**使用python编辑和读取word文档**



[weixin\_30682415](https://blog.csdn.net/weixin_30682415) 2018-07-12 17:23:00 https://csdnimg.cn/release/blogv2/dist/pc/img/articleReadEyes.png 3191 https://csdnimg.cn/release/blogv2/dist/pc/img/tobarCollect.png 收藏 18

文章标签： [python](https://www.csdn.net/tags/MtjaQg4sNDk0LWJsb2cO0O0O.html)

版权

python调用word接口主要用到的模板为python-docx，基本操作官方文档有说明。

[python-docx官方文档地址](https://python-docx.readthedocs.io/en/latest/index.html)

使用python新建一个word文档，操作就像文档里介绍的那样：

1 from docx import Document

2 from docx.shared import Inches

3

4 document = Document()

5

6 document.add\_heading('Document Title', 0) #插入标题

7

8 p = document.add\_paragraph('A plain paragraph having some ') #插入段落

9 p.add\_run('bold').bold = True

10 p.add\_run(' and some ')

11 p.add\_run('italic.').italic = True

12

13 document.add\_heading('Heading, level 1', level=1)

14 document.add\_paragraph('Intense quote', style='IntenseQuote')

15

16 document.add\_paragraph(

17 'first item in unordered list', style='ListBullet'

18 )

19 document.add\_paragraph(

20 'first item in ordered list', style='ListNumber'

21 )

22

23 document.add\_picture('monty-truth.png', width=Inches(1.25)) #插入图片

24

25 table = document.add\_table(rows=1, cols=3) #插入表格

26 hdr\_cells = table.rows[0].cells

27 hdr\_cells[0].text = 'Qty'

28 hdr\_cells[1].text = 'Id'

29 hdr\_cells[2].text = 'Desc'

30 for item in recordset:

31 row\_cells = table.add\_row().cells

32 row\_cells[0].text = str(item.qty)

33 row\_cells[1].text = str(item.id)

34 row\_cells[2].text = item.desc

35

36 document.add\_page\_break()

37

38 document.save('demo.docx') #保存文档

读取和编辑一个已有的word文档，只需在一开始添加上文件路径就行了，如下：

1 from docx import Document

2 from docx.shared import Inches

3

4 document = Document('demo.docx') #打开文件demo.docx

5 for paragraph in document.paragraphs:

6 print(paragraph.text) #打印各段落内容文本

7

8 document.add\_paragraph(

9 'Add new paragraph', style='ListNumber'

10 ) #添加新段落

11

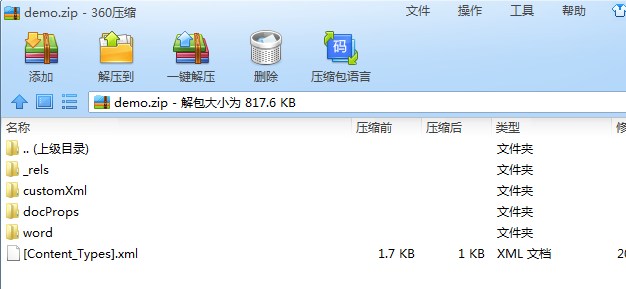
12 document.save('demo.docx') #保存文档

如果是想读取其中的图片或是更复杂地编辑，首先我们需要先来认识下docx文档的格式组成：

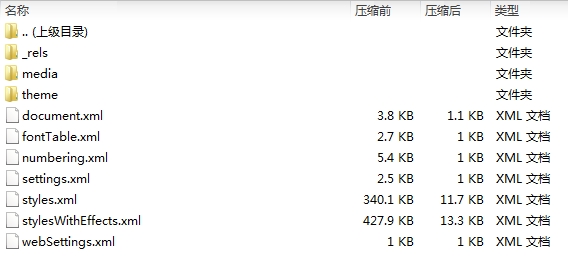
===docx是Microsoft Office2007之后版本使用的，用新的基于XML的压缩文件格式取代了其目前专有的默认文件格式，在传统的文件名扩展名后面添加了字母“x”（即“.docx”取代“.doc”、“.xlsx”取代“.xls”、“.pptx”取代“.ppt”）。===

docx格式的文件本质上是一个ZIP文件。将一个docx文件的后缀改为ZIP后是可以用解压工具打开或是解压的。事实上，Word2007的基本文件就是ZIP格式的，他可以算作是docx文件的容器。

docx 格式文件的主要内容是保存为XML格式的，但文件并非直接保存于磁盘。它是保存在一个ZIP文件中，然后取扩展名为docx。将.docx 格式的文件后缀改为ZIP后解压, 可以看到解压出来的文件夹中有word这样一个文件夹，它包含了Word文档的大部分内容。而其中的document.xml文件则包含了文档的主要文本内容。



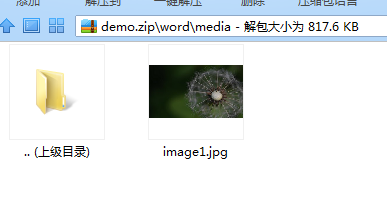
**word目录下：**



**document.xml文件内容：**



**media目录下存放word文档中插入的图片：**



所以，我们可以使用手工的方法编辑文件document.xml来对该word文档内容进行编辑，或是提取文档media中图片文件的方式来提取该word文档中所插入的所有图片。

1 import zipfile

2

3 f=zipfile.ZipFile('demo.docx','r')

4

5 for filename in f.namelist():

6 f.extract(filename)