重庆工商大学
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY
人工智能学院
（计算机科学与信息工程学院）

PDF 到 EPUB 电子书格式转换工具

· 详细设计文档 ·

专业	计算机科学与技术
开发成员	黄培, 李浩, 王子涵, 何先宇, 徐思杨
指导老师	朱超平
电话	18976758681
邮件	2398430768@qq.com
开发周期	2021.12.28-2022.4.1

目录

第一章.引言

1.1 编写目的

本报告的目的是对“PDF 到 EPUB 电子书格式转换工具”进行详细说明。以便用户及项目开发人员了解产品详细的设计与实现。为开发人员提供开发参考书。以下叙述将结合文字描述、伪代码，图表等来描述“PDF 到 EPUB 电子书格式转换工具”的详细设计和相关的模块描述。本报告的预期读者有服务外包大赛评委老师、开发人员以及跟该项目相关的其他竞争人员。

1.2 编写背景

目前，PDF 文档格式依旧作为印刷出版行业主流的电子文档保存格式，它的主要特点:可移植且可以保留任何原文档的字体、图像、图形和版面设置等;EPUB 文档格式逐渐成为新一代电子书保存格式，主要是因为它特有的可重排版特征，用户可以根据自己的设置，重新排版需要显示的内容。然而，在当今应用软件市场中，大部分的电子书阅读器软件仅仅提供对单一格式解析功能的支持,而在 PDF 到 EPUB 电子书格式转换功能方面的支持较少，如果企业的阅读软件能够做好这一功能的支持，不仅仅是在 PDF 格式到 EPUB 电子书格式之间建立起转换的桥梁，而且使得 PDF 文档具有 EPUB 文档可重排的特征，从而提升公司电子书阅读器的吸引力。

随着公司电子书阅读软件的不断升级，对于这两种电子书格式的解析

功能已经支持。但是，伴随着客户群体逐渐扩大，来自出版社相关的客户群体，建议公司应该在现有的阅读器中支持 PDF 到 EPUB 电子书格式转换的功能，原因在于，客户看中了 EPUB 格式可重排的特征,而现有的书籍资源长期使用 PDF 格式保存,若要制作 EPUB 版本的书籍，将增加客户的工作量，费时且费力。所以，这方面的客户群体，希望公司在技术上能够给予支持,实现自动转换的功能。解决客户的需求,提高用户的体验，是公司的职责所在。所以，PDF 到 EPUB 电子书格式转换的功能是时候在公司下一版本的电子书阅读器中提供支持。总之，研究与实现这一功能将是企业努力的方向。

研究 PDF 到 EPUB 电子书格式转换的意义在于:首先，进一步丰富应用市场中电子书阅读器在 PDF 到 EPUB 电子书格式转换方面的探索与实践，尤其是在资源排版对应方面解决方案的研究:其次，解决当今应用市场中电子书阅读器对 PDF 到 EPUB 电子书格式转换功能支持不完善的问题;最后,提高用户的工作效率和体验度,提高阅读器的实用性，是具有一定的经济价值的。

1.3 定义

- 所在文件:指相对于文件根目录的文件路径。
- 数据库：用来保存系统数据的后台应用软件。

1.4 参考资料

- 赤霓.ePub 指南——从入门到放弃(第 2 版)

- 张逸.PDF 到 EPUB 电子书格式转换工具的设计与实现
- Spire.PDF for Java 中文教程
- Apache PDFBox® - A Java PDF Library

1.4 支持环境

本系统软件运行环境包括：

- 操作系统:CentOS 7-2009
- 数据库:MySQL 5.7
- WebServer:tomcat(SpringBoot 内置)
- Java JDK:jdk 1.8.0_311

第二章.总体设计

2.1 需求概述

- 开发 PDF 文本格式到 EPUB 电子图书格式的自动转换工具，实现 PDF 到 EPUB 的批量转换服务开发完成的作品可以作为独立工具运行；也可作为服务运行，支持分布式任务调用
- 转换时需保留原有 PDF 文档的排版样式、标题格式和目录格式
- 转换后的文档支持保留原文件名和重新命名
- 转换过程有完整日志记录便于查看 转换完成进度

