# 提取PDF信息程序说明文档

## 撰写人： 易 品

## 制作时间：2018/7/19

### 简介

本程序是为了解决PDF文档导入EXCEL表格所面临的手工复制和粘贴繁琐的问题而开发，在技术规范的pdf文档中，由于都是三方提供，pdf有图片的也有文档形式的，对于录入信息到云平台十分不方便，大部分得靠纯手工复制粘贴，图片得额外找其他试图软件进行手动截图并转换，效率低下。

作为一个中间件程序，本项目是通过Java语言开发的maven项目，通过apache下的PdfBox工具包对PDF文档解析，取出每个pdf文档中的文本信息，并在java程序中对其进行逻辑处理，拆分成每章、节、条、和细节等，最终将整个pdf信息，按照约定的顺序，写入规范的EXCEL表格中进行，完成信息的提取和转化，存储整个过程。以下将对整个操作过程进行详细说明。

### PDF文档准备

针对PDF文档自身，可能是良好的能框选文本信息的pdf文档，也可能是已经打包成图片格式的pdf文档，针对后者无法框选文本时，本程序将失效；针对前者，不排除pdf文本局部关键文本信息可能也是以图片格式存储的（比如关键的章节名称等），那么本程序也将失效。不过这些都可以经过其他方法补救，经过笔者测试，由高级福昕PDF编辑器中提供的完整的解决方案，可以在保证较高的成功率的比例上有效的解决以上问题。

#### 软件安装和破解

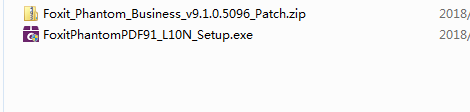


Figure 解压后的‘福昕高级PDF编辑器 v9.1.0 企业版及Crack’目录

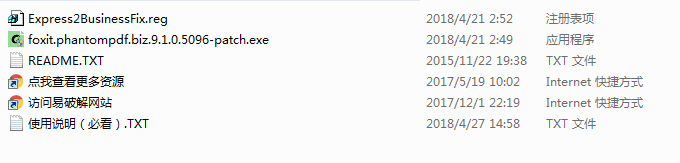


Figure 破解目录

打开图1安装软件，然后通过图2中的破解目录对文件进行破解，破解时需要将.reg注入进系统注册表，然后以管理员运行patch文件即可。随后打开福昕编辑器点击“关于福昕高级PDF编辑器”，继续点击授权信息，即可看到永久激活。



