PyPDF2

Mathieu Fenniak

May 22, 2022

USER GUIDE

PyPDF2 is a free and open source pure-python PDF library capable of splitting, merging, cropping, and transforming the pages of PDF files. It can also add custom data, viewing options, and passwords to PDF files. PyPDF2 can retrieve text and metadata from PDFs as well.

You can contribute to PyPDF2 on Github.

CHAPTER

ONE

INSTALLATION

There are several ways to install PyPDF2. The most common option is to use pip.

1.1 pip

PyPDF2 requires Python 2.7+ to run, but Python 2 is dead. Please use a recent version of Python 3 instead.

Typically Python comes with `pip`, a package installer. Using it you can install PyPDF2:

```
pip install PyPDF2
```

If you are not a super-user (a system adminstrator / root), you can also just install PyPDF2 for your current user:

```
pip install --user PyPDF2
```

1.2 Anaconda

Anaconda users can install PyPDF2 via conda-forge.

1.3 Development Version

In case you want to use the current version under development:

```
pip install git+https://github.com/py-pdf/PyPDF2.git
```

CHAPTER

TWO

ROBUSTNESS AND STRICT=FALSE

PDF is specified in various versions. The specification of PDF 1.7 has 978 pages. This length makes it hard to get everything right. As a consequence, a lot of PDF are not strictly following the specification.

If a PDF file does not follow the specification, it is not always possible to be certain what the intended effect would be. Think of the following broken Python code as an example:

```
# Broken
function (foo, bar):

# Potentially intendet:
def function(foo, bar):
    ...

# Also possible:
function = (foo, bar)
```

Writing a parser you can go two paths: Either you try to be forgiving and try to figure out what the user intendet, or you are strict and just tell the user that they should fix their stuff.

PyPDF2 gives you the option to be strict or not.

PyPDF2 has three core objects and all of them have a `strict` parameter:

- `PdfFileReader`
- `PdfFileWriter`
- `PdfFileMerger`

Choosing `strict=True` means that PyPDF2 will raise an exception if a PDF does not follow the specification.

Choosing `strict=False` means that PyPDF2 will try to be forgiving and do something reasonable, but it will log a warning message. It is a best-effort approach.

CHAPTER

THREE

METADATA

3.1 Reading metadata

```
from PyPDF2 import PdfFileReader

reader = PdfFileReader("example.pdf")

info = reader.getDocumentInfo()

print(reader.numPages)

# All of the following could be None!
print(info.author)
print(info.creator)
print(info.producer)
print(info.subject)
print(info.title)
```

3.2 Writing metadata

```
from PyPDF2 import PdfFileReader, PdfFileWriter

reader = PdfFileReader("example.pdf")
writer = PdfFileWriter()

# Add all pages to the writer
for i in range(reader.numPages):
    page = reader.pages[i]
    writer.addPage(page)

# Add the metadata
writer.addMetadata(
    {
        "/Author": "Martin",
        "/Producer": "Libre Writer",
    }
)
```

(continues on next page)

(continued from previous page)

```
# Save the new PDF to a file
with open("meta-pdf.pdf", "wb") as f:
    writer.write(f)
```

CHAPTER

FOUR

EXTRACT TEXT FROM A PDF

You can extract text from a PDF like this:

```
from PyPDF2 import PdfFileReader

reader = PdfFileReader("example.pdf")
page = reader.pages[0]
print(page.extractText())
```

4.1 Why Text Extraction is hard

Extracting text from a PDF can be pretty tricky. In several cases there is no clear answer what the expected result should look like:

1. **Paragraphs**: Should the text of a paragraph have line breaks at the same places where the original PDF had them or should it rather be one block of text?
2. **Page numbers**: Should they be included in the extract?
3. **Outlines**: Should outlines be extracted at all?
4. **Formatting**: If text is **bold** or *italic*, should it be included in the output?
5. **Tables**: Should the text extraction skip tables? Should it extract just the text? Should the borders be shown in some Markdown-like way or should the structure be present e.g. as an HTML table? How would you deal with merged cells?
6. **Captions**: Should image and table captions be included?
7. **Ligatures**: The Unicode symbol U+FB00 is a single symbol for two lowercase letters 'f'. Should that be parsed as the Unicode symbol '' or as two ASCII symbols 'ff'?

Then there are issues where most people would agree on the correct output, but the way PDF stores information just makes it hard to achieve that:

1. **Tables**: Typically, tables are just absolutely positioned text. In the worst case, ever single letter could be absolutely positioned. That makes it hard to tell where columns / rows are.
2. **Images**: Sometimes PDFs do not contain the text as it's displayed, but instead an image. You notice that when you cannot copy the text. Then there are PDF files that contain an image and a text layer in the background. That typically happens when a document was scanned. Although the scanning software (OCR) is pretty good today, it still fails once in a while. PyPDF2 is no OCR software; it will not be able to detect those failures. PyPDF2 will also never be able to extract text from images.

And finally there are issues that PyPDF2 will deal with. If you find such a text extraction bug, please share the PDF with us so we can work on it!

CHAPTER

FIVE

ENCRYPTION AND DECRYPTION OF PDFS

5.1 Encrypt

Add a password to a PDF (encrypt it):

```
from PyPDF2 import PdfFileReader, PdfFileWriter

reader = PdfFileReader("example.pdf")
writer = PdfFileWriter()

# Add all pages to the writer
for i in range(reader.numPages):
    page = reader.pages[i]
    writer.addPage(page)

# Add a password to the new PDF
writer.encrypt("my-secret-password")

# Save the new PDF to a file
with open("encrypted-pdf.pdf", "wb") as f:
    writer.write(f)
```

5.2 Decrypt

Remove the password from a PDF (decrypt it):

```
from PyPDF2 import PdfFileReader, PdfFileWriter

reader = PdfFileReader("encrypted-pdf.pdf")
writer = PdfFileWriter()

if reader.isEncrypted:
    reader.decrypt("my-secret-password")

# Add all pages to the writer
for i in range(reader.numPages):
    page = reader.pages[i]
    writer.addPage(page)
```

(continues on next page)

(continued from previous page)

```
# Save the new PDF to a file
with open("decrypted-pdf.pdf", "wb") as f:
    writer.write(f)
```

CHAPTER

SIX

MERGING PDF FILES

6.1 Basic Example

```
from PyPDF2 import PdfFileMerger

merger = PdfFileMerger()

for pdf in ["file1.pdf", "file2.pdf", "file3.pdf"]:
    merger.append(pdf)

merger.write("merged-pdf.pdf")
merger.close()
```

For more details, see an excellent answer on StackOverflow by Paul Rooney.

6.2 Showing more merging options

```
from PyPDF2 import PdfFileMerger

merger = PdfFileMerger()

input1 = open("document1.pdf", "rb")
input2 = open("document2.pdf", "rb")
input3 = open("document3.pdf", "rb")

# add the first 3 pages of input1 document to output
merger.append(fileobj=input1, pages=(0, 3))

# insert the first page of input2 into the output beginning after the second page
merger.merge(position=2, fileobj=input2, pages=(0, 1))

# append entire input3 document to the end of the output document
merger.append(input3)

# Write to an output PDF document
output = open("document-output.pdf", "wb")
merger.write(output)
```

(continues on next page)

(continued from previous page)

```
# Close File Descriptors
merger.close()
output.close()
```

CHAPTER

SEVEN

CROPPING AND TRANSFORMING PDFS

```
from PyPDF2 import PdfFileWriter, PdfFileReader

reader = PdfFileReader("example.pdf")
writer = PdfFileWriter()

# add page 1 from reader to output document, unchanged:
writer.addPage(reader.pages[0])

# add page 2 from reader, but rotated clockwise 90 degrees:
writer.addPage(reader.pages[1].rotateClockwise(90))

# add page 3 from reader, but crop it to half size:
page3 = reader.pages[2]
page3.mediaBox.upperRight = (
    page3.mediaBox.getUpperRight_x() / 2,
    page3.mediaBox.getUpperRight_y() / 2,
)
writer.addPage(page3)

# add some Javascript to launch the print window on opening this PDF.
# the password dialog may prevent the print dialog from being shown,
# comment the the encription lines, if that's the case, to try this out:
writer.addJS("this.print({bUI:true,bSilent:false,bShrinkToFit:true});")

# write to document-output.pdf
with open("PyPDF2-output.pdf", "wb") as fp:
    writer.write(fp)
```

CHAPTER

EIGHT

ADDING A WATERMARK TO A PDF

```
from PyPDF2 import PdfFileWriter, PdfFileReader


# Read the watermark
watermark = PdfFileReader("watermark.pdf")

# Read the page without watermark
reader = PdfFileReader("example.pdf")
page = reader.pages[0]

# Add the watermark to the page
page.mergePage(watermark.pages[0])

# Add the page to the writer
writer = PdfFileWriter()
writer.addPage(page)

# finally, write the new document with a watermark
with open("PyPDF2-output.pdf", "wb") as fp:
    output.write(fp)
```

CHAPTER

NINE

READING PDF ANNOTATIONS

PDF 1.7 defines 25 different annotation types:

- Text
- Link
- FreeText
- Line, Square, Circle, Polygon, PolyLine, Highlight, Underline, Squiggly, StrikeOut
- Stamp, Caret, Ink
- Popup
- FileAttachment
- Sound, Movie
- Widget, Screen
- PrinterMark
- TrapNet
- Watermark
- 3D

Reading the most common ones is described here.

