# tPMCrafty Getting Started Guide

*tPMCrafty* is a corpus "handcrafting" tool that allows you to open text files, perform some transformations and then output the results as plain text with UTF-8 and/or UTF-16 file encoding. It is a separate application from *The Prime Machine*, but was designed with *tPM*'s DIY Text Tools in mind.

Some of the things *tPMCrafty* can do:

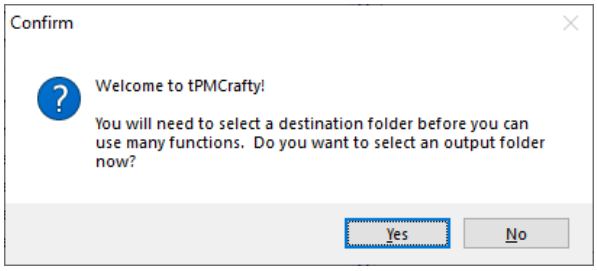
- Import PDF, DOCX, DOC, PPTX, PPT, RTF, EPUB and TXT files;
- Standardise line spacing;
- Remove HTML and XML tags;
- Split files;
- Process ebooks, splitting chapters into separate files;

tPMCrafty is available for Windows 64 bit and MacOS.

## Section 1: Start up and the output folder

### 1.1 Starting up

When you start *tPMCrafty* it will prompt you to select a destination folder for your processed files. This should typically be an empty folder where you want the new text files to be saved.

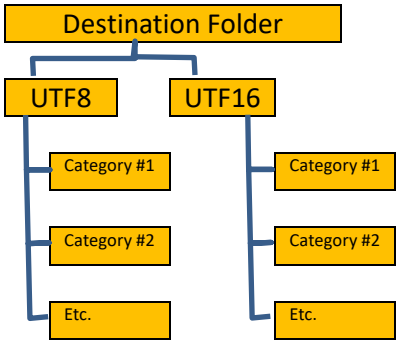



You can change the destination folder later by clicking "Select main corpus folder…" under the TEXT(S) OUT block on the right of the main application window.

### 1.2 Destination Folder

When you save your processed texts, they will be saved in folders inside the Destination Folder.

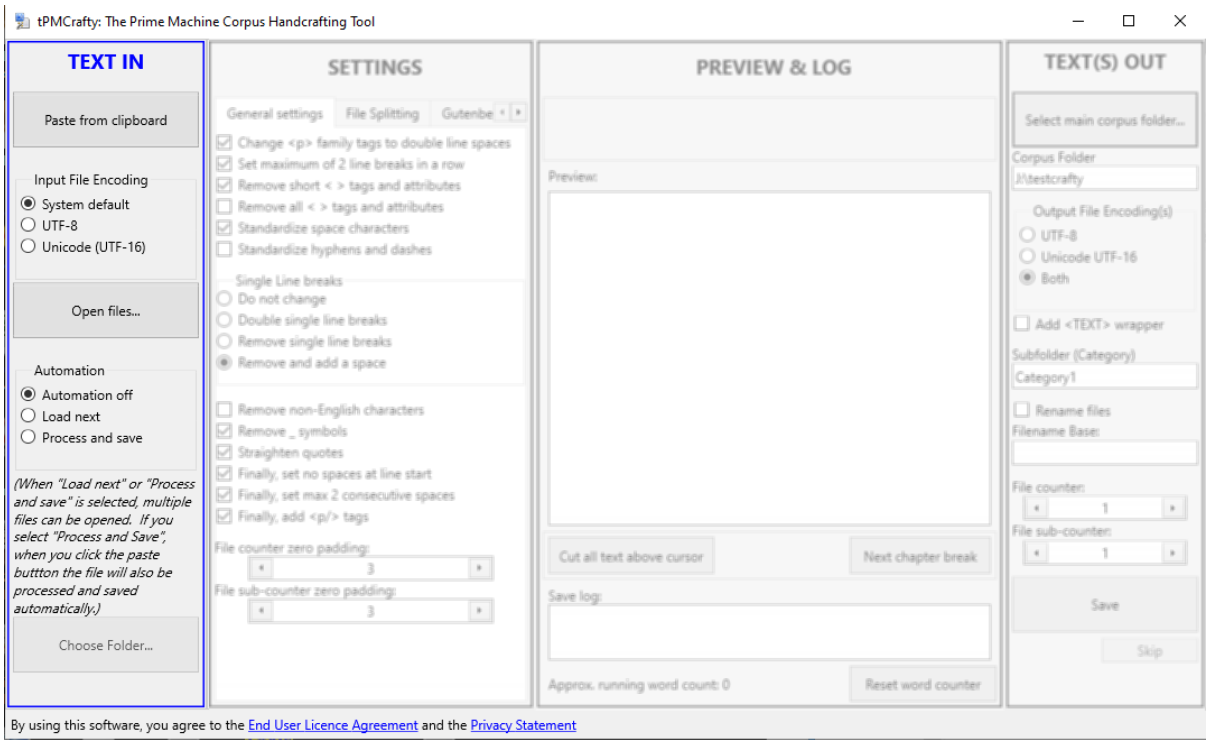
The path will look like this:




See Section 5 below for details about Category Subfolders.