Figure 界面工具栏



Figure 授权信息

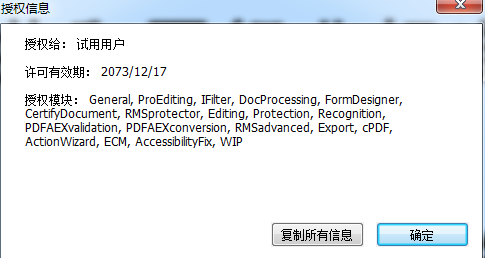


Figure 永久激活信息

#### 除水印

部分PDF文档是带水印的，如果不对文件进行除水印，那么提取出来的文档将有杂质文本信息，而这一步也是可以在福昕阅读器完成。

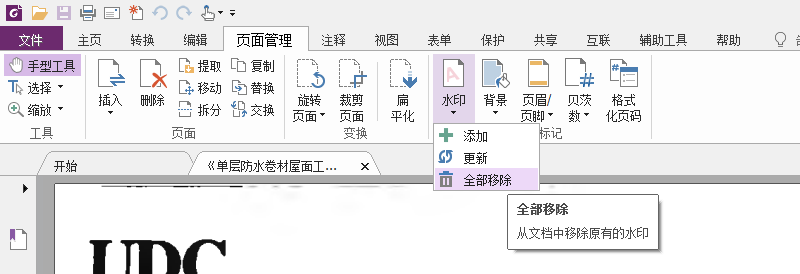
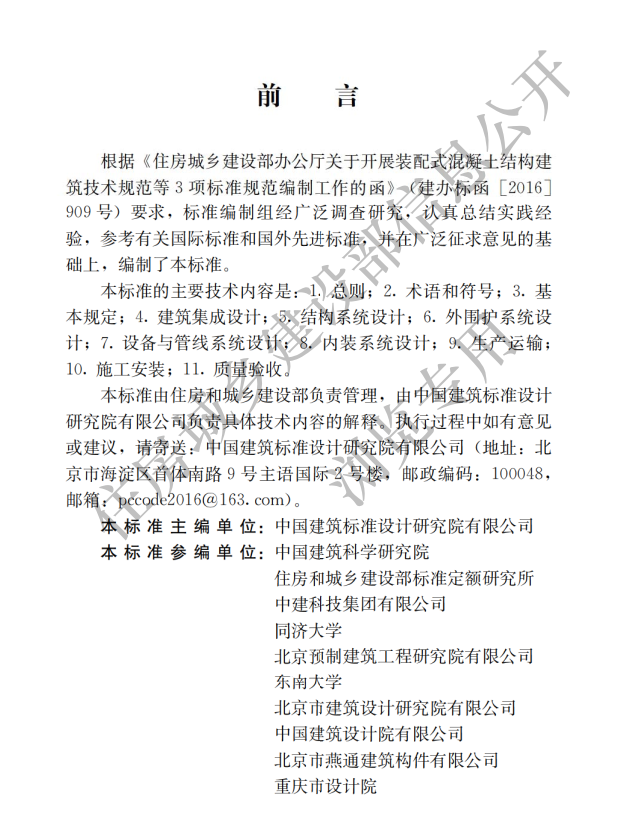
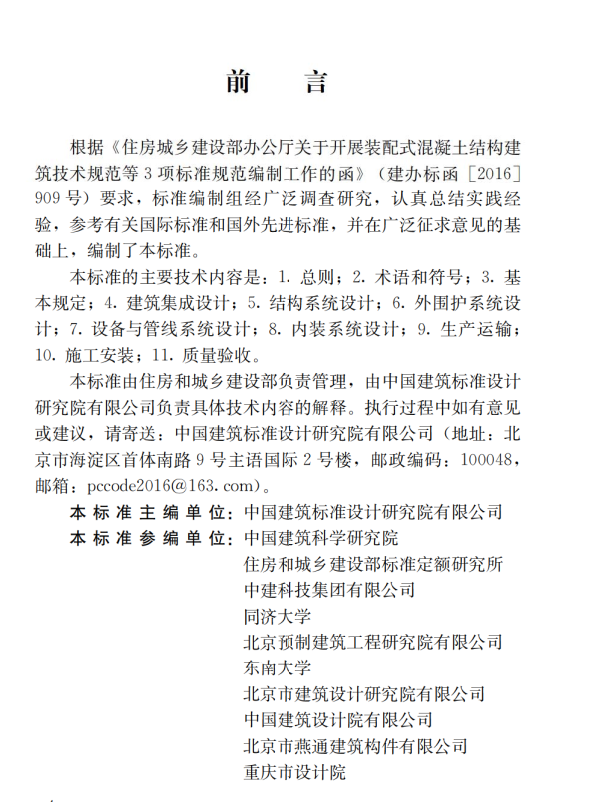