9.1 Text

```
from PyPDF2 import PdfFileReader

reader = PdfFileReader("example.pdf")

for page in reader.pages:
    if "/Annots" in page:
        for annot in page["/Annots"]:
            subtype = annot.getObject()["/Subtype"]
            if subtype == "/Text":
                print(annot.getObject()["/Contents"])
```

9.2 Highlights

```
from PyPDF2 import PdfFileReader

reader = PdfFileReader("commented.pdf")

for page in reader.pages:
    if "/Annots" in page:
        for annot in page["/Annots"]:
            subtype = annot.getObject()["/Subtype"]
            if subtype == "/Highlight":
                coords = annot.getObject()["/QuadPoints"]
                x1, y1, x2, y2, x3, y3, x4, y4 = coords
```

9.3 Attachments

```
from PyPDF2 import PdfFileReader

reader = PdfFileReader("example.pdf")

attachments = {}
for page in reader.pages:
    if "/Annots" in page:
        for annotation in page["/Annots"]:
            subtype = annot.getObject()["/Subtype"]
            if subtype == "/FileAttachment":
                fileobj = annotobj["/FS"]
                attachments[fileobj["/F"]] = fileobj["/EF"]["/F"].getData()
```

CHAPTER

TEN

ADDING PDF ANNOTATIONS

10.1 Attachments

```
from PyPDF2 import PdfFileWriter

writer = PdfFileWriter()
writer.addBlankPage(width=200, height=200)

data = b"any bytes - typically read from a file"
writer.addAttachment("smile.png", data)

with open("output.pdf", "wb") as output_stream:
    writer.write(output_stream)
```

CHAPTER

ELEVEN

INTERACTIONS WITH PDF FORMS

11.1 Reading form fields

```
from PyPDF2 import PdfFileReader

reader = PdfFileReader("form.pdf")
fields = reader.getFormTextFields()
fields == {"key": "value", "key2": "value2"}
```

11.2 Filling out forms

```
from PyPDF2 import PdfFileReader, PdfFileWriter

reader = PdfFileReader("form.pdf")
writer = PdfFileWriter()

page = reader.pages[0]
fields = reader.getFields()

writer.addPage(page)

writer.updatePageFormFieldValues(
    writer.getPage(0), {"fieldname": "some filled in text"}
)

# write "output" to PyPDF2-output.pdf
with open("filled-out.pdf", "wb") as output_stream:
    writer.write(output_stream)
```

CHAPTER

TWELVE

STREAMING DATA WITH PYPDF2

In some cases you might want to avoid saving things explicitly as a file to disk, e.g. when you want to store the PDF in a database or AWS S3.

PyPDF2 supports streaming data to a file-like object and here is how.

```
from io import BytesIO

# Prepare example
with open("example.pdf", "rb") as fh:
    bytes_stream = BytesIO(fh.read())

# Read from bytes_stream
reader = PdfFileReader(bytes_stream)

# Write to bytes_stream
writer = PdfFileWriter()
with BytesIO() as bytes_stream:
    writer.write(bytes_stream)
```

CHAPTER

THIRTEEN

REDUCE PDF SIZE

There are multiple ways to reduce the size of a given PDF file. The easiest one is to remove content (e.g. images) or pages.

13.1 Remove images

```
import PyPDF2

reader = PyPDF2.PdfFileReader("example.pdf")
writer = PyPDF2.PdfFileWriter()

for page in reader.pages:
    writer.addPage(page)

writer.removeImages()

with open("out.pdf", "wb") as f:
    writer.write(f)
```

13.2 Compression

```
import PyPDF2

reader = PyPDF2.PdfFileReader("example.pdf")
writer = PyPDF2.PdfFileWriter()

for page in reader.pages:
    page.compressContentStreams()
    writer.addPage(page)

with open("out.pdf", "wb") as f:
    writer.write(f)
```

CHAPTER

FOURTEEN

PDF VERSION SUPPORT

PDF comes in the following versions:

- 1993: 1.0
- 1994: 1.1
- 1996: 1.2
- 1999: 1.3
- 2001: 1.4
- 2003: 1.5
- 2004: 1.6
- 2006 - 2012: 1.7, ISO 32000-1:2008
- 2017: 2.0

The general format didn't change, but new features got added. It can be that PyPDF2 can do the operations you want on PDF 2.0 files without fully supporting all features of PDF 2.0.

14.1 PDF Feature Support by PyPDF2

Feature	PDF-Version	PyPDF2 Support
Transparent Graphics	1.4	?
CMaps	1.4	#201, #464, #805
Object Streams	1.5	?
Cross-reference Streams	1.5	?
Optional Content Groups (OCGs) - Layers	1.5	?
Content Stream Compression	1.5	?
AES Encryption	1.6	#749

See History of PDF for more features.

CHAPTER

FIFTEEN

THE PDFFILEREADER CLASS

class PyPDF2.pdf.**PdfFileReader**(*stream*, *strict=True*, *warndest=None*, *overwriteWarnings=True*)

Bases: `object`

Initialize a PdfFileReader object.

This operation can take some time, as the PDF stream's cross-reference tables are read into memory.

Parameters

- **stream** – A File object or an object that supports the standard read and seek methods similar to a File object. Could also be a string representing a path to a PDF file.
- **strict** (*bool*) – Determines whether user should be warned of all problems and also causes some correctable problems to be fatal. Defaults to `True`.
- **warndest** – Destination for logging warnings (defaults to `sys.stderr`).
- **overwriteWarnings** (*bool*) – Determines whether to override Python's `warnings.py` module with a custom implementation (defaults to `True`).

cacheGetIndirectObject(*generation*, *idnum*)

cacheIndirectObject(*generation*, *idnum*, *obj*)

decode_permissions(*permissions_code*)

decrypt(*password*)

When using an encrypted / secured PDF file with the PDF Standard encryption handler, this function will allow the file to be decrypted. It checks the given password against the document's user password and owner password, and then stores the resulting decryption key if either password is correct.

It does not matter which password was matched. Both passwords provide the correct decryption key that will allow the document to be used with this library.

Parameters **password** (*str*) – The password to match.

Returns `0` if the password failed, `1` if the password matched the user password, and `2` if the password matched the owner password.

Return type int

Raises **NotImplementedError** – if document uses an unsupported encryption method.

property documentInfo

Read-only property that accesses the *getDocumentInfo()* function.

getDestinationPageNumber(*destination*)

Retrieve page number of a given Destination object

Parameters **destination** (Destination) – The destination to get page number. Should be an instance of *Destination*

Returns the page number or -1 if page not found

Return type int

getDocumentInfo()

Retrieve the PDF file's document information dictionary, if it exists. Note that some PDF files use metadata streams instead of docinfo dictionaries, and these metadata streams will not be accessed by this function.

Returns the document information of this PDF file

Return type DocumentInformation or None if none exists.

getFields(*tree=None*, *retval=None*, *fileobj=None*)

Extracts field data if this PDF contains interactive form fields. The *tree* and *retval* parameters are for recursive use.

Parameters **fileobj** – A file object (usually a text file) to write a report to on all interactive form fields found.

Returns A dictionary where each key is a field name, and each value is a *Field* object. By default, the mapping name is used for keys.

Return type dict, or None if form data could not be located.

getFormTextFields()

Retrieves form fields from the document with textual data (inputs, dropdowns)

getIsEncrypted()

getNamedDestinations(*tree=None*, *retval=None*)

Retrieves the named destinations present in the document.

Returns a dictionary which maps names to *Destinations*.

Return type dict

getNumPages()

Calculates the number of pages in this PDF file.

Returns number of pages

Return type int

Raises **PdfReadError** – if file is encrypted and restrictions prevent this action.

getObject(*indirectReference*)

getOutlines(*node=None*, *outlines=None*)

Retrieve the document outline present in the document.

Returns a nested list of *Destinations*.

getPage(*pageNumber*)

Retrieves a page by number from this PDF file.

Parameters **pageNumber** (*int*) – The page number to retrieve (pages begin at zero)

Returns a PageObject instance.

Return type PageObject

getPageLayout()

Get the page layout.

See *setPageLayout()* for a description of valid layouts.

Returns Page layout currently being used.

