---

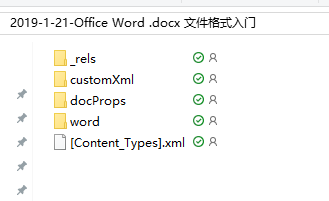
category: 技术

date: 2019-1-21 23:27

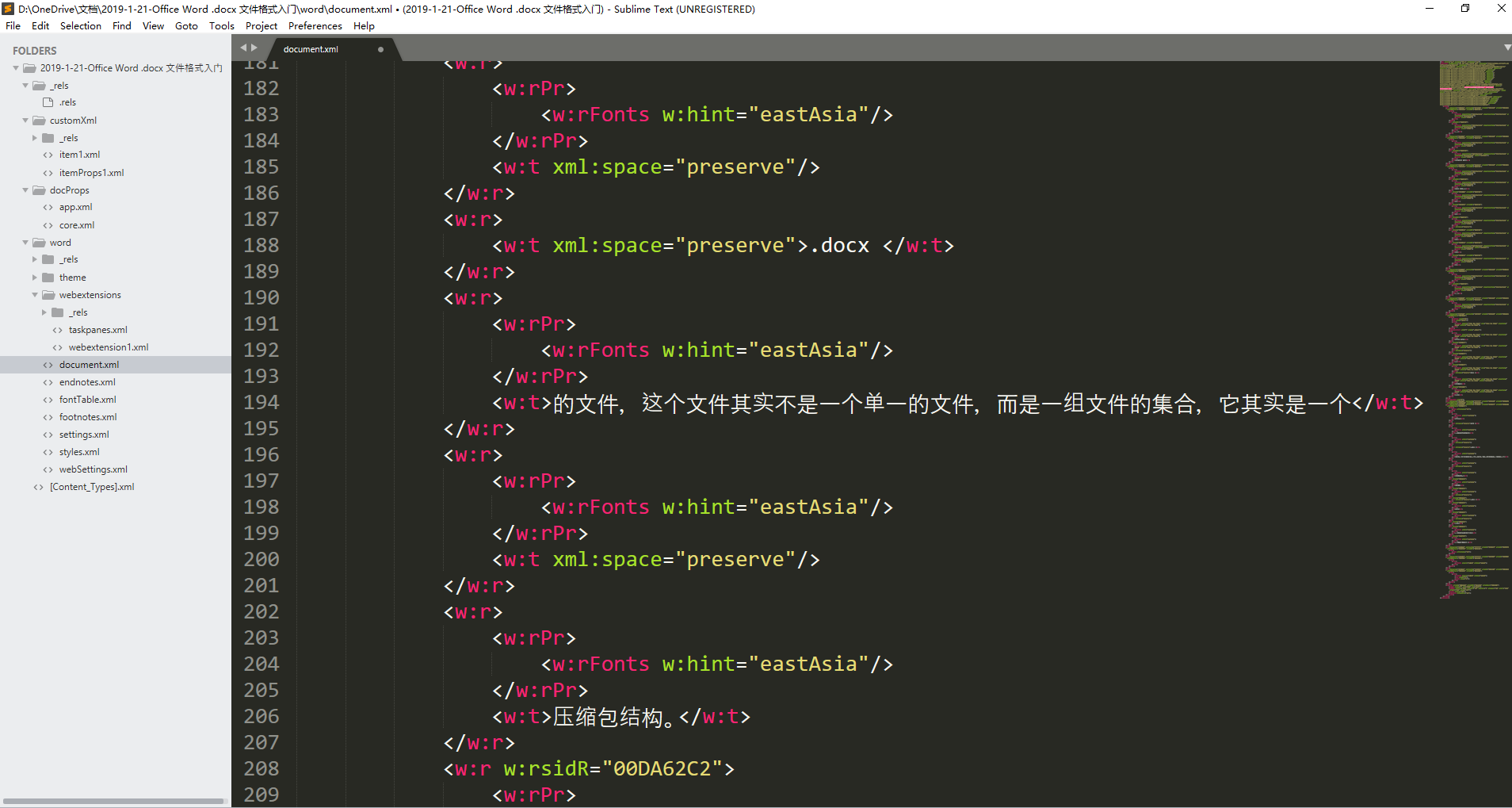
---

Office Word .docx 文件格式入门

Office Word 一般会保存为后缀名为 .docx 的文件，这个文件其实不是一个单一的文件，而是一组文件的集合，它其实是一个 压缩包结构。把后缀 .docx 改为 .zip，双击就可以解压这个文件，结构如下图所示：



打开word/document.xml 后，你可以看到如下内容：



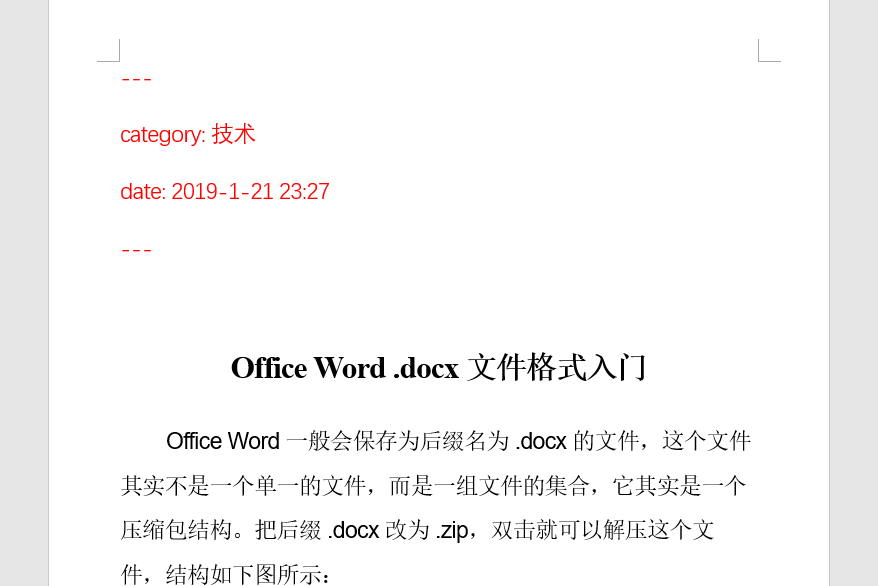
可以看到，docx 文件基本上是由很多个XML文件组成的，协议比较复杂，描述了字体、缩进、颜色等等Word里面的功能。

已经有很多开源的类库提供了读写Word文档的功能，如[docx](https://github.com/dolanmiu/docx)、[PHPOffice/PHPWord](https://github.com/PHPOffice/PHPWord)、[docxtemplater](https://github.com/open-xml-templating/docxtemplater)等等。

其实，如果只是简单的读取 docx 文件，只需要使用 XML 库，读取对应的标签即可。

读取Word文件中的文本

类比于Jekyll中的markdown文件，可以在文件头使用Yaml 格式设置一些文件信息，如下面所示：



那么，现在我们怎样可以把文件头中的文本内容读取出来呢？

我们知道一个.docx文件就是一个zip文件，所以我们需要先读取这个zip文件，解压出来document.xml文件，并从中解析出所有的<w:t></w:t>标签中的内容。