2.2 软件结构

2.2.1 总体流程图

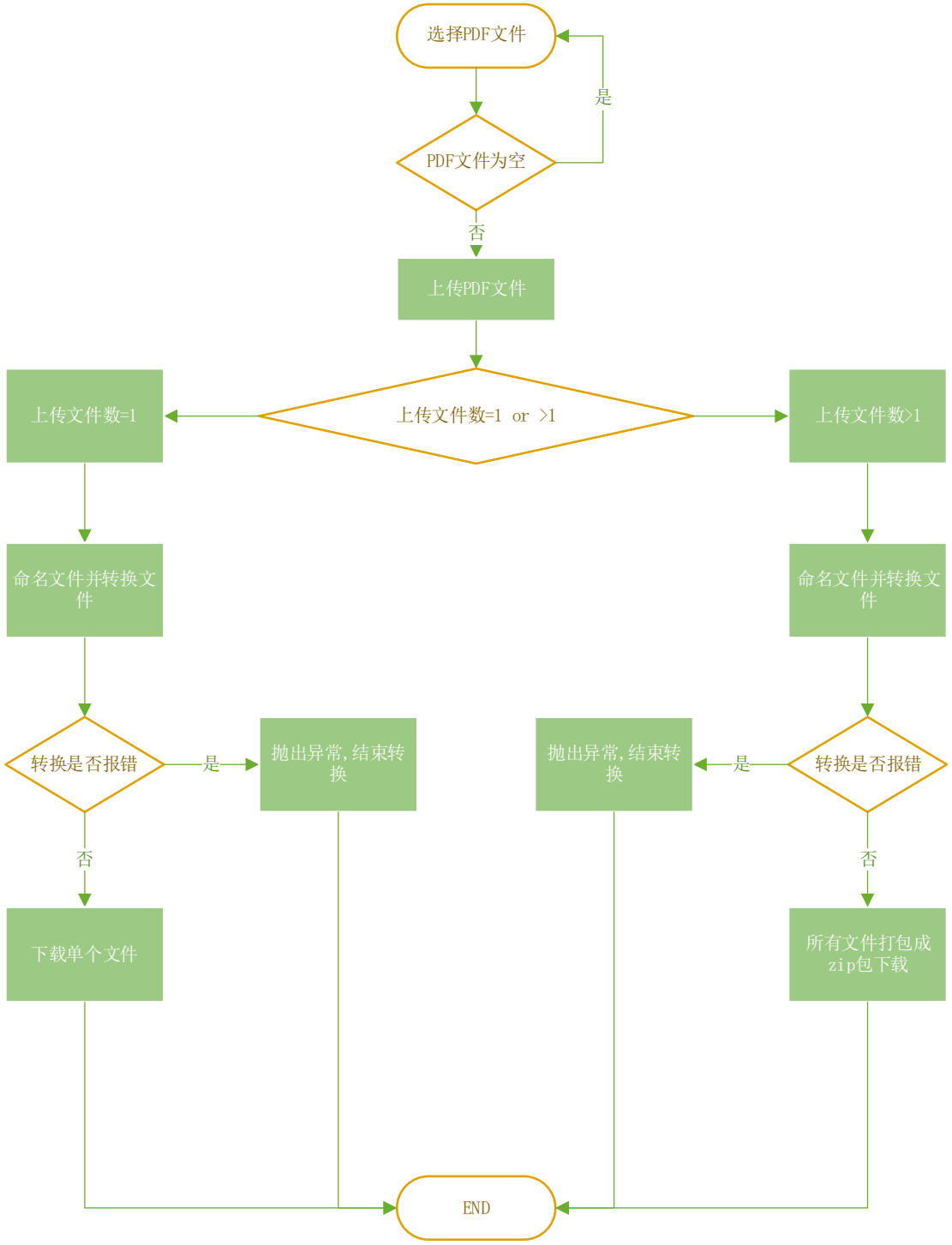

图 2.1 程序流程图

2.3.2 总体结构图

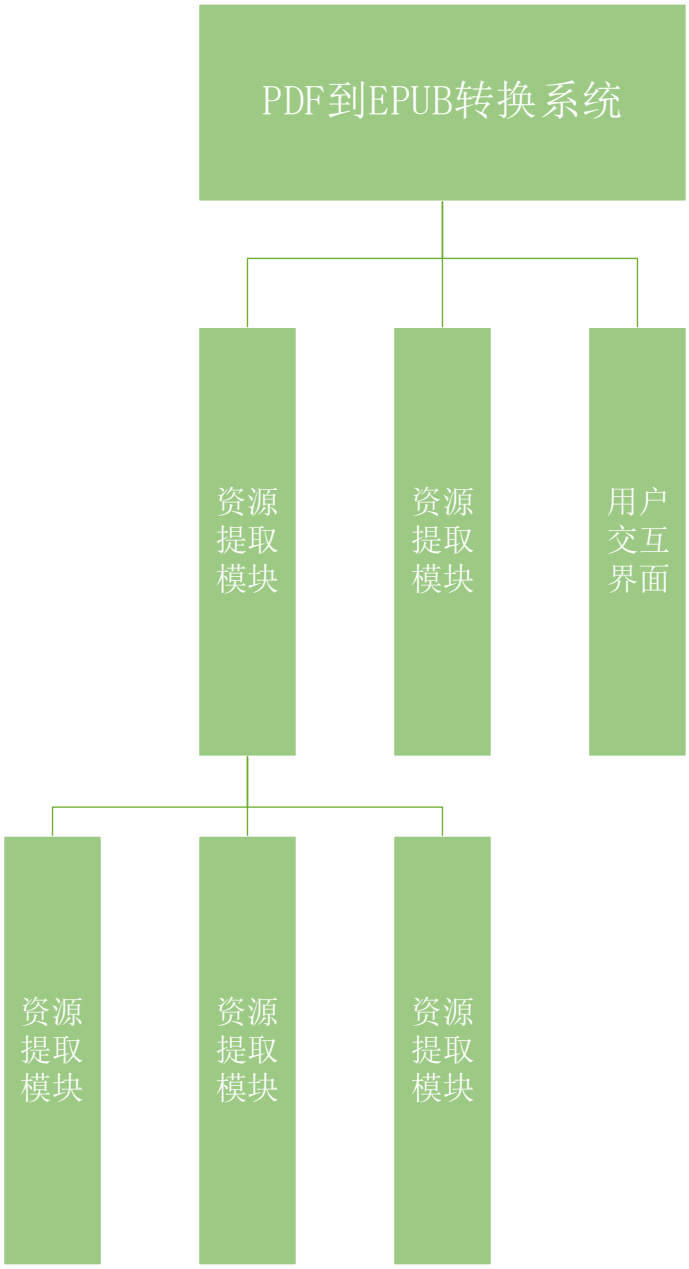

图 2.2 整体结构图

第三章.程序描述

3.1 模块基本信息

名称	编号	所在文件
图片信息提取模块	1.1	com/example/pdfconvertepub/convert/extract/ExtractPdf.java
文字信息提取模块	1.2	com/example/pdfconvertepub/convert/extract/PrintTextLocations.java
书签提取模块	1.3	com/example/pdfconvertepub/convert/extract/PrintBookmarks.java
资源写入打包模块	2.1	com/example/pdfconvertepub/convert/epub/EpubCreatorText.java
用户交互界面	2.3	templates/epub.html

表 3.1 模块基本信息表

3.2 功能概述

3.2.1 图片信息提取模块功能概述

1. **模块描述**
   该模块的功能主要是从 PDF 提取出图片的相关信息(图片左上角坐标，图片长宽等)并写入 txt 文件中保存。
2. **输入、输出**
   输入:PDF 文档文件
   输出:图片相关信息 txt

3.2.2 文字信息提取模块功能概述

1. **模块描述**
   该模块的功能主要是从 PDF 提取出文字的相关信息(文字坐标，大小，高宽等)并写入 txt 文件中保存。
2. **输入、输出**
   输入:PDF 文档文件
   输出:文字相关信息 txt

3.2.3 书签提取模块功能概述

1. **模块描述**
   该模块的功能主要是从 PDF 提取出目录的相关信息(目录名和对应页码)并写入 txt 文件中保存。
2. **输入、输出**
   输入:PDF 文档文件
   输出:目录相关信息 txt