Return type str, None if not specified

getPageMode()

Get the page mode. See *setPageMode()* for a description of valid modes.

Returns Page mode currently being used.

Return type str, None if not specified

getPageNumber(*page*)

Retrieve page number of a given PageObject

Parameters **page** (PageObject) – The page to get page number. Should be an instance of *PageObject*

Returns the page number or -1 if page not found

Return type int

getXmpMetadata()

Retrieve XMP (Extensible Metadata Platform) data from the PDF document root.

Returns a XmpInformation instance that can be used to access XMP metadata from the document.

Return type XmpInformation or None if no metadata was found on the document root.

property isEncrypted

Read-only boolean property showing whether this PDF file is encrypted. Note that this property, if true, will remain true even after the *decrypt()* method is called.

property namedDestinations

Read-only property that accesses the *getNamedDestinations()* function.

property numPages

Read-only property that accesses the *getNumPages()* function.

property outlines

Read-only property that accesses the *getOutlines()* function.

property pageLayout

Read-only property accessing the *getPageLayout()* method.

property pageMode

Read-only property accessing the *getPageMode()* method.

property pages

Read-only property that emulates a list based upon the *getNumPages()* and *getPage()* methods.

read(*stream*)

readNextEndLine(*stream*, *limit_offset=0*)

readObjectHeader(*stream*)

property xmpMetadata

Read-only property that accesses the *getXmpMetadata()* function.

CHAPTER

SIXTEEN

THE PDFFILEWRITER CLASS

class PyPDF2.pdf.**PdfFileWriter**

Bases: object

This class supports writing PDF files out, given pages produced by another class (typically *PdfFileReader*).

addAttachment(*fname*, *fdata*)

Embed a file inside the PDF.

Parameters

- **fname** (*str*) – The filename to display.
- **fdata** (*str*) – The data in the file.

Reference: https://www.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/PDF32000_2008.pdf Section 7.11.3

addBlankPage(*width=None*, *height=None*)

Append a blank page to this PDF file and returns it. If no page size is specified, use the size of the last page.

Parameters

- **width** (*float*) – The width of the new page expressed in default user space units.
- **height** (*float*) – The height of the new page expressed in default user space units.

Returns the newly appended page

Return type *PageObject*

Raises **PageSizeNotDefinedError** – if width and height are not defined and previous page does not exist.

addBookmark(*title*, *pagenum*, *parent=None*, *color=None*, *bold=False*, *italic=False*, *fit='/Fit'*, **args*)

Add a bookmark to this PDF file.

Parameters

- **title** (*str*) – Title to use for this bookmark.
- **pagenum** (*int*) – Page number this bookmark will point to.
- **parent** – A reference to a parent bookmark to create nested bookmarks.
- **color** (*tuple*) – Color of the bookmark as a red, green, blue tuple from 0.0 to 1.0
- **bold** (*bool*) – Bookmark is bold
- **italic** (*bool*) – Bookmark is italic
- **fit** (*str*) – The fit of the destination page. See *addLink()* for details.

addBookmarkDestination(*dest*, *parent=None*)

addBookmarkDict(*bookmark*, *parent=None*)

addJS(*javascript*)

Add Javascript which will launch upon opening this PDF.

Parameters **javascript** (*str*) – Your Javascript.

```
>>> output.addJS("this.print({bUI:true,bSilent:false,bShrinkToFit:true});")
# Example: This will launch the print window when the PDF is opened.
```

addLink(*pagenum*, *pagedest*, *rect*, *border=None*, *fit='/Fit'*, **args*)

Add an internal link from a rectangular area to the specified page.

Parameters

- **pagenum** (*int*) – index of the page on which to place the link.
- **pagedest** (*int*) – index of the page to which the link should go.
- **rect** – *RectangleObject* or array of four integers specifying the clickable rectangular area `[xLL, yLL, xUR, yUR]`, or string in the form `"[ xLL yLL xUR yUR ]"`.
- **border** – if provided, an array describing border-drawing properties. See the PDF spec for details. No border will be drawn if this argument is omitted.
- **fit** (*str*) – Page fit or 'zoom' option (see below). Additional arguments may need to be supplied. Passing `None` will be read as a null value for that coordinate.

Table 1: Valid `zoom` arguments (see Table 8.2 of the PDF 1.7 reference for details)

/Fit	No additional arguments
/XYZ	[left] [top] [zoomFactor]
/FitH	[top]
/FitV	[left]
/FitR	[left] [bottom] [right] [top]
/FitB	No additional arguments
/FitBH	[top]
/FitBV	[left]

addMetadata(*infos*)

Add custom metadata to the output.

Parameters **infos** (*dict*) – a Python dictionary where each key is a field and each value is your new metadata.

addNamedDestination(*title*, *pagenum*)

addNamedDestinationObject(*dest*)

addPage(*page*)

Add a page to this PDF file. The page is usually acquired from a *PdfFileReader* instance.

Parameters **page** (PageObject) – The page to add to the document. Should be an instance of *PageObject*

addURI(*pagenum*, *uri*, *rect*, *border=None*)

Add an URI from a rectangular area to the specified page. This uses the basic structure of AddLink

Parameters

- **pagenum** (*int*) – index of the page on which to place the URI action.
- **uri** (*int*) – string – uri of resource to link to.
- **rect** – *RectangleObject* or array of four integers specifying the clickable rectangular area `[xLL, yLL, xUR, yUR]`, or string in the form `"[ xLL yLL xUR yUR ]"`.
- **border** – if provided, an array describing border-drawing properties. See the PDF spec for details. No border will be drawn if this argument is omitted.

REMOVED FIT/ZOOM ARG -John Mulligan

appendPagesFromReader(*reader*, *after_page_append=None*)

Copy pages from reader to writer. Includes an optional callback parameter which is invoked after pages are appended to the writer.

Parameters

- **reader** – a PdfFileReader object from which to copy page annotations to this writer object. The writer's annots will then be updated
- **reference)** (*writer_pageref (PDF page*) – Reference to the page appended to the writer.

Callback after_page_append (function) Callback function that is invoked after each page is appended to the writer. Callback signature:

cloneDocumentFromReader(*reader*, *after_page_append=None*)

Create a copy (clone) of a document from a PDF file reader

Parameters **reader** – PDF file reader instance from which the clone should be created.

Callback after_page_append (function) Callback function that is invoked after each page is appended to the writer. Signature includes a reference to the appended page (delegates to appendPagesFromReader). Callback signature:

param writer_pageref (PDF page reference) Reference to the page just appended to the document.

cloneReaderDocumentRoot(*reader*)

Copy the reader document root to the writer.

Parameters **reader** – PdfFileReader from the document root should be copied.

Callback after_page_append

encrypt(*user_pwd*, *owner_pwd=None*, *use_128bit=True*, *permissions_flag=- 1*)

Encrypt this PDF file with the PDF Standard encryption handler.

Parameters

- **user_pwd** (*str*) – The "user password", which allows for opening and reading the PDF file with the restrictions provided.
- **owner_pwd** (*str*) – The "owner password", which allows for opening the PDF files without any restrictions. By default, the owner password is the same as the user password.
- **use_128bit** (*bool*) – flag as to whether to use 128bit encryption. When false, 40bit encryption will be used. By default, this flag is on.

- **permissions_flag** (*unsigned int*) – permissions as described in TABLE 3.20 of the PDF 1.7 specification. A bit value of 1 means the permission is grantend. Hence an integer value of -1 will set all flags. Bit position 3 is for printing, 4 is for modifying content, 5 and 6 control annotations, 9 for form fields, 10 for extraction of text and graphics.

getNamedDestRoot()

getNumPages()

Returns the number of pages.

Return type int

getObject(*ido*)

getOutlineRoot()

getPage(*pageNumber*)

Retrieve a page by number from this PDF file.

Parameters **pageNumber** (*int*) – The page number to retrieve (pages begin at zero)

Returns the page at the index given by *pageNumber*

Return type PageObject

getPageLayout()

Get the page layout.

See *setPageLayout()* for a description of valid layouts.

Returns Page layout currently being used.

Return type str, None if not specified

getPageMode()

Get the page mode. See *setPageMode()* for a description of valid modes.

Returns Page mode currently being used.

Return type str, None if not specified.

getReference(*obj*)

insertBlankPage(*width=None*, *height=None*, *index=0*)

Insert a blank page to this PDF file and returns it. If no page size is specified, use the size of the last page.

Parameters

- **width** (*float*) – The width of the new page expressed in default user space units.
- **height** (*float*) – The height of the new page expressed in default user space units.
- **index** (*int*) – Position to add the page.

Returns the newly appended page

Return type *PageObject*

Raises **PageSizeNotDefinedError** – if width and height are not defined and previous page does not exist.

insertPage(*page*, *index=0*)

Insert a page in this PDF file. The page is usually acquired from a *PdfFileReader* instance.

