[How To Get Text from an Image Only PDF File Using Linux](#)

This article is for technically savvy people, and especially for those who use Linux on their Desktop or Laptop PC as I do.

I wanted to post a 19th-century book, *The Black Pope*, but the text from the PDF file I downloaded is not extractable using copy and paste. I knew I needed to use OCR software to get the text. On your PC you need pdftoppm and tesseract installed. pdftoppm was already installed on my laptop by default. I just needed to install tesseract which is OCR software. I use Fedora Linux, but this proceedure will work in any distribution of Linux.

First I made a folder for Black_Pope.pdf and moved to file into the folder. Then I opened Terminal inside the folder.

These are the commands I used in Terminal to get the ASCII text.

pdftoppm -png Black_Pope.pdf black-pope

(This made PNG files of each page of the PDF. There were 404 in all.)

for i in black-pope-???.png; do tesseract "$i" "text-$i" -l eng; done;

(This command scans each of the PNG files and creates .txt files of each of them.

cat text-black-pope* > black-pope-complete.txt

(This combined all of the 404 text files into a single file.)

It only took a few minutes! The PC did all the work. Just think how long it would have taken me if I scanned each one of those PNG files one by one, and combined them all one by one. Probably an hour or more.

I wound up with a single file to proofread. After I proofread a section, I removed all the extra line returns with an online tool: [Remove line breaks with paragraph restoration](#) before I copied the text into [the WordPress post](#).