

ExtractText

pdftotext, java pdftotext, pdfbox, pdf to text

Table of contents

1 Description.....	2
--------------------	---

1. Description

This application will extract all text from the given PDF document.

usage: java org.apache.pdfbox.ExtractText [OPTIONS] <PDF file> [Text file]

Command Line Parameter	Type	Default Value	Description
-password <password>	string	None	The password to the PDF document.
-encoding <output encoding>	string	default encoding	The encoding type of the text file, e.g. ISO-8859-1, UTF-8, UTF-16BE.
-console	boolean	false	Send text to console instead of file.
-html	boolean	false	Output in HTML format instead of raw text.
-sort	boolean	false	Sort the text before writing
-ignoreBeads	boolean	false	Disables the separation by beads.
-force	boolean	false	Enables pdfbox to ignore corrupt objects.
-startPage <start page>	integer	1	The first page to extract, one based.
-endPage <end page>	integer	Integer.MAX_INT	The last page to extract, one based.