

# GABRIELE ROMANATO

Menu

## Estrarre testo da un documento PDF con Node.js

L'estrazione di testo da un file PDF è una delle operazioni più comuni e utili in vari contesti di sviluppo. Grazie alla potenza di Node.js e alla disponibilità di molte librerie open source, possiamo realizzare facilmente questa funzionalità. In questo articolo vedremo come estrarre il testo da un documento PDF utilizzando Node.js, attraverso librerie dedicate.

Esistono diverse librerie che permettono di lavorare con i PDF in Node.js. Due tra le più popolari per l'estrazione del testo sono:

1. **pdf-parse**: Una libreria che permette di analizzare e estrarre testo da file PDF in maniera semplice.
2. **pdf-lib**: Una libreria più potente e completa che consente di creare, modificare ed estrarre dati dai PDF.

Per questo articolo ci concentreremo sulla libreria pdf-parse in quanto è la soluzione più semplice ed efficace per estrarre testo da un PDF.

La libreria pdf-parse semplifica notevolmente il processo di estrazione del testo da un PDF. Vediamo un esempio di come utilizzarla.

```
const fs = require('fs');
const pdfParse = require('pdf-parse');

// Legge il file PDF
const dataBuffer = fs.readFileSync('document.pdf');

// Eseguo l'estrazione del testo
pdfParse(dataBuffer).then(function(data) {
  // Stampa il testo estratto
  console.log(data.text);
}).catch(function(error) {
  console.error('Errore durante l'estrazione del testo:', error);
});
```

Spiegazione del codice:

- **fs.readFileSync('document.pdf')**: Legge il contenuto del file PDF come un buffer binario.
- **pdfParse(dataBuffer)**: Passiamo il buffer alla funzione pdfParse, che restituisce una Promise che, una volta risolta, contiene il testo estratto dal PDF.
- **data.text**: La proprietà text contiene il testo estratto dal documento PDF.

Oltre al semplice estrarre testo, pdf-parse offre anche altre informazioni utili sul PDF, come:

- Numero di pagine (data.numpages).

- Dimensioni del file (`data.filesize`).
- Testo strutturato in blocchi o su singole pagine.

Puoi accedere a queste informazioni tramite l'oggetto `data` restituito dalla Promise. Ecco un esempio:

```
pdfParse(dataBuffer).then(function(data) {  
  console.log('Numero di pagine:', data.numpages);  
  console.log('Dimensioni del file:', data.filesize);  
  console.log('Testo estratto:', data.text);  
});
```

## Conclusion

In questo articolo abbiamo visto come estrarre testo da un documento PDF utilizzando Node.js e la libreria `pdf-parse`. Questo processo è rapido e può essere facilmente integrato in un'applicazione per automatizzare operazioni che riguardano l'estrazione e l'elaborazione di dati da PDF.

## Applicazioni Correlate



- 

### **Node.js Placeholder Image**

Applicazione per la generazione con Node.js di immagini segnaposto.  
Docker Docker Compose Node.js JavaScript ExpressJS



- 

## **Node.js URL Shortener**

Implementazione in Node.js di un sistema per l'abbreviazione degli URL.

Docker Docker Compose Node.js JavaScript Express JS MongoDB



# JS

- **JavaScript App Hash Change**

Applicazione che sfrutta gli hash degli URL per gestire contenuto dinamico in JavaScript.  
Docker Docker Compose Node.js JavaScript Express JS MySQL