Extract formatted text from PDF document for search and analysis

Written by Apitron Documentation Team

Introduction

Text in PDF documents is being drawn using individual text drawing and positioning commands and very often its initial formatting and logical structure doesn't get preserved because of this process. When you see a textual paragraph it doesn't mean that this paragraph is being stored or drawn as a whole thing. It can consist of many pieces each having its own unique properties or transformation and, most often, different fonts. So while it looks solid it's actually chunky and needs further processing to get back its logical structure and formatting, being *appearance dependent*.

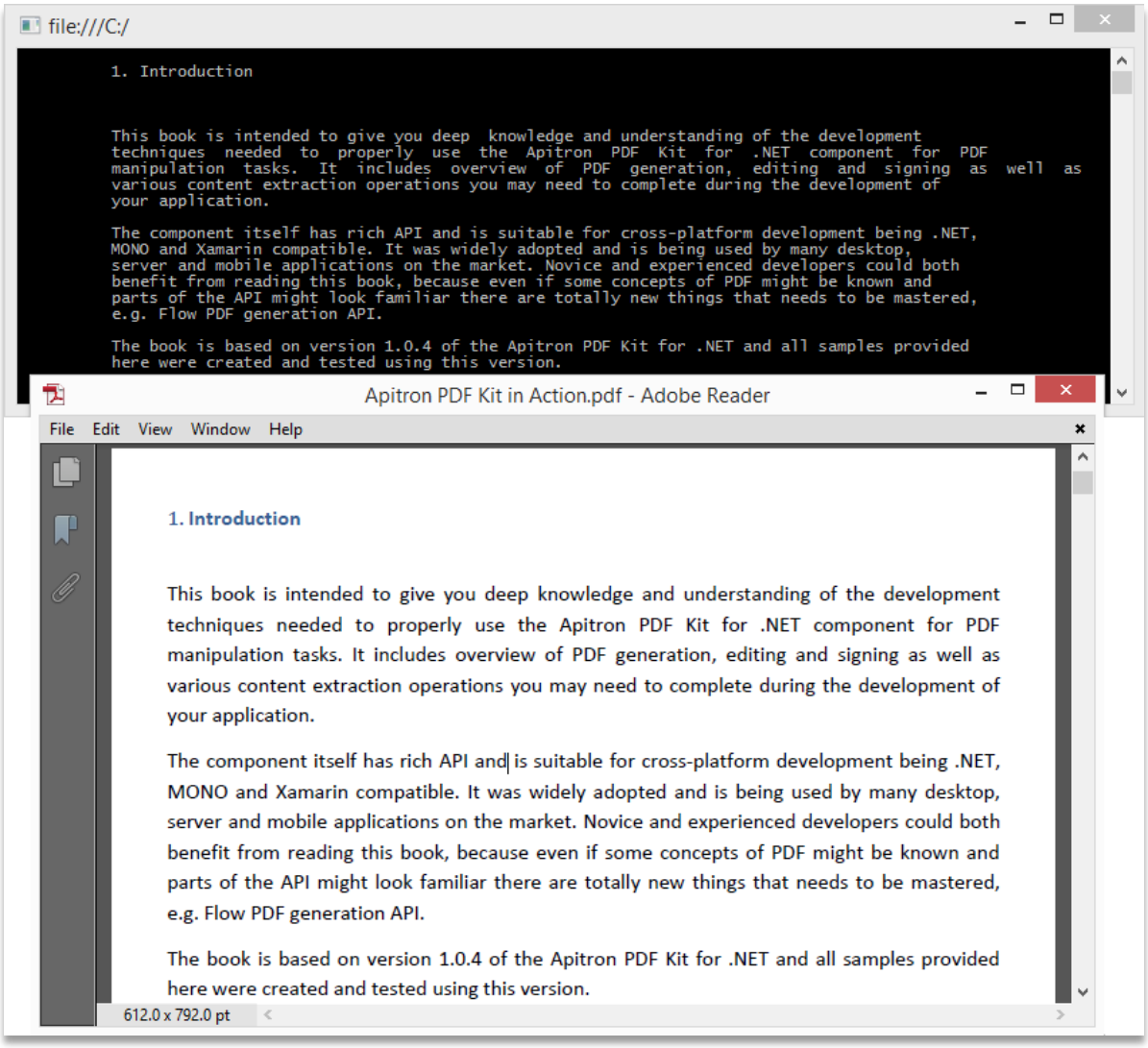
Apitron PDF Kit is a .NET component that provides very simple and easy way to get formatted text from PDF page, perform search in this text or analyze it in any desired way. Component features are described on our website and we have written a book describing it in action which you may download by the following link.

Getting the formatted text from PDF page

Code sample below shows how to extract *formatted* text from PDF page, the formatting is being applied intelligently using own algorithms which add necessary line breaks or spacing.

```
using (Stream stream = File.Open("Apitron PDF Kit in Action.pdf", FileMode.Open))
{
    FixedDocument document = new FixedDocument(stream);
    // extract formatted text
    string text = document.Pages[1].ExtractText(TextExtractionOptions.FormattedText);
    // set console window size and print text
    Console.SetWindowSize(116, 60);
    Console.WriteLine(text);
}
```

Compare the program output with the original PDF file on the image below:



Pic. 1 Formatted text extraction from PDF document

This C# code sample is pretty self-describing and demonstrates text extraction feature one may easily use to get text from PDF document. One thing should be noted though, if text in PDF file was created using embedded font subset with custom encoding, which doesn't have a Unicode mapping, then it couldn't be extracted.

The Apitron PDF Kit .NET component can be downloaded from our website. We'll be happy to answer your questions and welcome any feedback.