# Read or Extract Text from PDF with Python — A Comprehensive Guide

[[Alice Yang](https://medium.com/@alice.yang_10652?source=post_page-----eb22c440e22a--------------------------------)](https://medium.com/@alice.yang_10652?source=post_page-----eb22c440e22a--------------------------------)

[Alice Yang](https://medium.com/@alice.yang_10652?source=post_page-----eb22c440e22a--------------------------------)

·

[Follow](https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fsubscribe%2Fuser%2Fab67334728f3&operation=register&redirect=https%3A%2F%2Fmedium.com%2F%40alice.yang_10652%2Fwith-read-or-extract-text-from-pdf-with-python-a-comprehensive-guide-eb22c440e22a&user=Alice+Yang&userId=ab67334728f3&source=post_page-ab67334728f3----eb22c440e22a---------------------post_header-----------)

5 min read

·

Sep 5

3

1



Extract Text from PDF using Python

PDF documents such as research papers, legal documents, contracts, or reports often contain important textual information. By extracting the text from these documents, you can make the information accessible in a format that can be easily searched, copied, or modified for further analysis or reference. In this article, we will explore how to **read or extract text from PDF documents using Python**.

We’ll discuss the following topics:

* [Extract Text from an Entire PDF in Python](https://medium.com/@alice.yang_10652/with-read-or-extract-text-from-pdf-with-python-a-comprehensive-guide-eb22c440e22a#9d4a)
* [Extract Text from a Particular Page in PDF in Python](https://medium.com/@alice.yang_10652/with-read-or-extract-text-from-pdf-with-python-a-comprehensive-guide-eb22c440e22a#6cff)
* [Extract Text from a Rectangular Area of a Page in PDF in Python](https://medium.com/@alice.yang_10652/with-read-or-extract-text-from-pdf-with-python-a-comprehensive-guide-eb22c440e22a#4fb5)
* [Extract Highlighted Text from a PDF in Python](https://medium.com/@alice.yang_10652/with-read-or-extract-text-from-pdf-with-python-a-comprehensive-guide-eb22c440e22a#5bd7)

# Python Library for Text Extraction from PDF

To perform text extraction on PDF files with Python, we can use the [Spire.PDF for Python](https://www.e-iceblue.com/Introduce/pdf-for-python.html) library.

Spire.PDF for Python is a feature-rich and user-friendly library that enables creating, reading, editing, and converting PDF files within Python applications. With this library, you can perform a wide range of manipulations on PDFs, including adding text or images, extracting text or images, adding digital signatures, adding or deleting pages, merging or splitting PDFs, creating bookmarks, adding text or image watermarks, inserting fillable forms and many more. In addition, you are also able to convert PDF files to various file formats, such as Word, Excel, images, HTML, SVG, XPS, OFD, PCL, and PostScript.

Installing Spire.PDF for Python is incredibly easy. Just follow these simple steps:

1. Open your project’s terminal.
2. Execute this pip command: **pip install spire.pdf**

# Extract Text from an Entire PDF in Python

You can simply extract text from an entire PDF document by looping through the pages in the document and then calling the **PdfPageBase.ExtractText()** function to extract text from every page of the PDF document.

Here is a simple example that shows how to extract text from an entire PDF document using Python and Spire.PDF for Python:

from spire.pdf.common import \*  
from spire.pdf import \*  
  
# Create a PdfDocument object  
doc = PdfDocument()  
# Load a PDF document  
doc.LoadFromFile("Sample.pdf")  
  
# Create an empty list to store extracted text  
list = []  
  
# Loop through the pages in the document  
for i in range(doc.Pages.Count):  
 page = doc.Pages.get\_Item(i)  
 # Extract text from each page and keep white spaces  
 text = page.ExtractText(True)  
 # Append the extracted text to the list  
 list.append(text)  
  
# Write the extracted text into a text file  
with open("ExtractTextFromDocument.txt", "w", encoding = "utf-8") as text\_file:  
 for text in list:  
 text\_file.write(text + "\n")  
  
# Close the PdfDocument object  
doc.Close()

# Extract Text from a Particular Page in PDF in Python

To extract text from a particular page, you can access that page from the page collection of the document using **PdfDocument.Pages[pageindex]** property and then call the **PdfPageBase.ExtractText()** function to extract text from that page.

