How to read all the text from pdf document using PDFBox 2.0

In this PDFBox Tutorial, we shall learn to read all the text from pdf document using PDFBox 2.0 libraries in a Java Program.

Read all the text from pdf document using PDFBox 2.0

PDF document may contain text, embedded images etc., as its contents. PDFTextStripper class in PDFBox provides functions to extract all the text from PDF document. Following are the steps that are helpful in extracting the text from pdf :

Step 1 : Load PDF

Load the pdf file into PDDocument

```
PDDocument doc =
```

```
PDDocument doc = PDDocument.load(new File("sample.pdf"));
```

Step 2 : Use PDFTextStripper.getText method

Get the text from doc using PDFTextStripper

```
String text = new
```

```
String text = new PDFTextStripper().getText(doc);
```

PDFTextStripper ignores formatting and placement of text chunks in the pdf document. PDFTextStripper just strips out all the text from all the pages of pdf document. getText returns the text of the pdf document.

Complete Java Program

ExtractText.java program to extract all the text from PDF document

```
import java.io.File;
```

```
import java.io.File;
import java.io.IOException;

import org.apache.pdfbox.pdmodel.PDDocument;
import org.apache.pdfbox.text.PDFTextStripper;

public class ExtractText {

   public static void main(String[] args) {
      try {
         PDDocument doc = PDDocument.load(new File("sample.pdf"));
         String text = new PDFTextStripper().getText(doc);;
         System.out.println("Text in PDF\n--------------------------------");
         System.out.println(text);
         } catch (IOException e) {
         e.printStackTrace();
      }
   }
}
```

Output of ExtractText.java

```
Text in PDF
-------------------
```

```
Text in PDF
------------------
This is a sample PDF.
```

And pdf file used in the example is ? sample.pdf

Reference :

You may find more information about PDFTextStripper class in the java documentation of PDFTextStripper class, visit ? here[https://pdfbox.apache.org/docs/2.0.5/javadocs/org/apache/pdfbox/text/PDFTextStripper.html].

Conclusion :

We have learnt to read all the text from pdf document using PDFBox 2.0.