## Section 2: Loading text(s)

The left-hand block contains the buttons to load the text files.




The **Paste from clipboard** button is easy to use. Simply click on it and any text in the clipboard will be processed by *tPMCrafty*.

If you load TXT files from specific files or a folder, you need to select the correct **Input File Encoding**. The other file formats do not use this setting.

The **Automation** options allow you to quickly process multiple files with the same text transformation settings.

- If you select **Load next** or **Process and save**, you will be able to select multiple files using the **Open files..** or **Choose Folder…** buttons. *tPMCrafty* will add these files to a queue.
- If you select **Load next**, the next file in the queue will be loaded and processed immediately after you click the **Save** or **Skip** buttons on the right-hand TEXT(S) OUT block. See Section 5 for more details.
- If you select **Process and save**, pasted text will be automatically processed and saved and if you select multiple files all of these will be processed and saved automatically.

Using **Process and save** is especially useful if you are copying and pasting multiple pages (e.g. from the internet or from another application). You can copy text from the other application, click the **Paste from clipboard** button and *tPMCrafty* will automatically process it and create a new text file.

## Section 3: Processing text(s)

### 3.1 General Settings

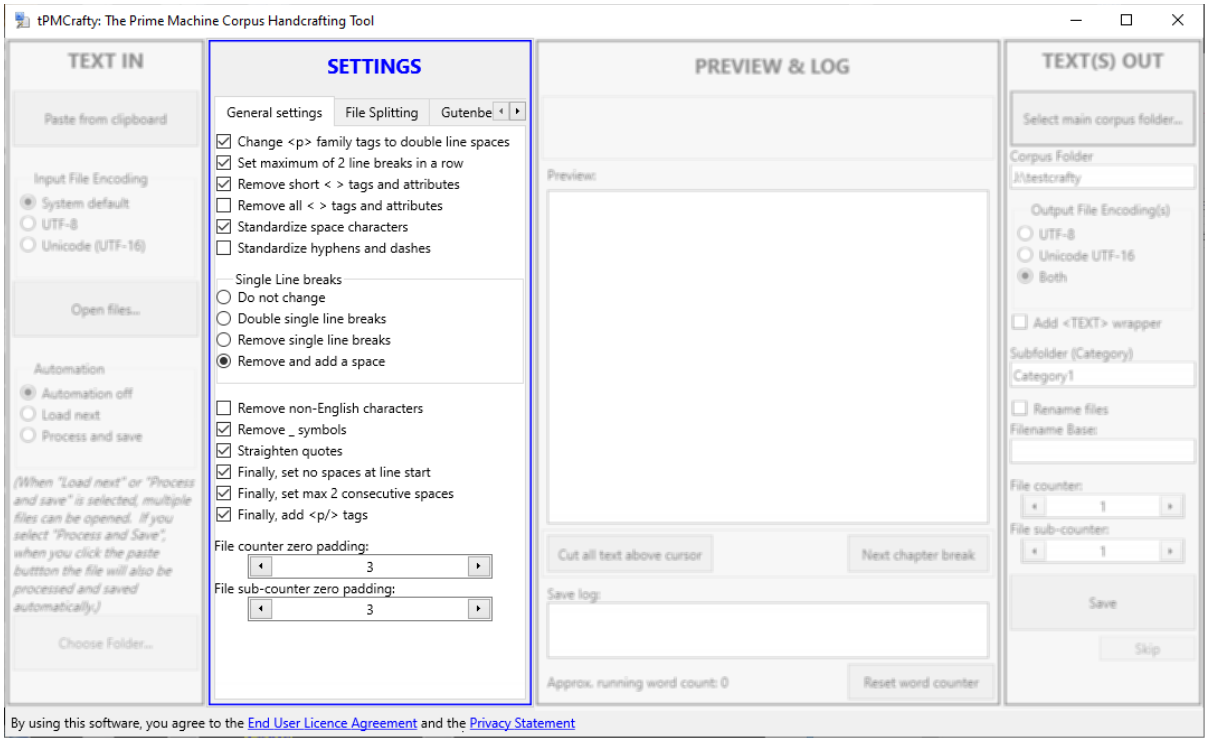
Each time you start *tPMCrafty* a number of text transformation processes will be automatically selected. Each time you tick or untick an option, the text will be re-processed. You will be able to see the effects in the **PREVIEW** box (in the third block).

Depending how you obtained the text files, line breaks can be problematic.

- Some files may have new line breaks at the end of each line of text; you may want to remove these single line breaks so paragraphs flow properly.
- Texts converted from PDF file may have new line breaks at the end of each line and not have spaces; you may want to remove these single line breaks and **add a space** so paragraphs flow properly and words from the ends of lines do not run into one-another (e.g. avoidthis).
- Some text files may have a single line break for each paragraph; you may want to **Double single line breaks** to create clearer-looking paragraphs.

If you have loaded text from HTML or XML, you may want to make use of the <p> tags (to break paragraphs) but you may not want to output the others for your corpus. Several options relate to this.

