

Word2PDF: A Python Tool to Convert and/or Combine RTF Files into a Single Bookmarked PDF/RTF with Table of Contents in a User-defined Order

Xianhua Zeng, Taimei Intelligence Pharmaceutical, Shanghai, China

ABSTRACT

As part of clinical trial reporting, a large number of RTF/PDF outputs are created, and at the completion of a milestone in a study, it is required to convert and/or combine all reports into a single, reviewer-friendly document for easy delivery and review. This paper presents a Windows desktop application built with Python to convert and/or combine multiple RTF/PDF files into a single bookmarked PDF/RTF/DOCX document with a table of contents in a user-defined order. The application, called Word2PDF, runs on Microsoft® Windows 10 or later with Microsoft Word installed. It also supports DOC and DOCX files as input. Key features include navigation heading insertion and per-section page numbering.

The app and test files are available on my GitHub:

<https://github.com/XianhuaZeng/Projects/tree/master/Word2PDF>

INTRODUCTION

In the pharmaceutical industry, statistical analyses are presented in multiple SAS® outputs (i.e., tables, listings, and figures (TLF)). These results are created from multiple SAS programs and are delivered to medical writers and clients for review. It may be impractical for clients to handle each output file individually, especially when there are many outputs. It is often more convenient for the client to review results in a single file rather than multiple files. This single file can be created in different user-friendly formats such as Microsoft Word, HTML, and Portable Document Format (PDF). This paper explores a scenario that requires creating a single bookmarked PDF file with a table of contents in a user-defined order.

Several methods have been developed to handle this task over the past decade, but each has its own drawback. For example:

1. PDF editor approach (Adobe Acrobat or Nuance's Power PDF): both are not free and support only adding filename as bookmark.
2. SAS approach: it must be executed in environment with SAS installed. Users are unable to use it without SAS programming knowledge.
3. Other programming language approaches: similarly, knowledge of programming skills is needed to use these approaches.

This paper introduces an easy-to-use Python application to convert and/or combine any type of RTF file into a single bookmarked PDF/RTF/DOCX document with a table of contents in a user-defined order.

WHY PYTHON?

Python is an open-source, general-purpose, module-based programming language. The base installation of Python offers powerful tools for basic programming, and many third-party libraries further extend its capabilities. Python can interpret different file types, including PDF and Microsoft Word files—file types that SAS cannot fully process in a meaningful way (Nathan Hammett Kosiba, 2019).

Python is also used to build user interfaces, making tools available to a broader audience. User interfaces in Python allow non-programmers to use tools without any coding knowledge. Tkinter is the standard GUI library for Python; it provides a fast and easy way to create GUI applications. Figure 1 shows the GUI of Word2PDF.

