JAVA读取WORD,EXCEL,POWERPOINT,PDF文件的方法Java认证考试 PDF转换可能丢失图片或格式，建议阅读原文 https://www.100test.com/kao_ti2020/564/2021_2022_JAVA_E8_AF_BB_E5_8F_96_c104_564748.htm 百考试题编辑整理：JAVA读取WORD,EXCEL,POWERPOINT,PDF文件的方法 OFFICE文档使用POI控件，PDF可以使用PDFBOX0.7.3控件，完全支持中文，用XPDF也行，不过感觉PDFBOX比较好，而且作者也在更新。水平有限，万望各位指正 WORD: import org.apache.lucene.document.Document. import org.apache.lucene.document.Field. import org.apache.poi.hwpf.extractor.WordExtractor. import java.io.File. import java.io.InputStream. import java.io.FileInputStream. import com.search.code.Index. public Document getDocument(Index index, String url, String title, InputStream is) throws DocCenterException { String bodyText = null. try { WordExtractor ex = new WordExtractor(is).//is是WORD文件的InputStream bodyText = ex.getText(). if(!bodyText.equals("")){ index.AddIndex(url, title, bodyText). } }catch (DocCenterException e) { throw new DocCenterException("无法从该Mocriosoft Word文档中提取内容", e). }catch(Exception e){ e.printStackTrace(). } } return null. } Excel: import org.apache.lucene.document.Document. import org.apache.lucene.document.Field. import org.apache.poi.hwpf.extractor.WordExtractor. import org.apache.poi.hssf.usermodel.HSSFWorkbook. import org.apache.poi.hssf.usermodel.HSSFSheet. import

org.apache.poi.hssf.usermodel.HSSFRow. import org.apache.poi.hssf.usermodel.HSSFCell. import java.io.File. import java.io.InputStream. import java.io.FileInputStream. import com.search.code.Index. public Document getDocument(Index index, String url, String title, InputStream is) throws DocCenterException { StringBuffer content = new StringBuffer(). try{ 100Test 下载频道开通，各类考试题目直接下载。详细请访问 www.100test.com