

Use Python to Combine LST Files into a PDF File with Bookmarks

By Hengwei Liu

The Purpose of the Python Programs

- ▶ Programmers may produce hundreds of LST files for the tables and listings in a clinical study.
- ▶ We would like to create a single PDF files with bookmarks. This makes it easier for the study team to review the output.

A sample LST file

XXXXXXXXXXXXXXXXX, Inc.
Protocol: XXX-XXX-XXX
Interim Analysis

Page 1 of 20
Program Name: demog.sas
Run datetime: <2019-03-31 12:43>

Table 14.1.2 Demographics and Baseline Characteristics
Safety Population

	Treatment A (N=xx)	Treatment B (N=xx)
Age		
n	xx	xx
Mean	xx.x	xx.x
Standard Deviation	xx.x	xx.x
Median	xx.x	xx.x
Min, Max	xx.x, xx.x	xx.x, xx.x
Race		
White	xx	xx
Black or African American	xx	xx
Asian	xx	xx
Other	xx	xx
Sex		
Male	xx	xx
Female	xx	xx

[illegible]

Requirement on the title of the LST files

- ▶ The table number is on the fourth line.
- ▶ The titles end above a horizontal line.
- ▶ The table numbering consists of numbers less than 99 separated by a dot; the last number can be followed by one of the letters from a to z.
- ▶ For example, 14.2.11, 14.3.1a.

Do it in three steps

- ▶ In the first step we rename each LST file using the numbering and title of the table.
- ▶ In the second step we produce a pdf file for each LST file.
- ▶ In the third step we combine all the pdf files into a single pdf file with bookmarks.

Step 1: rename the LST files

- ▶ The Python program `rename.py` reads the LST file from the fourth row to the row above the first horizontal line. This is the table number and all the titles for the table.
- ▶ The python program `rename.py` adds a new numbering in front.
- ▶ This new numbering is generated by replacing the dot in the table number by a padding zero if the number after the dot is single digit. For example, the file `t140101.lst` for Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42) is renamed as “140101 Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42).lst”.

Step 1: the Program

```
import glob
import os
import re

pattern1 = re.compile(r"[a-z]", re.IGNORECASE) # Compile a
case-insensitive regex
read_files = glob.glob("*.lst")

alp='abcdefghijklmnopqrstuvwxyz'

for f in read_files:
    with open(f, "rt") as infile:
        lines=infile.readlines()
        line=lines[3].strip().split()[1]
        g=' '
```

Step 1: the Program, cont.

```
for i in range(10,0, -1):
    if re.search('___', lines[i]) != None:
        title_line=i

title=lines[3]
if title_line>4:
    for i in range(4,title_line):
        title = title.strip() + ' ' + lines[i].strip()
if pattern1.search(line) != None:
    result=re.search(r"[a-zA-Z]", line).start()
    g=line[result:]
```

Step 1: the Program, cont.

```
nums=line.split(".")
lin=' '
for h in nums:
    h=h.lower()

    h=h.strip(alp)
    if len(h)>1:
        h2=h
    elif len(h)==1:
        h2='0'+h
    else:
        h2=''
    lin += h2
lin += g
lin1=lin.strip() + ' ' + title.strip() + '.1st'
os.rename(f,lin1)
```