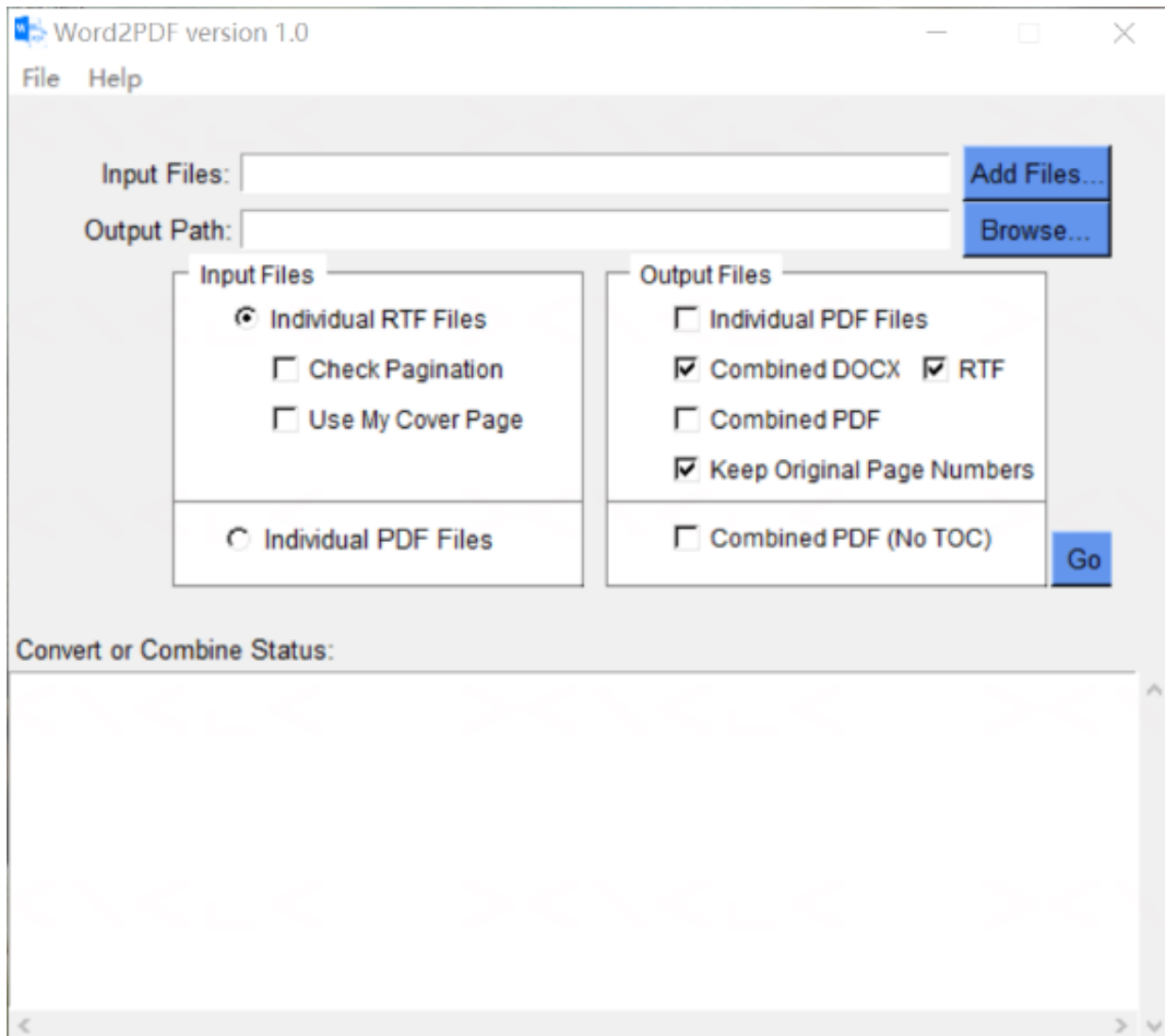


Figure 1. GUI of Word2PDF

FUNCTIONALITIES

Currently there are eight functionalities in this app:

1. "Check Pagination": checks for pagination issue. Note that usually the pagination issue is Physical Page > Logical Page. So Logical Page > Physical Page is fake positive, which can occur when an RTF section break is set to "Continuous" instead of "Next Page". It is strongly recommended to check pagination issue after combining to RTF.
2. "Use My Cover Page": inserts a cover page.
3. "Individual PDF Files": serves two purposes. As a standalone convert-only operation, it converts Word/RTF files to individual PDFs. When used together with "Combined PDF (No TOC)", it retains the individual PDF files after combining.
4. "Combined DOCX": controls whether Combined.docx is created. The tool creates Combined.docx when this option is checked (default). Unchecking it skips DOCX creation, which can save significant time when combining large files.
5. "RTF": controls whether Combined.rtf is created. Checked by default; uncheck to skip RTF creation

and save processing time.

6. "Combined PDF": controls whether Combined.pdf is created. Checked by default; uncheck to skip PDF creation.
7. "Keep Original Page Numbers": preserves the per-document "Page X of Y" pagination from each individual RTF file in the combined output. Without this option, Word rennumbers pages consecutively across the whole document. See the Methodology section for details.
8. "Combined PDF (No TOC)": combines existing PDF files (or RTF files converted to PDF) into a single bookmarked PDF without generating a table of contents. Suitable for ad-hoc deliveries or when the TOC is not required.