Figure 福昕自带的除水印功能

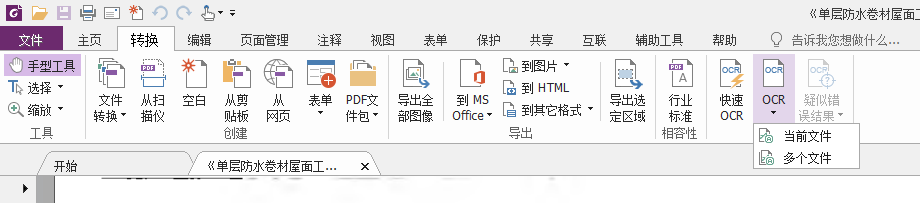
效果如下：

#### 图片转文本

对于图片式的pdf，无法框选其中的文本框，需要使用福昕中的识图功能，操作如下：

1. 首选点击OCR的当前文件按钮，对当前pdf文件进行转换。



Figure

1. 点选‘确定’按钮即可。

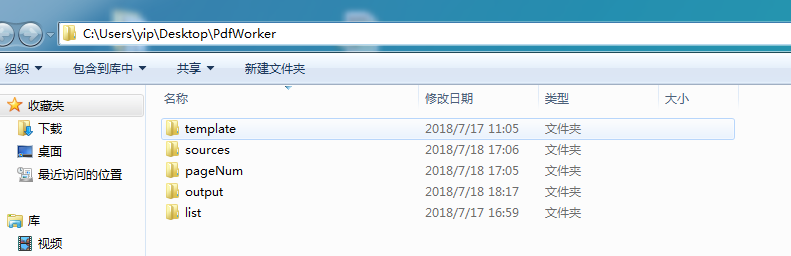


Figure

1. 等待几分钟就可以完成转换，该pdf文件就可以框选文本了。

### 项目目录约定

项目所需的目录必须严格的按照约束的目录放置相应的文件，本项目是基于IDEA平台开发的，程序运行时，会需要输入参数字符串，即初始目录，该目录为工作目录，举个例子，如下图所示：C:\Users\yip\Desktop\PdfWorker



Figure

其中PdfWorker目录可以随意命名，用户按照输入的初始路径：

C:\Users\yip\Desktop\PdfWorker 来定义，但是该目录下各个子目录必须严格按照如图所示的要求命名，下面分别说明：

* **子目录template：**存放的是模板excel文件，里面除了标题栏等，并无其他数据内容。内部只有一个子文件如下：

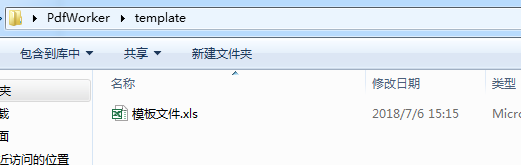


Figure 模板文件

注意：模板文件命名必须严格按照上要求，以及excel格式也必须是.xls格式。

* **子目录sources：**存放您所批量转换的pdf文档的目录，只需将已经能框选文本信息的pdf文档复制到这里即可，内部如下所示：

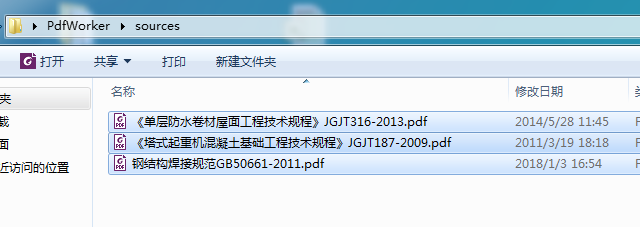


Figure 存放pdf目录

* **子目录：pageNum：**存放每个pdf文档正文页码的映射信息表，每个pdf文档信息，要精确提取正文信息，需要知道其中的技术规范和条文说明的正文内容（不包括附录信息）所在的起始和终止页码，共4个参数，如果不存在条文说明，那么只需要两个参数。如下：

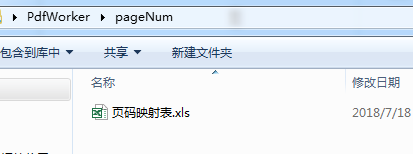


Figure 页码映射表

请注意，里面只需要这一个excel文件，而且一定注意。以下是excel文件内部，所需参数填写必须严格下图逻辑进行：



Figure 页码映射表内部书写信息图

注意：页码映射表的命名以及excel文件的格式等必须严格按照上图完成；而且每一行的pdf页码信息填写时，首先文件名不能填错，必须和你的pdf文件名相同，复制粘贴即可。技术规程两个页码参数必须要有，条文说明在pdf中有如果有也必须加上，没有可以不加。页码参数必须是偶数个！

