SPL-1 Project Report, 2019

PDF FILE CONVERTER

SE 305: Software Project Lab 1

Submitted by

Md Sabbir Hossain

BSSE Roll No.:1014

BSSE Session: 2017-18

Supervised by

Md. Mahbubul Alam Joarder

Designation: Professor

Institute of Information Technology

Institute of Information Technology

University of Dhaka

[29-05-2019]

Table of Contents

1. Introduction

Portable Document Format (PDF) is a file format used to present and exchange documents reliably, independent of software, hardware, or operating system. Invented by Adobe, PDF is now an open standard maintained by the International Organization for Standardization (ISO). PDFs can contain links and buttons, form fields, audio, video, and business logic.

In 1991, Adobe cofounder Dr. John Warnock launched the paper-to-digital revolution with an idea he called The Camelot Project. The goal was to enable anyone to capture documents from any application, send electronic versions of these documents anywhere, and view and print them on any machine. By 1992, Camelot had developed into PDF. Today, it is the format trusted by businesses around the world.

1.1. Background Study

In This journey with PDF file I need to study what PDF is. I need to study about the PDF specification to understand the nature of PDF file. This is a vast area. I tried hard but I can understand a little part of PDF file specification like collecting a stone from sea beach . Below I will mention about my background study.

Magic Number:

In computer science 'Magic Number' is the signature of file by which we can identify a file. PDF file has a specific magic number by which we can test a file PDF or not. The magic number of PDF file is

%PDF- = 25 50 44 46 2d

Here the value of right hand side is Hex value converted from ascii. And the value will associated in header secsion of PDF file.

A PDF file has four part. They are

- Header
- Body
- 'xref' Table
- Trailer

Header:

In header PDF carry file signature or magic number and the version number of PDF file.

The header of PDF looks like %PDF-1.

Body:

In the Body the PDF document carry all the object. There are 8 types of object. They are

- Boolean
- Integer and Real Number
- String
- Name
- Arrays
- Dictionaries
- Streams
- The null object

XREF Table:

This is the cross reference table, which contains the references to all the objects in the document. The purpose of a cross reference table is that it allows random access to objects in the file, so we don't need to read the whole PDF document to locate the particular object. Each object is represented by one entry in the cross reference table, which is always 20 bytes long. Let's show an example

```
xref
0 7
0000000000 65535 f
0000000009 00000 n
0000000074 00000 n
0000000120 00000 n
0000000179 00000 n
0000000300 00000 n
0000001532 00000 n
```

We can display the cross reference table of the PDF document by simply opening the PDF with a text editor and scrolling to the bottom of the document. In the example above, we can see that we have four subsections (note the four lines that only contain two numbers). The first number in those lines corresponds to the object number, while the second line states the number of objects in the current subsection. Each object is represented by one entry, which is 20 bytes long (including the CRLF). The first 10 bytes are the object's offset from the start of the PDF document to the beginning of that object. What follows is a space separator with another number specifying the object's generation number. After that there is another space separator followed by a letter 'f' or 'n' to indicate whether the object is free or in use.

Trailer:

The PDF trailer specifies how the application reading the PDF document should find the cross reference table and other special objects. All PDF readers should start reading a PDF from the end of the file. An example trailer is presented below:

```
trailer
   << /Size 7
      /Root 1 0 R
   >>
startxref
1556
%%EOF
```

1.2. Challenges

During the long journey I have faced many challenge. They are

Understanding the PDF file format

PDF file is a formatted file. That's why The main challenge of the project is understanding the PDF file specification. It is a huge area . How xref table work, how trailer section work ,how text are situated in PDF file, how graphics object are stay in PDF file .

Convert the file in doc file

Microsoft Document file is also a formatted file. So, if I want to create a doc file I also need to understand about document file format. But my level of knowledge is not enough to understand the file format within the short time. Understanding document file format is so long process. I will need minimum 1.5 years to correctly understand about document file format.

Almost all the PDF file are in encoded form

Which PDF file we see everywhere are in encoded form. So ,we can not work or convert encoded PDF file manually in text file or document file.

