

Read Word Documents with Python: Extract Data from Word



Alexander Stock · [Follow](#)

5 min read · Jun 20, 2024



Listen



Share



More



Extracting specific data, such as text, tables, images, or metadata, from Word documents programmatically for further analysis or integration into other systems. This can be useful for automating document processing tasks.

In this post, I'll guide you through the process of reading or extracting different types of data from a Word document using Python and the [Spire.Doc for Python](#) library.

- [Extract Text from a Specific Paragraph in Python](#)
- [Extract Text from an Entire Word Document in Python](#)

- [Extract Tables from a Word Document in Python](#)
- [Extract Images from a Word Document in Python](#)
- [Extract Metadata of a Word Document in Python](#)

Python Library for Reading Word Documents

Spire.Doc is a Python library that simplifies working with Microsoft Office Word documents. It allows you to read, write, and manipulate Word documents programmatically, making it easier to automate document-related tasks.

You can install the library from PyPI using the following command.

```
pip install Spire.Doc
```

Extract Text from a Specific Paragraph in Python

Spire.Doc makes it easy to work with specific parts of a Word document. You can access a section using `Document.Sections[index]`, then a paragraph within that section using `Section.Paragraphs[index]`. Finally, you can get the text of the paragraph using `Paragraph.Text`.

```
from spire.doc import *
from spire.doc.common import *

# Create a Document object
doc = Document()

# Load a Word document
doc.LoadFromFile("C:\\Users\\Administrator\\Desktop\\input.docx")

# Get a specific section
section = doc.Sections[0]

# Get a specific paragraph
paragraph = section.Paragraphs[3]

# Get text of the paragraph
str = paragraph.Text
```

```
# Print result
print(str)
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL

PS C:\Users\Administrator\Python> & C:/Users/Administrator/AppData/Local/Programs/Python/Python310/python.exe c:/Users/Administrator/Python/WordExamples/ExtractTextFromParagraph.py
Unless otherwise stated, this Policy describes and governs the information collection, use, and sharing practices of [COMPANY NAME] with respect to your use of our website ([WEBSITE URL]) and the services ("Services") we provide and/or host on our servers.
PS C:\Users\Administrator\Python>
```

Figure 1. Extract text from a paragraph.

Extract Text from an Entire Word Document in Python

To get text of an entire Word document, you can simply use the `Document.GetText()` method.

```
from spire.doc import *
from spire.doc.common import *

# Create a Document object
doc = Document()

# Load a Word file
doc.LoadFromFile("C:\\Users\\Administrator\\Desktop\\input.docx")

# Get text from the entire document
text = doc.GetText()

# Print result
print(text)
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL

* Data collection, storage, and processing should be simplified as much as possible to enhance security, ensure consistency, and make the practices easy for users to understand.
* Data practices should meet the reasonable expectations of users.
Information We Collect
We collect information in multiple ways, including when you provide information directly to us; when we passively collect information from you, such as from your browser or device; and from third parties.
Information You Provide Directly to Us
We will collect any information you provide to us. We may collect information from you in a variety of ways, such as when you: (a) create an online account, (b) make a donation or purchase, (c) contact us or provide feedback, (d) subscribe to our newsletter.
Information that Is Automatically Collected
Device/Usage Information
We may automatically collect certain information about the computer or devices (including mobile devices or tablets) you use to access the Services. As described further below, we may collect and analyze (a) device information such as IP addresses, location information (by country and city),

unique device identifiers, IMEI and TCP/IP address, browser types, browser language, operating system, mobile device carrier information, and (b) information related to the ways in which you interact with the Services, such as referring and exit web pages and URLs, platform type, the number of clicks, domain names, landing pages, pages and content viewed and the order of those pages, statistical information about the use of the Services, the amount of time spent on particular pages, the date and time you used the Services, the frequency of your use of the Services, error logs, and other similar information. As described further below, we may use third-party analytics providers and technologies, including cookies and similar tools, to assist in collecting this information.

PS C:\Users\Administrator\Python> █
```

Figure 2. Extract text from an entire document.

Extract Tables from a Word Document in Python

With Spire.Doc, you can access the tables in a section using `Section.Tables`. You can then get a specific table and retrieve the cells. The text content of each cell is available through `TableCell.Paragraphs.get_Item().Text`.

```
from spire.doc import *
from spire.doc.common import *

# Create a Document object
doc = Document()

# Load a Word document
doc.LoadFromFile("C:\\Users\\Administrator\\Desktop\\input.docx")

# Iterate through the sections
for i in range(doc.Sections.Count):

    # Get a specific section
    section = doc.Sections.get_Item(i)

    # Get tables from the section
    tables = section.Tables
```

```

# Iterate through the tables
for j in range(0, tables.Count):

    # Get a certain table
    table = tables.get_Item(j)

    # Declare a variable to store the table data
    tableData = ""

    # Iterate through the rows of the table
    for m in range(0, table.Rows.Count):

        # Iterate through the cells of the row
        for n in range(0, table.Rows.get_Item(m).Cells.Count):

            # Get a cell
            cell = table.Rows.get_Item(m).Cells.get_Item(n)


            # Get the text in the cell
            cellText = ""
            for para in range(cell.Paragraphs.Count):
                paragraphText = cell.Paragraphs.get_Item(para).Text
                cellText += (paragraphText + " ")

            # Add the text to the string
            tableData += cellText

        # Add a new line
        tableData += "\n"

    # Save the table data to a text file
    with open(f"output/WordTable_{i+1}_{j+1}.txt", "w", encoding="utf-8") as f:
        f.write(tableData)

```



Product	Description	Unit price	Amount	
Wireless Mouse	700mAh, Rechargeable, Black	\$13	\$1950	
Keyboard	Wired, Black	\$12	\$2400	
Monitor	27 Inch FHD, HDMI, VGA	\$150	\$12000	
Seagate HDD	1TB, Portable, USB 3.0	\$69	\$4380	

Figure 3. Extract tables from a Word document.

