Reading the PDF properties/metadata in Python

Ask Question

How can I read the properties/metadata like Title, Author, Subject and Keywords stored on a PDF file using Python?

python pdf metadata

edited Jun 2 at 20:10

wolφi
4,874 13 39

asked Jan 8 '13 at 6:13

Khaleel
1,301 3 16 33

4 Answers

Try pdfminer:

```
from pdfminer.pdfparser
from pdfminer.pdfdocume

fp = open('diveintopyth
parser = PDFParser(fp)
doc = PDFDocument(parse

print doc.info  # The "
```

Here's the output:

```
>>> [{'CreationDate': '
  'Creator': 'DocBook X
  'Keywords': 'Python,
documentation, book, fr
  'Producer': 'htmldoc
Rights Reserved.',
  'Title': 'Dive Into P
```

For more info, look at this tutorial: A lightweight XMP parser for extracting PDF metadata

ıswered Jan 8 '13 at 6:22

namit
4,505 2 22 36

1 A heads-up: the author of pdfminer says it's incompatible with Python 3, at least as of date of this post (link) – JSmyth Jan 5 '14 at 22:36

8 As of November 2013, the "PDFDocument class now takes a PDFParser object as an argument. PDFDocument.set_parser() and PDFParser.set_document() is removed." So you can just do doc=PDFDocument(parser), and skip the calls to set_document, set_parser, and initialize. – Derek Kurth Oct 14 '14 at 15:55

@JSmyth The PyPi Index currently lists three working `pdfminer` forks that are compatible with Python 3. `pip search pdfminer` – zero2cx Jan 19 '17 at 7:10

@zero2cx thanks for the update. I personally settled on pdfminer3k. Works well for my purposes. One has to read the API document in the repo though as the accepted here answer is not a valid API for pdfminer3k anymore. – JSmyth Jan 21 '17 at 23:33

1 There is now an official Python 3 fork of the project github.com/pdfminer/pdfminer.six – Harsh Jun 4 '17 at 12:46

For Python 3 see PyPDF2 with example code from @Khaleel updated to:

```
from PyPDF2 import PdfF
pdf_toread = PdfFileRea
pdf_info = pdf_toread.g
print(str(pdf_info))
```

Install using `pip install PyPDF2` .

nswered Oct 8 '16 at 11:31

Morten Zilmer
12.3k 2 16 37

I have implemented this using pyPdf. Please see the sample code below.

```
from pyPdf import PdfFi
pdf_toread = PdfFileRea
pdf_info = pdf_toread.g
print str(pdf_info)
```

Output:

```
{'/Title': u'Microsoft
'/CreationDate': u"D:20
10.0.0 (Windows)', '/Mo
u'PScript5.dll Version
```

Note: pyPdf homepage says it is no longer maintained.

dited Jan 22 at 14:22

nswered Jan 8 '13 at 8:49

Khaleel
1,301 3 16 33

1 Don't use `file` , use `open` instead. – Burhan Khalid Jan 9 '13 at 6:21

2 Note that the pyPdf is marked on the

For Python 3 and new pdfminer (pip install pdfminer3k):

```
import os
from pdfminer.pdfparser
from pdfminer.pdfparser

fp = open("foo.pdf", 'r
parser = PDFParser(fp)
doc = PDFDocument(parse
parser.set_document(doc
doc.set_parser(parser)
if len(doc.info) > 0:
    info = doc.info[0]
    print(info)
```

nswered Dec 19 '16 at 1:36

Rabash
695 8 10