Collecting simple PDF file for testing program

Because of being a beginner programmer it is impossible for me to work with encoded and compressed PDF file. But it is too difficult to find simple PDF file for working and testing my program. During this session I faced this major problem. Ans I am not able to find simple PDF file containing table data .

No Decisive way out

There was no decisive way out for me to convert PDF file in document file . So, I faces huge trouble for that.

Fear

Working with formatted file is like riding in mountain . That’s why it is very simple to being feared and confident less . I also feared about the huge project .

2.Project Overview

I have complete a small part of my project. I can successfully convert a simple a simple file into .txt file .

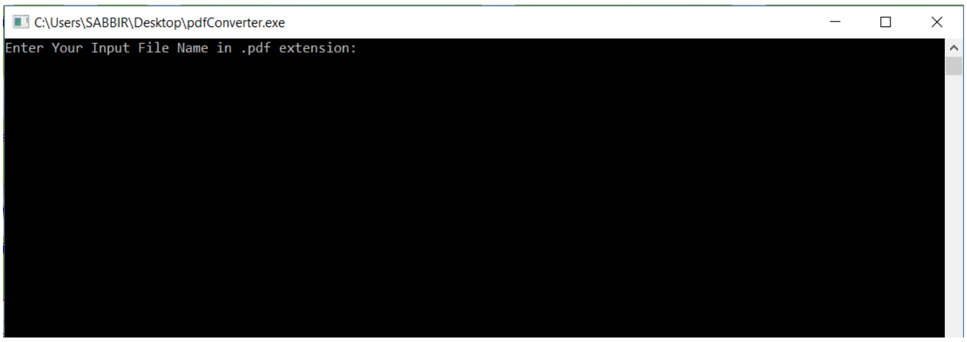
I also tried to draw graphics object using graphics.h library and wanted to save the output in a document file . But it was a vague idea . Because I can not create document file manually and this idea was not clear to me .

I am able to draw rectangle, line , Bezier curve in graphics window.

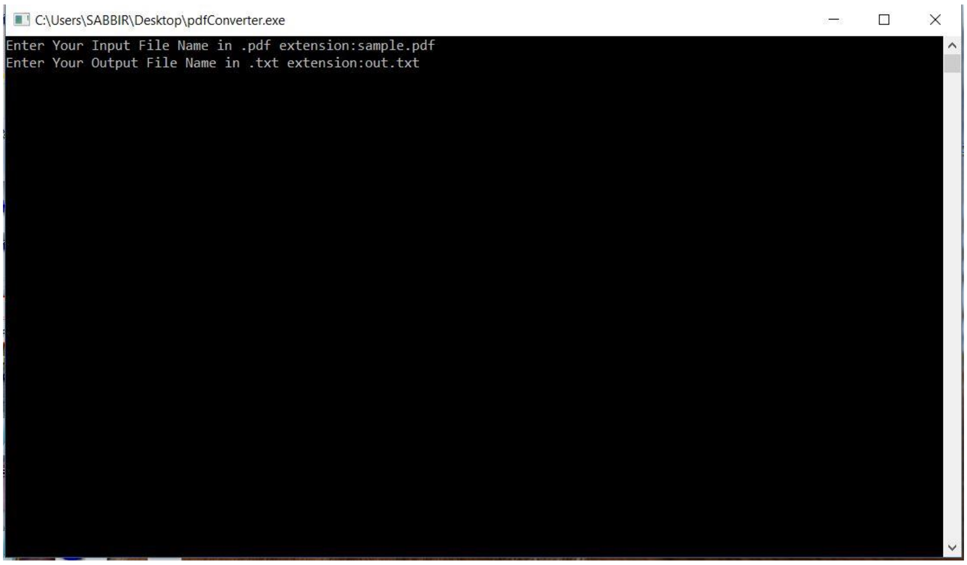
I still trying to convert simple PDF file in document file using API. But will not easy for me.

