Add comments to PDF files automagically with regular expressions [closed]

Asked 12 years, 7 months ago Modified 2 years, 10 months ago Viewed 8k times

17

Closed. This question needs to be more focused. It is not currently accepting answers.

Want to improve this question? Update the question so it focuses on one problem only by editing this post.

Closed 3 years ago.

Improve this question

I've been grading academic papers for a couple of years now and I've started to see numerous patterns in spelling and grammer mistakes. Also, I've noticed that less experienced academics tend to use certain constructs that immediately raise "smells" to more experienced researchers.

I would like to automagically recognize and annotate these in PDF files. Is anyone aware of a script that I could use to automagically annotate and comment PDF files? Perhaps it's dead simple, but I feel like I'm one of the first ones to ask this question.

Programming is no problem.

regex pdf annotations comments ghostscript

Share Follow

edited Jan 13, 2015 at 12:48
Kurt Pfeifle
86.3k 23 248 345

asked Dec 13, 2010 at 8:36
Slinger Jansen
237 2 14

2 Answers

Sorted by:
Highest score (default)

24

To solve this task, you need 3 things:

1. A good text extracting tool to get the contents from the PDFs (you're basically asking for this).
2. The knowledge about what keywords you want to use in order to create appropriate textual notes/comments and trigger a PDF annotation automatism (you say you have

3. A method to insert your comments into the PDF, preferably on the correct pages, or even on the exactly correct spot on the page (you're asking for this).

Text extraction

PDFlib's TET (text extraction toolkit) lets you extract text from any PDF. It's the most powerful of available PDF text extraction tools out there that allows you access via commandline and scripting. It can handle such weirdies (from the p.o.v. of text extraction) as ligatures as well as different text encodings. More important, it can tell you the exact page number and coordinates on the PDF page for any character or text string it extracted.

Inserting PDF annotations

After you parsed the text, and your logic decided which comment to add for which page, you can use PDFlib or Ghostscript to add comments ("annotations") to the original PDF.

I'm not delivering a tutorial about how to use PDFlib in order to add annotations to existing PDFs here. But I will leak some insider knowledge about how Ghostscript can do it:

Using Ghostscript for adding annotations to PDFs

To add an annotation with Ghostscript to an existing PDF, first create a text file called *my-pdfmarks.txt* (or whatever name you prefer). Now type into that textfile the content of your annotation, using the following syntax:

```
 [ /Title (Annotation experiments by -pipitas-)
   /Author (pipitas)
   /Subject (I'm trying to add annotations to existing PDFs with the help of
Ghostscript...)
   /Keywords (comma, separated, keywords, spelling mistakes, grammar mistakes,
raising "smells")
   /ModDate (D:20101219192842)
   /CreationDate (D:20101219092842)
   /Creator (pipitas' brainz)
   /Producer (Ghostscript under the direction of pipitas)
   /DOCINFO pdfmark

 [ /Contents (Smell: This statement was bloody well rebutted by decades of academic
research...)
   /Rect [10 10 50 50]
   /Subtype /Text
   /Name Note
   /SrcPg 2
   /Open true
   /ModDate (D:20101220193344)
   /Title (A Comment on Page 2)
   /Color [.5 .5 0]
   /ANN pdfmark
```

Then, run Ghostscript command like the following. I'm assuming Windows now -- for Linux/Unix/MacOSX use `gs` instead of `gswin32c.exe` for the executable, and use `\` instead of `^` for the line continuation marks:

```
gs ^
  -o original-annotated.pdf ^
  -sDEVICE=pdfwrite ^
  -dPDFSETTINGS=/prepress ^
   original.pdf ^
   my-pdfmarks.txt
```

Voila! Your output PDF now has an annotation on page 2.

Now you probably didn't understand what exactly you were doing:

- The first part of the *my-pdfmarks.txt* file manipulates the PDF's meta data. Just delete it if you don't want this.
- The second part adds an annotation ('*/Subtype /Text*' and '*/Name /Note*') on Page 2 ('*/SrcPg 2*') of the output PDF at the lower left corner, 10 points away from each page border ('*/Rect [10 10 50 50]*'), using a greenish DeviceRGB color ('*/Color [0.5 0.5 0]*'), and opening it by default ('*/Open true*') when accessing the page.

Tweakable parameter values (after each keyword) in the *my-annotations.txt* file are all **BUT** the following:

1. " `/DOCINFO pdfmark` "
2. " `/Subtype /Text` "
3. " `/Name /Note` "
4. " `/ANN pdfmark` "

For example, to make the annotation appear in pure red, use `/Color [1 0 0]` .

In order to fully understand the pdfmark syntax (and add more tweaks to your procedure), you'll need to google for Adobe's *pdfmark Reference Manual* and read that.

Since you said '*programming is no problem*' you now have all the building blocks to automate this with any scripting language of your choice.

Share Follow

edited Sep 7, 2020 at 12:08
malat
12.1k 13 88 157

answered Dec 19, 2010 at 17:52
Kurt Pfeifle
86.3k 23 248 345

This totally did the trick! Wonderful, I can now annotate pdfs automagically. Some problems I've run into that'll require some more work: (1) TETml can be output in two formats, being words and lines. Words are annotated with an X and Y coordinate, whereas lines are... Not. In some cases however (such as with "it's" and ", which") I need to know the context of a piece of text. I still need to write the code to connect these two formats... Ugh. (2) I see that if a student makes one mistake frequently, the comments get kind of repetitive too. Thanks a bundle! – Slinger Jansen Dec 30, 2010 at 10:54

1 After two weeks of working with my own tool, I've noticed lots of glitches in the system (mostly Ghostview related). Especially pdf files generated by MS Word, but certainly not only these, will

I got an empty rectangle with no text. If I change `/Subtype` to `/FreeText` it works. – Jesse Aldridge Oct 11, 2019 at 0:24

Regardless what I do, ModDate won't show on my annotation. I'm on Windows 10 and using GS 9.27. I'd prefer to pass it in as a variable, but regardless - it won't show. – Ben Rice May 7, 2021 at 21:32

@BenRice: Do the other annotation items (Creator, Producer,...) show? – Kurt Pfeifle May 8, 2021 at 9:10

3

If I were you I would start with the PDF Library SDK which supports the things you're looking for:

- Extract content
- Add comments to documents

One drawback is that you have to apply for it and Adobe may refuse your request.

EDIT:

PDFedit seems promising. It's an open source GUI application that allows you to modify PDF manually or by scripting.

Share Follow

edited Dec 13, 2010 at 10:48

answered Dec 13, 2010 at 8:41

VVS

19.4k 5 46 65

Really? That seems serious, isn't there some cool open source toolkit I can use? Adobe is a dinosaur that I would love to throw my mini spear at... And is it scriptable? – Slinger Jansen Dec 13, 2010 at 8:49

There are plenty of libraries that allow you to create PDF but I don't of know any open library that allows you to read or modify PDF. – VVS Dec 13, 2010 at 9:03

Perhaps you can define that your students have to send you papers in a more open format like ODF. OpenOffice.Org and current versions of Microsoft Word are able to save in this format and its structure ist well defined XML. – VVS Dec 13, 2010 at 9:06

Interesting, but not an option. Many academic papers are delivered in specific conference formats (IEEE, ACM). Please note, these are mostly other academics, not students. If no more answers come in I will explore the SDK option. Thanks! – Slinger Jansen Dec 13, 2010 at 10:10

PDFedit may do the trick! Thanks, will come back with feedback. – Slinger Jansen Dec 13, 2010 at 10:53