

C

博客

Q

☆

C

≡

李青锋的专栏

≡ 目录视图

≡ 摘要视图

RSS 订阅

个人资料



i_wooden

+ 加关注

发私信

访问：143409次

积分：1377

等级：BLOG > 4

排名：千里之外

原创：25篇

转载：29篇

译文：0篇

评论：57条

文章搜索

Q

文章分类

数据结构与算法 (8)

开发工具 (6)

项目开发 (9)

opencv (2)

windows编程 (1)

linux (4)

图像处理 (2)

C++ (3)

文档 (0)

android (1)

python (1)

notepad++源码学习 (1)

文章存档

2016年10月 (1)

2016年07月 (1)

2016年06月 (2)

2016年01月 (1)

2015年08月 (1)

展开

阅读排行

如何编译生成dll (33844)

在windows下配置pthrea

原

用python处理MS Word

标签：python word MS-WORD DOC DOCX

2016-06-03 22:04 4450人阅读 评论(6) 收藏 举报

≡ 分类：python

快速回复

我要收藏

版权声明：本文为博主原创文章，未经博主允许不得转载。

目录(?)

[+]

使用Python工具读写MS Word文件（docx与doc文件），主要利用了python-docx包。本文给出一些常用的操作，并完成一个样例，帮助大家快速入手。

安装

pyhton处理docx文件需要使用python-docx包，可以利用pip工具很方便的安装，pip工具在python安装路径下的Scripts文件夹中

1 pip install python-docx

当然你也可以选择使用easy_install或者手动方式进行安装

写入文件内容

此处我们直接给出一个样例，根据自己的需要摘取有用的内容

1 #coding=utf-8

2

3 from docx import Document

4 from docx.shared import Pt

5 from docx.shared import Inches

6 from docx.oxml.ns import qn

7 #打开文档

8 document = Document()

9 #加入不同等级的标题

10 document.add_heading(u' MS WORD写入测试',0)

11 document.add_heading(u' 一级标题',1)

12 document.add_heading(u' 二级标题',2)

13 #添加文本

14 paragraph = document.add_paragraph(u' 我们在做文本测试! ')

15 #设置字号

16 run = paragraph.add_run(u' 设置字号、')

17 run.font.size = Pt(24)

18

19 #设置字体

20 run = paragraph.add_run(' Set Font,')

21 run.font.name = 'Consolas'

22

23 #设置中文字体

24 run = paragraph.add_run(u' 设置中文字体、')

25 run.font.name=u' 宋体'

26 r = run._element

27 r.rPr.rFonts.set(qn('w: eastAsia'), u' 宋体')

- 【opencv】图像细化 (21072)
- 如何将土豆优酷视频插入 (13115)
- 【opencv】两条平行线 (9483)
- 用python处理MS Word (6997)
- linux下编译动态链接库的 (4435)
- 判断两个矩形是否重叠 (4336)
- 关于统计变换 (CT/MCT, (3903)
- C++中 int i 与 int &i 注意 (3546)
- C++中 int i 与 int &i 注意 (3284)

评论排行

- 【opencv】图像细化 (29)
- 在windows下配置pthreads (8)
- 用python处理MS Word (6)
- 【opencv】两条平行线 (5)
- Android NDK读取文件 (3)
- 关于统计变换 (CT/MCT, (2)
- notepad++ v1.0源码学习 (1)
- 判断两个矩形是否重叠 (1)
- emacs在win8乱码 (1)
- 并查集算法介绍 (1)

推荐文章

- * 【《Real-Time Rendering 3rd》提炼总结】(一) 全书知识点总览
- * CSDN日报20170409 ——《扯蛋的密码规则》
- * Shader2D: 一些2D效果的Shader实现
- * 一个屌丝程序员的人生 (六十一)
- * 自定义控件三部曲视图篇 (三) ——瀑布流容器 WaterFallLayout实现
- * 面向服务的体系架构 (SOA) ——架构篇