3.2.4 资源写入打包模块功能概述

1. **模块描述**
   该模块的功能是将所有提取出的信息整合并写入 EPUB 模板中，再用压缩算法打包成 EPUB 文件。
2. **输入、输出**
   输入:相关信息 txt
   输出:EPUB 文档文件

3.2.5 用户交互界面概述

1. **模块描述**
   该模块的功能是连接后端功能，给用户提供一个更直观的转换过程。
2. **输入、输出**
   输入:单个 PDF 或多个 PDF
   输出:单个 EPUB 文档(单个 PDF)或 ZIP 包(多个 PDF)

3.3 模块处理逻辑

3.3.1 图片信息提取模块处理逻辑

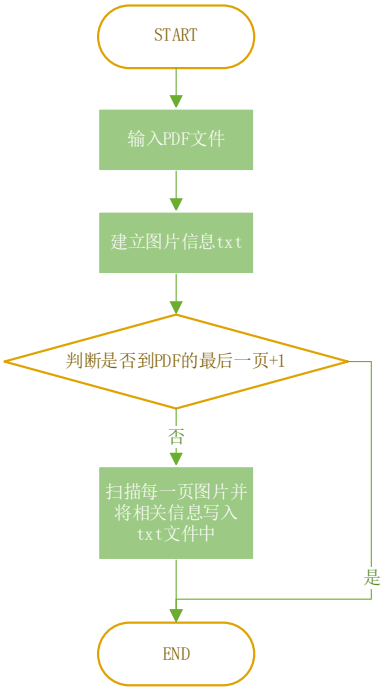

图 3.1 图片信息提取模块流程图

3.3.2 文字信息提取模块处理逻辑

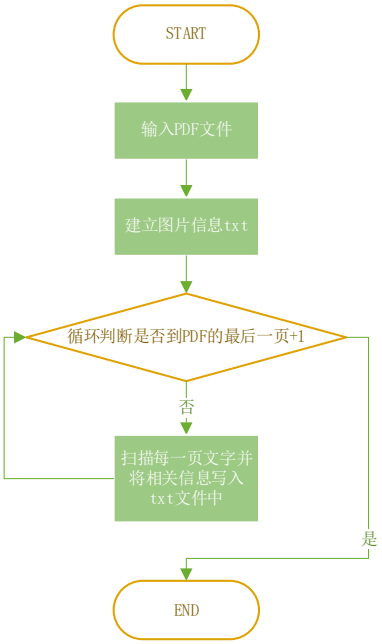

图 3.2 文字信息提取模块流程图

3.3.3 书签提取模块处理逻辑

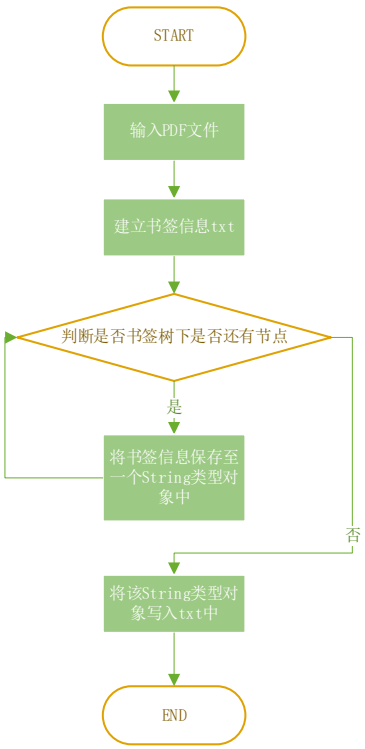

图 3.3 书签信息提取模块流程图

3.3.4 资源写入打包模块处理逻辑

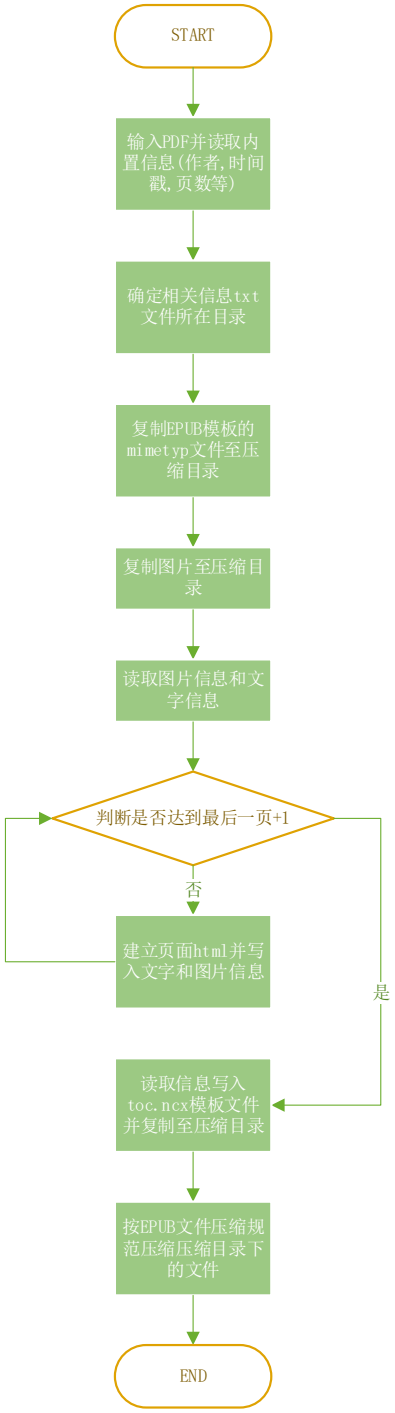

图 3.4 资源写入打包模块流程图

3.3.5 用户交互界面处理逻辑

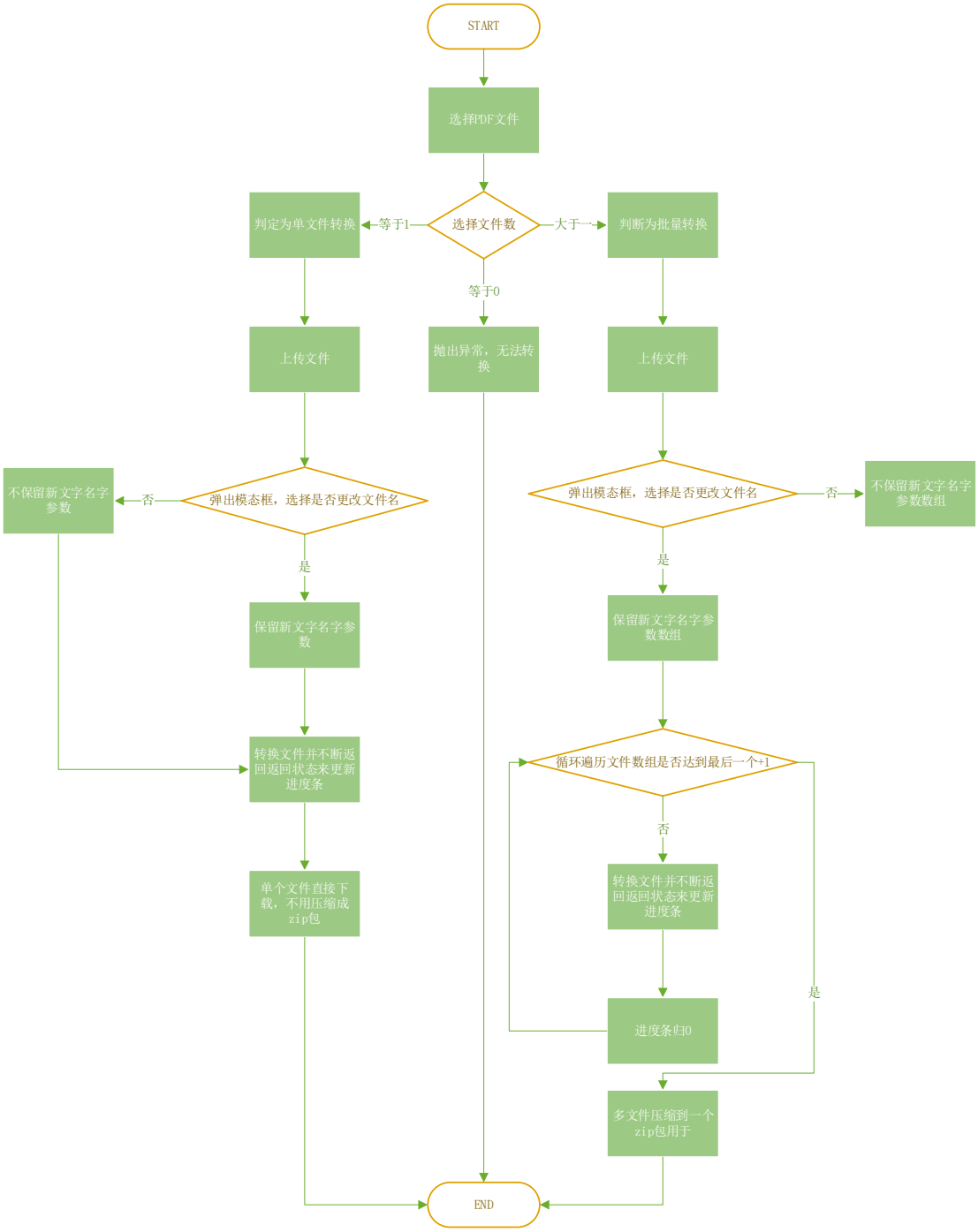

