pdf2textbox Documentation

Release 0.1.1

Oliver Stapel

Sep 17, 2018

Contents:

A PDF-to-text converter based on pdfminer2 (which is based on pdfminer.six which is based on pdfminer). Converts PDF files with up to 3 columns and a header (optional) to text and avoids many caveats that multi-columned PDF files have in store for PDF conversion.

Allows command line parameter -s (–slice) to indicate that only part of the PDF document is of interest. Start and end page will then be either retreived from the document's name using '_' or 'l' as delimiters or - if start and end page cannot be found - user input is requested.

See also:

Extract PDF-files with pdfminer2

CHAPTER 1

What's the difference?

While pdfminer2 works well for plain text there will be issues when that text is in two or more columns, when there are indented quotes, headers on every page, pagenumbers, dates, and more. The solutions that can be found (i.e. pyPDF, pdfminer2, pdfx, . . .)

- Will include headers into the text flow
- Will return small or very small (DetailledAggregator) text units
- Will return text from left to right without acknowledging the columns
- Will mix text from different columns into one text flow

See also:

Stackoverflow hottest answer tagged pdfminer

Using PDFx

Using PyPDF2

pdfminer using a DetailedAggregator

pdf2textbox returns the extracted text in two meta structures: a dict that differentiates between 'header' and 'columns' and boxes similar to the paragraphs of the original text which makes the final conversion into a coherent and meaningful text document easier:

- Split text into coherent sentences
- Differentiate dots at the end of a sentence from dots behind dates, titles, etc.
- Allows simple regex routines to combine words seperated by hyphens (at linebreaks)
- Allign information from header to parts of text (i.e. page number, date, . . .)

Graphs, tables, and other visual or graphic 2D blocks, and tocs will cause havock.

CHAPTER 2

Features

Convert PDF to text in the original order. This works well for PDF-files without tables, graphs, and other stuff.

After extracting the text in boxes, there still has to be run another function to strip the text of special signs, zeroes and the like.

CHAPTER 3

Note

Often the textboxes will NOT be identical with paragraphs of the PDF-file. There might be cases when a textbox ends mid-sentence just to be continued with the next box. This is due to the PDF file's graphic-oriented organization of content. However, the order of text will be correct.

CHAPTER 4

Indices and tables

- genindex
- modindex
- search