Parsing PDF files in Node.js

Parsing a PDF file in Node.js is a common operation when you want to extract text, metadata, or other information from documents. In this article, we'll see how to use the `pdf-parse` module, a simple and effective library for extracting text content from PDFs.

Installation

To get started, create a new Node.js project or use an existing one, then install the `pdf-parse` package:

```
npm install pdf-parse
```

Example Code

The following script reads a PDF file from the filesystem and extracts its text content:

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

const dataBuffer = fs.readFileSync('document.pdf');

pdf(dataBuffer).then(function(data) {
    console.log('PDF Content:');
    console.log(data.text);
}).catch(function(error) {
    console.error('Error parsing PDF:', error);
});
```

Explanation

- **fs.readFileSync** reads the binary content of the PDF file.
- **pdf()** takes a buffer and returns a `Promise` containing the text and other data.
- In the `then` block, you access `data.text`, which contains the extracted text from the PDF.

Additional Information

The `pdf-parse` package also returns other useful details, such as:

```
{
  numpages: 2,
  numrender: 2,
  info: {
    PDFFormatVersion: '1.3',
    IsAcroFormPresent: false,
    IsXFAPresent: false,
    Title: 'Document Title',
    Author: 'Author',
    CreationDate: 'D:20230717120000Z'
  },
  metadata: null,
  text: 'Extracted text...'
}
```

Conclusion

With `pdf-parse`, parsing PDF files in Node.js is simple and straightforward. For more advanced applications, you might consider libraries like `pdfjs-dist` or `pdf-lib`, which offer greater control over the document structure.