图 3.4 用户界面处理逻辑

3.4 数据库设计

3.4.1 标识符和状态

数据库软件的名称：Mysql 5.7

数据库的名称：labtest

表名	标识符或名称	描述信息	状态
错误表	错误 id	用来保存程序运行时产生的错误信息	使用

表 3.1 标识符和状态表

3.4.2 使用它的模块

模块名称	访问的数据表
图片信息提取模块算法	错误表
文字信息提取模块	错误表
书签提取模块	错误表
资源写入打包模块	错误表
用户交互界面	错误表

表 3.1 模块使用表

3.4.3 逻辑结构设计

字段名	数据类型	长度	主键	非空	描述
id	bigint	100	是	是	主键 id

call_name	varchar	255			名称
call_time	timestamp	0			调用时间
call_function_full_name	varchar	255			调用方法名称
exception_message	longtext	0			异常信息

3.5 测试计划

模块名称	输入数据	预期结果
图片信息提取模块	带有图片的 PDF 文档	输出图片相关信息 txt
文字信息提取模块	带有文字的 PDF 文档	输出每一页文字的坐标的 txt 集合
书签提取模块	带有目录的 PDF 文档	输出目录信息 txt
资源写入打包模块	资源目录	EPUB 文件
用户交互界面	单个 PDF 文档	单个 EPUB 文件