3.User Manual

This is the user manual for my project

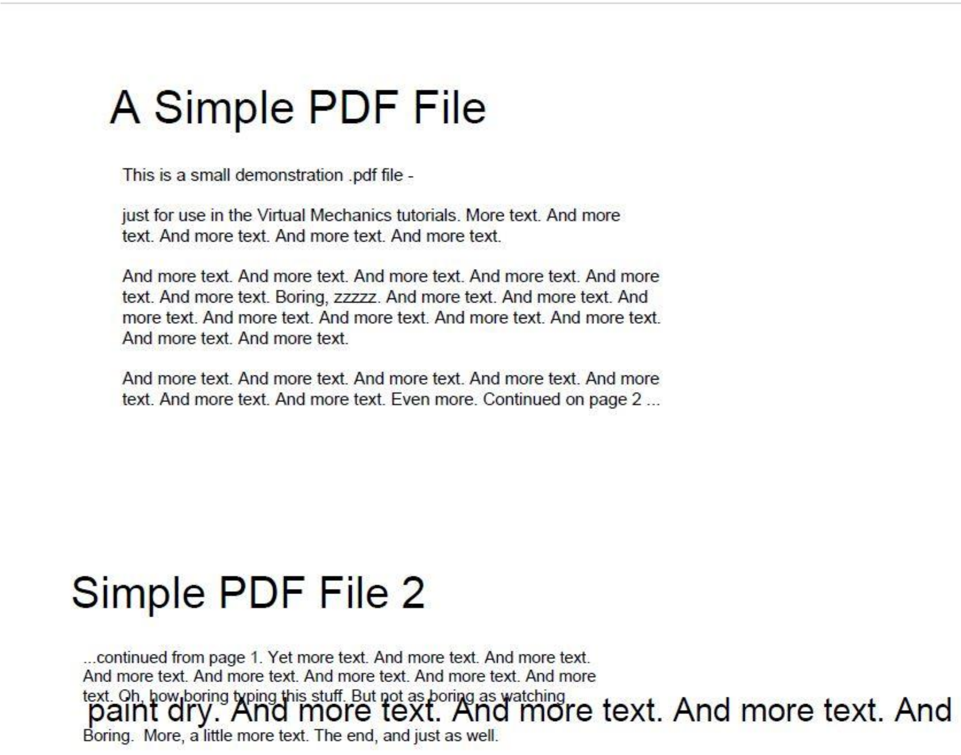


```
Enter Your Input File Name in .pdf extension:
```



```
Enter Your Input File Name in .pdf extension:sample.pdf
Enter Your Output File Name in .txt extension:out.txt
```

Input:



A Simple PDF File

This is a small demonstration .pdf file -

just for use in the Virtual Mechanics tutorials. More text. And more text. And more text. And more text. And more text.

And more text. And more text. And more text. And more text. And more text. And more text. Boring, zzzzz. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text.

And more text. And more text. And more text. And more text. And more text. And more text. And more text. Even more. Continued on page 2 ...

Simple PDF File 2

...continued from page 1. Yet more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. Oh, how boring typing this stuff. But not as boring as watching paint dry. And more text. And more text. And more text. And Boring. More, a little more text. The end, and just as well.

Output:

```
A Simple PDF File
This is a small demonstration .pdf file -
just for use in the Virtual Mechanics tutorials. More text. And more
text. And more text. And more text. And more text.
And more text. And more text. And more text. And more text. And more
text. And more text. Boring, zzzzz. And more text. And more text. And
more text. And more text. And more text. And more text. And more text.
And more text. And more text.
And more text. And more text. And more text. And more text. And more
text. And more text. And more text. Even more. Continued on page 2 ..
Simple PDF File 2
...continued from page 1. Yet more text. And more text. And more text.
And more text. And more text. And more text. And more text. And more
text. Oh, how boring typing this stuff. But not as boring as watching
paint dry. And more text. And more text. And more text. And more text.
Boring.  More, a little more text. The end, and just as well.
```

4.Concluding

This is my first project . I wasn’t very comfortable with the project. But I have learn many new and interesting and this experience will help me in future working.

5.Appendix

I can’t progress with the project very much. I am still trying to create doc file and next I want to work with encoded PDF file.

