pdfminer.six

Release 20240706

Yusuke Shinyama, Philippe Guglielmetti & Pieter Marsman

Jan 15, 2026

CONTENTS

1 Content **3**
- 1.1 Tutorials . . . 3
- 1.2 How-to guides . . . 6
- 1.3 Topics . . . 15
- 1.4 API Reference . . . 17
- 1.5 Frequently asked questions . . . 24

2 Features **27**

3 Installation instructions **29**

4 Contributing **31**

Index **33**

We fathom PDF.

Pdfminer.six is a python package for extracting information from PDF documents.

Check out the source on github.

CHAPTER

ONE

CONTENT

This documentation is organized into four sections (according to the Diátaxis documentation framework). The *Tutorials* section helps you setup and use pdfminer.six for the first time. Read this section if this is your first time working with pdfminer.six. The *How-to guides* offers specific recipies for solving common problems. Take a look at the *Topics* if you want more background information on how pdfminer.six works internally. The *API Reference* provides detailed api documentation for all the common classes and functions in pdfminer.six.

1.1 Tutorials

Tutorials help you get started with specific parts of pdfminer.six.

1.1.1 Install pdfminer.six as a Python package

To use pdfminer.six for the first time, you need to install the Python package in your Python environment.

This tutorial requires you to have a system with a working Python and pip installation. If you don't have one and don't know how to install it, take a look at The Hitchhiker's Guide to Python!.

Install using pip

Run the following command on the commandline to install pdfminer.six as a Python package:

```
pip install pdfminer.six
```

Test pdfminer.six installation

You can test the pdfminer.six installation by importing it in Python.

Open an interactive Python session from the commandline import pdfminer .six:

```
>>> import pdfminer
>>> print(pdfminer.__version__)
'<installed version>'
```

Now you can use pdfminer.six as a Python package. But pdfminer.six also comes with a couple of useful commandline tools. To test if these tools are correctly installed, run the following on your commandline:

```
$ pdf2txt.py --version
pdfminer.six <installed version>
```

1.1.2 Extract text from a PDF using the commandline

pdfminer.six has several tools that can be used from the command line. The command-line tools are aimed at users that occasionally want to extract text from a pdf.

Take a look at the high-level or composable interface if you want to use pdfminer.six programmatically.

Examples

pdf2txt.py

```
$ pdf2txt.py example.pdf
all the text from the pdf appears on the command line
```

The *pdf2txt.py* tool extracts all the text from a PDF. It uses layout analysis with sensible defaults to order and group the text in a sensible way.

dumppdf.py

```
$ dumppdf.py -a example.pdf
<pdf><object id="1">
...
</object>
...
</pdf>
```

The *dumppdf.py* tool can be used to extract the internal structure from a PDF. This tool is primarily for debugging purposes, but that can be useful to anybody working with PDF's.

1.1.3 Extract text from a PDF using Python

The high-level API can be used to do common tasks.

The most simple way to extract text from a PDF is to use *extract_text*:

```
>>> from pdfminer.high_level import extract_text
>>> text = extract_text('samples/simple1.pdf')
>>> print(repr(text))
'Hello \n\nWorld\n\nHello \n\nWorld\n\nH e l l o  \n\nW o r l d\n\nH e l l o  \n\nW o␣
↪r l d\n\n\x0c'
>>> print(text)
...
Hello

World

Hello

World

H e l l o

W o r l d
```

(continues on next page)

(continued from previous page)

```
H e l l o

W o r l d
```

To read text from a PDF and print it on the command line:

```
>>> from io import StringIO
>>> from pdfminer.high_level import extract_text_to_fp
>>> output_string = StringIO()
>>> with open('samples/simple1.pdf', 'rb') as fin:
...     extract_text_to_fp(fin, output_string)
>>> print(output_string.getvalue().strip())
Hello WorldHello WorldHello WorldHello World
```

Or to convert it to html and use layout analysis:

```
>>> from io import StringIO
>>> from pdfminer.high_level import extract_text_to_fp
>>> from pdfminer.layout import LAParams
>>> output_string = StringIO()
>>> with open('samples/simple1.pdf', 'rb') as fin:
...     extract_text_to_fp(fin, output_string, laparams=LAParams(),
...                        output_type='html', codec=None)
```

1.1.4 Extract text from a PDF using Python - part 2

The command line tools and the high-level API are just shortcuts for often used combinations of pdfminer.six components. You can use these components to modify pdfminer.six to your own needs.

For example, to extract the text from a PDF file and save it in a python variable:

```
from io import StringIO

from pdfminer.converter import TextConverter
from pdfminer.layout import LAParams
from pdfminer.pdfdocument import PDFDocument
from pdfminer.pdfinterp import PDFResourceManager, PDFPageInterpreter
from pdfminer.pdfpage import PDFPage
from pdfminer.pdfparser import PDFParser

output_string = StringIO()
with open('samples/simple1.pdf', 'rb') as in_file:
    parser = PDFParser(in_file)
    doc = PDFDocument(parser)
    rsrcmgr = PDFResourceManager()
    device = TextConverter(rsrcmgr, output_string, laparams=LAParams())
    interpreter = PDFPageInterpreter(rsrcmgr, device)
    for page in PDFPage.create_pages(doc):
        interpreter.process_page(page)

print(output_string.getvalue())
```

1.1.5 Extract elements from a PDF using Python

The high level functions can be used to achieve common tasks. In this case, we can use *extract_pages*:

```
from pdfminer.high_level import extract_pages
for page_layout in extract_pages("test.pdf"):
    for element in page_layout:
        print(element)
```

Each `element` will be an `LTTextBox`, `LTFigure`, `LTLine`, `LTRect` or an `LTImage`. Some of these can be iterated further, for example iterating though an `LTTextBox` will give you an `LTTextLine`, and these in turn can be iterated through to get an `LTChar`. See the diagram here: *Layout analysis algorithm*.

