PDF (autore, titolo, ecc.)
<code>reader.get_page(n)</code>	Ottiene la pagina <code>n</code> del PDF
<code>reader.is_encrypted</code>	Restituisce <code>True</code> se il PDF è cifrato
<code>reader.decrypt(password)</code>	Decifra un PDF protetto da password

Estrarre testo da PDF

Come funziona PyPDF2

Esempio di Lettura

```
from PyPDF2 import PdfReader

# Leggo il documento pdf
reader = PdfReader("documento.pdf")

# attraverso reader.pages posso sapere le pagine presenti
num_pagine = len(reader.pages)
print(f"Numero di pagine: {num_pagine}")

# stampo ogni pagina
for page in reader.pages:
    print(page.extract_text())
```

Estrarre testo da PDF

Domande



Estrarre testo da PDF



1. Creare un programma python che fornito in input un documento pdf estrae il testo contenuto all'interno.
2. Creare un programma python che fornito un testo ed una chiave da cercare, recupera la prima informazione associata ad essa.
3. Creare un esercizio che fornito in input un documento pdf ed una chiave da cercare mi fornisca la prima informazione associata ad essa se presente.

Sfruttare le funzioni di: split, normalizzazione del testo, strip, sub, etc... per rendere il testo più possibile leggibile ed individuare facilmente le key.

