PharmaSUG China 2022 – Paper AD-116

Concatenate RTF files with python

Yingnan Guo, Xin Jiang, Jiangsu Simcere Pharmaceutical Co. Ltd

ABSTRACT

In pharmaceutical company Rich Text Format (RTF) files are usually created when using SAS to generate tables, listings or figures. Usually, we need to concatenate RTF files into one file and get the directories and hyperlinks. Although the current approach we use can solve this problem, we are trying to find a more efficient approach. This paper introduces a tool created with Python, which can not only solve the problem efficiently and easily, but also solve the problem of Chinese characters.

INTRODUCTION

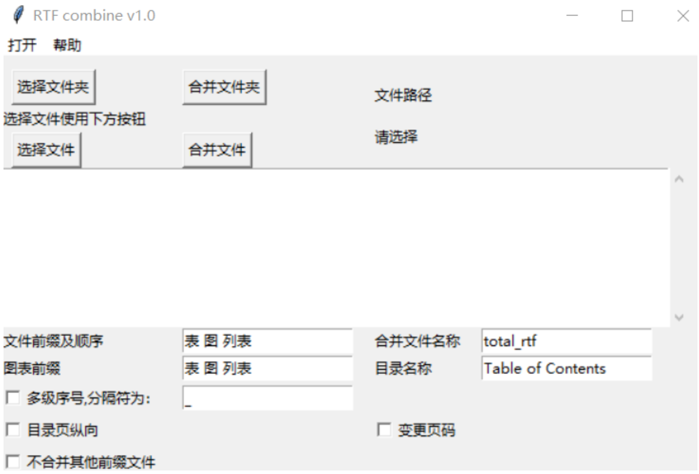
Rich Text Format (RTF) files are usually created in statistical reports using ODS RTF with SAS software. Usually there are many individual RTF files within one project. However, in daily work we need to concatenate all tables, listings and figures together for reviewing or archiving. There are several methods can handle this work like Visual Basic, SAS macro and so on. Each of them has have their limitations. For example, Visual Basic program is time consuming to handle a large number of files, while SAS macro need to iterate through all the original SAS programs for relevant information and many parameters need to be predefined. There, we looked for a way to concatenate RTF files easily and efficiently with Python.

Python is a popular programming language. It has many libraries called modules, which can help us deal with many affairs in life. Python can deal with RTF file itself directly, which is very efficient. There have been some papers introduce how to deal with RTF code, however, few papers have pointed out how to deal with Chinese word in rtf files. Here we provide a simple tool to concatenate rtf files easily with different request, which was created with Python.

APPROACH

This tool was constructed mainly with "tkinter" module in Python, which allows the tool to have a clean, interactive interface which makes it easier for users. The interface of the tool is shown in the figure below (Figure 1 shows Chinese version), all the wording in the interface can be easily changed, also more buttons or labels can be easily added by designer. This tool currently provides functions such as files choosing, files order assigning, and other basic options for concatenated file. This tool was packaged into an executable file using "pyinstaller" module, allowing people whose computer without Python to use. The detailed working process of this tool is as followed:

Figure 1 The Interface of the Tool Made by Python.



文件前缀及顺序 表 图 列表

合并文件名称 total_rtf

图表前缀 表 图 列表

目录名称 Table of Contents

□ 多级序号,分隔符为: _

□ 目录页纵向

□ 变更页码

□ 不合并其他前缀文件

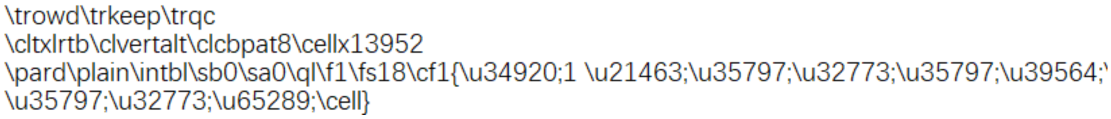
CHOOSE FILES NEED TO BE CONCATENATED

There are two buttons in the tool that user can choose folder or files. The Python module "*tkinter.filedialog*" provides some functions to help user to get filename list. The function "*askdirectory*" can help user to choose a folder which can later help us to get all files under this path. The function "*askopenfilenames*" can help user to choose a series of files we wanted to concatenate. Meanwhile, user can reorder the filename list in the interface which will affect final merged RTF file output.

EXTRACT USEFUL INFORMATION FROM RTF FILES

Python can extract information in RTF file similarly as .txt file with "open" function. Meanwhile, RTF files with similar structure makes it easy to extract information. This tool used predefined text to help extract useful information by using regular expression in Python. In RTF files, Chinese words were converted to special encoding form. For example, "表" in Chinese maybe demonstrated as "\u34920; " in a rtf file. "\u34920;" can split to "\u", "34920" and ";" in which "\u" represents the text has been transformed to Unicode code form and "34920" can be converted to "8868" in hexadecimal and ";" represent for ending. "\u8868" is the Unicode encoding form representing then Chinese character "表". Since we have the ability to translate all code into a readable expression, then we can easily extract useful information like title and page number and so on. For example, the regular expression "r'{\u34920;.*?}" can help us to find the tile in a RTF file with code in Figure 2. With similar way, we can get other information like page number, TFL type and any other information we want.

Figure 2 An Example of Title in RTF File.



```
\trowd\trkeep\trqc
\cltxlrtb\clvertalt\clcbpat8\cellx13952
\pard\plain\intbl\sb0\sa0\ql\f1\fs18\cf1{\u34920;1 \u21463;\u35797;\u32773;\u35797;\u39564;\
\u35797;\u32773;\u65289;\cell}
```

MODIFY AND CONCATENATE FILES

There has been a paper reported how RTF files can be concatenated (Reference 1), so we would not show the details in this paper. In our tool, title and page information can be got by regular expression with Python as described above, which makes the tool with better applicability. Some documents (Reference 2) provide how various encodings work in RTF files which can help us understand the encodings of RTF files. With the understanding of encodings and the help of Python, we can easily modify the merged RTF file's format. Also, we can add table of contents and bookmarks and so on in the merged file.

CONCLUSION

This tool is very convenient for everyone to use as user can change the predefined parameters easily. Also, the flexibility of Python module makes is easy to add more custom functions in this tool if needed. The important advantage of the tool is that it can quickly deal with large number of files. When testing, 800 files can be concatenated in less than 10 seconds, which largely enhances our work. Besides this tool is language friendly as it can deal with not only English file, but also Chinese characters or other special characters.

Now we are exploring how to get the title when it is embedded in the image, maybe the OCR technology provided by Python is a proper way.

REFERENCES

1.Zhiping Yan. 2015. "A Fully Automated Approach to Concatenate RTF outputs and Create TOC." PharmaSUG China 2015.

2."Rich Text Format(RTF) Version 1.5 Specification". http://www.biblioscape.com/rtf15_spec.htm

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Yingnan Guo
Jiangsu Simcere Pharmaceutical Co. Ltd
+86-18210096610
gyn734374927@126.com