Let's say we want to extract all of the text. We could do:

```
from pdfminer.high_level import extract_pages
from pdfminer.layout import LTTextContainer
for page_layout in extract_pages("test.pdf"):
    for element in page_layout:
        if isinstance(element, LTTextContainer):
            print(element.get_text())
```

Or, we could extract the fontname or size of each individual character:

```
from pdfminer.high_level import extract_pages
from pdfminer.layout import LTTextContainer, LTChar
for page_layout in extract_pages("test.pdf"):
    for element in page_layout:
        if isinstance(element, LTTextContainer):
            for text_line in element:
                for character in text_line:
                    if isinstance(character, LTChar):
                        print(character.fontname)
                        print(character.size)
```

1.2 How-to guides

How-to guides help you to solve specific problems with pdfminer.six.

1.2.1 How to extract images from a PDF

Before you start, make sure you have *installed pdfminer.six*. The second thing you need is a PDF with images. If you don't have one, you can download this research paper with images of cats and dogs and save it as *example.pdf*:

```
$ curl https://www.robots.ox.ac.uk/~vgg/publications/2012/parkhi12a/parkhi12a.pdf --
↪output example.pdf
```

Then run the *pdf2txt* command:

```
$ pdf2txt.py example.pdf --output-dir cats-and-dogs
```

This command extracts all the images from the PDF and saves them into the *cats-and-dogs* directory.

1.2.2 How to extract AcroForm interactive form fields from a PDF using PDFMiner

Before you start, make sure you have *installed pdfminer.six*.

The second thing you need is a PDF with AcroForms (as found in PDF files with fillable forms or multiple choices). There are some examples of these in the GitHub repository under *samples/acroform*.

Only AcroForm interactive forms are supported, XFA forms are not supported.

```
from pdfminer.pdfparser import PDFParser
from pdfminer.pdfdocument import PDFDocument
from pdfminer.pdftypes import resolve1
from pdfminer.psparser import PSLiteral, PSKeyword
from pdfminer.utils import decode_text


data = {}


def decode_value(value):

    # decode PSLiteral, PSKeyword
    if isinstance(value, (PSLiteral, PSKeyword)):
        value = value.name

    # decode bytes
    if isinstance(value, bytes):
        value = decode_text(value)

    return value


with open(file_path, 'rb') as fp:
    parser = PDFParser(fp)

    doc = PDFDocument(parser)
    res = resolve1(doc.catalog)

    if 'AcroForm' not in res:
        raise ValueError("No AcroForm Found")

    fields = resolve1(doc.catalog['AcroForm'])['Fields']  # may need further resolving

    for f in fields:
        field = resolve1(f)
        name, values = field.get('T'), field.get('V')

        # decode name
        name = decode_text(name)

        # resolve indirect obj
        values = resolve1(values)

        # decode value(s)
        if isinstance(values, list):
            values = [decode_value(v) for v in values]
        else:
            values = decode_value(values)
```

(continues on next page)

(continued from previous page)

```
        data.update({name: values})

        print(name, values)
```

This code snippet will print all the fields' names and values and save them in the "data" dictionary.

How it works:

- Initialize the parser and the PDFDocument objects

```
parser = PDFParser(fp)
doc = PDFDocument(parser)
```

- Get the Catalog

  (the catalog contains references to other objects defining the document structure, see section 7.7.2 of PDF 32000-1:2008 specs: https://opensource.adobe.com/dc-acrobat-sdk-docs/pdflsdk/index.html#pdf-reference)

```
res = resolve1(doc.catalog)
```

- Check if the catalog contains the AcroForm key and raise ValueError if not

  (the PDF does not contain Acroform type of interactive forms if this key is missing in the catalog, see section 12.7.2 of PDF 32000-1:2008 specs)

```
if 'AcroForm' not in res:
    raise ValueError("No AcroForm Found")
```

- Get the field list resolving the entry in the catalog

```
fields = resolve1(doc.catalog['AcroForm'])['Fields']
for f in fields:
    field = resolve1(f)
```

- Get field name and field value(s)

```
name, values = field.get('T'), field.get('V')
```

- Decode field name.

```
name = decode_text(name)
```

- Resolve indirect field value objects

```
values = resolve1(value)
```

- Call the value(s) decoding method as needed

  (a single field can hold multiple values, for example, a combo box can hold more than one value at a time)

```
if isinstance(values, list):
    values = [decode_value(v) for v in values]
else:
    values = decode_value(values)
```

(the decode_value method takes care of decoding the field's value, returning a string)

- Decode PSLiteral and PSKeyword field values

```
if isinstance(value, (PSLiteral, PSKeyword)):
    value = value.name
```

- Decode bytes field values

```
if isinstance(value, bytes):
    value = utils.decode_text(value)
```

1.2.3 How to resolve the target page of ToC entries

pdfminer.six allows to access the Table of Contents (or "Outlines" as called in the PDF internal structure) of a document through the method `PDFDocument.get_outlines()`.

