# 合并多个RTF文件，转换成PDF并添加书签

Python脚本和PDFtk

**John Wang @dMed, 2016-11-12**

**jun.wang AT dmedglobal.com**

## 所需工具：

1. Python v2.7, pandas模块

安装pandas:

在命令行中，运行：pip install pandas

在PYTHON环境下运行：import pandas as pd

如果没有错误，则说明安装成功。

1. PDFtk 免费版本 <https://www.pdflabs.com/tools/pdftk-the-pdf-toolkit/>

## 步骤：

1. 生成空白TOC文件：toc.sas
2. 生成各个RTF报告文件，注意需要开启ODS RTF TOC\_DATA选项，并使用ODS PROCLABEL。dummy\_rtf\_files.sas
3. 使用combineRTF.py脚本，合并多个RTF文件及TOC文件，到一个合并后的文件，如：combined\_tables\_2016-11-13.rtf。
4. 打开这个合并后的文件，更新首页的目录（全选第一页，然后按F9功能键即可）。然后利用WORD内置的功能或者ADOBE ACROBAT，将其另存为PDF格式。此时在PDF文件首页，应该有TOC目录，并且可以点击跳转。但是没有PDF BOOKMARK书签。

另存为PDF格式的这两种方式，ADOBE ACROBAT速度更快。且ACROBAT生成的PDF文件在后面导入书签步骤时速度也更快。所以此步骤，**推荐使用ADOBE ACROBAT**。

1. 运行combineRTF.py脚本后半段，生成PDFtk格式的书签文本。中文书签的支持没有问题，需要使用HTML numerical entities字符。此步骤还可以结合外部EXCEL文件（如tocfile.xlsx），对每个RTF文件给表格标题的描述。此步骤使用了pandas模块读取Excel文件。

## 中文书签支持

## https://www.pdflabs.com/blog/export-and-import-pdf-bookmarks/

## http://stackoverflow.com/questions/24646723/convert-chinese-characters-into-xml-html-style-numerical-entities-and-into-unico

text = u"中文支持如何"

ee = text.encode('ascii','xmlcharrefreplace')

1. 运行命令行脚本import\_bookmark.cmd，将PDFtk格式的书签文本导入到PDF文件中。所有步骤完成。该脚本使用了PDFtk命令导入书签。

使用combineRTF.py脚本，示例：

comb = combineRTF()

td = datetime.date.today()

tddtc = datetime.date.isoformat(td)

comb.pathin = "C:/TEMP/"

comb.pathout = "C:/TEMP/Combined/"

comb.rtffilepattern = "class\_\*.rtf"

comb.fileout = "combined\_tables\_%s.rtf" % tddtc

comb.getfilelist()

ttfiles = comb.filelist

ttfiles.sort(key=lambda x:(int(x.split('\\class\_')[1][0:-4])))

comb.filelist = ttfiles

comb.combineRTF()

**toc = pd.read\_excel("c:/temp/tocfile.xlsx",sheetname="Sheet1")**

tdesc = dict([(f,desc) for f, desc in zip(toc.filename, toc.titledescription)])

comb.genPDFtkBookmark(pageoffset=5,titledescription=tdesc)

## 测试结果

合并200个RTF文档，用时7秒；

**更新TOC并另存为PDF文件，用时1分41秒；共404页。**

导入书签，用时3秒。

## 结论

使用PYTHON脚本结合PDFtk工具，可以完美支持中英文PDF书签的生成。而且速度上很快，程序运行时间很短。主要取决于我们操作电脑的时间和将WORD文件更新目录及另存PDF的时间。

PDFtk导入书签的方式很简单易用。在制作aCRF的书签时也可以使用该方法。

程序代码及相关文档可到Github下载：

<https://github.com/wangjun-sh/combineRTF>