pdf2textbox Documentation

Release 0.1.1

Oliver Stapel

Mar 11, 2022

Contents:

A PDF-to-text converter based on pdfminer2.

Converts PDF-files to text and avoids caveats (headers, pagenumbers, indented quotes, . . .) that PDF-files have in store for the unsuspecting PDF converter.

Extracts two-columned PDF-files with a good chance to get text in the original order.

See also:

Extract PDF-files with pdfminer2

CHAPTER 1

What's the tweak?

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. detailled aggregator, pyPDF, . . .) break down the PDF document into lines which in consequence leads to problems when parsing the resulting text.

See also:

Stackoverflow hottest answer tagged pdfminer

pdfminer using a DetailedAggregator

pdf2textbox returns the extracted text within boxes which makes the final conversion into a coherent and meaningful text document easier and prone to success. Graphs, tables, and other visual or 2D blocks will still 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

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 coninued 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.

When a document contains indexes, tocs, footnotes and other stuff, there is still a good chance to get hold of the text in a meaningful order, however the task to strip and reconnect the boxes will become tedious.

CHAPTER 4

Indices and tables

- genindex
- modindex
- search