A minimal example would be:

```
from pathlib import Path
from pdfminer.pdfparser import PDFParser, PDFSyntaxError
from pdfminer.pdfdocument import PDFDocument, PDFNoOutlines


file_name = Path("...")

with open(file_name, "rb") as fp:
    try:
        parser = PDFParser(fp)
        document = PDFDocument(parser)
        outlines = document.get_outlines()
        for (level, title, dest, a, se) in outlines:
            ...  # do something
    except PDFNoOutlines:
        print("No outlines found.")
    except PDFSyntaxError:
        print("Corrupted PDF or non-PDF file.")
    finally:
        parser.close()
```

But what do the different fields of each outline entry mean? To answer this question we can refer to the section *12.3.3 Document Outline* of the PDF Reference:

- **Level** (`int`): This is, unsurprisingly, the level at which the entry is. Entries at the top level will have level `1`. Entries nested within those ones (i.e., their children), will have level `2`, and so on.
- **Title** (`str`): Again, quite self-explanatory, this field contains the name of the entry. For example: "1. Introduction".
- **Dest** (`Union[list, bytes]`, *optional*): This is where things start to get interesting. First thing to mention is that if a **Dest** entry is present, the **A** entry shall not be present. Both of them allow to specify the object the entry targets (this could be a page or any other object). Destinations can be specified in multiple ways. In order to not paraphrase what the PDF Reference states, we refer the reader to the chapter *12.3.2 Destinations* for more information on the topic.
- **A** (`pdfminer.pdftypes.PDFObjRef`, *optional*): Alternatively to using a destination, the target of an entry can also be specified as an action. Again, actions can get somewhat complicated, so we refer the reader to the chapter *12.6 Actions* of the reference.
- **SE** (`pdfminer.pdftypes.PDFObjRef`, *optional*): This field contains the structure element the entry points at. More information about structure elements can be found in the chapter *14.7.2 Structure Hierar-*

chy. It is worth mentioning that most PDFs will not include this field, using **Dest** or **A** instead, or if they do, they might still include a destination (**Dest**) to keep compatibility with PDF versions previous to 1.3.

Unfortunately, pdfminer.six doesn't expose the page number that each of the entries targets. However, once we know what each of the fields above mean, we can implement a ToC-entry page number resolver ourselves:

```
from enum import Enum, auto
from pathlib import Path
from typing import Any, Optional
from pdfminer.pdfdocument import PDFDocument, PDFNoOutlines
from pdfminer.pdfpage import PDFPage, LITERAL_PAGE
from pdfminer.pdfparser import PDFParser, PDFSyntaxError
from pdfminer.pdftypes import PDFObjRef


class PDFRefType(Enum):
    """PDF reference type."""

    PDF_OBJ_REF = auto()
    DICTIONARY = auto()
    LIST = auto()
    NAMED_REF = auto()
    UNK = auto()  # fallback


class RefPageNumberResolver:
    """PDF Reference to page number resolver.

    .. note::

       Remote Go-To Actions (see 12.6.4.3 in
       `https://www.adobe.com/go/pdfreference/`__)
       are out of the scope of this resolver.

    Attributes:
        document (:obj:`pdfminer.pdfdocument.PDFDocument`):
            The document that contains the references.
        objid_to_pagenum (:obj:`dict[int, int]`):
            Mapping from an object id to the number of the page that contains
            that object.
    """

    def __init__(self, document: PDFDocument):
        self.document = document
        # obj_id -> page_number
        self.objid_to_pagenum: dict[int, int] = {
            page.pageid: page_num
            for page_num, page in enumerate(PDFPage.create_pages(document), 1)
        }

    @classmethod
    def get_ref_type(cls, ref: Any) -> PDFRefType:
        """Get the type of a PDF reference."""
        if isinstance(ref, PDFObjRef):
            return PDFRefType.PDF_OBJ_REF
        elif isinstance(ref, dict) and "D" in ref:
            return PDFRefType.DICTIONARY
        elif isinstance(ref, list) and any(isinstance(e, PDFObjRef) for e in ref):
```

(continues on next page)

(continued from previous page)

```
            return PDFRefType.LIST
        elif isinstance(ref, bytes):
            return PDFRefType.NAMED_REF
        else:
            return PDFRefType.UNK

    @classmethod
    def is_ref_page(cls, ref: Any) -> bool:
        """Check whether a reference is of type '/Page'.

        Args:
            ref (:obj:`Any`):
                The PDF reference.

        Returns:
            :obj:`bool`: :obj:`True` if the reference references
            a page, :obj:`False` otherwise.
        """
        return isinstance(ref, dict) and "Type" in ref and ref["Type"] is LITERAL_PAGE

    def resolve(self, ref: Any) -> Optional[int]:
        """Resolve a PDF reference to a page number recursively.

        Args:
            ref (:obj:`Any`):
                The PDF reference.

        Returns:
            :obj:`Optional[int]`: The page number or :obj:`None`
            if the reference could not be resolved (e.g., remote Go-To
            Actions or malformed references).
        """
        ref_type = self.get_ref_type(ref)

        if ref_type is PDFRefType.PDF_OBJ_REF and self.is_ref_page(ref.resolve()):
            return self.objid_to_pagenum.get(ref.objid)
        elif ref_type is PDFRefType.PDF_OBJ_REF:
            return self.resolve(ref.resolve())

        if ref_type is PDFRefType.DICTIONARY:
            return self.resolve(ref["D"])

        if ref_type is PDFRefType.LIST:
            # Get the PDFObjRef in the list (usually first element).
            return self.resolve(next(filter(lambda e: isinstance(e, PDFObjRef), ref)))

        if ref_type is PDFRefType.NAMED_REF:
            return self.resolve(self.document.get_dest(ref))

        return None  # PDFRefType.UNK
```

The class `PDFRefType` is just a helper to categorize the type of reference we are dealing with. Due to the fact that a reference can point to another reference, in some cases we will have to recursively call `RefPageNumberResolver.resolve()` until we finally reach a page object. Then, we can get the page number by accessing the dictionary `RefPageNumberResolver.objid_to_pagenum`, which maps the page object id to the page number.