* **子目录：list**：对于有目录丢失的文档，如果程序运行中抛出了ListMissingException那么请单独为该文档添加一份目录如下：

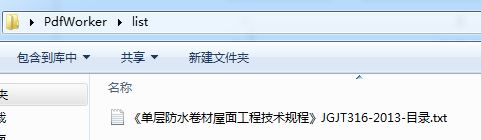


Figure 单独添加的目录

注意：该目录的命名格式是：pdf文件名-目录，如果命名错误，则无法顺利生成目录，所以请严格按照要求进行。关于目录的制作，详情规范如下：

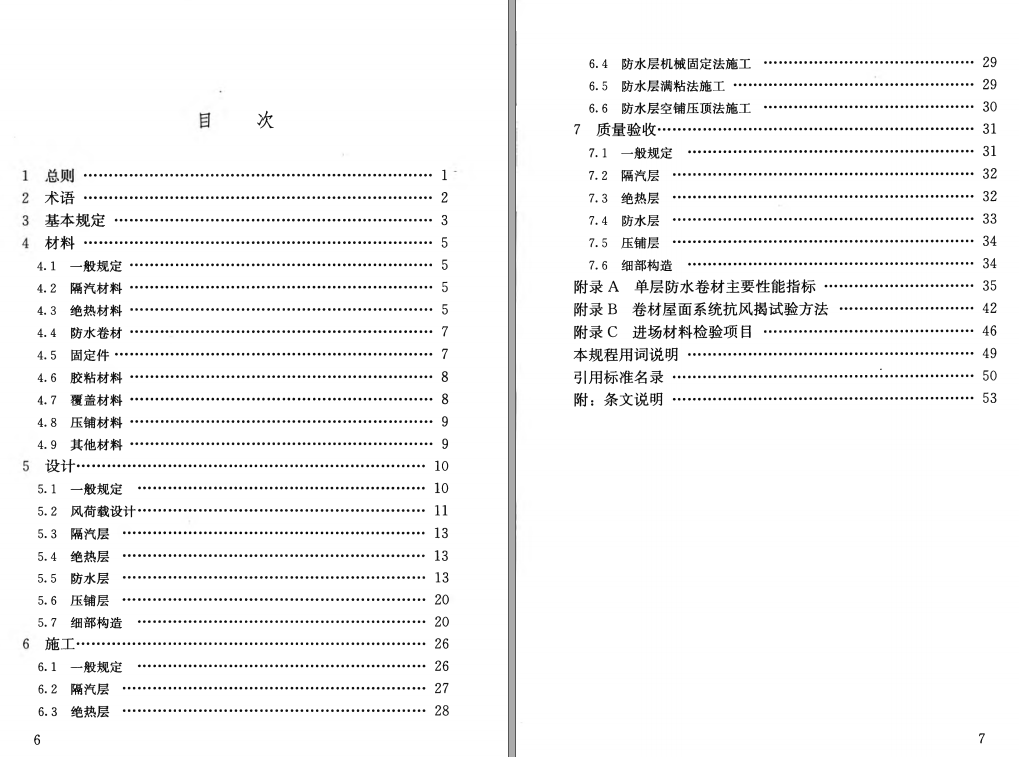


Figure pdf文档中的目录截图

手动输入的目录如下排布：

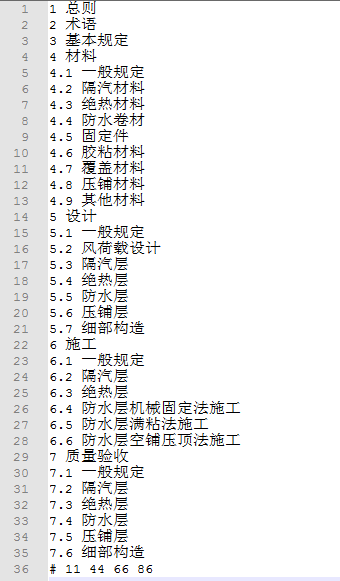


Figure 手动输入的目录（notepad++打开）

请注意：章节的序号与章节的标题之间用一个空格隔开即可，每一章、节之间必须换行；

关于目录页的获取，如果是图片式pdf，建议先使用OneNote截图，然后粘贴进OneNote进行文本提取，准确率可达85%，然后手动修改即可。

* **子目录：output**：默认为**空**，最终程序将生成所有的excel文件；

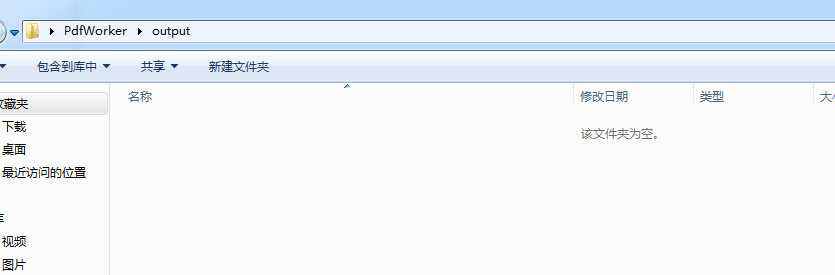


Figure 文件默认为空

请注意：如果里面已经包含了对应pdf的excel输出文件，那么程序将报错，必须手动清楚里面的所有文件先。如果对于无法成功提取的pdf文档，那么程序将再生成相应的txt文档，方便复制和粘贴。最终达到效果类似下图：