图片信息提取模块	不带有图片的 PDF 文档	输出图片相关信息 txt 但内容为空
文字信息提取模块	不带有文字的 PDF 文档	输出文字相关信息 txt 集合但所有内容为空
书签提取模块	不带有目录的 PDF 文档	输出目录信息 txt 但内容为空
资源写入打包模块	资源目录为空	报错并抛出异常
用户交互界面	多个 PDF 文档	当个 ZIP 打包文件
用户交互界面	空	无法上传和转换

第四章.项目改进

4.1 OCR 技术(开发中)

4.1.1 技术介绍

此项目采用的 OCR 技术是百度的开源项目 PaddleOCR，PaddleOCR 基于深度学习技术实现的，使用该项技术，可以将扫描版本的 PDF 文字识别并提取出来，然后转换成 epub，提高用户的阅读体验。同时该项目具有很优秀的识别性能：支持多语言识别，目前能够支持 80 多种语言；除了能对中文、英语、数字识别之外，还能应对字体倾斜、文本中含有小数点字符等复杂情况。因此对于版本较老的扫描版 pdf，都提供了较好的转换效果。

以下为 PaddleOCR 的识别效果：

图 4.1 PaddleOCR 对于英文文档的识别效果

图 4.1 PaddleOCR 对于中文文档的识别效果

4.1.3 ocr 技术部署方案

PaddleOCR 在服务端采用 Paddle Serving 进行部署，同时提供了 API 以供调用，所以此项目的 ocr 识别模块和文档转换模块是分离的，大致关系如下图所示：