Using this page number resolver, we can for example print the Table of Contents of a document in a human-readable format with the following code:

```
def print_outlines(file: str) -> dict[int, int]:
    """Pretty print the outlines (ToC) of a PDF document."""
    with open(file, "rb") as fp:
        try:
            parser = PDFParser(fp)
            document = PDFDocument(parser)

            ref_pagenum_resolver = RefPageNumberResolver(document)

            outlines = list(document.get_outlines())
            if not outlines:
                print("No outlines found.")
            for (level, title, dest, a, se) in outlines:
                if dest:
                    page_num = ref_pagenum_resolver.resolve(dest)
                elif a:
                    page_num = ref_pagenum_resolver.resolve(a)
                elif se:
                    page_num = ref_pagenum_resolver.resolve(se)
                else:
                    page_num = None

                # Calculate leading spaces and filling dots for formatting.
                leading_spaces = (level-1) * 4
                fill_dots = 80 - len(title) - leading_spaces

                print(
                    f"{' ' * leading_spaces}"
                    f"{title}",
                    f"{'.' * fill_dots}",
                    f"{page_num:>3}"
                )
        except PDFNoOutlines:
            print("No outlines found.")
        except PDFSyntaxError:
            print("Corrupted PDF or non-PDF file.")
        finally:
            try:
                parser.close()
            except NameError:
                pass  # nothing to do


def main():
    file_name = Path("...")
    print_outlines(file_name)


if __name__ == "__main__":
    main()
```

1.2.4 How to extract font names and sizes from PDF's

Before you start, make sure you have *installed pdfminer.six*.

The following code sample shows how to extract font names and sizes for each of the characters. This uses the [simple1.pdf](https://raw.githubusercontent.com/pdfminer/pdfminer.six/master/samples/simple1.pdf).

```
from pathlib import Path
from typing import Iterable, Any

from pdfminer.high_level import extract_pages


def show_ltitem_hierarchy(o: Any, depth=0):
    """Show location and text of LTItem and all its descendants"""
    if depth == 0:
        print('element                        font                  stroking color␣
↪text')
        print('------------------------------ --------------------- --------------  --
↪--------')

    print(
        f'{get_indented_name(o, depth):<30.30s} '
        f'{get_optional_fontinfo(o):<20.20s} '
        f'{get_optional_color(o):<17.17s}'
        f'{get_optional_text(o)}'
    )

    if isinstance(o, Iterable):
        for i in o:
            show_ltitem_hierarchy(i, depth=depth + 1)


def get_indented_name(o: Any, depth: int) -> str:
    """Indented name of class"""
    return '  ' * depth + o.__class__.__name__


def get_optional_fontinfo(o: Any) -> str:
    """Font info of LTChar if available, otherwise empty string"""
    if hasattr(o, 'fontname') and hasattr(o, 'size'):
        return f'{o.fontname} {round(o.size)}pt'
    return ''

def get_optional_color(o: Any) -> str:
    """Font info of LTChar if available, otherwise empty string"""
    if hasattr(o, 'graphicstate'):
        return f'{o.graphicstate.scolor}'
    return ''


def get_optional_text(o: Any) -> str:
    """Text of LTItem if available, otherwise empty string"""
    if hasattr(o, 'get_text'):
        return o.get_text().strip()
    return ''
```

(continues on next page)

(continued from previous page)

```
path = Path('samples/simple1.pdf').expanduser()
pages = extract_pages(path)
show_ltitem_hierarchy(pages)
```

The output looks like below. Note that it shows the hierarchical structure of the layout elements. The layout algorithm groups characters into lines and lines into boxes. And boxes appear on a page. The pages, boxes and lines do not have font information because this can change for each character. The stroking color is always *None* in this example, but it will contain the color if the PDF does specify colors.

```
element                        font                  stroking color  text
------------------------------ --------------------- --------------  ----------
generator
  LTPage
    LTTextBoxHorizontal                                              Hello
      LTTextLineHorizontal                                           Hello
        LTChar                 Helvetica 24pt        None            H
        LTChar                 Helvetica 24pt        None            e
        LTChar                 Helvetica 24pt        None            l
        LTChar                 Helvetica 24pt        None            l
        LTChar                 Helvetica 24pt        None            o
        LTChar                 Helvetica 24pt        None
        LTAnno
    LTTextBoxHorizontal                                              World
      LTTextLineHorizontal                                           World
        LTChar                 Helvetica 24pt        None            W
        LTChar                 Helvetica 24pt        None            o
        LTChar                 Helvetica 24pt        None            r
        LTChar                 Helvetica 24pt        None            l
        LTChar                 Helvetica 24pt        None            d
        LTAnno
    LTTextBoxHorizontal                                              Hello
      LTTextLineHorizontal                                           Hello
        LTChar                 Helvetica 24pt        None            H
        LTChar                 Helvetica 24pt        None            e
        LTChar                 Helvetica 24pt        None            l
        LTChar                 Helvetica 24pt        None            l
        LTChar                 Helvetica 24pt        None            o
        LTChar                 Helvetica 24pt        None
        LTAnno
    LTTextBoxHorizontal                                              World
      LTTextLineHorizontal                                           World
        LTChar                 Helvetica 24pt        None            W
        LTChar                 Helvetica 24pt        None            o
        LTChar                 Helvetica 24pt        None            r
        LTChar                 Helvetica 24pt        None            l
        LTChar                 Helvetica 24pt        None            d
        LTAnno
    LTTextBoxHorizontal                                              H e l l o
      LTTextLineHorizontal                                           H e l l o
        LTChar                 Helvetica 24pt        None            H
        LTAnno
        LTChar                 Helvetica 24pt        None            e
        LTAnno
        LTChar                 Helvetica 24pt        None            l
        LTAnno
        LTChar                 Helvetica 24pt        None            l
```

