PDFBox - PDF Highlighting

Java PDF Library, highlight, highlight pdf, highlight pdf text, java

Table of contents

1. Highlighting text in a PDF

There are cases when you might want to highlight text in a PDF document. For example, if the PDF is the result of a search request you might want to highlight the word in the resulting PDF document. There are several ways this can be achieved, each method varying in complexity and flexibility.

1.1. 1. Use the 'search' open parameter

Acrobat supports passing is various parameters that tell it what to do once the PDF is open. See PDF Open Parameters for documentation on all the open parameters. One of the parameters is the 'search' parameter, this will automatically run the search functionality inside of Acrobat once the PDF is open. For example:
http://www.pdfbox.org/userguide/text_extraction.pdf#search="check"

> **Note:**
> The words must be enclosed in quotes and separated by spaces; for example:#search="pdfbox rocks"

This is a great solution because of its simplicity! It doesn't require PDFBox at all, but it is a potential solution that many developers are not aware of.

1.2. 2. Generate a highlight XML document

Acrobat also allows you to tell it to highlight specific words in the PDF document. It does this by passing an XML document to Acrobat when opening the PDF. See the PDF Highlight File Format for more detailed documentation.
Basically the document allows you to tell it the characters to highlight in the PDF by using character offsets on a page. As this is just an XML document, there are many ways you could create it but PDFBox does have a utility to make it easier. Take a look at the javadoc for the PDFHighlighter class. This will allow you specify a set of words that you want have highlighted and generate the XML document for you.
PDFBox also ships with a complete web application example of using this class, take a look at the pdfbox.war directory in your PDFBox installation.
You pass the xml to acrobat through a URL (or command line) parameter like this:
http://www.pdfbox.org/userguide/text_extraction.pdf#xml=http://www.pdfbox.org/highlight.xml

> **Note:**
> The value of the xml parameter must be a full URL to the XML document.
> http://www.pdfbox.org/userguide/text_extraction.pdf#xml=highlight.xml will not work
> http://www.pdfbox.org/userguide/text_extraction.pdf#xml=http://www.pdfbox.org/highlight.xml is correct!

The one drawback to this solution is that you must parse the PDF and then generate an XML document, which is a time consuming operation.

1.3. 3. Alter pdf contents to highlight specific text

Using PDFBox it is possible to regenerate the appearance stream to add highlighting to specific areas. While this is possible, it will require recreating a new PDF for every search request. There is nothing prebuilt in PDFBox to do this automatically for you and will require a significant coding effort.

You would need to

1. Find all locations of the text, determine x/y coordinates, width/height
2. Regenerate the PDF appearance stream and draw a highlighted box behind the text. Yellow would be easiest, if you want an inverted black/white, then you would need to change the color of the text to be white and draw a black box.
3. Stream the PDF back to the user

This is the most flexible but is also the most work to implement and is also more resource intensive.