1. doc/docx/wps/wpt/docm/dotx
2. doc/wps/wpt
3. 文本提取

使用到的库：

import org.apache.poi.hwpf.HWPFDocument;

import org.apache.poi.hwpf.extractor.WordExtractor;

具体流程：

WordExtractor wordExtractor = new WordExtractor(bis);

//bis是文件输入流，这里是定义一个文本提取器

String[]paragraphs=wordExtractor.getParagraphText();

//wordExtractor获取每个段落的文本信息

for(int i=0;i< paragraphs.length;i++){

result=paragraphs[i]+"```";

bos.write(result.getBytes());

}//循环将每一个段落中的文本写入到文本输出流中

1. 图像提取

使用到的库:

import org.apache.poi.hwpf.model.PicturesTable;

import org.apache.poi.hwpf.usermodel.Picture;

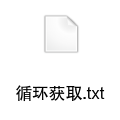
具体流程:

HWPFDocument doc = new HWPFDocument(fis);//读取文档

PicturesTable picturesTable = doc.getPicturesTable();

List<Picture> pictures = picturesTable.getAllPictures();

/\*这个方法返回一个XWPFPictureData类的集合，每个 XWPFPictureData对象代表文档中的一个图片，并包含了图片的二进制 数据和相关属性.getAllPictures 方法返回一个包含所有图片数据的列 表，然后可以遍历这个列表来处理每个图片。\*/



picture.getContent()的作用

这个方法通常用于获取图片的内容，即图片的二进制数据

1. docx/docm/dotx
2. 文本提取

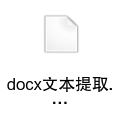
使用到打的库：

import org.apache.poi.xwpf.usermodel.XWPFDocument;

import org.apache.poi.xwpf.usermodel.XWPFParagraph;

具体流程：

//与doc的提取流程相同



1. 图像提取