METHODOLOGY: NAVIGATION HEADING INSERTION

SAS-generated RTF files do not always carry a navigation bookmark. Word2PDF detects each file's structure and inserts an invisible navigation heading so every section appears in the combined document's navigation pane and table of contents. The inserted paragraph uses the RTF control sequence:

```
\pard\lsN\pagebb0\sb0\sa0\sl1\slmult0\fl\fs1\cf8\outlinelevel1 {title}\cf0\par
```

The \sN token matches the Heading style number found in the first navigation-enabled file in the batch, so inserted headings align with any pre-existing navigation entries at the same level. Six structural patterns are recognized, evaluated in the order below.

Case	Trigger Condition	Action
Case 1	Stylesheet has Heading style applied, or paragraph has \outlinelevel	File is left unmodified; existing navigation is preserved.
Case 2	Title is a \qc body paragraph outside any table (no \intbl). Typical: t_14 series.	Inserts separator row (full-width bottom-border cell) before column header; places hidden heading between them; removes \clbrdr from column-header row.
Case 2c	Title is in a bold \intbl cell of a single-column table with no \trhdr.	Same as Case 2.
Case 3a	Title in page header; body table has two \trhdr rows (separator + column header). Typical: Boehringer format.	Inserts hidden heading between the two \trhdr rows.
Case 3b	Title in page header; body table has one \trhdr row or none. Typical: figure files.	Inserts hidden heading immediately after the \bkmkend IDX bookmark. If no \footer block exists, injects an empty footer to prevent inheritance from the prior section.
Case 3c	Page header is empty; title is composed of bold centred \intbl rows in the body.	Concatenates bold-row text as title; then follows Case 3b insertion logic.

Table 1. Structural Patterns

METHODOLOGY: RTF CONCATENATION ALGORITHM

RTF is a plain-text format in which all content and formatting are expressed as control words enclosed in curly braces. Because all RTF files share the same hierarchical grammar, multiple files can be merged into one document by reusing the header from the first file and appending content sections from the rest.

RTF FILE STRUCTURE

Every RTF file consists of three regions:

1. Opening section – font table, color table, stylesheet, and document-wide settings. Begins with {\rtf1 and ends just before the first \sectd.
2. Content section – one or more sections (\sectd) containing the actual tables, paragraphs, and figures.
3. Closing section – a single } that terminates the outermost RTF group.

MERGING RULES

Files are processed in the order defined in Bookmark.xls when that file is present; otherwise files are sorted by type (tables first, then figures, then listings) and by number within each type. Before writing begins, font tables from all files are extracted and merged into a single \fonttbl block. To avoid conflicts, each file's font references are remapped: the n-th file's \f0, \f1, ... are renamed to \fn00, \fn01, ... (e.g. file 1 uses \f100, \f101; file 2 uses \f200, \f201). The rule for writing content is:

1. First file: write the merged opening section (including the combined font table) + content section into the output buffer.
2. Every subsequent file: discard the opening section, then replace its leading \sectd inline with \pard\sect\sectd (which embeds the section break directly into the existing line rather than prepending a separate line), then append the content section.
3. After all files: append a single closing }.

METHODOLOGY: KEEP ORIGINAL PAGE NUMBERS

When multiple RTF files are merged, Word rennumbers pages consecutively across the whole document. The Keep Original Page Numbers feature restores per-document "Page X of Y" pagination by rewriting footer field codes in each file before assembly, using RTF bookmark references as page-offset anchors.

PAGE NUMBER FIELD REWRITING

Once the footer line is located, the tool replaces the standard PAGE / NUMPAGES field codes with computed RTF field expressions. The expressions implement the following arithmetic:

Current page within this document:

```
PAGE - PageRef(<start_bookmark>) + 1
```

Total pages in this document:

1. For every document except the last:

```
PageRef(<next_bookmark>) - PageRef(<start_bookmark>)
```

2. For the final document in the combined file:

```
NUMPAGES - PageRef(<start_bookmark>) + 1
```

PageRef is a built-in RTF/Word field that returns the physical page number of the page on which a named bookmark appears. Because bookmarks are placed at the first page of each section, `PageRef(<start_bookmark>)` gives the absolute page number where the section begins, and the subtraction yields a document-relative offset.

