The pdftotext Utility

The pdftotext command-line utility provides capabilities for converting PDF fils into TXT files for further processing.

References:

- Documentation
- Installing on Windows
- Example invocation from a Python script, by @rcy222

Installation

First see if Git is already installed (it may come pre-installed):

```
# Mac Terminal:
pdftotext --help #> pdftotext version 0.65.0 ...
which pdftotext #> /usr/local/bin/pdftotext

# Windows Command Prompt, Anaconda Prompt, or Git Bash:
# first navigate to the folder where you downloaded the "pdftotext.exe" file,
then...
pdftotext --help #> pdftotext version 0.65.0 ...
where pdftotext #> /path/to/pdftotext.exe
```

If these commands produce a version-looking output and a filepath-looking output, respectively, then the utility is already installed and you can skip down to the "Usage" section. Otherwise, follow the OS-specific sections below to install it.

Installation on Mac

Mac users can install pdftotext via homebrew:

```
brew install pkg-config poppler
```

After installing, restart your terminal application, where you should now be able to execute pdftotext commands (like pdftotext --version).

Installation on Windows

Windows users can install pdftotext by visiting https://www.xpdfreader.com/download.html and clicking "Download the Xpdf tools". After downloading onto your local computer, locate the zip file and unzip / extract it, then move the unzipped folder to a location like the Desktop or the Programs directory. Inside the unzipped folder, observe the absolute filepath location of the executable file called "bin64/pdftotext.exe". When you need to use the `pdftotext` utility in the future, either reference it from this location (e.g. `/path/to/pdftotext --version`) or add an alias to that location via your "~/.bash_profile", or move a copy of that file into any project repository you'd like to reference it from.

Usage

Download a PDF file onto your Desktop or some other location (e.g. "/path/to/my_document.pdf"), then navigate there from the command-line. Then process the PDF into a new TXT file (e.g. "/path/to/my_document.txt"):

```
pdftotext /path/to/my_document.pdf /path/to/my_document.txt
```

Then examine the contents of the TXT file in your text editor to see how well it was able to parse the original PDF document:

```
code /path/to/my_document.txt
```