UNIX & LINUX

How to OCR a PDF file and get the text stored within the PDF?

Asked 7 years, 9 months ago Modified 1 year, 10 months ago Viewed 69k times

91

First, apologies if this has been asked before - I searched for a while through the existing posts, but could not find support.

I am interested in a solution for Fedora to OCR a multipage non-searchable PDF and to turn this PDF into a new PDF file that contains the text layer on top of the image. On Mac OSX or Windows we could use Adobe Acrobat, but is there a solution on Linux, specifically on Fedora?

This seems to describe a solution - but unfortunately I am already lost when retrieving exact-image.

command-line pdf ocr

Share Improve this question Follow

edited Jan 23, 2020 at 12:22

asked Aug 4, 2016 at 15:39

ingli
1,859 1 16 33

1 There is a problem with the nice pdfocr script that the page you are linking to recommends: it relies upon pdftk which is essentially deprecated (for two reasons, its dependence on libgcj and on iText5+). So a different solution is needed anyway... – Maxim Mar 14, 2017 at 6:04

4 Answers

Sorted by: Highest score (default)

116

`ocrmypdf` does a good job and can be used like this:

```
ocrmypdf in.pdf out.pdf
```

To install:

```
pip install ocrmypdf
```

or

```
sudo apt install ocrmypdf     # ubuntu
sudo dnf -y install ocrmypdf  # fedora
```

Share Improve this answer Follow

edited Jun 18, 2022 at 11:13

ingli
1,859 1 16 33

answered Feb 3, 2018 at 19:23

Eduard Florinescu
11.7k 19 57 68

6 Used `ocrmypdf` on Fedora 30 (via `dnf install`) - worked like a charm. – Heinrich Ulbricht Jan 23, 2020 at 13:02

3 very good, thanks. Unlike the other ocr proposed in this thread, this ocr gives an output only *slighlty bigger* than the original (image pdf). It would even better if it could give an output *smaller* (only text): it is possible? – Duns Apr 16, 2020 at 16:36

@Duns if you will do that the resulting text layout will have a lot of errors, so you cannot using this tool, there is no way to have a decent looking OCRed without the original layout output, another better dimension and better looking solution would be Adobe Clearscan or alternative to clearscan softwarerecs.stackexchange.com/questions/10242/… – Eduard Florinescu Apr 16, 2020 at 17:14

7 OCRmyPDF worked like a dream for me too. It's based on Tesseract under the hood, so (among other things) handles many languages well: I just used it for a document in a mixture of English and Georgian (ქართული ენა) and got near-perfect results. – PLL Feb 17, 2021 at 9:33

1 Ubuntu 20.04: When creating an ocr pdf, `ocrmypdf` states that `jbig2enc` is not installed and is needed for compressing and higher quality PDF files. `jbig2enc` must be built from source, but it has dependencies of `libtool` [that contains both `libtoolize` and `glibtoolize`] to be installed with `sudo apt install libtool` , and `libleptonica-dev` (which contains `Leptonica`): `sudo apt install libleptonica-dev` . Then, follow the instructions for git-cloning `jbig2enc` at `git clone https://github.com/agl/jbig2enc` and running `./autogen.sh` / `./config` / `make` / `sudo make install` – lawlist Sep 5, 2022 at 1:00

22

After learning that Tesseract can now also produce searchable PDFs, I found the script sandwich: http://www.tobias-elze.de/pdfsandwich/

after installing dependencies (this might not be the complete list)

```
sudo dnf install svn ocaml unpaper tesseract
```

I followed the script's guide for compiling from source

> Compile from sources
>
> pdfsandwich is open source software (license: GPL). You can download the sources either as .tar.bz2 package from the download area on the project website or check them out by subversion:

```
svn checkout svn://svn.code.sf.net/p/pdfsandwich/code/trunk/src pdfsandwich
```

> If OCaml is installed on your system, you can compile and install as follows:

```
cd pdfsandwich
./configure
make
sudo make install
```

and this now allows me to run

```
sandwich multipaged-non-searchable.pdf
```

resulting in a searchable PDF.

Here is a list of repositories (e.g., Debian Stable, AUR, Homebrew) containing pdfsandwich.