Parameters

- **page** (PageObject) – The page to add to the document. This argument should be an instance of PageObject.
- **index** (*int*) – Position at which the page will be inserted.

property pageLayout

Read and write property accessing the *getPageLayout()* and *setPageLayout()* methods.

property pageMode

Read and write property accessing the *getPageMode()* and *setPageMode()* methods.

removeImages(*ignoreByteStringObject=False*)

Remove images from this output.

Parameters **ignoreByteStringObject** (*bool*) – optional parameter to ignore ByteString Objects.

removeLinks()

Remove links and annotations from this output.

removeText(*ignoreByteStringObject=False*)

Remove text from this output.

Parameters **ignoreByteStringObject** (*bool*) – optional parameter to ignore ByteString Objects.

setPageLayout(*layout*)

Set the page layout.

Parameters **layout** (*str*) – The page layout to be used.

Table 2: Valid layout arguments

/NoLayout	Layout explicitly not specified
/SinglePage	Show one page at a time
/OneColumn	Show one column at a time
/TwoColumnLeft	Show pages in two columns, odd-numbered pages on the left
/TwoColumn-Right	Show pages in two columns, odd-numbered pages on the right
/TwoPageLeft	Show two pages at a time, odd-numbered pages on the left
/TwoPageRight	Show two pages at a time, odd-numbered pages on the right

setPageMode(*mode*)

Set the page mode.

Parameters **mode** (*str*) – The page mode to use.

Table 3: Valid mode arguments

/UseNone	Do not show outlines or thumbnails panels
/UseOutlines	Show outlines (aka bookmarks) panel
/UseThumbs	Show page thumbnails panel
/FullScreen	Fullscreen view
/UseOC	Show Optional Content Group (OCG) panel
/UseAttachments	Show attachments panel

`set_need_appearances_writer`()

`updatePageFormFieldValues`(*page*, *fields*, *flags=0*)

Update the form field values for a given page from a fields dictionary. Copy field texts and values from fields to page. If the field links to a parent object, add the information to the parent.

Parameters

- **page** – Page reference from PDF writer where the annotations and field data will be updated.
- **fields** – a Python dictionary of field names (/T) and text values (/V)
- **flags** – An integer (0 to 7). The first bit sets ReadOnly, the second bit sets Required, the third bit sets NoExport. See PDF Reference Table 8.70 for details.

`write`(*stream*)

Write the collection of pages added to this object out as a PDF file.

Parameters **stream** – An object to write the file to. The object must support the write method and the tell method, similar to a file object.

CHAPTER

SEVENTEEN

THE PDFFILEMERGER CLASS

class PyPDF2.merger.**PdfFileMerger**(*strict=True*, *overwriteWarnings=True*)

Bases: `object`

Initializes a `PdfFileMerger` object. `PdfFileMerger` merges multiple PDFs into a single PDF. It can concatenate, slice, insert, or any combination of the above.

See the functions *merge()* (or *append()*) and *write()* for usage information.

Parameters

- **strict** (*bool*) – Determines whether user should be warned of all problems and also causes some correctable problems to be fatal. Defaults to `True`.
- **overwriteWarnings** (*bool*) – Determines whether to override Python's `warnings.py` module with a custom implementation (defaults to `True`).

addBookmark(*title*, *pagenum*, *parent=None*, *color=None*, *bold=False*, *italic=False*, *fit='/Fit'*, **args*)

Add a bookmark to this PDF file.

Parameters

- **title** (*str*) – Title to use for this bookmark.
- **pagenum** (*int*) – Page number this bookmark will point to.
- **parent** – A reference to a parent bookmark to create nested bookmarks.
- **color** (*tuple*) – Color of the bookmark as a red, green, blue tuple from 0.0 to 1.0
- **bold** (*bool*) – Bookmark is bold
- **italic** (*bool*) – Bookmark is italic
- **fit** (*str*) – The fit of the destination page. See `addLink()` for details.

addMetadata(*infos*)

Add custom metadata to the output.

Parameters **infos** (*dict*) – a Python dictionary where each key is a field and each value is your new metadata. Example: `{u'/Title': u'My title'}`

addNamedDestination(*title*, *pagenum*)

Add a destination to the output.

Parameters

- **title** (*str*) – Title to use
- **pagenum** (*int*) – Page number this destination points at.

append(*fileobj*, *bookmark=None*, *pages=None*, *import_bookmarks=True*)

Identical to the *merge()* method, but assumes you want to concatenate all pages onto the end of the file instead of specifying a position.

Parameters

- **fileobj** – A File Object or an object that supports the standard read and seek methods similar to a File Object. Could also be a string representing a path to a PDF file.
- **bookmark** (*str*) – Optionally, you may specify a bookmark to be applied at the beginning of the included file by supplying the text of the bookmark.
- **pages** – can be a *PageRange* or a `(start, stop[, step])` tuple to merge only the specified range of pages from the source document into the output document.
- **import_bookmarks** (*bool*) – You may prevent the source document's bookmarks from being imported by specifying this as `False`.

close()

Shuts all file descriptors (input and output) and clears all memory usage.

findBookmark(*bookmark*, *root=None*)

merge(*position*, *fileobj*, *bookmark=None*, *pages=None*, *import_bookmarks=True*)

Merges the pages from the given file into the output file at the specified page number.

Parameters

- **position** (*int*) – The *page number* to insert this file. File will be inserted after the given number.
- **fileobj** – A File Object or an object that supports the standard read and seek methods similar to a File Object. Could also be a string representing a path to a PDF file.
- **bookmark** (*str*) – Optionally, you may specify a bookmark to be applied at the beginning of the included file by supplying the text of the bookmark.
- **pages** – can be a *PageRange* or a `(start, stop[, step])` tuple to merge only the specified range of pages from the source document into the output document.
- **import_bookmarks** (*bool*) – You may prevent the source document's bookmarks from being imported by specifying this as `False`.

setPageLayout(*layout*)

Set the page layout

Parameters **layout** (*str*) – The page layout to be used

Table 1: Valid `layout` arguments

/NoLayout	Layout explicitly not specified
/SinglePage	Show one page at a time
/OneColumn	Show one column at a time
/TwoColumnLeft	Show pages in two columns, odd-numbered pages on the left
/TwoColumn-Right	Show pages in two columns, odd-numbered pages on the right
/TwoPageLeft	Show two pages at a time, odd-numbered pages on the left
/TwoPageRight	Show two pages at a time, odd-numbered pages on the right

setPageMode(*mode*)

Set the page mode.

Parameters **mode** (*str*) – The page mode to use.

Table 2: Valid mode arguments

/UseNone	Do not show outlines or thumbnails panels
/UseOutlines	Show outlines (aka bookmarks) panel
/UseThumbs	Show page thumbnails panel
/FullScreen	Fullscreen view
/UseOC	Show Optional Content Group (OCG) panel
/UseAttachments	Show attachments panel

write(*fileobj*)

Writes all data that has been merged to the given output file.

Parameters **fileobj** – Output file. Can be a filename or any kind of file-like object.

CHAPTER

EIGHTEEN

THE PAGEOBJECT CLASS

class PyPDF2.pdf.**PageObject**(*pdf=None*, *indirectRef=None*)

Bases: PyPDF2.generic.DictionaryObject

PageObject represents a single page within a PDF file.

Typically this object will be created by accessing the getPage() method of the PdfFileReader class, but it is also possible to create an empty page with the *createBlankPage()* static method.

Parameters

- **pdf** – PDF file the page belongs to.
- **indirectRef** – Stores the original indirect reference to this object in its source PDF

addTransformation(*ctm*)

Apply a transformation matrix to the page.

Parameters **ctm** (*tuple*) – A 6-element tuple containing the operands of the transformation matrix.

property artBox

A *RectangleObject*, expressed in default user space units, defining the extent of the page's meaningful content as intended by the page's creator.

property bleedBox

A *RectangleObject*, expressed in default user space units, defining the region to which the contents of the page should be clipped when output in a production enviroment.

compressContentStreams()

Compress the size of this page by joining all content streams and applying a FlateDecode filter.

However, it is possible that this function will perform no action if content stream compression becomes "automatic" for some reason.

static createBlankPage(*pdf=None*, *width=None*, *height=None*)

Return a new blank page.

If width or height is None, try to get the page size from the last page of *pdf*.

Parameters

- **pdf** – PDF file the page belongs to
- **width** (*float*) – The width of the new page expressed in default user space units.
- **height** (*float*) – The height of the new page expressed in default user space units.