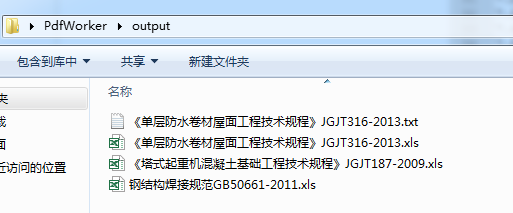
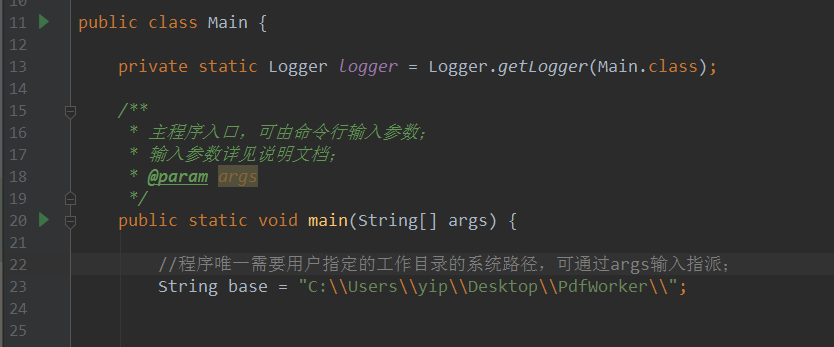


