**语法**

1. Pdf语法基于什么字符集？

PDF文件就是一个8位字节序列。这些字节序列可以根据如下语义规则分组成若干token。一个或多个token组成一个更高一级的语义实体（主要是构成PDF文档基本数据的对象）。PDF文件可全部由可见可打印ASCII字符子集和空白符（例如Space、Tab、回车、换行）对应的8位字节来表示。但PDF文件并不局限于ASCII字符集，实际上它可以包含任意8位字节，只要遵循如下要求即可：

a.分隔对象和描述文件结构的token全以ASCII表示，所有保留关键字和标准字典中的键值名也以ASCII表示。

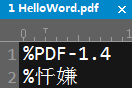
b.特定类型对象的数据值可以但不必全为ASCII，但出于可读性，最好用ASCII表示。在实际中，一些本身就是

二进制的数据，例如图像，出于压缩和效率，最好直接以二进制表示。

c.包含二进制的PDF文件必须在传输和存储过程中保留所有原始字。

2. 基于HelloWord.pdf说明Pdf文件结构？

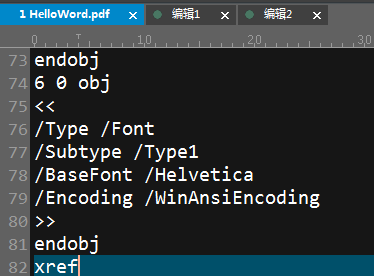
a.文件头：



这个文件头表明了这个pdf文件使用的版本号，后面的乱码则表面了在该pdf文件当中包含有二进制数据，所以在版本号之后紧跟了4个字符的二进制数据，由于转码的问题，所以显示的是表示中文字符的乱码。

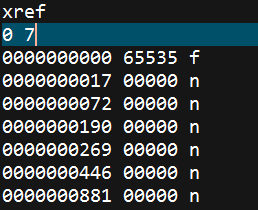
b.文件体

除了文件头，在xref标志之前的所有内容都属于文件体



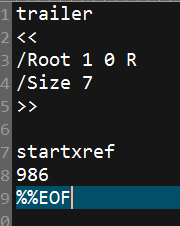
c.交叉表

交叉应用表包含一些信息，使得应用程序可以随机访问文件中的任何间接对象，而无需读取整个文件。helloword.pdf示例中包含了一个交叉应用节，其含有一个子节，子节中有7个项，其中6个项是有效项，1个项是空闲项。

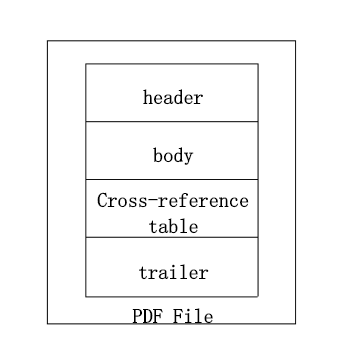


d.文件尾

一个PDF文件的尾部使得一个PDF应用程序快速定位它的交叉应用表和某些特定对象的位置。PDF应用程序应该从PDF文件的尾部开如解析。PDF文件的最后一行只包含一个文件结尾标志%%EOF。结尾标志的前面两行分别是关键词startxref和最后一个交叉应用节的偏移地址（从文件开始算起）。关键词startxref前一行是一个字典对象，其由一个关键词trailer和紧随其后的由《》括起来的一系列键值对组成。如下图helloword.pdf所示的文件尾：

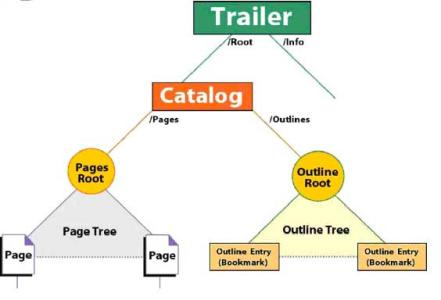


pdf大致的文件结构如下图所示：

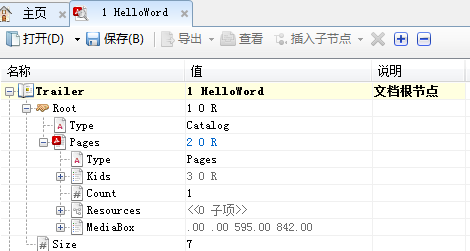


3. 基于HelloWord.pdf说明Pdf文档结构？

文件结构反映的是二进制数据在存储介质上的存放格式，是一个物理概念；而文档结构反映的是信息的一种逻辑上的层级关系，是一个逻辑概念。一个PDF文档是由一些称为“对象”的模块组成的。并且每个对象都有数字标号，这样的话可以这些对象就可以北其他的对象所引用。这些对象不需要按照顺序出现在PDF文档里面，出现的顺序可以是任意的，比如一个PDF文件有3页，第3页可以出现在第1页以前，对象按照顺序出现唯一的好处就是能够增加文件的可读性。正是因为页与页之间的不相关性，就可以对PDF文件的页码进行随机的访问。整个pdf在文档解析的时候，首先从文件尾（Trailer）开始解析，文件中的Trailer里面包含了根对象(/Root)的对象号（和相应的/Info信息，如果有的话），并且说明交叉引用表的位置，通过对交叉引用表的查询可以目录对象(Catalog)。这个目录对象是该PDF文档的根对象，包含PDF文档的大纲(/Outline)和页面组对象（/Pages）引用。大纲对象(/Outline)是指PDF文件的书签树；页面组对象（/Pages）包含该文件的页面数，各个页面对象(page)的对象号。整体上如下图所示：