Returns the new blank page:

Return type *PageObject*

Raises **PageSizeNotDefinedError** – if pdf is None or contains no page

property cropBox

A *RectangleObject*, expressed in default user space units, defining the visible region of default user space. When the page is displayed or printed, its contents are to be clipped (cropped) to this rectangle and then imposed on the output medium in some implementation-defined manner. Default value: same as *mediaBox*.

extractText(*Tj_sep=''*, *TJ_sep=''*)

Locate all text drawing commands, in the order they are provided in the content stream, and extract the text. This works well for some PDF files, but poorly for others, depending on the generator used. This will be refined in the future. Do not rely on the order of text coming out of this function, as it will change if this function is made more sophisticated.

Returns a unicode string object.

getContents()

Access the page contents.

Returns the /Contents object, or None if it doesn't exist. /Contents is optional, as described in PDF Reference 7.7.3.3

property mediaBox

A *RectangleObject*, expressed in default user space units, defining the boundaries of the physical medium on which the page is intended to be displayed or printed.

mergePage(*page2*)

Merge the content streams of two pages into one.

Resource references (i.e. fonts) are maintained from both pages. The mediabox/cropbox/etc of this page are not altered. The parameter page's content stream will be added to the end of this page's content stream, meaning that it will be drawn after, or "on top" of this page.

Parameters **page2** (*PageObject*) – The page to be merged into this one. Should be an instance of *PageObject*.

mergeRotatedPage(*page2*, *rotation*, *expand=False*)

mergeRotatedPage is similar to mergePage, but the stream to be merged is rotated by appling a transformation matrix.

Parameters

- **page2** (*PageObject*) – the page to be merged into this one. Should be an instance of *PageObject*.
- **rotation** (*float*) – The angle of the rotation, in degrees
- **expand** (*bool*) – Whether the page should be expanded to fit the dimensions of the page to be merged.

mergeRotatedScaledPage(*page2*, *rotation*, *scale*, *expand=False*)

mergeRotatedScaledPage is similar to mergePage, but the stream to be merged is rotated and scaled by appling a transformation matrix.

Parameters

- **page2** (*PageObject*) – the page to be merged into this one. Should be an instance of *PageObject*.
- **rotation** (*float*) – The angle of the rotation, in degrees

- **scale** (*float*) – The scaling factor
- **expand** (*bool*) – Whether the page should be expanded to fit the dimensions of the page to be merged.

mergeRotatedScaledTranslatedPage(*page2*, *rotation*, *scale*, *tx*, *ty*, *expand=False*)

mergeRotatedScaledTranslatedPage is similar to mergePage, but the stream to be merged is translated, rotated and scaled by appling a transformation matrix.

Parameters

- **page2** (PageObject) – the page to be merged into this one. Should be an instance of *PageObject*.
- **tx** (*float*) – The translation on X axis
- **ty** (*float*) – The translation on Y axis
- **rotation** (*float*) – The angle of the rotation, in degrees
- **scale** (*float*) – The scaling factor
- **expand** (*bool*) – Whether the page should be expanded to fit the dimensions of the page to be merged.

mergeRotatedTranslatedPage(*page2*, *rotation*, *tx*, *ty*, *expand=False*)

mergeRotatedTranslatedPage is similar to mergePage, but the stream to be merged is rotated and translated by appling a transformation matrix.

Parameters

- **page2** (PageObject) – the page to be merged into this one. Should be an instance of *PageObject*.
- **tx** (*float*) – The translation on X axis
- **ty** (*float*) – The translation on Y axis
- **rotation** (*float*) – The angle of the rotation, in degrees
- **expand** (*bool*) – Whether the page should be expanded to fit the dimensions of the page to be merged.

mergeScaledPage(*page2*, *scale*, *expand=False*)

mergeScaledPage is similar to mergePage, but the stream to be merged is scaled by appling a transformation matrix.

Parameters

- **page2** (PageObject) – The page to be merged into this one. Should be an instance of *PageObject*.
- **scale** (*float*) – The scaling factor
- **expand** (*bool*) – Whether the page should be expanded to fit the dimensions of the page to be merged.

mergeScaledTranslatedPage(*page2*, *scale*, *tx*, *ty*, *expand=False*)

mergeScaledTranslatedPage is similar to mergePage, but the stream to be merged is translated and scaled by appling a transformation matrix.

Parameters

- **page2** (PageObject) – the page to be merged into this one. Should be an instance of *PageObject*.

- **scale** (*float*) – The scaling factor
- **tx** (*float*) – The translation on X axis
- **ty** (*float*) – The translation on Y axis
- **expand** (*bool*) – Whether the page should be expanded to fit the dimensions of the page to be merged.

mergeTransformedPage(*page2*, *ctm*, *expand=False*)

mergeTransformedPage is similar to mergePage, but a transformation matrix is applied to the merged stream.

Parameters

- **page2** (PageObject) – The page to be merged into this one. Should be an instance of *PageObject*.
- **ctm** (*tuple*) – a 6-element tuple containing the operands of the transformation matrix
- **expand** (*bool*) – Whether the page should be expanded to fit the dimensions of the page to be merged.

mergeTranslatedPage(*page2*, *tx*, *ty*, *expand=False*)

mergeTranslatedPage is similar to mergePage, but the stream to be merged is translated by appling a transformation matrix.

Parameters

- **page2** (PageObject) – the page to be merged into this one. Should be an instance of *PageObject*.
- **tx** (*float*) – The translation on X axis
- **ty** (*float*) – The translation on Y axis
- **expand** (*bool*) – Whether the page should be expanded to fit the dimensions of the page to be merged.

rotateClockwise(*angle*)

Rotate a page clockwise by increments of 90 degrees.

Parameters **angle** (*int*) – Angle to rotate the page. Must be an increment of 90 deg.

rotateCounterClockwise(*angle*)

Rotate a page counter-clockwise by increments of 90 degrees.

Parameters **angle** (*int*) – Angle to rotate the page. Must be an increment of 90 deg.

scale(*sx*, *sy*)

Scale a page by the given factors by appling a transformation matrix to its content and updating the page size.

Parameters

- **sx** (*float*) – The scaling factor on horizontal axis.
- **sy** (*float*) – The scaling factor on vertical axis.

scaleBy(*factor*)

Scale a page by the given factor by appling a transformation matrix to its content and updating the page size.

Parameters **factor** (*float*) – The scaling factor (for both X and Y axis).

scaleTo(*width*, *height*)

Scale a page to the specified dimentions by appling a transformation matrix to its content and updating the page size.

Parameters

- **width** (*float*) – The new width.
- **height** (*float*) – The new heigth.

property trimBox

A *RectangleObject*, expressed in default user space units, defining the intended dimensions of the finished page after trimming.

CHAPTER

NINETEEN

THE DOCUMENTINFORMATION CLASS

class PyPDF2.pdf.**DocumentInformation**

Bases: PyPDF2.generic.DictionaryObject

A class representing the basic document metadata provided in a PDF File. This class is accessible through *getDocumentInfo()*

All text properties of the document metadata have *two* properties, eg. author and author_raw. The non-raw property will always return a TextStringObject, making it ideal for a case where the metadata is being displayed. The raw property can sometimes return a ByteStringObject, if PyPDF2 was unable to decode the string's text encoding; this requires additional safety in the caller and therefore is not as commonly accessed.

property author

Read-only property accessing the document's **author**. Returns a unicode string (TextStringObject) or None if the author is not specified.

property author_raw

The "raw" version of author; can return a ByteStringObject.

property creator

Read-only property accessing the document's **creator**. If the document was converted to PDF from another format, this is the name of the application (e.g. OpenOffice) that created the original document from which it was converted. Returns a unicode string (TextStringObject) or None if the creator is not specified.

property creator_raw

The "raw" version of creator; can return a ByteStringObject.

getText(*key*)

property producer

Read-only property accessing the document's **producer**. If the document was converted to PDF from another format, this is the name of the application (for example, OSX Quartz) that converted it to PDF. Returns a unicode string (TextStringObject) or None if the producer is not specified.

property producer_raw

The "raw" version of producer; can return a ByteStringObject.

property subject

Read-only property accessing the document's **subject**. Returns a unicode string (TextStringObject) or None if the subject is not specified.

property subject_raw

The "raw" version of subject; can return a ByteStringObject.

property title

Read-only property accessing the document's **title**. Returns a unicode string (`TextStringObject`) or `None` if the title is not specified.

property title_raw

The "raw" version of title; can return a `ByteStringObject`.

CHAPTER

TWENTY

THE XMPINFORMATION CLASS

class PyPDF2.xmp.**XmpInformation**(*stream*)

Bases: PyPDF2.generic.PdfObject

An object that represents Adobe XMP metadata. Usually accessed by getXmpMetadata()

property custom_properties

Retrieves custom metadata properties defined in the undocumented pdfx metadata schema.

Returns a dictionary of key/value items for custom metadata properties.

