Optical Character Recognition for Apache PDFBox

Short Description

Apache PDFBox is a widely used library to extract text from PDF files, manipulate existing PDF files and create new PDF files. The current functionality of text extraction is done by fetching data stream of the particular PDF file and passing through several filters and algorithms of PFDBox. This method does not work in situations where PDF file has malformed character encoding and text embedded in images. PDFBox currently have tools to create images from a given input PDF file. In this GSoC project, a new approach is proposed to extract text from a PDF file through Optical Character Recognition by using the image generated using PDF box.

Proposal Title : Optical Character Recognition for Apache PDFBox
Student Name: Dimuthu Upeksha
Student Email : dimuthu.upeksha2@gmail.com
JIRA Issue : https://issues.apache.org/jira/browse/PDFBOX-1912

Deliverables

- This improves the acceptance of PDFBox among community because there are lots of PDF files that has scanned images which can not be extracted using normal text extracting libraries. Specially Governmental documents
- Accuracy of the text extraction can be improved.
- Provide the flexibility to users to choose best suiting method (OCR or normal method) or both in their applications.
- Even documents with wrong character encoding can be used to extract its text using OCR algorithms because OCR doesn't depend on encoding of the characters but the shape of characters.

Detailed Description

Apache PDFBox currently can extract text by analyzing it's COS model. A PDF consists of a COS model which includes different variable types like Boolean, Number, String and etc. Current text extraction algorithm go through this model and fetch text + location data. Coming up with an accurate text extraction is very complex to automate because model of PDF files are defined not in machine readable manner. It's structured in human readable way. So it needs a lot of processing to do a correct text extraction considering placement of each word and character of the PDF file. What PDFBox does is mapping whose word + location data into meaningful output text. But this method does not work in some use cases.