Frequently Asked Questions

Table of contents

1. General Questions.....2

1.1. When will the next version of PDFBox be released? 2

1.2. I am getting the below Log4J warning message, how do I remove it? 2

1.3. Is PDFBox thread safe? 2

1.4. Why do I get a "Warning: You did not close the PDF Document"? 2

2. Text Extraction.....3

2.1. How come I am not getting any text from the PDF document? 3

2.2. How come I am getting gibberish(G38G43G36G51G5) when extracting text? 3

2.3. What does "java.io.IOException: Can't handle font width" mean? 3

2.4. Why do I get "You do not have permission to extract text" on some documents? 3

2.5. Can't we just extract the text without parsing the whole document or extract text as it is parsed..... 3

1. General Questions

1.1. When will the next version of PDFBox be released?

As fixes are made and integrated into the repository these changes are documented in the release notes. An estimate will be given of when the next version will be released.
Of course, this is only an estimate and could change.

1.2. I am getting the below Log4J warning message, how do I remove it?

```
log4j:WARN No appenders could be found for logger (org.pdfbox.util.ResourceLoader).
log4j:WARN Please initialize the log4j system properly.
```

This message means that you need to configure the log4j logging system. See the log4j documentation for more information.

PDFBox comes with a sample log4j configuration file. To use it you set a system property like this

java -Dlog4j.configuration=log4j.xml org.pdfbox.ExtractText <PDF-file> <output-text-file>

If this is not working for you then you may have to specify the log4j config file using a URL path, like this:
log4j.configuration=file:///<path to config file>
Please see this forum thread for more information.

1.3. Is PDFBox thread safe?

No! Only one thread may access a single document at a time. You can have multiple threads each accessing their own PDDocument object.

1.4. Why do I get a "Warning: You did not close the PDF Document"?

You need to call close() on the PDDocument inside the finally block, if you don't then the document will not be closed properly. Also, you must close all PDDocument objects that get created. The following code creates two PDDocument objects; one from the "new PDDocument()" and the second by the load method.

PDDocument doc = new PDDocument(); try { doc = PDDocument.load("my.pdf"); } finally { if(doc != null) { doc.close(); } }

2. Text Extraction

2.1. How come I am not getting any text from the PDF document?

Text extraction from a pdf document is a complicated task and there are many factors involved that effect the possibility and accuracy of text extraction. It would be helpful to the PDFBox team if you could try a couple things.

- Open the PDF in Acrobat and try to extract text from there. If Acrobat can extract text then PDFBox should be able to as well and it is a bug if it cannot. If Acrobat cannot extract text then PDFBox 'probably' cannot either.
- It might really be an image instead of text. Some PDF documents are just images that have been scanned in. You can tell by using the selection tool in Acrobat, if you can't select any text then it is probably an image.

2.2. How come I am getting gibberish(G38G43G36G51G5) when extracting text?

This is because the characters in a PDF document can use a custom encoding instead of unicode or ASCII. When you see gibberish text then it probably means that a meaningless internal encoding is being used. The only way to access the text is to use OCR. This may be a future enhancement.

2.3. What does "java.io.IOException: Can't handle font width" mean?

This probably means that the "Resources" directory is not in your classpath. The Resources directory is included in the PDFBox jar so this is only a problem if you are building PDFBox yourself and not using the binary.

2.4. Why do I get "You do not have permission to extract text" on some documents?

PDF documents have certain security permissions that can be applied to them and two passwords associated with them, a user password and a master password. If the "cannot extract text" permission bit is set then you need to decrypt the document with the master password in order to extract the text.

2.5. Can't we just extract the text without parsing the whole document or extract text as it is parsed.

Not really, for a couple reasons.

1. If the document is encrypted then you need to parse at least until the encryption dictionary before you can decrypt.
2. Sometimes the PDFont contains vital information needed for text extraction.

3. Text on a page does not have to be drawn in reading order. For example; if the page said "Hello World", the pdf could have been written such that "World" gets drawn and then the cursor moves to the left and the word "Hello" is drawn.