How to batch convert pdf files to text

Last updated on Aug 9, 2020 · 3 min read · · Course-related, Quantitative Methods

Frequently I am asked: I have a bunch of pdf files, how can I convert them to plain text so that analyze them using quantitative techniques? Here is my recommendation.

1. Download the xpdf suite of tools for your platform. This includes the part we will use, **pdftotext**.
   Alternatives are the Apache PDFBox Java pdf library, and the Python-based PDFminer.

2. [Windows only – Mac and Linux/Unix have this built in to the Terminal or shell already]: You will need a bash shell for your platform. (It is possible to do what I suggest below using the Windows shell, but it's been so long since I programmed in the Windows DOS/command line script language that I won't even attempt it now.) The main options seem to be win-bash and Cygwin.

3. Create a folder called pdfs in your home folder (for this example – of course it can be elsewhere). Copy your pdf files to this folder.

4. In a text edtor, create a text file called `convertmyfiles.sh` with the following contents:

```
#!/bin/bash
FILES=~//pdfs/*.pdf
for f in $FILES
do
 echo "Processing $f file..."
 pdftotext -enc UTF-8 $f
done
```

(I am not providing a link because if you cannot create a text file and copy this text to it — and crucially edit it slightly for your own needs — then you probably won't have much luck with these steps anyway.)

```
 * Open the bash shell (Terminal.app or win-bash or equivalent) and execute the following: cd pdfs
```

```
./convertmyfiles.sh
```

```
Now you will have a set of text files (ending with .txt) converted as a set. These will probably need tidying up, as the conversion te

    pdftotext -h

Note that in the file provided, the extracted text is given a UTF-8 (Unicode) character encoding, which is what you should be using wh
Example: (from Terminal.app on my Mac)

    Last login: Thu Jul 31 11:29:44 on ttys001
    KBs-MBP13:~ kbenoit$ cd pdfs
    KBs-MBP13:pdfs kbenoit$ pwd
    /Users/kbenoit/pdfs
    KBs-MBP13:pdfs kbenoit$ rm *txt
    KBs-MBP13:pdfs kbenoit$ ls
    11centerpartiet2004.pdf
    11folkpartiet2004.pdf
    11kristdemokraterna2004.pdf
    11kristdemokraterna2004_300k.pdf
    11miljopartiet_de_grone2004.pdf
    13radikale_venste2004_ENGL.pdf
    13socialdemokraterne2004.pdf
    21Ecolo_programme_2004.pdf
    21Mouvement_Reformateur_100_propositions_pour_2_0lect_Vlaams_en_europe.PDF
    21SPA_europeesprogramma2004.pdf
    convertmyfiles.sh
    KBs-MBP13:pdfs kbenoit$ ./convertmyfiles.sh
    Processing /Users/kbenoit//pdfs/11centerpartiet2004.pdf file...
    Processing /Users/kbenoit//pdfs/11folkpartiet2004.pdf file...
    Processing /Users/kbenoit//pdfs/11kristdemokraterna2004.pdf file...
    Processing /Users/kbenoit//pdfs/11kristdemokraterna2004_300k.pdf file...
    Processing /Users/kbenoit//pdfs/11miljopartiet_de_grone2004.pdf file...
    Processing /Users/kbenoit//pdfs/13radikale_venste2004_ENGL.pdf file...
    Processing /Users/kbenoit//pdfs/13socialdemokraterne2004.pdf file...
    Processing /Users/kbenoit//pdfs/21Ecolo_programme_2004.pdf file...
    Processing /Users/kbenoit//pdfs/21SPA_europeesprogramma2004.pdf file...
    KBs-MBP13:pdfs kbenoit$ ls
    11centerpartiet2004.pdf
    11centerpartiet2004.txt
    11folkpartiet2004.pdf
    11folkpartiet2004.txt
    11kristdemokraterna2004.pdf
    11kristdemokraterna2004.txt
    11kristdemokraterna2004_300k.pdf
    11kristdemokraterna2004_300k.txt
    11miljopartiet_de_grone2004.pdf
    11miljopartiet_de_grone2004.txt
    13radikale_venste2004_ENGL.pdf
    13radikale_venste2004_ENGL.txt
    13socialdemokraterne2004.pdf
    13socialdemokraterne2004.txt
    21Ecolo_programme_2004.pdf
    21Ecolo_programme_2004.txt
    21Mouvement_Reformateur_100_propositions_pour_2_0lect_Vlaams_en_europe.PDF
    21SPA_europeesprogramma2004.pdf
    21SPA_europeesprogramma2004.txt
    convertmyfiles.sh
    KBs-MBP13:pdfs kbenoit$
```

Update 12 November 2015 for Windows (thanks Thomas)

For Windows, one way to do the is to use Windows PowerShell ISE (Integrated scripting environment) in Programs/Accessories as follows:

```
cd mypdffolder
$FILES= ls *.pdf
foreach ($f in $FILES) {
    C:\Program` Files\xpdf\bin32\pdftotext -enc UTF-8 $f
}
```

Ken Benoit

Professor of Computational Social Science

cv

Related

- Quantitative Text Analysis (TCD 2016)
- Quantitative Text Analysis 2E, Essex 2014
- Quantitative Text Analysis (TCD)
- ME104 Linear Regression Analysis, 2012
- Computer-Assisted Text Analysis (Essex Summer School)