(continues on next page)

(continued from previous page)

```
      LTAnno
      LTChar                  Helvetica 24pt       None             o
      LTAnno
      LTChar                  Helvetica 24pt       None
      LTAnno
    LTTextBoxHorizontal                                            W o r l d
      LTTextLineHorizontal                                         W o r l d
        LTChar                Helvetica 24pt       None             W
        LTAnno
        LTChar                Helvetica 24pt       None             o
        LTAnno
        LTChar                Helvetica 24pt       None             r
        LTAnno
        LTChar                Helvetica 24pt       None             l
        LTAnno
        LTChar                Helvetica 24pt       None             d
        LTAnno
    LTTextBoxHorizontal                                            H e l l o
      LTTextLineHorizontal                                         H e l l o
        LTChar                Helvetica 24pt       None             H
        LTAnno
        LTChar                Helvetica 24pt       None             e
        LTAnno
        LTChar                Helvetica 24pt       None             l
        LTAnno
        LTChar                Helvetica 24pt       None             l
        LTAnno
        LTChar                Helvetica 24pt       None             o
        LTAnno
        LTChar                Helvetica 24pt       None
        LTAnno
    LTTextBoxHorizontal                                            W o r l d
      LTTextLineHorizontal                                         W o r l d
        LTChar                Helvetica 24pt       None             W
        LTAnno
        LTChar                Helvetica 24pt       None             o
        LTAnno
        LTChar                Helvetica 24pt       None             r
        LTAnno
        LTChar                Helvetica 24pt       None             l
        LTAnno
        LTChar                Helvetica 24pt       None             d
        LTAnno
```

1.3 Topics

1.3.1 Converting a PDF file to text

Most PDF files look like they contain well-structured text. But the reality is that a PDF file does not contain anything that resembles paragraphs, sentences or even words. When it comes to text, a PDF file is only aware of the characters and their placement.

This makes extracting meaningful pieces of text from PDF files difficult. The characters that compose a paragraph are no different from those that compose the table, the page footer or the description of a figure. Unlike other document

formats, like a *.txt* file or a word document, the PDF format does not contain a stream of text.

A PDF document consists of a collection of objects that together describe the appearance of one or more pages, possibly accompanied by additional interactive elements and higher-level application data. A PDF file contains the objects making up a PDF document along with associated structural information, all represented as a single self-contained sequence of bytes.[1]

Layout analysis algorithm

PDFMiner attempts to reconstruct some of those structures by using heuristics on the positioning of characters. This works well for sentences and paragraphs because meaningful groups of nearby characters can be made.

The layout analysis consists of three different stages: it groups characters into words and lines, then it groups lines into boxes and finally it groups textboxes hierarchically. These stages are discussed in the following sections. The resulting output of the layout analysis is an ordered hierarchy of layout objects on a PDF page.

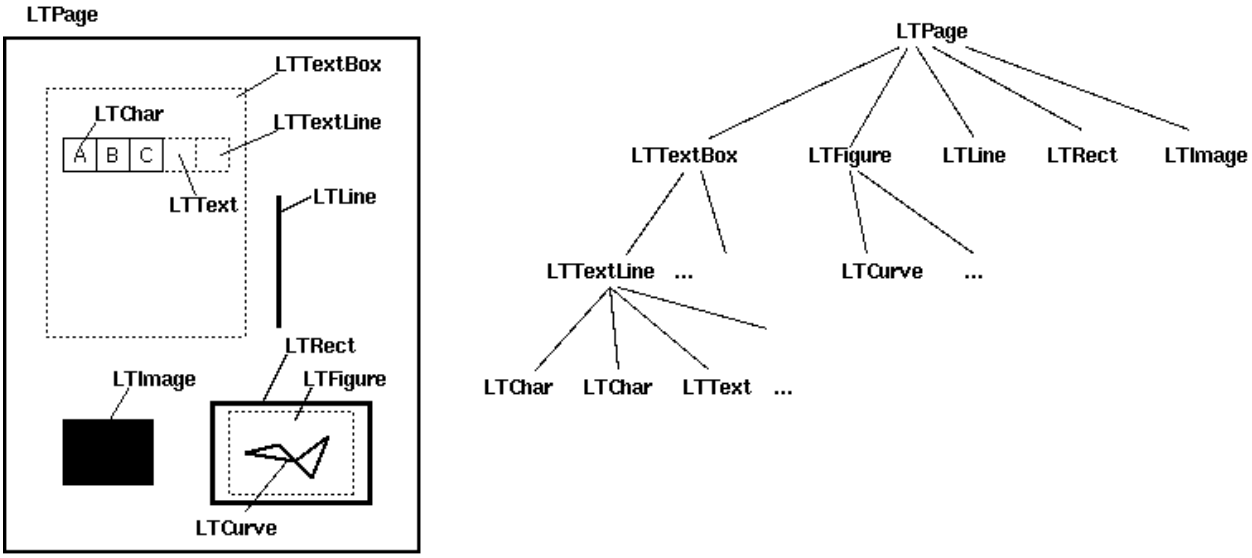


Fig. 1: The output of the layout analysis is a hierarchy of layout objects.

The output of the layout analysis heavily depends on a couple of parameters. All these parameters are part of the *LAParams* class.

Grouping characters into words and lines

The first step in going from characters to text is to group characters in a meaningful way. Each character has an x-coordinate and a y-coordinate for its bottom-left corner and upper-right corner, i.e. its bounding box. Pdfminer.six uses these bounding boxes to decide which characters belong together.

