链滴

Python 读取 PDF

作者：mubai

原文链接：https://ld246.com/article/1484587761781

来源网站：链滴

```
from urllib.request import urlopen
from pdfminer.pdfinterp import PDFResourceManager, process_pdf
from pdfminer.converter import TextConverter
from pdfminer.layout import LAParams
from io import StringIO
from io import open

def readPDF(pdfFile):
    rsrcmgr = PDFResourceManager()
    retstr = StringIO()
    laparams = LAParams()
    device = TextConverter(rsrcmgr, retstr, laparams=laparams)

    process_pdf(rsrcmgr, device, pdfFile)
    device.close()

    content = retstr.getvalue()
    retstr.close()
    return content

pdfFile = urlopen("http://pythonscraping.com/pages/warandpeace/chapter1.pdf")
# pdfFile = open("../chapter1.pdf", 'rb')
outputString = readPDF(pdfFile)
print(outputString)
pdfFile.close()
```

参考：《Python网络数据采集》