Lendo documentos

Arquivos DOCX

(Microsoft Word)

Lendo documentos

Arquivos DOCX

Documento criado pelo Microsoft Word, um programa comum de processamento de texto; contém texto do documento, imagens, formatação, estilos, objetos desenhados e outras configurações do documento; comumente usado para criar documentos em ambientes comerciais e acadêmicos.

Ao contrário dos arquivos .DOC, que armazenam dados de documentos em um único arquivo binário, os arquivos DOCX são criados usando o formato Open XML, que armazena documentos como uma coleção de arquivos e pastas separados em um pacote zip compactado. Os arquivos DOCX contêm arquivos XML e três pastas, docProps, Word e _rels, que contêm as propriedades, o conteúdo e os relacionamentos do documento entre os arquivos.

Lendo documentos

Arquivos DOCX

Arquivos DOCX são projetados para tornar o conteúdo do documento acessível. Por exemplo, o texto do documento é salvo usando arquivos de texto simples e as imagens do documento são armazenadas como arquivos de imagem individuais no arquivo DOCX.

Arquivos DOCX podem ser abertos pelo Word 2007 ou posterior para Windows, ou com o Word 2008 ou posterior para Mac OS X. Eles também podem ser abertos com versões anteriores do Word para Mac e Windows por meio de suporte a documentos Open XML. Arquivos DOCX também podem ser abertos pelo LibreOffice Writer, que é um editor de textos similar ao Microsoft Word, porém de código aberto.

OBSERVAÇÃO: Para explorar o conteúdo de um arquivo DOCX manualmente, renomeie a extensão ".docx" para ".zip" e, em seguida, descompacte o arquivo resultante com qualquer utilitário de descompactação de zip.

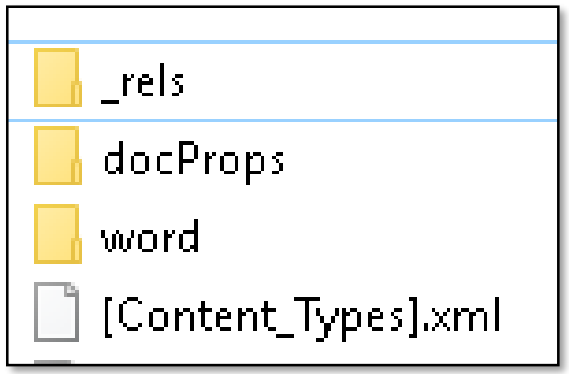
Fonte: http://fileformat.wikia.com/wiki/DOCX

Lendo documentos

Arquivos DOCX

Veja o conteúdo de um arquivo DOCX descompactado:

```
|   [Content_Types].xml
|
----docProps
|       app.xml
|       core.xml
|
----word
|   |   document.xml
|   |   fontTable.xml
|   |   settings.xml
|   |   styles.xml
|   |   webSettings.xml
|   |
|   ----media
|   |   |       image1.png
|   |   |       image2.png
|   |   |       image3.png
|   |   |       image4.png
|   |   |       image5.png
|   |   |
|   ----theme
|   |   |       theme1.xml
|   |   |
|   ----_rels
|           document.xml.rels
|
----_rels
        .rels
```



Lendo documentos

Arquivos DOCX

Para trabalharmos com arquivos DOCX podemos utilizar a biblioteca *python-docx* que pode ser instalada utilizando-se o *pip*.

$ pip install python-docx

Lendo documentos

Arquivos DOCX

Vamos ler um documento DOCX e imprimir seu conteúdo.

```
import docx

doc = docx.Document('ArquivoWord.docx')

for a in doc.paragraphs:
    print(a.text)
```

Para importar a biblioteca Python-Docx usamos “import docx”.

Lendo documentos

Arquivos DOCX

Comparado ao texto simples, os arquivos *.docx* têm muita estrutura. Essa estrutura é representada por três tipos de dados diferentes no *Python-Docx*. No nível mais alto, um objeto *Document* representa o documento inteiro. O objeto *Document* contém uma lista de objetos *Paragraph* para os parágrafos do documento. (Um novo parágrafo começa sempre que o usuário pressiona *ENTER* ou *RETURN* enquanto digita em um documento do *Word*.)

Lendo documentos

Arquivos DOCX

Cada um desses objetos *Paragraph* contém uma lista de um ou mais objetos *Run*. Um objeto *Run* é uma execução contígua de texto com o mesmo estilo. Um novo objeto *Run* é necessário sempre que o estilo de texto é alterado.

O parágrafo a seguir possui 7 objetos *Run*.

Olá, estamos trabalhando com arquivos **DOCX**, que são documentos do *Microsoft Word*. Adeus!

7 Objetos Run porque ocorreram seis mudanças de estilo no texto.

```
(auladocx) C:\Users\evaldo\testedocx>python arquivo_docx3.py
Olá, estamos trabalhando com arquivos
DOCX
, que são documentos do
Microsoft Word
.
Adeus
!
```

Texto normal.
Negrito
Texto normal
Itálico
Texto normal
Sublinhado
Texto normal

Lendo documentos

Arquivos DOCX

Imprimindo o conteúdo *text* do objeto *Run*.

```
import docx

doc = docx.Document('ArquivoWord2.docx')

total_runs = len(doc.paragraphs[0].runs)
contador = 0

while contador < total_runs:
    print(doc.paragraphs[0].runs[contador].text)
    contador += 1
```

Lendo documentos

Arquivos DOCX

Salvando o conteúdo do arquivo DOCX em um arquivo de texto.

```
import docx

doc = docx.Document('ArquivoWord.docx')

for a in doc.paragraphs:
    with open('resultado.txt', 'a') as arquivo:
        arquivo.write(a.text)
```

Lendo documentos

Arquivos DOCX

Criando uma função para ler um arquivo DOCX.

Vamos salvar

este arquivo com

o nome lerDocx.py

```
import docx


def ler(arquivo):
    try:
        doc = docx.Document(arquivo)
        texto = []
        for paragrafo in doc.paragraphs:
            texto.append(paragrafo.text)
        return '\n'.join(texto)
    except Exception as erro:
        return "Ocorreu um erro:", erro
```

Lendo documentos

Arquivos DOCX

Vamos agora implementar um outro programa que vai importar o arquivo lerDocx.py

```
import lerDocx

print(lerDocx.ler('ArquivoWord.docx'))
```

Lendo documentos

Arquivos DOCX

Identificando o estilo do parágrafo

```
import docx

doc = docx.Document('ArquivoWord3.docx')

paragrafos = len(doc.paragraphs)
contador = 0

while contador < paragrafos:
    print(doc.paragraphs[contador].style.name)
    if doc.paragraphs[contador].style.name == 'Heading 4':
        doc.paragraphs[contador].style.name = 'Normal'
    contador += 1

doc.save('arquivo_salvo.docx')
```

FIM