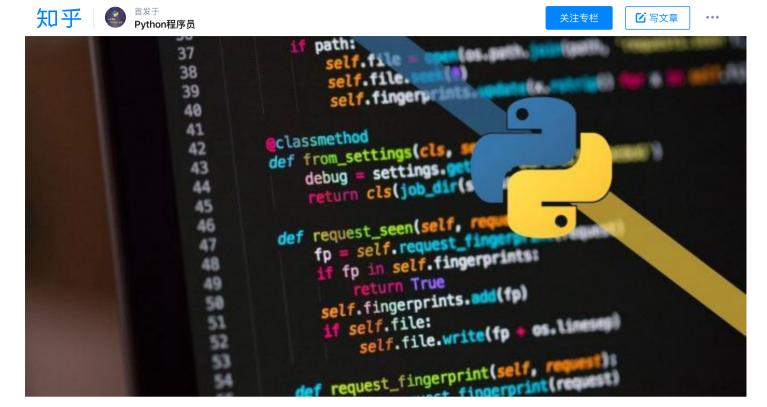
我们检测到你可能使用了 AdBlock 或 Adblock Plus,它的部分策略可能会影响到正常功能的使用(如关注)。 你可以设定特殊规则或将知乎加入白名单,以便我们更好地提供服务。(为什么?)





python玩转PDF文档



Python语...

公众号:深度学习与python

+ 关注他

19 人赞同了该文章

python作为一种具有相对简单语法的高级解释语言,即使对于那些没有编程经验的人来说,Python也是简单易操作的。强大的Python库让你事半功倍。

在处理文本信息时,通常我们需要从word、PDF文档中提取出信息,而PDF是最重要和最广泛使用的用来呈现和交换文件的数字媒体之一,。PDF包含有用的信息,链接和按钮,表单域,音频,视频和业务逻辑。python库很好地集成并提供处理非结构化数据源。运用python可以轻松从PDF中提取有用信息后,您可以轻松地将该数据用于任何机器学习或自然语言处理模型。

常见的Python库

以下是可用于处理PDF文件的一些Python库

- 1. **PDFMiner**: 一个从PDF文档中提取信息的工具。与其他PDF相关工具不同,它完全专注于获取和分析文本数据。
- 2. **PyPDF2**: 一个纯python PDF库,能够分割,合并,裁剪和转换PDF文件的页面。它还可以向PDF文件添加自定义数据,查看选项和密码。它可以从PDF中检索文本和元数据,以及将整个文件合并在一起。
- 3. **Tabula-py:** 一个 tabula-java的简单Python包装器,它可以读取PDF表。您可以从PDF读取表格并转换为pandas的DataFrame。tabula-py还允许您将PDF文件转换为CSV / TSV / JSON文件。
- 4. Slate: PDFMiner的包装器实现
- 5. **PDFQuery**: pdfminer, lxml和pyquery的轻量级包装器。它旨在使用尽可能少的代码可靠地从PDF集合中提取数据。
- 6. xpdf: xpdf的 Python包装器(目前只是"pdftotext"实用程序)

从pdf中提取文本

▲ 赞同 19

▼

● 1条评论

✔ 分享

● 喜欢

使用PyPDF2从pdf中提取简单文本,示例代码如下:

```
1
```

```
import PyPDF2
# pdf file object
# you can find find the pdf file with complete code in below
pdfFileObj = open('example.pdf', 'rb')
# pdf reader object
pdfReader = PyPDF2.PdfFileReader(pdfFileObj)
# number of pages in pdf
print(pdfReader.numPages)
# a page object
pageObj = pdfReader.getPage(0)
# extracting text from page.
# this will print the text you can also save that into String
print(pageObj.extractText())
```

从pdf中读取表格数据

使用Pdf中的Table数据,我们可以使用Tabula-py,示例代码如下:

```
import tabula
# readinf the PDF file that contain Table Data
# you can find find the pdf file with complete code in below
# read_pdf will save the pdf table into Pandas Dataframe
df = tabula.read_pdf("offense.pdf")
# in order to print first 5 lines of Table
df.head()
```

如果您的Pdf文件包含多个表,可以进行如下设置:

```
df = tabula.read_pdf ("crime.pdf", multiple_tables = True)
```

还可以从任何特定PDF页面的特定部分提取信息

```
tabula.read_pdf ("crime.pdf", area = (126,149,212,462) , pages = 1)
```

设置读取输出为JSON格式

```
tabula.read_pdf ("crime.pdf", output_format ="json")
```

将Pdf导出到Excel

使用以下代码将PDF数据转换为Excel或CSV

```
tabula.convert_into ("crime.pdf", "crime_testing.xlsx", output_format ="xlsx")
```

更多参考资料

python提取pdf信息:



https://towardsdatascience.com/python-for-pdf-ef0fac2808b0

@ towardsdatascience.com



1

PyPDF2库文档:



公众号:深度学习与Python,专注于深度学习、机器学习前沿知识与资讯

发布于 2019-01-27

Python PDF 文本分析

文章被以下专栏收录



关注专栏

推荐阅读

Python数据挖掘——文本分析

作者 | zhouyue65 来源 | 君泉计量原文 | Python数据挖掘——文本分析文本挖掘: 从大量文本数据中抽取出有价值的知识,并且利用这些知识重新组织信息的过程。一、语料库(Corpus)语料库是...

CDA数据... 发表于大数据分析...



实例8: 用Python暴力破解PDF 密码

TT 11-2

Pytho...

发表于Pytho...

Python网络爬虫和文本分析!

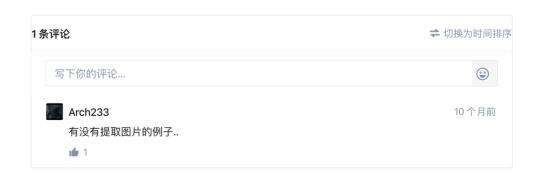
文本大数据分析在社科学术研究中的应用方兴未艾。本文以搜集长沙市历年政府工作报告,并统计其中与环境规制相关的词汇出现频次这一问题为例,基于python3,介绍网络爬虫和文本分析的基本工...

一枚程序媛

从Excel至 个Panda

本文涉及P 数,通过证据生成和导理,以及引 筛选,分享操作。生月

一枚程序類



▲ **赞同 19** ▼ ● 1条评论 **7** 分享 ● 喜欢 ★ 收藏 …