Share Improve this answer Follow

edited Sep 7, 2020 at 8:22 — Matthias Braun — 8,289 7 48 59

answered Aug 4, 2016 at 15:39 — ingli — 1,859 1 16 33

for a related, but separate question, building on this one, see unix.stackexchange.com/questions/306051/… – ingli Aug 27, 2016 at 18:25

3 FWIW: pdfsandwich is also available in Ubuntu's apt package repository. Other distros might have it as well. – Laurence Gonsalves Mar 14, 2018 at 6:25

Just came across fedoramagazine.org/4-cool-new-projects-try-copr-october-2018 showing a COPR package for fedora that packages pdfsandwich – ingli Oct 26, 2018 at 8:59

8

An easy tool available in Ubuntu is 'ocrfeeder' it allows the generation of PDFs with OCR text overlaid on the original documents. It makes use of Tesseract plus other OCR engines (not sure which) and provides for image rotation/'unpaper', etc, as well.

- http://live.gnome.org/OCRFeeder
- https://github.com/GNOME/ocrfeeder

Share Improve this answer Follow

answered Oct 18, 2018 at 4:14

jdpipe
181 1 4

6

I had this same problem so I wrote this over the weekend. Give it a shot; it works great! It is a simple wrapper around `tesseract`. It uses `pdftoppm` to convert a PDF into a bunch of TIFF files, then it uses `tesseract` to perform OCR (Optical Character Recognition) on them and produce a searchable PDF as output. All intermediate temporary files are automatically deleted when the script completes.

Source code: https://github.com/ElectricRCAircraftGuy/PDF2SearchablePDF

Instructions to install & use `pdf2searchablepdf`:

Tested on **Ubuntu 18.04** on 11 Nov 2019 and on **Ubuntu 20.04** Nov. 2020.

Install:

```
git clone https://github.com/ElectricRCAircraftGuy/PDF2SearchablePDF.git
./PDF2SearchablePDF/install.sh

sudo apt update
sudo apt install tesseract-ocr
```

Use:

```
# General:
pdf2searchablepdf [options] <input.pdf|dir_of_imgs> [lang]

# Make a PDF searchable:
pdf2searchablepdf mypdf.pdf

# Make an entire directory of images into a single searchable PDF:
pdf2searchablepdf directory_of_imgs
```

You'll now have a pdf called **mypdf_searchable.pdf**, which contains searchable text!

Done. It has no python dependencies, as it's currently written entirely in bash.

See `pdf2searchablepdf -h` for the help menu and more options and examples.

References or Related Resources:

1. **PDF2SearchablePDF**: https://github.com/ElectricRCAircraftGuy/PDF2SearchablePDF
2. https://askubuntu.com/questions/473843/how-to-turn-a-pdf-into-a-text-searchable-pdf/1187881#1187881
3. https://askubuntu.com/questions/16268/whats-the-best-simplest-ocr-solution
4. https://askubuntu.com/questions/150100/extracting-embedded-images-from-a-pdf/1187844#1187844
5. **pdfsandwich**: Alternative software wrapper I just discovered, that is worth checking out too! http://www.tobias-elze.de/pdfsandwich/

Share Improve this answer Follow

edited Jan 20, 2022 at 7:07

answered Nov 11, 2019 at 9:22

Gabriel Staples
2,612 2 28 43

Good utility. One thing you might do is add support for file names with spaces in them. Right now, that doesn't work (you get a usage message for `pdftoppm`). Just adding a few quotation marks in some of the commands should do it. – Wilson F Jan 4, 2020 at 1:01

1 Thanks for the feedback! I'll see when I can make the change and test it. I opened an issue here: github.com/ElectricRCAircraftGuy/PDF2SearchablePDF/issues/6 – Gabriel Staples Jan 5, 2020 at 8:51

1 @WilsonF, done! v0.4.0 just released to resolve this issue. github.com/ElectricRCAircraftGuy/PDF2SearchablePDF/releases – Gabriel Staples Mar 15, 2020 at 3:54

1 Excellent utility. Faster and more stable than my brief experiments with ocrmypdf and pdfsandwich. This on Ubuntu 18.04 and only a couple of PDF scanned documents as images. My issues arose when having relatively high resolutions (300dpi). – Patrick Refondini Dec 1, 2021 at 14:57

1 @GabrielStaples `pdf2searchablepdf` is fast and stable. I had issue with `ocrmypdf` and `pdfsandwich` when doing only a couple of tests with resolutions >= 300dpi. – Patrick Refondini Dec 3, 2021 at 15:13