Step 1, screenshot

```
Activities Terminal Jul 11 09:24
hengweiliu@new-host:~/Downloads/lst2pdf
[hengweiliu@new-host ~]$ cd Downloads/lst2pdf
[hengweiliu@new-host lst2pdf]$ ls -l
total 596
-rw-rw-r--. 1 hengweiliu hengweiliu 485 Jun 4 23:56 lst2pdf.py
-rwxrwxr--. 1 hengweiliu hengweiliu 1971 Jun 4 23:56 pdfmerger.py
-rw-rw-r--. 1 hengweiliu hengweiliu 1174 Jul 11 09:18 rename.py
-rw-rw-r--. 1 hengweiliu hengweiliu 2563 Jul 11 09:23 t140101.lst
-rw-rw-r--. 1 hengweiliu hengweiliu 3225 Jul 11 09:23 t14_04_01.lst
-rw-rw-r--. 1 hengweiliu hengweiliu 2563 Jul 11 09:23 table14_02_01.lst
-rw-rw-r--. 1 hengweiliu hengweiliu 3367 Jul 11 09:23 table14_04_02_a.lst
-rw-rw-r--. 1 hengweiliu hengweiliu 3357 Jul 11 09:23 table14_04_02.lst
-rw-rw-r--. 1 hengweiliu hengweiliu 3367 Jul 11 09:23 table14_04_11.lst
-rw-rw-r--. 1 hengweiliu hengweiliu 572662 Jul 11 09:23 table14_04_12.lst
[hengweiliu@new-host lst2pdf]$ python3 rename.py
[hengweiliu@new-host lst2pdf]$ ls -l
total 596
-rw-rw-r--. 1 hengweiliu hengweiliu 2563 Jul 11 09:23 '140101 Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42).lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 2563 Jul 11 09:23 '140201 Table 14.2.1 Maturity of DoR for Cohort 2 subjects with prior platinum therapy (N=38).lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 3225 Jul 11 09:23 '140401 Table 14.4.1 Time to Response (Months) by Independent Central Review Full Analysis Set.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 3367 Jul 11 09:23 '140402a Table 14.4.2a Time to Response (Weeks) by Independent Central Review Full Analysis Set Cohort 2 Subject
s with prior platinum-based therapy.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 3357 Jul 11 09:23 '140402 Table 14.4.2 Time to Response (Weeks) by Independent Central Review Full Analysis Set Subjects with prio
r platinum-based therapy.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 3367 Jul 11 09:23 '140411 Table 14.4.11 Time to Response (Months) by Independent Central Review Full Analysis Set Cohort 1 Subject
s with prior platinum-based therapy.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 572662 Jul 11 09:23 '140412 Table 14.4.12 Treatment-Emergent Adverse Events by System Organ Class, Preferred Term, and Worst CTCAE G
rade Safety Analysis Set.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 485 Jun 4 23:56 lst2pdf.py
-rwxrwxr--. 1 hengweiliu hengweiliu 1971 Jun 4 23:56 pdfmerger.py
-rw-rw-r--. 1 hengweiliu hengweiliu 1174 Jul 11 09:18 rename.py
[hengweiliu@new-host lst2pdf]$
```

Step 2: from LST to PDF

- ▶ This is done through the FPDF module in Python.
- ▶ `pdf.cell(250, 3.4, txt = x, ln = 1, align = 'R')`

We got the w=250 and h=3.4 through trial and error. This works fine for the LST files generated with SAS options LS=132 and PS=52.

Step 2: from LST to PDF, the program

```
import glob
import os
from fpdf import FPDF

read_files2 = glob.glob("*.lst")
read_files2.sort()

for f in read_files2:
    lin=os.path.basename(f)
    lin2=lin.replace('lst','pdf')
    pdf = FPDF(orientation='L')
    pdf.add_page()
    pdf.set_font("Courier", size = 8)
    infi=open(f, 'r')
    for x in infi:
        pdf.cell(250, 3.4, txt = x, ln = 1, align = 'R')
    infi.close()
    pdf.output(lin2)
```