Extract Images from a Word Document in Python

To extract image from a Word document, you need first to iterate through the child objects in the document. Then, determine if a child object is a `DocPicture`. If yes, you can access the image data using `DocPicture.ImageBytes` and save it as an image file.

```
import queue
from spire.doc import *
from spire.doc.common import *

# Create a Document object
doc = Document()

# Load a Word file
doc.LoadFromFile("C:\\Users\\Administrator\\Desktop\\input2.docx")

# Create a Queue object
nodes = queue.Queue()
nodes.put(doc)

# Create a list
images = []

while nodes.qsize() > 0:
    node = nodes.get()

    # Loop through the child objects in the document
    for i in range(node.ChildObjects.Count):
        child = node.ChildObjects.get_Item(i)

        # Detect if a child object is a picture
        if child.DocumentObjectType == DocumentObjectType.Picture:
            picture = child if isinstance(child, DocPicture) else None
            dataBytes = picture.ImageBytes

            # Add the image data to the list
            images.append(dataBytes)

        elif isinstance(child, ICompositeObject):
            nodes.put(child if isinstance(child, ICompositeObject) else None)

# Loop through the images in the list
for i, item in enumerate(images):
    fileName = "Image-{}.png".format(i)
    with open("ExtractedImages/"+fileName, 'wb') as imageFile:
```



```
# Write the image to a specified path
imageFile.write(item)
```

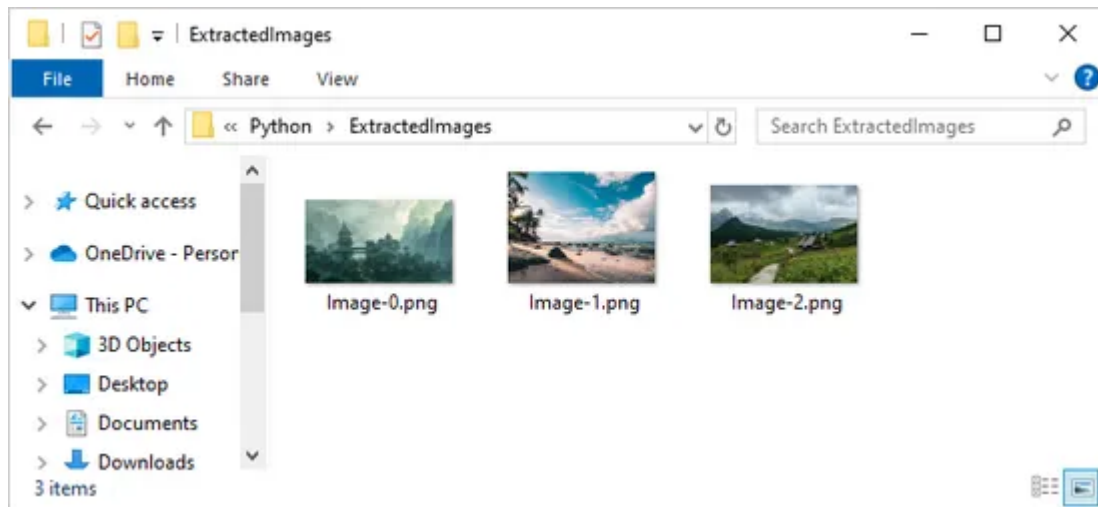


Figure 4. Extract images from a Word document.

Extract Metadata of a Word Document in Python

With Spire.Doc, you can access the built-in document properties of a Word document using the `Document.BuiltinDocumentProperties` property. This provides access to standard properties like Author, Company, Title, and Subject.

```
from spire.doc import *
from spire.doc.common import *

# Create a Document object
doc = Document()

# Load a Word document
doc.LoadFromFile("C:\\Users\\Administrator\\Desktop\\input.docx")

# Get the built-in properties of the document
builtinProperties = doc.BuiltinDocumentProperties

# Get the value of the built-in properties
properties = [
    "Author: " + builtinProperties.Author,
    "Company: " + builtinProperties.Company,
    "Title: " + builtinProperties.Title,
    "Subject: " + builtinProperties.Subject,
    "Keywords: " + builtinProperties.Keywords,
    "Category: " + builtinProperties.Category,
    "Manager: " + builtinProperties.Manager,
    "Comments: " + builtinProperties.Comments,
```

```
"Hyperlink Base: " + builtinProperties.HyperLinkBase,  
"Word Count: " + str(builtinProperties.WordCount),  
"Page Count: " + str(builtinProperties.PageCount),  
]  
  
# Print result  
for i in range(0, len(properties)):  
    print(properties[i])
```

Conclusion

In this blog post, we've explored how to extract text from specific paragraphs or the entire document, access the tables within the document, retrieve embedded images, and read standard document properties like the author, title, and subject, using Python. Hopefully, you can find this post informative and helpful.

See Also

[Create a Word Document with Python](#)

[Convert Word to Images in Python](#)

[Convert Word to PDF in Python](#)

[Find and Replace Text in Word in Python](#)

Python

Read Word Document

Extract Text From Word

Extract Images From Word

Extract Tables From Word



Follow

Written by Alexander Stock

160 Followers · 4 Following

Experienced software development consultant and blogger with more than 10 years in the field, focusing on office document tools and knowledge sharing.