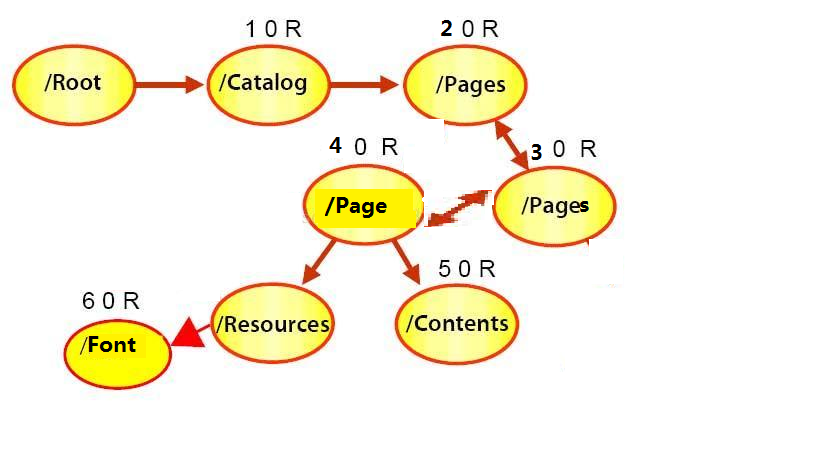
1. 文档目录



从上面的图可以看出，上述的helloword.pdf中并没有outline对象，只有pages对象作为Catalog对象的子对象。

1. Page树

Page对象作为PDF中的核心对象，包含有如何显示该页面的信息，例如使用的字体，包含的内容（文字，图片等），页面的大小。当然里面的子项也可以是其他对象的引用。页面中包含的信息是包含在一个称为流（stream）的对象里，这个流的长度（字节数）必须直接给出或指向另外一个对象。



Helloword.pdf中的文件结构如上图所示。

1. 名称字典

PDF文件中的某些类别的对象可以按名称而不是按对象来引用。名称和对象之间的对应关系由文档名称字典确定，该名称字典通过文档目录中的名称条目进行定位。

4. 基于HelloWord.pdf说明支持的基本类型？

pdf里面一共支持的基本数据类型有8种。

1.booleam

用关键字true或false表示。

2.numeric

包括整形和实型,不支持非十进制数字,不支持指数形式的数字。



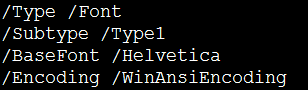
3.string

由一系列0-255之间的字节组成,一个string总长度不能超过65535。



4.name

由一个前导/和后面一系列字符组成,最大长度为127。和string不同的是,name是不可分割的和唯一的。



5.array

用[]包含的一组对象,可以是任何pdf对象(包括array).虽然pdf只支持一维array,但可以通过array的嵌套实现任意维数的array。

例如：

1.

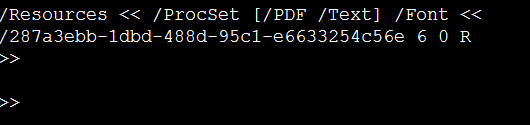
2. 

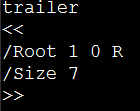
3. 

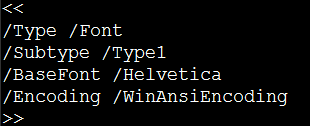
6.Dictionary

用"<<"和">>"包含的若干组条目,每组条目都由key和value组成,其中key必须是name对象,并且一个dictionary内的key是唯一的;value可以是任何pdf的合法对象(包括dictionary对象)。

例:



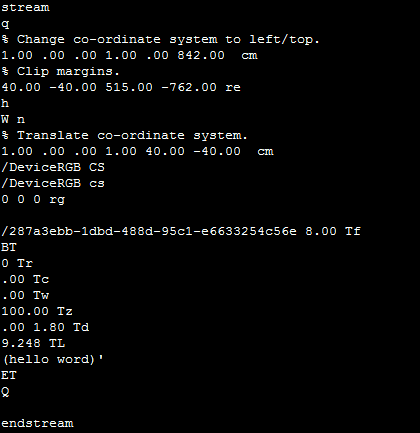




7.stream

由关键字stream和endstream包含一系列字节.内容和string很相似,但有区别:stream可以分几次读取,分开使用不同的部分,string必须作为一个整体一次全部读取使用;string有长度限制,但stream却没有这个限制。一般较大的数据都用stream表示。

例:



8.NULL

用null表示，代表空。如果一个key的值为null，则这个key可以被忽略；如果引用一个不存在的object则等价于引用一个空对象。

5. 工具使用PdfPatcher、UltraEdit