Figure 生成目录效果图

### 程序运行

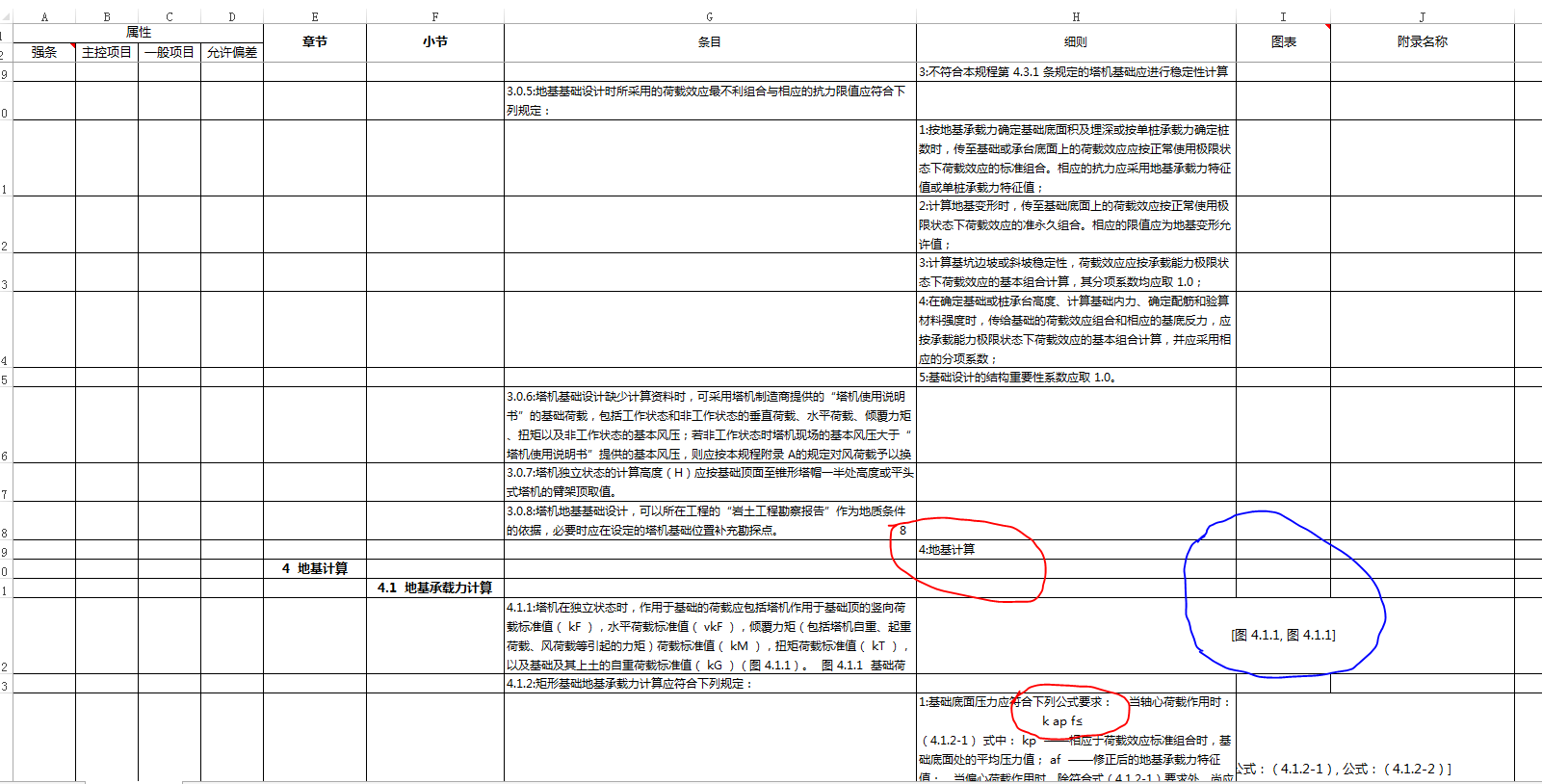
#### 参数输入



上图的base须有用户自己指派，不同的电脑上对应的路径也不一样，所以参数如果不对或者未加修改，则将抛出未知异常。

#### 程序输出

1. **正常输出：**



**请注意：**图表列提取出的信息格式不对，只需要在excel中调整格式即可，上面的红色方框文本由于程序bug，需要手动删除；下面的红色方框，由于pdfbox提取的文本里公式出现的乱码，程序无法处理；以及秋雅姐要求的一些值得优化的方面，由于时间的关系，也只做了部分优化。

1. **异常输出：**当output目录里如果生成了txt目录，那么生成的excel文档必定是有所缺失的，这时候，录入员需手动的对txt文档进行复制和粘贴操作，