最新评论

- 用python处理MS Word
h_rongy: 如何为表格设置边框线
table =
document.add_table(rows=1,cols=
- notepad++ v1.0源码学习 (一)
Happy-Worm: 请问哪里可以下在
v1.0的源代码
- Android Studio使用OpenCV
i_wooden: 真很抱歉各位, 由于
最近工作太忙了, 而且项目方向
转了, 一直没来得及补齐这篇博
文, 在此表示歉意, 浪费你们...
- Android Studio使用OpenCV
Jobs_Huang: 然后呢? (‘_.’)
- 在windows下配置pthread
flames60: @qianchenglenger:
谢谢, 解决了
- 在windows下配置pthread
i_wooden: @wdear0401:没有试
过, 不过您可以试试这个
https://sourceforge.net/p...
- 在windows下配置pthread
i_wooden: @flames60:百度一
下, 试试这个
http://bbs.csdn.net/topics/3919...
- 在windows下配置pthread
flames60: 作者你好, 为什么
vs2015comuunity在最后测试的

```
28
29 #设置斜体
30 run = paragraph.add_run(u'斜体、')
31 run.italic = True
32
33 #设置粗体
34 run = paragraph.add_run(u'粗体').bold = True
35
36 #增加引用
37 document.add_paragraph('Intense quote', style='Intense Quote')
38
39 #增加无序列表
40 document.add_paragraph(
41     u'无序列表元素1', style='List Bullet'
42 )
43 document.add_paragraph(
44     u'无序列表元素2', style='List Bullet'
45 )
46 #增加有序列表
47 document.add_paragraph(
48     u'有序列表元素1', style='List Number'
49 )
50 document.add_paragraph(
51     u'有序列表元素2', style='List Number'
52 )
53 #增加图像（此处用到图像image.bmp，请自行添加脚本所在目录中）
54 document.add_picture('image.bmp', width=Inches(1.25))
55
56 #增加表格
57 table = document.add_table(rows=1, cols=3)
58 hdr_cells = table.rows[0].cells
59 hdr_cells[0].text = 'Name'
60 hdr_cells[1].text = 'Id'
61 hdr_cells[2].text = 'Desc'
62 #再增加3行表格元素
63 for i in xrange(3):
64     row_cells = table.add_row().cells
65     row_cells[0].text = 'test'+str(i)
66     row_cells[1].text = str(i)
67     row_cells[2].text = 'desc'+str(i)
68
69 #增加分页
70 document.add_page_break()
71
72 #保存文件
73 document.save(u'测试.docx')
```

快速回复

我要收藏

时候发生了错误：C2011
"timesp...

【opencv】两条平行线之间的距
Cesar-Wu: 从打印的结果看 你的
直线不是平行线，他们的起始点
是一样的

在windows下配置pthread
wdear0401: 你好~我在windows
用vs2015+64位编译linux项目，
用到文中链接给的pthread，反...