Step 2, Screenshot

```
Activities Terminal Jul 11 09:25
hengweiliu@new-host:~/Downloads/lst2pdf
[hengweiliu@new-host ~]$ cd Downloads/lst2pdf
[hengweiliu@new-host lst2pdf]$ python3 lst2pdf.py
[hengweiliu@new-host lst2pdf]$ ls -l
total 756
-rw-rw-r--. 1 hengweiliu hengweiliu 2563 Jul 11 09:23 '140101 Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42).lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 1719 Jul 11 09:25 '140101 Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42).pdf'
-rw-rw-r--. 1 hengweiliu hengweiliu 2563 Jul 11 09:23 '140201 Table 14.2.1 Maturity of DoR for Cohort 2 subjects with prior platinum therapy (N=38).lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 1737 Jul 11 09:25 '140201 Table 14.2.1 Maturity of DoR for Cohort 2 subjects with prior platinum therapy (N=38).pdf'
-rw-rw-r--. 1 hengweiliu hengweiliu 3225 Jul 11 09:23 '140401 Table 14.4.1 Time to Response (Months) by Independent Central Review Full Analysis Set.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 1886 Jul 11 09:25 '140401 Table 14.4.1 Time to Response (Months) by Independent Central Review Full Analysis Set.pdf'
-rw-rw-r--. 1 hengweiliu hengweiliu 3367 Jul 11 09:23 '140402a Table 14.4.2a Time to Response (Weeks) by Independent Central Review Full Analysis Set Cohort 2 Subject
s with prior platinum-based therapy.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 1921 Jul 11 09:25 '140402a Table 14.4.2a Time to Response (Weeks) by Independent Central Review Full Analysis Set Cohort 2 Subject
s with prior platinum-based therapy.pdf'
-rw-rw-r--. 1 hengweiliu hengweiliu 3357 Jul 11 09:23 '140402 Table 14.4.2 Time to Response (Weeks) by Independent Central Review Full Analysis Set Subjects with prio
r platinum-based therapy.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 1910 Jul 11 09:25 '140402 Table 14.4.2 Time to Response (Weeks) by Independent Central Review Full Analysis Set Subjects with prio
r platinum-based therapy.pdf'
-rw-rw-r--. 1 hengweiliu hengweiliu 3367 Jul 11 09:23 '140411 Table 14.4.11 Time to Response (Months) by Independent Central Review Full Analysis Set Cohort 1 Subject
s with prior platinum-based therapy.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 1925 Jul 11 09:25 '140411 Table 14.4.11 Time to Response (Months) by Independent Central Review Full Analysis Set Cohort 1 Subject
s with prior platinum-based therapy.pdf'
-rw-rw-r--. 1 hengweiliu hengweiliu 572662 Jul 11 09:23 '140412 Table 14.4.12 Treatment-Emergent Adverse Events by System Organ Class, Preferred Term, and Worst CTCAE G
rade Safety Analysis Set.lst'
-rw-rw-r--. 1 hengweiliu hengweiliu 136554 Jul 11 09:25 '140412 Table 14.4.12 Treatment-Emergent Adverse Events by System Organ Class, Preferred Term, and Worst CTCAE G
rade Safety Analysis Set.pdf'
-rw-rw-r--. 1 hengweiliu hengweiliu 466 Jul 11 09:24 lst2pdf.py
-rwxrwxr--. 1 hengweiliu hengweiliu 1971 Jun 4 23:56 pdfmerger.py
-rw-rw-r--. 1 hengweiliu hengweiliu 1174 Jul 11 09:18 rename.py
[hengweiliu@new-host lst2pdf]$
```

Step 3 combine many PDF files into a single PDF file with bookmarks

- ▶ We use a program from github
- ▶ Original author Nicholas Kim, modified by Yan Pashkovsky.
- ▶ Available at <https://gist.github.com/dhwajraj/f330ed64915ac98e29ab46966dfc326d>