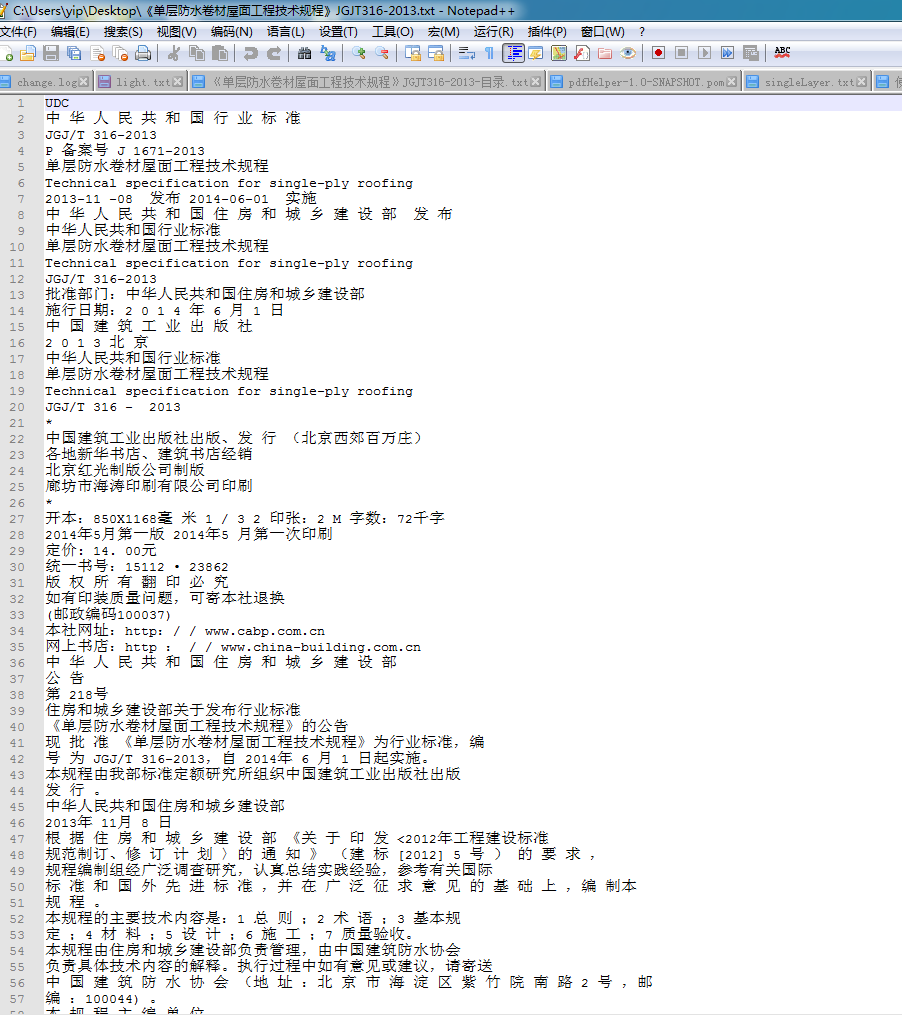
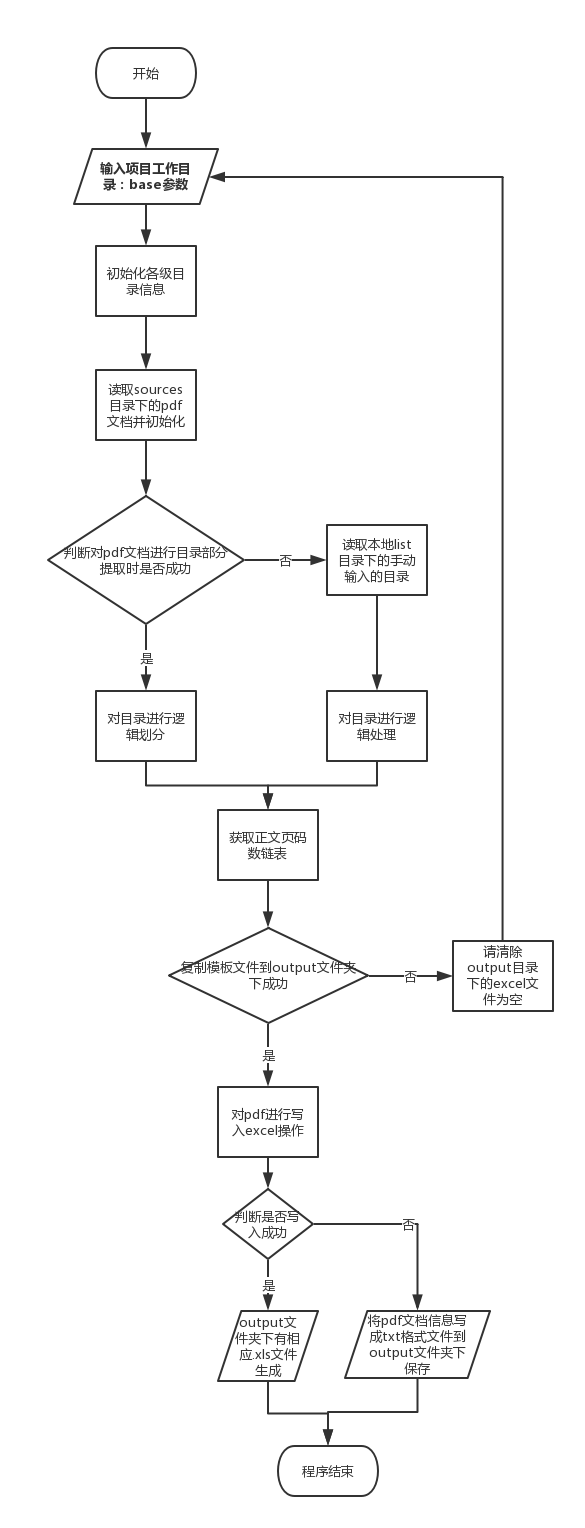


Figure 格式为txt文档的pdf文本内容

#### 逻辑说明



### 程序打包

* 本程序由于时间关系并未打包成可运行的jar包，麻烦后续交接者继续执行此项工作，同时base路径依然封装在main函数里，如需使用请自行修改。
* 本程序尚存在不少bug，而且参数耦合性依然很高，提取方法依然有不少难以做到通用，希望后续接替者可以继续进一步优化。
* 本程序在IDEA中运行，故最终会将整个项目压缩成压缩包形式发送给杰哥和二哥。
* 由于时间紧迫，本说明书难免会有遗漏的地方，如有不明白地方，[可以通过以下邮箱 2058197702@qq.com](mailto:可以通过以下邮箱%202058197702@qq.com)并发送相应邮件把问题反馈过去，笔者会及时回复的；也可找到通过阅读代码解决。
* 代码逻辑有不少混乱的地方，因为是第一次实习，希望海涵，谢谢！