该段代码生成的文档样式如下

MS·WORD 写入测试

一级标题

二级标题

我们在做文本测试！设置字号、Set·Font,设置中文字体、斜体、粗体


Intense·quote

无序列表元素 1

无序列表元素 2

有序列表元素 1

有序列表元素 2



Name	Id	Desc
test0	0	desc0
test1	1	desc1
test2	2	desc2

分页符

快速回复

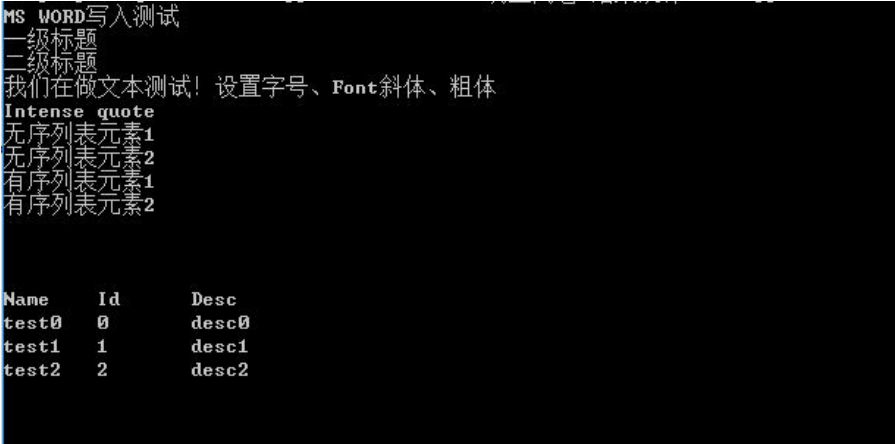
我要收藏

注：有一个问题没找到如何解决，即如何为表格设置边框线。如果您知道，还请能够指教。

读取文件内容

```
1 #coding=utf-8
2
3 from docx import Document
4 #打开文档
5 document = Document(u'测试.docx')
6 #读取每段资料
7 l = [ paragraph.text.encode('gb2312') for paragraph in document.paragraphs];
8 #输出并观察结果，也可以通过其他手段处理文本即可
9 for i in l:
10     print i
11 #读取表格材料，并输出结果
12 tables = [table for table in document.tables];
13 for table in tables:
14     for row in table.rows:
15         for cell in row.cells:
16             print cell.text.encode('gb2312'),'\\t',
17             print
18             print '\\n'
```

我们仍然使用刚才我们生成的文件，可以看到，输出的结果为



注意：此处我们使用gb2312编码方式读取，主要是保证中文的读写正确。一般情况下，使用的utf-8编码方式。另外，python-docx主要处理docx文件，在加载doc文件时，会出现问题，如果有大量doc文件，建议先将doc文件批量转换为docx文件，例如利用工具[doc2doc](#)

参考资料

- [python-docx](#)
- [Best way to extract text from a Word doc without using COM/automation?](#)
- [Extracting text from MS word files in python](#)
- [Add a w:ea:EastAsia attribute](#)

致谢

感谢zanlaoli在文章评论中帮助指出如何设置中文字体的问题。
感谢[四路组相联](#) 指出包引入缺失的问题。

快速回复

+ ☆ 我要收藏

顶
2

踩
0

- ▲ 上一篇 [Android NDK读取文件](#)
- ▼ 下一篇 [通过SSH在Ubuntu下安装TeamViewer软件](#)

参考知识库



Python知识库
22433 关注 | 1580 收录



软件测试知识库
4359 关注 | 318 收录

猜你在找

- 使用决策树算法对测试数据进行分类实战
- Python自动化开发基础 列表-字典-IO处理 day2
- 使用python操作Oracle
- python自动化运维开发【零基础】快速入门
- 软件测试基础
- 批处理中关于errorlevel 和 % 判断依据 及 编程中程
- 深度学习DL与卷积神经网络CNN学习笔记随笔-04-基于
- undefined与null的区别
- Python爬虫利器之PhantomJS的用法
- 各大机器学习包汇总python版持续更新

查看评论

3楼 [h_rongy](#) 2017-02-04 16:28发表 



如何为表格设置边框线
`table = document.add_table(rows=1,cols=3,style = 'Table Grid')`

2楼 [四路组相联](#) 2016-09-22 14:43发表 



请问设置中文字体那部分中的qn是从哪里来的？
编译器提示未定义的变量

Re: [四路组相联](#) 2016-09-22 14:51发表 



回复四路组相联：是不是少了这个？`from docx.xml.ns import qn`

Re: [i_wooden](#) 2016-09-23 10:58发表 



回复四路组相联：已添加上，谢谢。另外修正了一个style lookup by style_id is deprecated的问题

1楼 [zanlaoli](#) 2016-06-19 01:41发表 

中文字体设置的问题 python-docx原本只支持对ascii和hAnsi字体的设置 我们给它补上eastAsia字体的设置即可



Re: i_wooden 2016-06-24 21:29发表



回复zanlaoli：感谢帮忙指出，已经在文章中补充。

发表评论

用户名： weixin_38266467

评论内容：

提交

快速回复

我要收藏

* 以上用户言论只代表其个人观点，不代表CSDN网站的观点或立场

核心技术类目

- 全部主题
- Hadoop AWS 移动游戏 Java Android iOS Swift 智能硬件 Docker OpenStack VPN Spark ERP IE10 Eclipse CRM JavaScript 数据库 Ubuntu NFC WAP jQuery BI HTML5 Spring Apache .NET API HTML SDK IIS Fedora XML LBS Unity Splashtop UML components Windows Mobile Rails QEMU KDE Cassandra CloudStack FTC coremail OPhone CouchBase 云计算 iOS6 Rackspace Web App SpringSide Maemo Compuware 大数据 aptech Perl Tornado Ruby Hibernate ThinkPHP HBase Pure Solr Angular Cloud Foundry Redis Scala Django Bootstrap