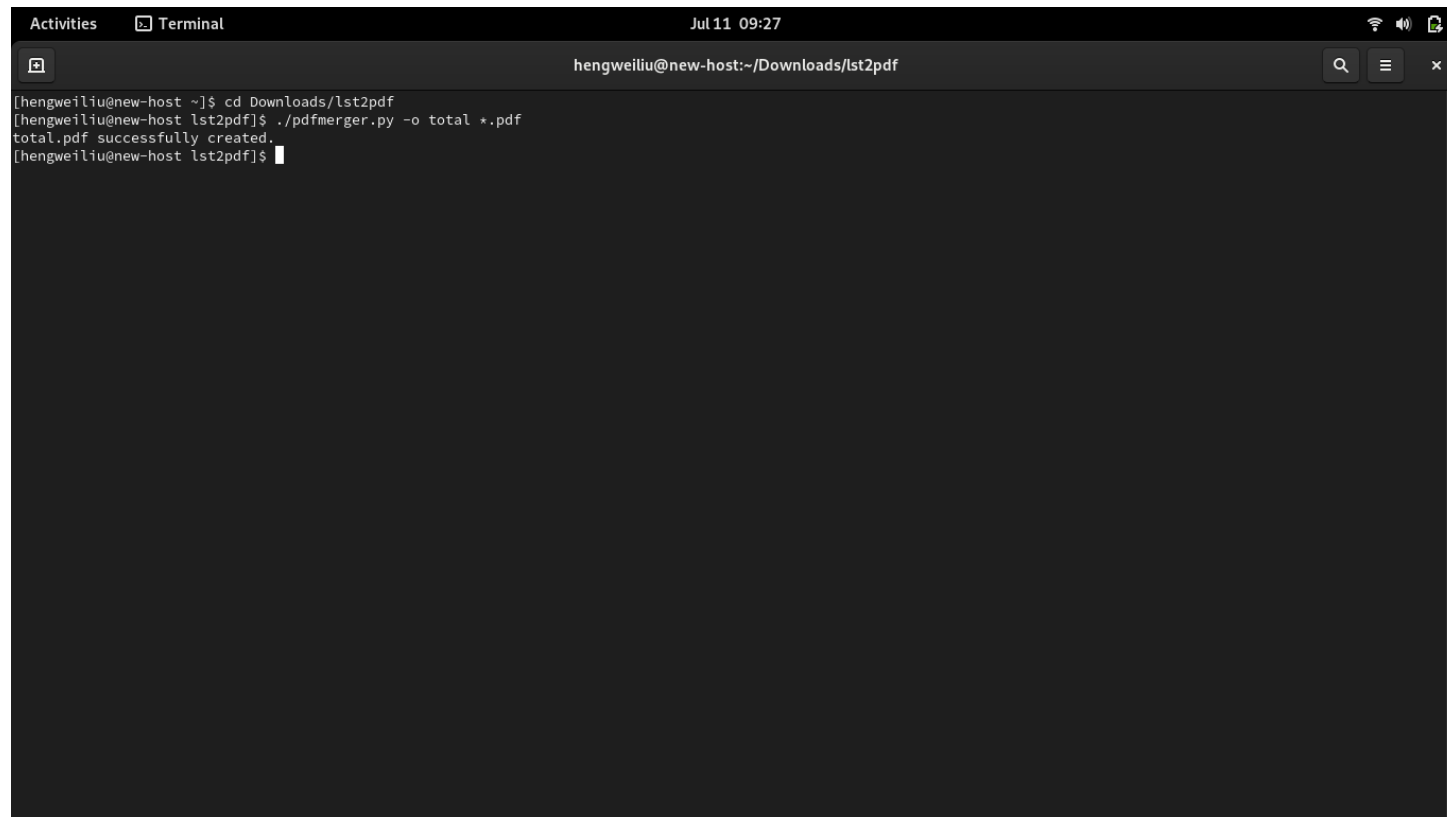
Step 3, the code that was modified

```
# merge page by page
output_pdf_stream = PdfFileWriter()
j=0
k=0
for f in files_to_merge:
    for i in range(f.numPages):
        output_pdf_stream.addPage(f.getPage(i))
        if i==0:
            output_pdf_stream.addBookmark(str(filenamees[k]).split(' ',1)[1][::-1].split('.',1)[1][::-1],j)
            j = j + 1
    k += 1
```

Step 3: the codes modified

- ▶ We modified the code to remove the numbering created in step 1 and the file extension .pdf
- ▶ Example of file name: 140101 Table 14.1.1 Maturity of DoR for Cohort 2 subjects (N=42).pdf
- ▶ The codes modified: `output_pdf_stream.addBookmark(str(filenamees[k]).split(' ',1)[1][::-1].split('.',1)[1][::-1],j)`