Characters that are both horizontally and vertically close are grouped onto one line. How close they should be is determined by the *char_margin* (M in the figure) and the *line_overlap* (not in figure) parameter. The horizontal *distance* between the bounding boxes of two characters should be smaller than the *char_margin* and the vertical *overlap* between the bounding boxes should be smaller than the *line_overlap*.

The values of *char_margin* and *line_overlap* are relative to the size of the bounding boxes of the characters. The *char_margin* is relative to the maximum width of either one of the bounding boxes, and the *line_overlap* is relative to the minimum height of either one of the bounding boxes.

[1] Adobe System Inc. (2007). *Pdf reference: Adobe portable document format, version 1.7.*

Spaces need to be inserted between characters because the PDF format has no notion of the space character. A space is inserted if the characters are further apart than the *word_margin* (W in the figure). The *word_margin* is relative to the maximum width or height of the new character. Having a smaller *word_margin* creates smaller words. Note that the *word_margin* should at least be smaller than the *char_margin* otherwise none of the characters will be separated by a space.

The result of this stage is a list of lines. Each line consists of a list of characters. These characters are either original *LTChar* characters that originate from the PDF file or inserted *LTAnno* characters that represent spaces between words or newlines at the end of each line.

Grouping lines into boxes

The second step is grouping lines in a meaningful way. Each line has a bounding box that is determined by the bounding boxes of the characters that it contains. Like grouping characters, pdfminer.six uses the bounding boxes to group the lines.

Lines that are both horizontally overlapping and vertically close are grouped. How vertically close the lines should be is determined by the *line_margin*. This margin is specified relative to the height of the bounding box. Lines are close if the gap between the tops (see L_1 in the figure) and bottoms (see L_2) in the figure) of the bounding boxes are closer together than the absolute line margin, i.e. the *line_margin* multiplied by the height of the bounding box.

The result of this stage is a list of text boxes. Each box consists of a list of lines.

Grouping textboxes hierarchically

The last step is to group the text boxes in a meaningful way. This step repeatedly merges the two text boxes that are closest to each other.

The closeness of bounding boxes is computed as the area that is between the two text boxes (the blue area in the figure). In other words, it is the area of the bounding box that surrounds both lines, minus the area of the bounding boxes of the individual lines.

Working with rotated characters

The algorithm described above assumes that all characters have the same orientation. However, any writing direction is possible in a PDF. To accommodate for this, pdfminer.six allows detecting vertical writing with the *detect_vertical* parameter. This will apply all the grouping steps as if the pdf was rotated 90 (or 270) degrees

References

1.4 API Reference

1.4.1 Command-line API

pdf2txt.py

A command line tool for extracting text and images from PDF and output it to plain text, html, xml or tags.

```
usage: python tools/pdf2txt.py [-h] [--version] [--debug] [--disable-caching]
                               [--page-numbers PAGE_NUMBERS [PAGE_NUMBERS ...]]
                               [--pagenos PAGENOS] [--maxpages MAXPAGES]
                               [--password PASSWORD] [--rotation ROTATION]
                               [--no-laparams] [--detect-vertical]
                               [--line-overlap LINE_OVERLAP]
                               [--char-margin CHAR_MARGIN]
                               [--word-margin WORD_MARGIN]
                               [--line-margin LINE_MARGIN]
                               [--boxes-flow BOXES_FLOW] [--all-texts]
                               [--outfile OUTFILE] [--output_type OUTPUT_TYPE]
                               [--codec CODEC] [--output-dir OUTPUT_DIR]
                               [--layoutmode LAYOUTMODE] [--scale SCALE]
                               [--strip-control]
                               files [files ...]
```

Positional Arguments

files One or more paths to PDF files.

Named Arguments

--version, -v show program's version number and exit

--debug, -d Use debug logging level.

Default: False

--disable-caching, -C If caching or resources, such as fonts, should be disabled.

Default: False

Parser

Used during PDF parsing

--page-numbers A space-seperated list of page numbers to parse.

--pagenos, -p A comma-separated list of page numbers to parse. Included for legacy applications, use –page-numbers for more idiomatic argument entry.

--maxpages, -m The maximum number of pages to parse.

Default: 0

--password, -P The password to use for decrypting PDF file.

Default: “”

--rotation, -R The number of degrees to rotate the PDF before other types of processing.

Default: 0

Layout analysis

Used during layout analysis.

--no-laparams, -n If layout analysis parameters should be ignored.

Default: False

--detect-vertical, -V If vertical text should be considered during layout analysis

Default: False

--line-overlap If two characters have more overlap than this they are considered to be on the same line. The overlap is specified relative to the minimum height of both characters.

Default: 0.5

--char-margin, -M If two characters are closer together than this margin they are considered to be part of the same line. The margin is specified relative to the width of the character.

Default: 2.0

--word-margin, -W If two characters on the same line are further apart than this margin then they are considered to be two separate words, and an intermediate space will be added for readability. The margin is specified relative to the width of the character.

Default: 0.1

--line-margin, -L If two lines are close together they are considered to be part of the same paragraph. The margin is specified relative to the height of a line.

Default: 0.5

--boxes-flow, -F Specifies how much a horizontal and vertical position of a text matters when determining the order of lines. The value should be within the range of -1.0 (only horizontal position matters) to +1.0 (only vertical position matters). You can also pass *disabled* to disable advanced layout analysis, and instead return text based on the position of the bottom left corner of the text box.

Default: 0.5

--all-texts, -A If layout analysis should be performed on text in figures.

Default: False

Output

Used during output generation.

--outfile, -o Path to file where output is written. Or "-" (default) to write to stdout.

Default: "-"

--output_type, -t Type of output to generate {text,html,xml,tag}.

