一、pdfminer

PDF 格式不是规范格式. 尽管它被叫做"PDF 文档"，但并不像 word 或者 html 文档。PDF 的表现更像一张图片。PDF 更像是在一张纸的各个准确的位置上把内容都摆放出来。大部分情况下，没有逻辑结构，比如句子或段落，并且不能自适应页面大小的调整。PDFMiner 尝试通过猜测它们的布局来重建它们的结构，但是不保证一定能工作。我知道这样很难看，但是，PDF 确实不够规范。

更多关于 PDF 内部结构的技术详情，请见《如何手工提取 PDF 内容》。
http://www.youtube.com/watch?v=k34wRxaxA_c
http://www.youtube.com/watch?v=_A1M4OdNsiQ
http://www.youtube.com/watch?v=sfV_7cWPgZE

模块	说明	备注
PDFParser	从文件中获取数据	无
PDFDocument	存储文档数据结构到内存中	
PDFPageInterpreter	解析 page 内容	
PDFDevice	把解析到的内容转化为你需要的东西	
PDFResourceManager	存储共享资源，例如字体或图片	共享

下图显示了 PDFMiner 中各个类之间的关系。