Return type dict

property dc_contributor

Contributors to the resource (other than the authors). An unsorted array of names.

property dc_coverage

Text describing the extent or scope of the resource.

property dc_creator

A sorted array of names of the authors of the resource, listed in order of precedence.

property dc_date

A sorted array of dates (datetime.datetime instances) of significance to the resource. The dates and times are in UTC.

property dc_description

A language-keyed dictionary of textual descriptions of the content of the resource.

property dc_format

The mime-type of the resource.

property dc_identifier

Unique identifier of the resource.

property dc_language

An unordered array specifying the languages used in the resource.

property dc_publisher

An unordered array of publisher names.

property dc_relation

An unordered array of text descriptions of relationships to other documents.

property dc_rights

A language-keyed dictionary of textual descriptions of the rights the user has to this resource.

property dc_source

Unique identifier of the work from which this resource was derived.

property dc_subject

An unordered array of descriptive phrases or keywrods that specify the topic of the content of the resource.

property dc_title

A language-keyed dictionary of the title of the resource.

property dc_type

An unordered array of textual descriptions of the document type.

getElement(*aboutUri*, *namespace*, *name*)

getNodesInNamespace(*aboutUri*, *namespace*)

property pdf_keywords

An unformatted text string representing document keywords.

property pdf_pdfversion

The PDF file version, for example 1.0, 1.3.

property pdf_producer

The name of the tool that created the PDF document.

writeToStream(*stream*, *encryption_key*)

property xmp_createDate

The date and time the resource was originally created. The date and time are returned as a UTC datetime.datetime object.

property xmp_creatorTool

The name of the first known tool used to create the resource.

property xmp_metadataDate

The date and time that any metadata for this resource was last changed. The date and time are returned as a UTC datetime.datetime object.

property xmp_modifyDate

The date and time the resource was last modified. The date and time are returned as a UTC datetime.datetime object.

property xmpmm_documentId

The common identifier for all versions and renditions of this resource.

property xmpmm_instanceId

An identifier for a specific incarnation of a document, updated each time a file is saved.

CHAPTER

TWENTYONE

THE DESTINATION CLASS

class PyPDF2.pdf.**Destination**(*title*, *page*, *typ*, **args*)

Bases: PyPDF2.generic.TreeObject

A class representing a destination within a PDF file. See section 8.2.1 of the PDF 1.6 reference.

Parameters

- **title** (*str*) – Title of this destination.
- **page** (*IndirectObject*) – Reference to the page of this destination. Should be an instance of IndirectObject.
- **typ** (*str*) – How the destination is displayed.
- **args** – Additional arguments may be necessary depending on the type.

Raises **PdfReadError** – If destination type is invalid.

Table 1: Valid typ arguments (see PDF spec for details)

/Fit	No additional arguments
/XYZ	[left] [top] [zoomFactor]
/FitH	[top]
/FitV	[left]
/FitR	[left] [bottom] [right] [top]
/FitB	No additional arguments
/FitBH	[top]
/FitBV	[left]

property **bottom**

Read-only property accessing the bottom vertical coordinate.

Return type int, or None if not available.

getDestArray()

property **left**

Read-only property accessing the left horizontal coordinate.

Return type int, or None if not available.

property **page**

Read-only property accessing the destination page number.

Return type int

property right

Read-only property accessing the right horizontal coordinate.

Return type int, or None if not available.

property title

Read-only property accessing the destination title.

Return type str

property top

Read-only property accessing the top vertical coordinate.

Return type int, or None if not available.

property typ

Read-only property accessing the destination type.

Return type str

writeToStream(*stream*, *encryption_key*)

property zoom

Read-only property accessing the zoom factor.

Return type int, or None if not available.

CHAPTER

TWENTYTWO

THE RECTANGLEOBJECT CLASS

class PyPDF2.generic.**RectangleObject**(*arr*)

Bases: PyPDF2.generic.ArrayObject

This class is used to represent *page boxes* in PyPDF2. These boxes include:

- *artBox*
- *bleedBox*
- *cropBox*
- *mediaBox*
- *trimBox*

ensureIsNumber(*value*)

getHeight()

getLowerLeft()

getLowerLeft_x()

getLowerLeft_y()

getLowerRight()

getLowerRight_x()

getLowerRight_y()

getUpperLeft()

getUpperLeft_x()

getUpperLeft_y()

getUpperRight()

getUpperRight_x()

getUpperRight_y()

getWidth()

property lowerLeft

Property to read and modify the lower left coordinate of this box in (x,y) form.

property lowerRight

Property to read and modify the lower right coordinate of this box in (x,y) form.

setLowerLeft(*value*)

setLowerRight(*value*)

setUpperLeft(*value*)

setUpperRight(*value*)

property upperLeft

Property to read and modify the upper left coordinate of this box in (x,y) form.

property upperRight

Property to read and modify the upper right coordinate of this box in (x,y) form.

CHAPTER

TWENTYTHREE

THE FIELD CLASS

class PyPDF2.pdf.**Field**(*data*)

Bases: PyPDF2.generic.TreeObject

A class representing a field dictionary. This class is accessed through getFields()

property additionalActions

Read-only property accessing the additional actions dictionary. This dictionary defines the field's behavior in response to trigger events. See Section 8.5.2 of the PDF 1.7 reference.

property altName

Read-only property accessing the alternate name of this field.

property defaultValue

Read-only property accessing the default value of this field.

property fieldType

Read-only property accessing the type of this field.

property flags

Read-only property accessing the field flags, specifying various characteristics of the field (see Table 8.70 of the PDF 1.7 reference).

property kids

Read-only property accessing the kids of this field.

property mappingName

Read-only property accessing the mapping name of this field. This name is used by PyPDF2 as a key in the dictionary returned by getFields()

property name

Read-only property accessing the name of this field.

property parent

Read-only property accessing the parent of this field.

property value

Read-only property accessing the value of this field. Format varies based on field type.

CHAPTER

TWENTYFOUR

THE PAGERANGE CLASS

class PyPDF2.pagerange.**PageRange**(*arg*)

Bases: object

A slice-like representation of a range of page indices, i.e. page numbers, only starting at zero.

The syntax is like what you would put between brackets [ ]. The slice is one of the few Python types that can't be subclassed, but this class converts to and from slices, and allows similar use.

- PageRange(str) parses a string representing a page range.
- PageRange(slice) directly "imports" a slice.
- to_slice() gives the equivalent slice.
- str() and repr() allow printing.
- indices(n) is like slice.indices(n).

indices(*n*)

n is the length of the list of pages to choose from. Returns arguments for range(). See help(slice.indices).

to_slice()

Return the slice equivalent of this page range.

static valid(*input*)

True if input is a valid initializer for a PageRange.

CHAPTER

TWENTYFIVE

DEVELOPER INTRO

PyPDF2 is a library and hence its users are developers. This document is not for the users, but for people who want to work on PyPDF2 itself.

25.1 Installing Requirements

```
pip install -r requirements/dev.txt
```

25.2 Running Tests

```
pytest .
```

We have the following pytest markers defined:

- `no_py27`: Flag for tests that fail under Python 2.7 only
- `external`: Tests which use files from the `sample-files` git submodule

You can locally choose not to run those via `pytest -m "not external"`.

25.3 The sample-files git submodule

The reason for having the submodule `sample-files` is that we want to keep the size of the PyPDF2 repository small while we also want to have an extensive test suite. Those two goals contradict each other.

The `Resources` folder should contain a select set of core examples that cover most cases we typically want to test for. The `sample-files` might cover a lot more edge cases, the behavior we get when file sizes get bigger, different PDF producers.

25.4 Tools: git and pre-commit

Git is a command line application for version control. If you don't know it, you can play ohmygit to learn it.

Github is the service where the PyPDF2 project is hosted. While git is free and open source, Github is a paid service by Microsoft - but for free in lot of cases.

pre-commit is a command line application that uses git hooks to automatically execute code. This allows you to avoid style issues and other code quality issues. After you entered `pre-commit install` once in your local copy of PyPDF2, it will automatically be executed when you `git commit`.

25.5 Commit Messages

Having a clean commit message helps people to quickly understand what the commit was about, witout actually looking at the changes. The first line of the commit message is used to auto-generate the CHANGELOG. For this reason, the format should be:

```
PREFIX: DESCRIPTION

BODY
```

The `PREFIX` can be:

- `BUG`: A bug was fixed. Likely there is one or multiple issues. Then write in the `BODY`: `Closes #123` where 123 is the issue number on Github. It would be absolutely amazing if you could write a regression test in those cases. That is a test that would fail without the fix.
- `ENH`: A new feature! Describe in the body what it can be used for.
- `DEP`: A deprecation - either marking something as "this is going to be removed" or actually removing it.
- `ROB`: A robustness change. Dealing better with broken PDF files.
- `DOC`: A documentation change.
- `TST`: Adding / adjusting tests.
- `DEV`: Developer experience improvements - e.g. pre-commit or setting up CI
- `MAINT`: Quite a lot of different stuff. Performance improvements are for sure the most interesting changes in here. Refactorings as well.
- `STY`: A style change. Something that makes PyPDF2 code more consistent. Typically a small change.