References

URL
https://www.pdftron.com/documentation/samples/cpp/TextExtractTest [21/1/19]
https://www.oreilly.com/library/view/developing-with-pdf/9781449327903/ch01.html [25/1/19]
http://www.planetpdf.com/developer/article.asp?ContentID=navigating_the_internal_struct&gid=6802 [25/5/19]
http://brendanzagaeski.appspot.com/0005.html [25/1/19]
http://dl.icdst.org/pdfs/files/c07aea28bea2cd04bb952ac9d6d4d263.pdf [25/1/19]
https://stackoverflow.com/questions/18098400/how-to-get-raw-text-from-pdf-file-using-java [25/1/19]
https://www.leadtools.com/help/leadtools/v20/dh/to/document-file-formats-portable-document-format-pdf.html [5/2/19]
https://www.adobe.com/content/dam/acom/en/devnet/pdf/pdfs/pdf_reference_archives/PDFReference.pdf [5/2/19]
https://stackoverflow.com/questions/58730/open-source-pdf-library-for-c-c-application [20/2/19]
https://stackoverflow.com/questions/30066727/c-program-for-reading-doc-docx-pdf [20/2/19]
http://www.idconline.com/technical_references/pdfs/information_technology/File_handling_in_C_Programming.pdf [20/2/19]
https://www.codeproject.com/Articles/12445/%2FArticles%2F12445%2FConverting-PDF-to-Text-in-C [2/3/19]
http://www.planetpdf.com/planetpdf/pdfs/pdf2k/01W/rosenthol_intro2pdfprog.pdf [2/3/19]
https://www.albany.edu/faculty/dsaha/teach/2017Spring_CEN360/slides/lec18.pdf [10/3/19]
https://www.prepressure.com/pdf/basics/convert-pdf [10/3/19]
http://www.printmyfolders.com/understanding-pdf [10/3/19]
https://idrh.ku.edu/pdf-text-extraction [10/3/19]
https://www.google.com/search?q=windows.h+in+c%2B%2B+tutorial&oq=windows.h+in+&aqs=chrome.5.0j69i57j0l4.12840j0j7&sourceid=chrome&ie=UTF-8 [10/3/19]
https://www.quora.com/How-can-I-learn-about-windows-h-header-file-and-its-function-and-uses [20/4/19]
https://en.wikipedia.org/wiki/Windows.h [20/4/19]
https://hownot2code.com/2016/08/16/stdafx-h/ [20/4/19]
https://stackoverflow.com/questions/10302004/windows-h-or-stdafx-h-the-first-window-example-in-petzolds-book-programming-w [20/4/19]
https://en.wikipedia.org/wiki/Category:Windows_APIs [20/4/19]
https://en.wikipedia.org/wiki/Category:Microsoft_application_programming_interfaces [20/4/19]
https://en.wikipedia.org/wiki/Windows_API [25/5/19]
https://en.wikipedia.org/wiki/Microsoft_Windows_library_files [25/5/19]
https://stackoverflow.com/questions/145573/creating-opening-and-printing-a-word-file-from-c [25/5/19]
https://docs.microsoft.com/en-us/office/troubleshoot/automate-word-mail-merge-using-visual-c-mfc [25/5/19]
https://sebsauvage.net/wiki/doku.php?id=word_document_generation [20/5/19]
https://support.microsoft.com/en-us/help/196776/office-automation-using-visual-c [25/5/19]
https://stackoverflow.com/questions/30066727/c-program-for-reading-doc-docx-pdf [25/5/19]
https://stackoverflow.com/questions/2350621/write-a-doc-or-rtf-file-from-a-c-c-application [25/5/19]
https://stackoverflow.com/questions/21953000/can-we-write-to-a-word-file-using-c [25/5/19]
https://www.leadtools.com/help/leadtools/v20/dh/to/file-formats-microsoft-word-document-docx-doc.html [25/5/19]
https://www.sciencedirect.com/topics/computer-science/formatted-file [25/5/19]
https://stackoverflow.com/questions/22446637/decoding-a-flatedecoded-section-of-text-in-a-pdf-document [25/5/19]

https://www.programmingsimplified.com/c/graphics.h/circle[25/5/19]
Reference Book
https://www.adobe.com/content/dam/acom/en/devnet/pdf/pdfs/pdf_reference_archives/PDFReference.pdf