Default: "text"

--codec, -c Text encoding to use in output file.

Default: "utf-8"

--output-dir, -O The output directory to put extracted images in. If not given, images are not extracted.

--layoutmode, -Y Type of layout to use when generating html {normal,exact,loose}. If normal,each line is positioned separately in the html. If exact, each character is positioned separately in the html. If loose, same result as normal but with an additional newline after each text line. Only used when output_type is html.

Default: "normal"

--scale, -s The amount of zoom to use when generating html file. Only used when output_type is html.

Default: 1.0

--strip-control, -S Remove control statement from text. Only used when output_type is xml.

Default: False

dumppdf.py

Extract pdf structure in XML format

```
usage: python tools/dumppdf.py [-h] [--version] [--debug]
                               [--extract-toc | --extract-embedded EXTRACT_EMBEDDED]
                               [--page-numbers PAGE_NUMBERS [PAGE_NUMBERS ...]]
                               [--pagenos PAGENOS] [--objects OBJECTS] [--all]
                               [--show-fallback-xref] [--password PASSWORD]
                               [--outfile OUTFILE]
                               [--raw-stream | --binary-stream | --text-stream]
                               files [files ...]
```

Positional Arguments

files One or more paths to PDF files.

Named Arguments

--version, -v show program's version number and exit

--debug, -d Use debug logging level.

Default: False

--extract-toc, -T Extract structure of outline

Default: False

--extract-embedded, -E Extract embedded files

Parser

Used during PDF parsing

--page-numbers A space-seperated list of page numbers to parse.

--pagenos, -p A comma-separated list of page numbers to parse. Included for legacy applications, use –page-numbers for more idiomatic argument entry.

--objects, -i Comma separated list of object numbers to extract

--all, -a If the structure of all objects should be extracted

Default: False

--show-fallback-xref Additionally show the fallback xref. Use this if the PDF has zero or only invalid xref's. This setting is ignored if –extract-toc or –extract-embedded is used.

Default: False

--password, -P The password to use for decrypting PDF file.

Default: “”

Output

Used during output generation.

--outfile, -o Path to file where output is written. Or “-” (default) to write to stdout.

Default: “-“

--raw-stream, -r Write stream objects without encoding

Default: False

--binary-stream, -b Write stream objects with binary encoding

Default: False

--text-stream, -t Write stream objects as plain text

Default: False

1.4.2 High-level functions API

extract_text

pdfminer.high_level.**extract_text**(*pdf_file: Union[pathlib.PurePath, str, io.IOBase]*, *password: str = ''*, *page_numbers: Optional[Container[int]] = None*, *maxpages: int = 0*, *caching: bool = True*, *codec: str = 'utf-8'*, *laparams: Optional[*pdfminer.layout.LAParams*] = None*) → str

Parse and return the text contained in a PDF file.

Parameters

- **pdf_file** – Either a file path or a file-like object for the PDF file to be worked on.
- **password** – For encrypted PDFs, the password to decrypt.
- **page_numbers** – List of zero-indexed page numbers to extract.

- **maxpages** – The maximum number of pages to parse
- **caching** – If resources should be cached
- **codec** – Text decoding codec
- **laparams** – An LAParams object from pdfminer.layout. If None, uses some default settings that often work well.

Returns a string containing all of the text extracted.

extract_text_to_fp

pdfminer.high_level.**extract_text_to_fp**(*inf: BinaryIO*, *outfp: Union[TextIO, BinaryIO]*, *output_type: str = 'text'*, *codec: str = 'utf-8'*, *laparams: Optional[pdfminer.layout.LAParams] = None*, *maxpages: int = 0*, *page_numbers: Optional[Container[int]] = None*, *password: str = ''*, *scale: float = 1.0*, *rotation: int = 0*, *layoutmode: str = 'normal'*, *output_dir: Optional[str] = None*, *strip_control: bool = False*, *debug: bool = False*, *disable_caching: bool = False*, ***kwargs: Any*) → None

Parses text from inf-file and writes to outfp file-like object.

Takes loads of optional arguments but the defaults are somewhat sane. Beware laparams: Including an empty LAParams is not the same as passing None!

Parameters

- **inf** – a file-like object to read PDF structure from, such as a file handler (using the builtin *open()* function) or a *BytesIO*.
- **outfp** – a file-like object to write the text to.
- **output_type** – May be 'text', 'xml', 'html', 'hocr', 'tag'. Only 'text' works properly.
- **codec** – Text decoding codec
- **laparams** – An LAParams object from pdfminer.layout. Default is None but may not layout correctly.
- **maxpages** – How many pages to stop parsing after
- **page_numbers** – zero-indexed page numbers to operate on.
- **password** – For encrypted PDFs, the password to decrypt.
- **scale** – Scale factor
- **rotation** – Rotation factor
- **layoutmode** – Default is 'normal', see pdfminer.converter.HTMLConverter
- **output_dir** – If given, creates an ImageWriter for extracted images.
- **strip_control** – Does what it says on the tin
- **debug** – Output more logging data
- **disable_caching** – Does what it says on the tin
- **other** –

Returns nothing, acting as it does on two streams. Use StringIO to get strings.

extract_pages

pdfminer.high_level.**extract_pages**(*pdf_file: Union[pathlib.PurePath, str, io.IOBase]*, *password: str = ''*, *page_numbers: Optional[Container[int]] = None*, *maxpages: int = 0*, *caching: bool = True*, *laparams: Optional[pdfminer.layout.LAParams] = None*) → Iterator[pdfminer.layout.LTPage]

Extract and yield LTPage objects

Parameters

- **pdf_file** – Either a file path or a file-like object for the PDF file to be worked on.
- **password** – For encrypted PDFs, the password to decrypt.
- **page_numbers** – List of zero-indexed page numbers to extract.
- **maxpages** – The maximum number of pages to parse
- **caching** – If resources should be cached
- **laparams** – An LAParams object from pdfminer.layout. If None, uses some default settings that often work well.