25.6 Benchmarks

We need to keep an eye on performance and thus we have a few benchmarks.

See py-pdf.github.io/PyPDF2/dev/bench

CHAPTER

TWENTYSIX

THE PDF FORMAT

It's recommended to look in the PDF specification for details and clarifications. This is only intended to give a very rough overview of the format.

26.1 Overall Structure

A PDF consists of:

1. Header: Contains the version of the PDF, e.g. `%PDF-1.7`
2. Body: Contains a sequence of indirect objects
3. Cross-reference table (xref): Contains a list of the indirect objects in the body
4. Trailer

26.2 The xref table

A cross-reference table (xref) is a table of the indirect objects in the body. It allows quick access to those objects by pointing to their location in the file.

It looks like this:

```
xref 42 5
0000001000 65535 f
0000001234 00000 n
0000001987 00000 n
0000011987 00000 n
0000031987 00000 n
```

Let's go through it step-by-step:

- `xref` is justa keyword that specifies the start of the xref table.
- `42` is TODO; `6` is the number of entries in the xref table.
- Now every object has 3 entries `nnnnnnnnnn ggggg n`: The 10-digit byte offset, a 5-digit generation number, and a literal keyword which is either `n` or `f`.
  - `nnnnnnnnnn` is the byte offset of the object. It tells the reader where the object is in the file.
  - `ggggg` is the generation number. It tells the reader how old the object is.
  - `n` means that the object is a normal in-use object, `f` means that the object is a free object.

    * The first free object always has a generation number of 65535. It forms the head of a linked-list of all free objects.
    * The generation number of a normal object is always 0. The generation number allows the PDF format to contain multiple versions of the same object. This is a version history mechanism.

26.3 The body

The body is a sequence of indirect objects:

`counter generationnumber << the_object >> endobj`

- `counter` (integer) is a unique identifier for the object.
- `generationnumber` (integer) is the generation number of the object.
- `the_object` is the object itself. It can be empty. Starts with `/Keyword` to specify which kind of object it is.
- `endobj` marks the end of the object.

A concrete example can be found in `test_reader.py::test_get_images_raw`:

```
1 0 obj << /Count 1 /Kids [4 0 R] /Type /Pages >> endobj
2 0 obj << >> endobj
3 0 obj << >> endobj
4 0 obj << /Contents 3 0 R /CropBox [0.0 0.0 2550.0 3508.0]
 /MediaBox [0.0 0.0 2550.0 3508.0] /Parent 1 0 R
 /Resources << /Font << >> >>
 /Rotate 0 /Type /Page >> endobj
5 0 obj << /Pages 1 0 R /Type /Catalog >> endobj
```

26.4 The trailer

The trailer looks like this:

```
trailer << /Root 5 0 R
           /Size 6
        >>
startxref 1234
%%EOF
```

Let's go through it:

- `trailer` << indicates that the *trailer dictionary* starts. It ends with >>.
- `startxref` is a keyword followed by the byte-location of the `xref` keyword. As the trailer is always at the bottom of the file, this allows readers to quickly find the xref table.
- `%%EOF` is the end-of-file marker.

The trailer dictionary is a key-value list. The keys are specified in Table 3.13 of the PDF Reference 1.7, e.g. `/Root` and `/Size` (both are required).

- `/Root` (dictionary) contains the document catalog.
    - The `5` is the object number of the catalog dictionary

  - `0` is the generation number of the catalog dictionary
  - `R` is the keyword that indicates that the object is a reference to the catalog dictionary.
- `/Size` (integer) contains the total number of entries in the files xref table.

26.5 Reading PDF files

Most PDF files are compressed. If you want to read them, first uncompress them:

```
pdftk crazyones.pdf output crazyones-uncomp.pdf uncompress
```

Then rename `crazyones-uncomp.pdf` to `crazyones-uncomp.txt` and open it in our favorite IDE / text editor.

CHAPTER

TWENTYSEVEN

CMAPS

Looking at the cmap of "crazyones":

```
pdftk crazyones.pdf output crazyones-uncomp.pdf uncompress
```

You can see this:

```
begincmap
/CMapName /T1Encoding-UTF16 def
/CMapType 2 def
/CIDSystemInfo <<
  /Registry (Adobe)
  /Ordering (UCS)
  /Supplement 0
>> def
1 begincodespacerange
<00> <FF>
endcodespacerange
1 beginbfchar
<1B> <FB00>
endbfchar
endcmap
CMapName currentdict /CMap defineresource pop
```

27.1 codespacerange

A codespacerange maps a complete sequence of bytes to a range of unicode glyphs. It defines a starting point:

```
1 beginbfchar
<1B> <FB00>
```

That means that `1B` (Hex for 27) maps to the unicode character `FB00` - the ligature (two lowercase f's).

The two numbers in `begincodespacerange` mean that it starts with an offset of 0 (hence from `1B` `FB00`) upt to an offset of FF (dec: 255), hence 1B+FF = 282 FBFF.

Within the text stream, there is

```
(The)-342(mis\034ts.)
```

`\034` is octal for 28 decimal.

CHAPTER

TWENTYEIGHT

PROJECT GOVERNANCE

This document describes how the PyPDF2 project is managed. It describes the different actors, their roles, and the responsibilities they have.

28.1 Terminology

- The **project** is PyPDF2 - a free and open-source pure-python PDF library capable of splitting, merging, cropping, and transforming the pages of PDF files. It includes the code, issues, and disscussions on GitHub, and the documentation on ReadTheDocs, the package on PyPI, and the website on GitHub.
- A **maintainer** is a person who has technical permissions to change one or more part of the projects. It is a person who is driven to keep the project running and improving.
- A **contributor** is a person who contributes to the project. That could be through writing code - in the best case through forking and creating a pull request, but that is up to the maintainer. Other contributors describe issues, help to ask questions on existing issues to make them easier to answer, participate in discussions, and help to improve the documentation. Contributors are similar to maintainers, but without technial permissions.
- A **user** is a person who imports PyPDF2 into their code. All PyPDF2 users are developers, but not developers who know the internals of PyPDF2. They only use the public interface of PyPDF2. They will likely have less knowledge about PDF than contributors.
- The **community** is all of that - the users, the contributors, and the maintainers.

28.2 Governance, Leadership, and Steering PyPDF2 forward

PyPDF2 is a free and open source project with over 100 contributors and likely (way) more than 1000 users.

As PyPDF2 does not have any formal relationship with any company and no funding, all the work done by the community are voluntary contributions. People don't get paid, but choose to spend their free time to create software of which many more are profiting. This has to be honored and respected.

Despite such a big community, the project was dormant from 2016 to 2022. There were still questions asked, issues reported, and pull requests created. But the maintainer didn't have the time to move PyPDF2 forward. During that time, nobody else stepped up to become the new maintainer.

For this reason, PyPDF2 has the **Benevolent Dictator** governance model. The benevolent dictator is a maintainer with all technical permissions - most importantly the permission to push new PyPDF2 versions on PyPI.

Being benevolent, the benevolent dictator listens for decisions to the community and tries their best to make decisions from which the overall community profits - the current one and the potential future one. Being a dictator, the benevolent dictator always has the power and the right to make decisions on their own - also against some members of the community.

As PyPDF2 is free software, parts of the community can split off (fork the code) and create a new community. This should limit the harm a bad benevolent dictator can do.

28.3 Project Language

The project language is (american) English. All documentation and issues must be written in English to ensure that the community can understand it.

We appreciate the fact that large parts of the community don't have English as their mother tongue. We try our best to understand others - automatic translators might help.

28.4 Expectations

The community can expect the following:

- The **benevolent dictator** tries their best to make decisions from which the overall community profits. The benevolent dictator is aware that his/her decisons can shape the overall community. Once the benevolent dictator notices that she/he doesn't have the time to advance PyPDF2, he/she looks for a new benevolent dictator. As it is expected that the benevolent dictator will step down at some point of their choice (hopefully before their death), it is NOT a benevolent dictator for life (BDFL).
- Every **maintainer** (including the benevolent dictator) is aware of their permissions and the harm they could do. They value security and ensure that the project is not harmed. They give their technical permissions back if they don't need them any longer. Any long-time contributor can become a maintainer. Maintainers can - and should! - step down from their role when they realize that they can no longer commit that time. Their contribution will be honored in the *History of PyPDF2*.
- Every **contributor** is aware that the time of maintainers and the benevolent dictator is limited. Short pull requests that briefly describe the solved issue and have a unit test have a higher chance to get merged soon - simply because it's easier for maintainers to see that the contribution will not harm the overall project. Their contributions are documented in the git history and in the public issues. Let us know if you would appriciate something else!
- Every **community member** uses a respectful language. We are all human, we get upset about things we care and other things than what's visible on the internet go on in our live. PyPDF2 does not pay its contributors - keep all of that in mind when you interact with others. We are here because we want to help others.

