International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 06 | June 2018 www.irjet.net p-ISSN: 2395-0072

Study on Libraries for Text Extraction from PDF Document

Utkarsh Rastogi[1]

[1]*P.G Student, Dept. of MCA, VESIT, Mumbai, Maharashtra, India*

Abstract - *Documents in PDF format are called Universal Document Format. Text extraction is the primary step from PDF to do further processing of text, which we want to use in other place. In this paper we will see the different available libraries for text extraction. The aim of this paper is to give overview of libraries and comparative study between them.*

***Key Words*: PDF, Text Extraction, Text Extraction Libraries**

1. INTRODUCTION

PDF stands for portable document format. It was introduced by Adobe Systems in 1993. It has the following features [1][2][3]

- It is file format for representation of document.
- Independent of the application software, hardware and operating system.
- We can store wide range of data as PDF file.
- It supports various data formats like text, graphics, and images.
- It works in similar way with any operating system.

Retrieving text from PDF is a huge issue especially when you have to build a report or analysis. As the time has passed, the research has also gradually progressed. The problems that arises while retrieving a data is related to the font, display layout and tabular display.

2. TEXT EXTRACTION

We can Store data in PDF as String Object (Object is as a sequence of literal characters). Each object will coordinate with Font and characters for representing into visual characters.

There are few advantages to use PDF document

- PDF contains wide range of data
- Data is in rich and more clearer format
- Accuracy of data compare to acquired from OCR

There are challenges as well in terms of extracting data from PDF, reason behind is Tools, which are different in terms of object, font size, alignment, Gap between text etc.

Overall we can consider PDF documentation is secure way to keeping data in secure form. As said earlier extracting data from PDF is bit challenging, here are few libraries which are available for it.

3. AVAILABLE LIBRARIES FOR TEXT EXTRACTION

Various libraries are available for text extraction under different technology stack. Few common libraries are listed below:

1) Apache PDFBox® - A Java PDF Library
2) iText
3) PDFMiner
4) PDF.js
5) PDFxStream(PDFTextSream)

3.1 Apache PDFBox - A Java PDF Library

This library is an open source java tool that can be used with PDF documents. It was introduced in year 2002[4].

Features of PDFBox

- Data Extraction
- Large PDF can be subdivided into smaller PDF
- Formation of new PDF documents.
- PDF can store in format of Image.
- Multiple PDF documents can be merged into a single PDF document.

PDFBox is an open source library which is available free for use.

3.2 iText

This library is made available to use under technology stack of Java and .NET. It is recognized as iTextSharp, in .NET Framework where it is written in C# language.

Features of iText

- Text can be extracted
- Formation of new PDF documents
- Filling PDF Forms
- PDF can store in format of Image.

- Help in adding other contents in PDF like bookmarks, page numbers
- We can draw shapes in PDF document as like drawing book

Its code is open source but it is distributed under the AGPL License and can work freely only for applications which come under that license. Otherwise for commercial purpose we require a license to use it.

3.3 PDFxStream

PDFTextStream (part of PDFxStream v3.x) is available under the technology stack of Java and .NET Framework [5].

Features of PDFxStream

- It can be used for the decryption of data in PDF document
- Supports Bookmarking
- Retrieving of Image Data which includes image format.
- Multiple sub files can be combined into single large file.
- We can retain Meta data of any document which includes key value pairs or XML.
- Retrieving of raw data.

PDFxStream library is restricted around for 500 PDFs so for commercial use we have to buy the authorized license.

3.4 PDF.js

This is a JavaScript library which allows the conversion of PDF document data so that can be viewed into web browser in the form of HTML [6].

Features of PDF.js

- Text Extraction
- Larger PDF document can be splitted into many smaller documents.
- Different Documents can be merged into one.
- Creation of New Documents as PDF.
- Pdf.js is available for free to use.

3.5 PDFMiner

This library is available under the technology Stack of Python and different from others in a sense that it focuses on getting and analyzing text data from PDF document. Through this library user can get the other related information like exact location of text, information about font used in that document. Even user can convert PDF into other formats like HTML [7].

Features of PDFMiner

- Helps in analyze and conversion of PDF document.
- It gives feature of transformation from PDF to HTML.
- It provides Chinese, Japanese, and Korean languages and vertical writing script support.
- It gives the Strength for various font types (Type1, TrueType, Type3, and CID).
- Rebuild the original layout by grouping text chunks.
- It can be used for pulling out Table of Contents from PDF document.
- It provides support for PDF-1.7 specification.

4. COMPARISON BETWEEN PDFBOX, ITEXT AND PDFxSTREAM LIBRARIES

Difference between PDFBox and iText is that PDFBox always processes text glyph by glyph while iText normally processes it chunk by chunk. This feature in iText reduces the use of resources compare to PDFBox. iText architecture is event oriented. Therefore text parsing is much easier than compare to PDFBox. Thus reducing the demand for resources [8].

Tabular Content

Roll No	Name	Marks
15	XYZ	85
25	PQR	70

Fig -1: Actual PDF Contains Data in Tabular Content

Both Apache PDFBox and iText do not hold any textual content layout while retrieving textual content from PDF. The outcome of this is the retrieve data using these two libraries have inconsistent spacing between tabular data. This could distort the readable tabular data with empty cells as shown in Fig -2, Fig -3 and Fig -4.

```
******** Data Obtained After Extraction Uisng Snowtide PDFxStream ********

Tabular Content
Roll No     Name     Marks
15     XYZ     85
25     PQR     70
```

Fig -2: Data Extracted From PDF Using PDFTextStream

```
******** Data Obtained After Extraction Using PDFBox ********

Tabular Content
Roll No Name Marks
15 XYZ 85
25 PQR 70
```

Fig -3: Data Extracted From PDF Using PDFBox

```
******** Data Obtained After Extraction Using iText ********

Tabular Content
Roll No Name Marks
15 XYZ 85
25 PQR 70
```

Fig 4: Data Extracted From PDF Using iText

5. CONCLUSIONS

This paper simply shows the list of available libraries for text extraction from PDF document. It shows none of the libraries strictly subsumes another. There are many other paid libraries which we can use. Enhancements are still going on in each library by launching new versions. We expect the researchers to do more advancement in each library for text extraction that make extracting the text regions smooth in complex document.

REFERENCES

[1] https://www.w3.org/TR/WCAG-TECHS/pdf.html

[2] https://techterms.com/definition/pdf

[3] https://www.webopedia.com/TERM/P/PDF.html

[4] https://pdfbox.apache.org/

[5] https://www.snowtide.com/

[6] https://en.wikipedia.org/wiki/PDF.js

[7] https://media.readthedocs.org/pdf/pdfminer-docs/latest/pdfminer-docs.pdf

[8] https://stackoverflow.com/questions/22340674/performance-itext-vs-pdfbox