https://github.com/huilam/hl-doc-extractor-pdf

hl-doc-extractor-pdf

A simple lightweight (250 KB on disk) Java PDF extractor based on PDFBox.

Sample Code:

```
//## Initialize PDF Extract with a file
PDFExtractor extractor = new PDFExtractor(new File("file.pdf"));
extractor.setStartPageNo(0);
extractor.setEndPageNo(0);

//## Extract all selected pages
ExtractedContent data = extractor.extractAll();

//## Export to JSON
JSONObject jsonData = data.toJsonFormat(true); //## true to include image base64

//## Export to Plain Text & Images
JSONObject jsonData = data.toPlainTextFormat(true); //## true to indicate page number
Map<String,BufferedImage> mapImages = data.getExtractedBufferedImages(); //## <FileName, BufferedImage>
```