28.4.1 Issues and Discussions

An issue is any technical description that aims at bringing PyPDF2 forward:

- Bugs tickets: Something went wrong because PyPDF2 developers made a mistake.
- Feature requests: PyPDF2 does not support all features of the PDF specifications. There are certainly also convenience methods that would help users a lot.
- Robustness requests: There are many broken PDFs around. In some cases, we can deal with that. It's kind of a mixture between a bug ticket and a feature request.
- Performance tickets: PyPDF2 could be faster - let us know about your specific scenario.

Any comment that is in those technial descriptions which is not helping the discussion can be deleted. This is especially true for "me too" comments on bugs or "bump" comments for desired features. People can express this with / reactions.

Discussions are open. No comments will be deleted there - except if they are clearly unrelated spam or only try to insult people (luckily, the community was very respectful so far)

28.4.2 Releases

The maintainers follow semantic versioning. Most importantly, that means that breaking changes will have a major version bump.

Be aware that unintentional breaking changes might still happen. The PyPDF2 maintainers do their best to fix that in a timely manner - please report such issues!

28.5 People

- Martin Thoma is benevolent dictator since April 2022.
- Maintainers:
  - Matthew Stamy (mstamy2) was the benevolent dictator for a long time. He still is around on Github once in a while and has permissions on PyPI and Github.
  - Matthew Peveler (MasterOdin) is a maintainer on Github.

CHAPTER

TWENTYNINE

HISTORY OF PYPDF2

29.1 The Origins: pyPdf (2005-2010)

In 2005, Mathieu Fenniak launched pyPdf "as a PDF toolkit…" focused on

- document manipulation: by-page splitting, concatenation, and merging;
- document introspection;
- page cropping; and
- document encryption and decryption.

The last release of PyPI was pyPdf 1.13 in 2010.

29.2 PyPDF2 is born (2011-2016)

At the end of 2011, after consultation with Mathieu and others, Phaseit sponsored PyPDF2 as a fork of pyPdf on GitHub. The initial impetus was to handle a wider range of input PDF instances; Phaseit's commercial work often encounters PDF instances "in the wild" that it needs to manage (mostly concatenate and paginate), but that deviate so much from PDF standards that pyPdf can't read them. PyPDF2 reads a considerably wider range of real-world PDF instances.

Neither pyPdf nor PyPDF2 aims to be universal, that is, to provide all possible PDF-related functionality. Note that the similar-appearing pyfpdf of Mariano Reingart is most comparable to ReportLab, in that both ReportLab and pyfpdf emphasize document generation. Interestingly enough, pyfpdf builds in a basic HTML→PDF converter while PyPDF2 has no knowledge of HTML.

So what is PyPDF2 truly about? Think about popular pdftk for a moment. PyPDF2 does what pdftk does, and it does so within your current Python process, and it handles a wider range of variant PDF formats [explain]. PyPDF2 has its own FAQ to answer other questions that have arisen.

The Reddit /r/python crowd chatted obliquely and briefly about PyPDF2 in March 2012.

29.3 PyPDF3 and PyPDF4 (2018 - 2022)

Two approaches were made to get PyPDF2 active again: PyPDF3 and PyPDF4.

PyPDF3 had it's first release in 2018 and its last one in February 2022. It never got the user base from PyPDF2.

PyPDF4 only had one release in 2018.

29.4 PyPDF2: Reborn (2022-Today)

Martin Thoma took over maintenance of PyPDF2 in April 2022.

CHAPTER

THIRTY

PYPDF2 VS X

PyPDF2 is a free and open source pure-python PDF library capable of splitting, merging, cropping, and transforming the pages of PDF files. It can also add custom data, viewing options, and passwords to PDF files. PyPDF2 can retrieve text and metadata from PDFs as well.

30.1 PyMuPDF and PikePDF

PyMuPDF is a Python binding to MuPDF and PikePDF is the Python binding to QPDF.

While both are excellent libraries for various use-cases, using them is not always possible even when they support the use-case. Both of them are powered by C libraries which makes installation harder and might cause security concerns. For MuPDF you might also need to buy a commercial license.

A core feature of PyPDF2 is that it's pure Python. That means there is no C dependency. It has been used for over 10 years and for this reason a lot of support via StackOverflow and examples on the internet.

30.2 pyPDF

PyPDF2 was forked from pyPDF. pyPDF has been unmaintained for a long time.

30.3 PyPDF3 and PyPDF4

Developing and maintaining open source software is extremely time-intensive and in the case of PyPDF2 not paid at all. Having a continuous support is hard.

PyPDF2 was initially released in 2012 on PyPI and received releases until 2016. From 2016 to 2022 there was no update - but people were still using it.

As PyPDF2 is free software, there were attempts to fork it and continue the development. PyPDF3 was first released in 2018 and still receives updates. PyPDF4 has only one release from 2018.

I, Martin Thoma, the current maintainer of PyPDF2, hope that we can bring the community back to one path of development. Let's see.

30.4 pdfrw and pdfminer

I don't have experience with either of those libraries. Please add a comparison if you know PyPDF2 and `pdfrw` or `pdfminer.six`!

Please be aware that there is also `pdfminer` which is not maintained. Then there is `pdfrw2` which doesn't have a large community behind it.

And there are more:

- `pdfplumber`

30.5 Document Generation

There are (Python) tools to generate PDF documents. PyPDF2 is not one of them.

CHAPTER

THIRTYONE

FREQUENTLY-ASKED QUESTIONS

31.1 How is PyPDF2 related to pyPdf?

PyPDF2 is a fork from the no-longer-maintained pyPdf approved by the latter's founder.

31.2 Which Python versions are supported?

As Matthew writes, "… the intention is for PyPDF2 to work with Python 2 as well as Python 3." (source)

In January 2014, the main branch works with 2.6-2.7 and 3.1-3.3 [and maybe 2.5?]. Notice that 1.19–the latest in PyPI as of this writing–(mostly) did not work with 3.x.

I often merge [concatenate] various PDF instances, and my application 'craters' with certain files produced by {AutoCAD, my departmental scanner, … }, even though the original files display OK. What do I do now? Crucial ideas we want you to know:

- All of us contend with this sort of thing. Vendors often produce PDF with questionable syntax, or at least syntax that isn't what PyPDF2 expects.
- We're committed to resolving all these problems, so that your applications (and ours) can handle any PDF instances that come their way. Write whenever you have a problem a GitHub issue.
- In the meantime, while you're waiting on us, you have at least a couple of choices: you can debug PyPDF2 yourself; or use Acrobat or Preview or a similar consumer-grade PDF tool to 'mollify' [explain] your PDF instances so you get the results you are after.

CHAPTER

THIRTYTWO

PDFCAT

PyPDF2 contains a growing variety of sample programs meant to demonstrate its features. It also contains useful scripts such as `pdfcat`, located within the `Scripts` folder. This script makes it easy to concatenate PDF files by using Python slicing syntax. Because we are slicing PDF pages, we refer to the slices as *page ranges*.

Deprecation Discussion

We are thinking about moving pdfcat to a separate package. Please participate in the discussion.

Page range expression examples:

:	all pages	-1	last page
22	just the 23rd page	:-1	all but the last page
0:3	the first three pages	-2	second-to-last page
:3	the first three pages	-2:	last two pages
5:	from the sixth page onward	-3:-1	third & second to last

The third stride or step number is also recognized:

::2	0 2 4 . . . to the end
1:10:2	1 3 5 7 9
::-1	all pages in reverse order
3:0:-1	3 2 1 but not 0
2::-1	2 1 0

Usage for pdfcat is as follows:

```
$ pdfcat [-h] [-o output.pdf] [-v] input.pdf [page_range...] ...
```

You can add as many input files as you like. You may also specify as many page ranges as needed for each file.

Optional arguments:

-h	–help	Show the help message and exit
-o	–output	Follow this argument with the output PDF file. Will be created if it doesn't exist.
-v	–verbose	Show page ranges as they are being read

32.1 Examples

```
$ pdfcat -o output.pdf head.pdf content.pdf :6 7: tail.pdf -1
```

Concatenates all of head.pdf, all but page seven of content.pdf, and the last page of tail.pdf, producing output.pdf.

```
$ pdfcat chapter*.pdf >book.pdf
```

You can specify the output file by redirection.

```
$ pdfcat chapter?.pdf chapter10.pdf >book.pdf
```

In case you don't want chapter 10 before chapter 2.

Thanks to **Steve Witham** for this script!

CHAPTER

THIRTYTHREE

INDICES AND TABLES

- genindex
- modindex
- search

INDEX

E

F

G

T

U

V

W

X

Z