```
>>> with open(r'H:\示例\第6章\格式研究\HelloWorld.doc','rb') as f:
...     txt=f.read()
...
>>> txt[2560:2560+12]
b'Hello,World!'
```

2. 从.doc 文档中提取文本

和.xls 文档不同，目前还没有直接解析.doc 文档的第三方库，部分可用的第三方库包括 Compoundfiles、Olefile、Oletools。

Microsoft 公司官方提供了从二进制 Word 文档中提取文本内容的步骤，下面使用 Compoundfiles 库实现提取过程。

（1）将.doc 文档读入数据流。

```
>>> import compoundfiles
>>> file=r'H:\示例\第6章\格式研究\HelloWorld.doc'
>>> reader=compoundfiles.CompoundFileReader(file)
>>> reader
<compoundfiles.reader.CompoundFileReader object at 0x000000000289B400>
```

我们看到，变量 reader 就是 CompoundFileReader 类的一个实例化对象，用 dir 函数查看对象的属性和方法，主要包括：close、open、root。

使用 root 属性。

```
>>> reader.root
['<CompoundFileEntity name='Data'>',
 '<CompoundFileEntity name='1Table'>',
 '<CompoundFileEntity name='\x01CompObj'>',
 '<CompoundFileEntity name='WordDocument'>',
 '<CompoundFileEntity name='\x05SummaryInformation'>',
 '<CompoundFileEntity name='\x05DocumentSummaryInformation'>']
>>> type(reader.root)
<class 'compoundfiles.entities.CompoundFileEntity'>
```

我们可以使用 CompoundFileReader 对象的 open(filename_or_entity)方法，继续打开各种 entity，它返回 CompoundFileStream 对象。

```
>>> reader.open('WordDocument')
<compoundfiles.streams.CompoundFileNormalStream object at 0x0000000002CD32E8>
```

（2）在 Word 文件流（WordDocument）的偏移 0 处读取文件信息块（FIB）。

（3）在 FIB 内，找到 FibRgFcLcb97 结构。此结构从 FIB 的第 154 个字节开始，由一系列 4 字节字段组成。

（4）在第 268 个字节处读取 FibRgFcLcb97.fcClx 字段，在第 272 个字节处读取 FibRgFcLcb97.lcbClx 字段。这些字段指定 Clx 的偏移位置和大小。

由于 FibRgFcLcb97 之前有 154 字节，268+154=422，且 fcClx 结构的单位是 4 字节，因此按小端（Little-Endian）从右到左排列，即从 422 字节处从右往左读，得到 418～422 字节才是存放 fcClx 的位置。同样地，lcbClx 应该位于整个 Word 文件流的 422～426 字节处。

```
>>> wordDocument=reader.open('WordDocument').read()
>>> from struct import unpack
>>> fcClx=unpack('L', wordDocument[418:422])[0]
>>> lcbClx=unpack('L', wordDocument[422:426])[0]
>>> fcClx,lcbClx
(5217, 21)
```

（5）在 FibRgFcLcb97.fcClx 字段指定的偏移处开始，从表格流中读取 Clx 结构。