Here is a simple example that shows how to extract text from a particular page of a PDF document using Python and Spire.PDF for Python:

from spire.pdf.common import \*  
from spire.pdf import \*  
  
# Create a PdfDocument object  
doc = PdfDocument()  
# Load a PDF document  
doc.LoadFromFile("Sample.pdf")  
  
# Get the first page of the document  
page = doc.Pages[0]  
  
# Extract text from the page and keep white spaces  
text = page.ExtractText(True)  
  
# Write the extracted text into a text file  
with open("ExtractTextFromPage.txt", "w", encoding = "utf-8") as text\_file:  
 text\_file.write(text)  
  
# Close the PdfDocument object  
doc.Close()

# Extract Text from a Rectangular Area of a Page in PDF in Python

In addition to extracting text from an entire document or a particular page, you are also able to extract text from a rectangular area of a page by passing a **RectangleF** object to the **PdfPageBase.ExtractText()** function as a parameter.

Here is a simple example that shows how to extract text from a rectangular area of a page of a PDF document using Python and Spire.PDF for Python:

from spire.pdf.common import \*  
from spire.pdf import \*  
  
# Create a PdfDocument object  
doc = PdfDocument()  
# Load a PDF document  
doc.LoadFromFile("Sample.pdf")  
  
# Get the first page of the document  
page = doc.Pages[0]  
  
# Define a rectangle to specify the area for text extraction  
rectangle = RectangleF(0.0, 180.0, 500.0, 200.0)  
  
# Extract text from the specified rectangular area of the page  
text = page.ExtractText(rectangle)  
  
# Write the extracted text into a text file  
with open("ExtractTextFromRectangularPageArea.txt", "w", encoding = "utf-8") as text\_file:  
 text\_file.write(text)  
  
# Close the PdfDocument object  
doc.Close()

# Extract Highlighted Text from a PDF in Python

When a section of text is highlighted in a PDF, a highlight annotation is created to represent that highlight. The annotation includes information about the position and extent of the highlighted text, as well as the appearance properties such as color and opacity.

To extract highlighted text from a PDF page, you need to find the highlight annotations on the page, then pass their locations (**RectangleF** objects) to the **PdfPageBase.ExtractText()** function to get the text marked by the highlight annotations.

Here is a simple example that shows how to extract highlighted text on a page of a PDF document using Python and Spire.PDF for Python:

from spire.pdf.common import \*  
from spire.pdf import \*  
  
# Create a PdfDocument object  
doc = PdfDocument()  
# Load a PDF document  
doc.LoadFromFile("Sample.pdf")  
  
# Get the first page of the document  
page = doc.Pages[0]  
  
# Create an empty list to store extracted text  
list = []  
  
# Get the annotation collection of the page  
annotations = page.AnnotationsWidget  
  
textMarkupAnnotation = None  
  
#Loop through the annotations in the collection  
if annotations.Count > 0:  
 for i in range(annotations.Count):  
 textMarkupAnnotation = annotations.get\_Item(i)  
 # Check if the annotation is of type PdfTextMarkupAnnotationWidget   
 if isinstance(textMarkupAnnotation, PdfTextMarkupAnnotationWidget):  
 # Extract the text marked by the annotation  
 text = page.ExtractText(textMarkupAnnotation.Bounds)  
 # Append the extracted text to the list  
 list.append(text)  
  
# Write the extracted text into a text file  
with open("ExtractHighlightedText.txt", "w", encoding = "utf-8") as text\_file:  
 for text in list:  
 text\_file.write(text + "\n")  
  
# Close the PdfDocument object  
doc.Close()

# Conclusion

This article demonstrated various scenarios to extract text from PDF documents using Python and Spire.PDF for Python. We hope you can find it helpful.

# Related Topics

[Convert PDF to Word DOCX or DOC with Python](https://medium.com/@alice.yang_10652/convert-pdf-to-word-docx-or-doc-with-python-d623fa917283)