TESTING CONVERTER/COMBINER

To run the convert and combine function:

1. Prepare Bookmark.xls: Create an Excel file named Bookmark.xls in the same folder as the input Word/PDF files. The file uses a two-column layout:
 - Column A: output identifier – the RTF/DOCX/PDF filename stem (e.g., Table_1_1) used to match and order the input files.
 - Column B: bookmark title – the label displayed in the PDF bookmark panel (left-side navigation) when using the "Combined PDF (No TOC)" function (e.g., Table 1.1: Demographic

Characteristics). This column is not used by the "Combine with TOC" function, which derives navigation headings directly from the RTF file content.

This step is optional for "Combined PDF (No TOC)". If Bookmark.xls exists, files are combined in the order listed in Column A. If not found, files are sorted by type (tables first, then figures, then listings) and by number within each type. The bookmark list is normally provided by the statistician before TFL production begins. If it has not been provided, BookmarkGen.py (available on GitHub, see above) can generate it automatically by scanning the RTF output files and extracting TFL titles.

Output Identifier	Title
t_1_1_1.rtf	Table 1.1:1 Disposition of Patients (Randomized Set)
f_4_3_3_1_1.rtf	Figure 4.3:3.1.1 Systolic Blood Pressure by Visit - Treatment A (Treated Set)
l_1_1_1.rtf	Listing 1.1:1 Disposition of Patients (Screened Set)

Figure 2. Bookmark File

2. Select the RTF/PDF files to be converted and/or combined using the "Add Files..." button.
3. Specify the output path where the combined file will be generated using the "Browse..." button.. If left blank, the file will be saved in the input file directory.
4. Choose the desired output options (Combined PDF, Combined DOCX, Combined RTF, Keep Original Page Numbers, etc.).
5. Click the "Go" button.

If everything runs successfully, the execution status is displayed in the status panel. Figure 3 shows the GUI after a successful combine operation.