Returns LTPage objects

1.4.3 Composable API

LAParams

class pdfminer.layout.**LAParams**(*line_overlap: float = 0.5*, *char_margin: float = 2.0*, *line_margin: float = 0.5*, *word_margin: float = 0.1*, *boxes_flow: Optional[float] = 0.5*, *detect_vertical: bool = False*, *all_texts: bool = False*)

Parameters for layout analysis

Parameters

- **line_overlap** – If two characters have more overlap than this they are considered to be on the same line. The overlap is specified relative to the minimum height of both characters.
- **char_margin** – If two characters are closer together than this margin they are considered part of the same line. The margin is specified relative to the width of the character.
- **word_margin** – If two characters on the same line are further apart than this margin then they are considered to be two separate words, and an intermediate space will be added for readability. The margin is specified relative to the width of the character.
- **line_margin** – If two lines are are close together they are considered to be part of the same paragraph. The margin is specified relative to the height of a line.
- **boxes_flow** – Specifies how much a horizontal and vertical position of a text matters when determining the order of text boxes. The value should be within the range of -1.0 (only horizontal position matters) to +1.0 (only vertical position matters). You can also pass *None* to disable advanced layout analysis, and instead return text based on the position of the bottom left corner of the text box.
- **detect_vertical** – If vertical text should be considered during layout analysis
- **all_texts** – If layout analysis should be performed on text in figures.

Todo:

- ***PDFDevice***
  - *TextConverter*
  - *PDFPageAggregator*
- *PDFPageInterpreter*

1.5 Frequently asked questions

1.5.1 Why is it called pdfminer.six?

Pdfminer.six is a fork of the original pdfminer created by Euske. Almost all of the code and architecture are in -fact created by Euske. But, for a long time, this original pdfminer did not support Python 3. Until 2020 the original pdfminer only supported Python 2. The original goal of pdfminer.six was to add support for Python 3. This was done with the *six* package. The *six* package helps to write code that is compatible with both Python 2 and Python 3. Hence, pdfminer.six.

As of 2020, pdfminer.six dropped the support for Python 2 because it was end-of-life. While the .six part is no longer applicable, we kept the name to prevent breaking changes for existing users.

The current punchline "We fathom PDF" is a whimsical reference to the six. Fathom means both deeply understanding something, and a fathom is also equal to six feet.

1.5.2 How does pdfminer.six compare to other forks of pdfminer?

Pdfminer.six is now an independent and community-maintained package for extracting text from PDFs with Python. We actively fix bugs (also for PDFs that don't strictly follow the PDF Reference), add new features and improve the usability of pdfminer.six. This community separates pdfminer.six from the other forks of the original pdfminer. PDF as a format is very diverse and there are countless deviations from the official format. The only way to support all the PDFs out there is to have a community that actively uses and improves pdfminer.

Since 2020, the original pdfminer is dormant, and pdfminer.six is the fork which Euske recommends if you need an actively maintained version of pdfminer.

1.5.3 Why are there *(cid:x)* values in the textual output?

One of the most common issues with pdfminer.six is that the textual output contains raw character id's *(cid:x)*. This is often experienced as confusing because the text is shown fine in a PDF viewer and other text from the same PDF is extracted properly.

The underlying problem is that a PDF has two different representations of each character. Each character is mapped to a glyph that determines how the character is shown in a PDF viewer. And each character is also mapped to its unicode value that is used when copy-pasting the character. Some PDF's have incomplete unicode mappings and therefore it is impossible to convert the character to unicode. In these cases pdfminer.six defaults to showing the raw character id *(cid:x)*

A quick test to see if pdfminer.six should be able to do better is to copy-paste the text from a PDF viewer to a text editor. If the result is proper text, pdfminer.six should also be able to extract proper text. If the result is gibberish, pdfminer.six will also not be able to convert the characters to unicode.

References:

1. Chapter 5: Text, PDF Reference 1.7
2. Text: PDF, Wikipedia

CHAPTER

TWO

FEATURES

- Parse all objects from a PDF document into Python objects.
- Analyze and group text in a human-readable way.
- Extract text, images (JPG, JBIG2 and Bitmaps), table-of-contents, tagged contents and more.
- Support for (almost all) features from the PDF-1.7 specification
- Support for Chinese, Japanese and Korean CJK) languages as well as vertical writing.
- Support for various font types (Type1, TrueType, Type3, and CID).
- Support for RC4 and AES encryption.
- Support for AcroForm interactive form extraction.

CHAPTER

THREE

INSTALLATION INSTRUCTIONS

- Install Python 3.8 or newer.
- Install pdfminer.six.

:: $ pip install pdfminer.six`

- (Optionally) install extra dependencies for extracting images.

:: $ pip install 'pdfminer.six[image]'``

- Use the command-line interface to extract text from pdf.

:: $ pdf2txt.py example.pdf`

- Or use it with Python.

```
from pdfminer.high_level import extract_text

text = extract_text("example.pdf")
print(text)
```

CHAPTER

FOUR

CONTRIBUTING

We welcome any contributors to pdfminer.six! But, before doing anything, take a look at the contribution guide.

INDEX

E

extract_pages() (*in module pdfminer.high_level*), 23
extract_text() (*in module pdfminer.high_level*), 21
extract_text_to_fp() (*in module pdfminer.high_level*), 22

L

LAParams (*class in pdfminer.layout*), 23