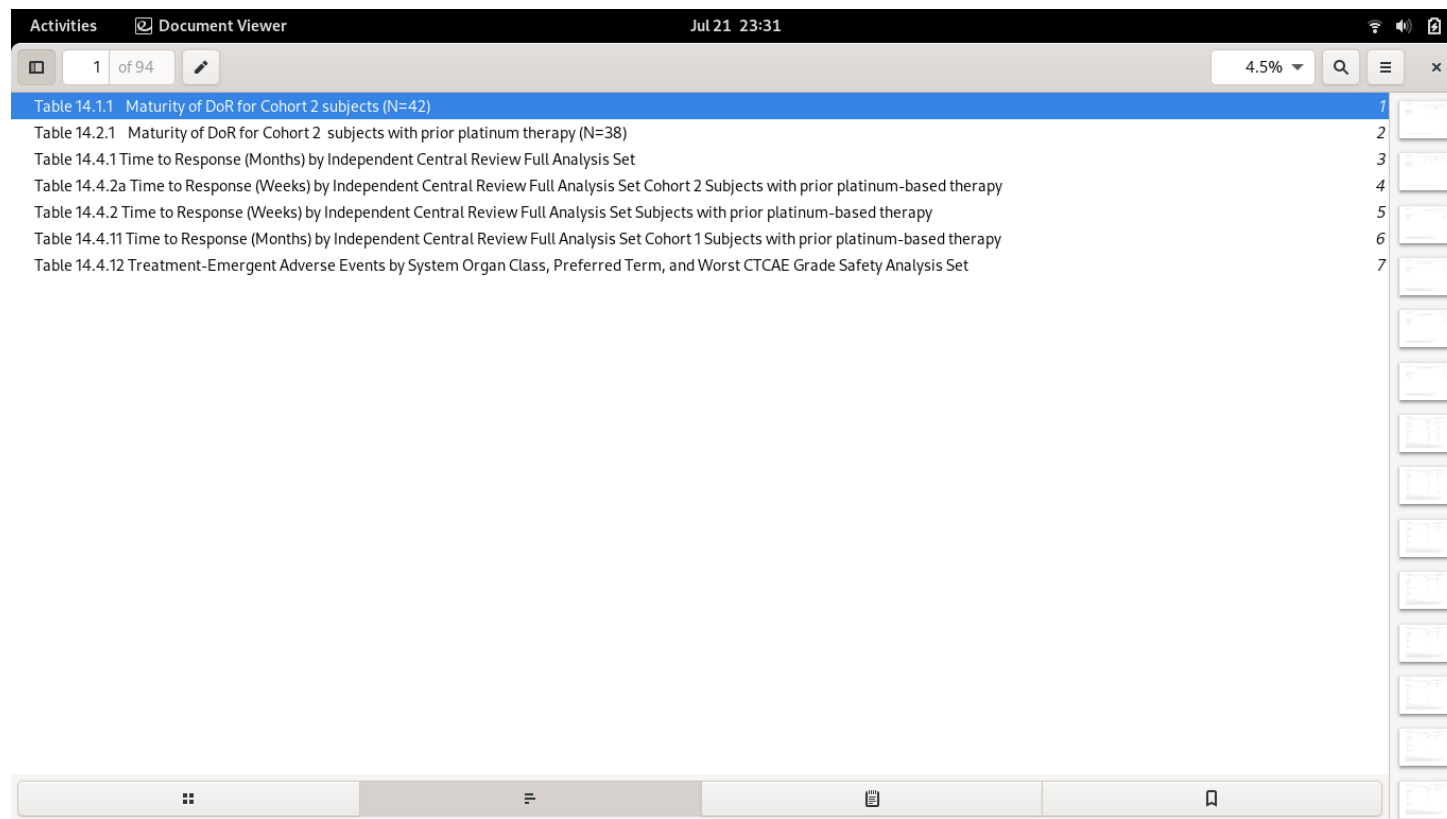
Step 3 Screenshot



A screenshot of a Linux terminal window. The window title bar shows 'Activities', 'Terminal', and the date/time 'Jul 11 09:27'. The terminal content shows a user named 'hengweiliu' at a 'new-host' machine. The user has navigated to the directory '~/Downloads/lst2pdf' and executed the command './pdfmerger.py -o total *.pdf'. The output of the command is 'total.pdf successfully created.', followed by a new prompt line.

```
hengweiliu@new-host:~/Downloads/lst2pdf
[hengweiliu@new-host ~]$ cd Downloads/lst2pdf
[hengweiliu@new-host lst2pdf]$ ./pdfmerger.py -o total *.pdf
total.pdf successfully created.
[hengweiliu@new-host lst2pdf]$
```

Sample for the final PDF file



Conclusion

- ▶ Combining multiple LST files into a single pdf file with bookmarks is an important task for programmers.
- ▶ We have shown that this can be achieved with Python in three steps.

About the Author

Hengwei Liu is director, Statistical programming at Hengrui USA in Princeton, New Jersey. He has been working in the industry since 2003. He is interested in developing tools for statistical programmers with SAS, R and Python.

Contact Information:

Hengwei Liu

Hengrui USA

400 Alexander Park

Princeton, NJ 08540

Email: Hengwei_liu@yahoo.com