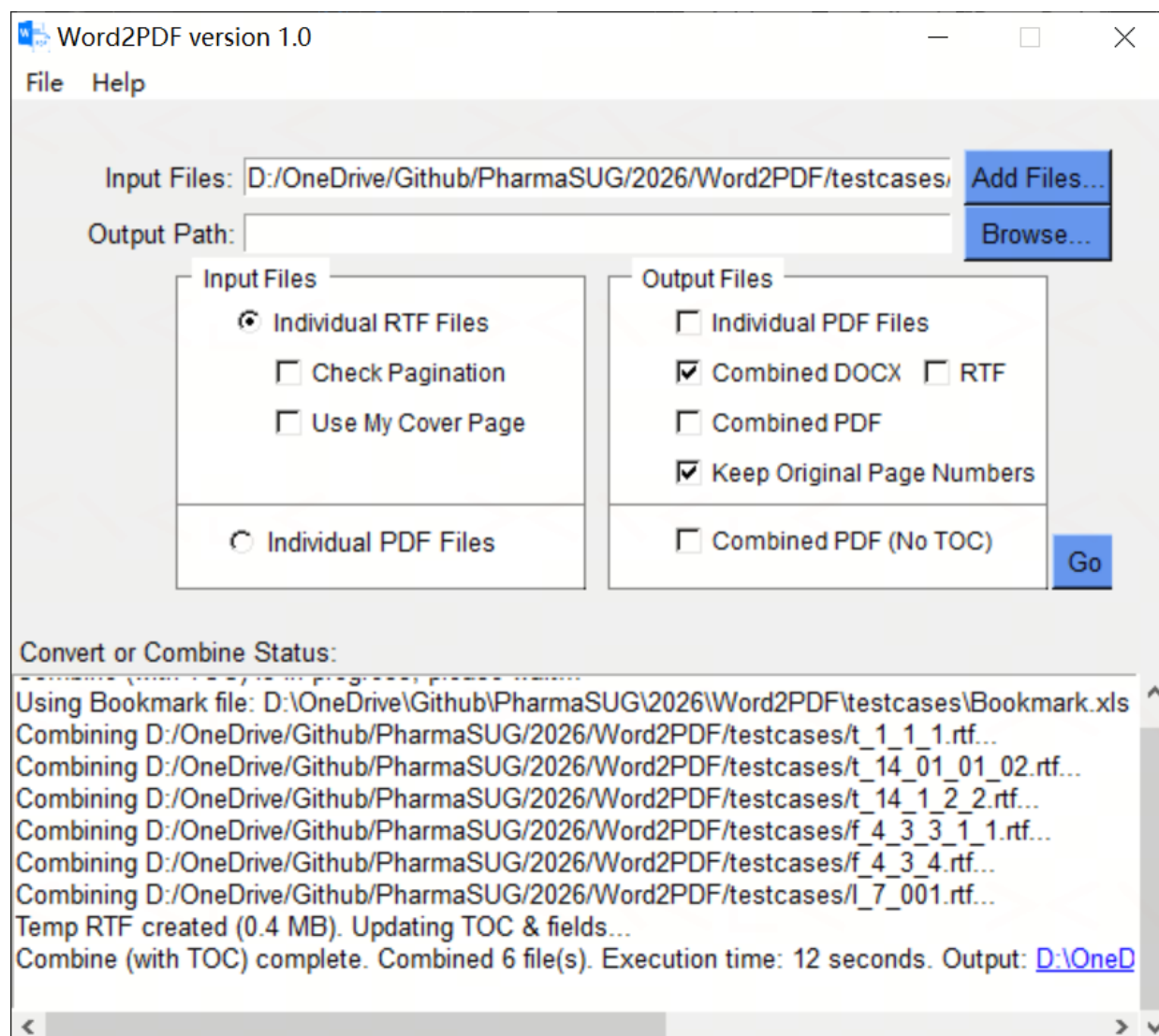


Figure 3. Execution Status of Word2PDF

Below are some test cases and results.

Test case	Case Description / Findings
Deliverable run - ranging from 6KB to 960KB	101 outputs in ~5 mins to create Combined.rtf and Combined.pdf
Large output run – ranging from 81KB to 75MB	4 outputs in ~12 mins to create Combined.rtf and Combined.pdf
Medium to Large output - 35MB	1 output in ~3 mins
Large output - 48MB	1 output in ~5 mins to create Combined_noTOC.pdf
Large Output - 75MB	1 output in ~7 mins to create Combined_noTOC.pdf
Large Output - 98MB	1 output in ~9 mins to create Combined_noTOC.pdf
Huge output run – ranging from 940KB to 499MB	4 outputs in ~1 hours, 17 mins to create Combined_noTOC.pdf

Table 2. Testing Result of Word2PDF

CONCLUSION

This paper presents an easy-to-use tool to convert and/or combine multiple RTF/PDF files into a single bookmarked PDF/RTF/DOCX document with a table of contents in a user-defined order. It runs on Microsoft Windows 10 or later without requiring SAS, though Microsoft Word must be installed. Non-programmers can use the application without any coding knowledge. Bookmark information is maintained in an external two-column Excel file (Bookmark.xls), making it simple to reorder or relabel sections.

Two key enhancements are described in the Methodology section. Navigation Heading Insertion automatically detects the structure of each RTF file and inserts an invisible heading so every section appears correctly in the combined document's navigation pane and table of contents. Keep Original Page Numbers rewrites the footer field codes in each constituent file before assembly, so each section retains its own standalone "Page X of Y" footer—a common requirement in clinical trial deliverables.

One known limitation: the Combine (with TOC) operation requires the combined RTF file to be under 512 MB, due to a constraint in Microsoft Word. Files exceeding this threshold can still be combined using the "Combined PDF (No TOC)" option.

REFERENCE

Guido van Rossum. "Python". Available at <https://www.python.org/>

Fredrik Lundh. "Tkinter". Available at <https://docs.python.org/3/library/tk.html>

Nathan Hammett Kosiba. 2019. "Automate the Mundane: Using Python for Text Mining". Available at <https://www.pharmasug.org/proceedings/2019/AD/PharmaSUG-2019-AD-129.pdf>

CONTACT INFORMATION

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

Name: Xianhua Zeng

Enterprise: Taimei Intelligence Pharmaceutical

Address: 6th Floor, Building 24, Phase 3, Caohejing Technology Oasis, No. 1999 Yishan Road, Minhang District, Shanghai, China, 200233

E-mail: xianhua.zeng@taimei.com

Web: <https://www.taimei.com/>

<http://www.xianhuazeng.com/>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.