Note: the last option **Finally, add <p/> tags** will allow software like *tPM* to identify paragraph breaks easily. You may want to switch this option off for some other corpus tools.



### 3.2 File Splitting

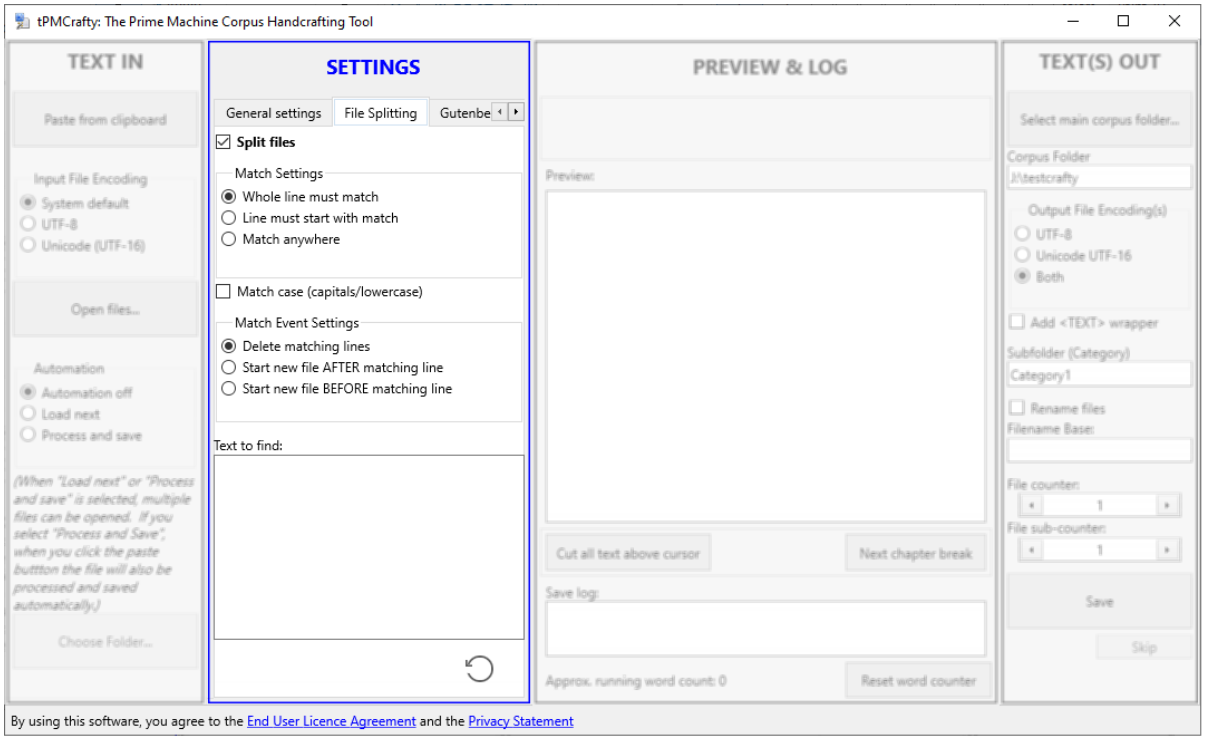
The File Splitting tab allows you to find a specific string of characters (or several strings) and split a single text file into multiple text files. If these settings are activated, *tPMCrafty* will create multiple files for each text file, numbering these using a sub-counter.

First, you need to decide whether the whole line must match (with no additional characters), or whether the string should occur at the beginning of a line, or whether the match could occur anywhere.

For example, if you want to split a file each time the string “Section ” occurs at the beginning of a line, you could select the second option.

The **Match Event Settings** allow you to decide whether the string that makes a match is removed when the file is split, or whether it should be the last line of the previous section or first line of the next.

You can enter different strings to match in the **Text to find:** box. This is the only setting that does not automatically re-process the text as you change it. After you have changed the strings, you can click the refresh button to apply the changes. The new strings will also be used for any future text processing events.

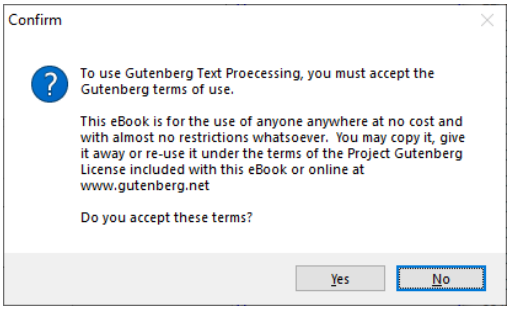



In version 1.06 there is a new tab called “Cut Strings” which looks similar; with the new tab you can simply remove lines from the text matching, containing or starting with a specific string.

### 3.3 Ebook Settings

*tPMCrafty* provides some helpful ways to transform ebooks into plain text for your corpus.

These features are intended to help student researchers create corpora for their own personal use.




If you select the option to clip the Gutenberg header/footer, you will be bound by the Gutenberg terms of use associated with the ebook files you have downloaded.

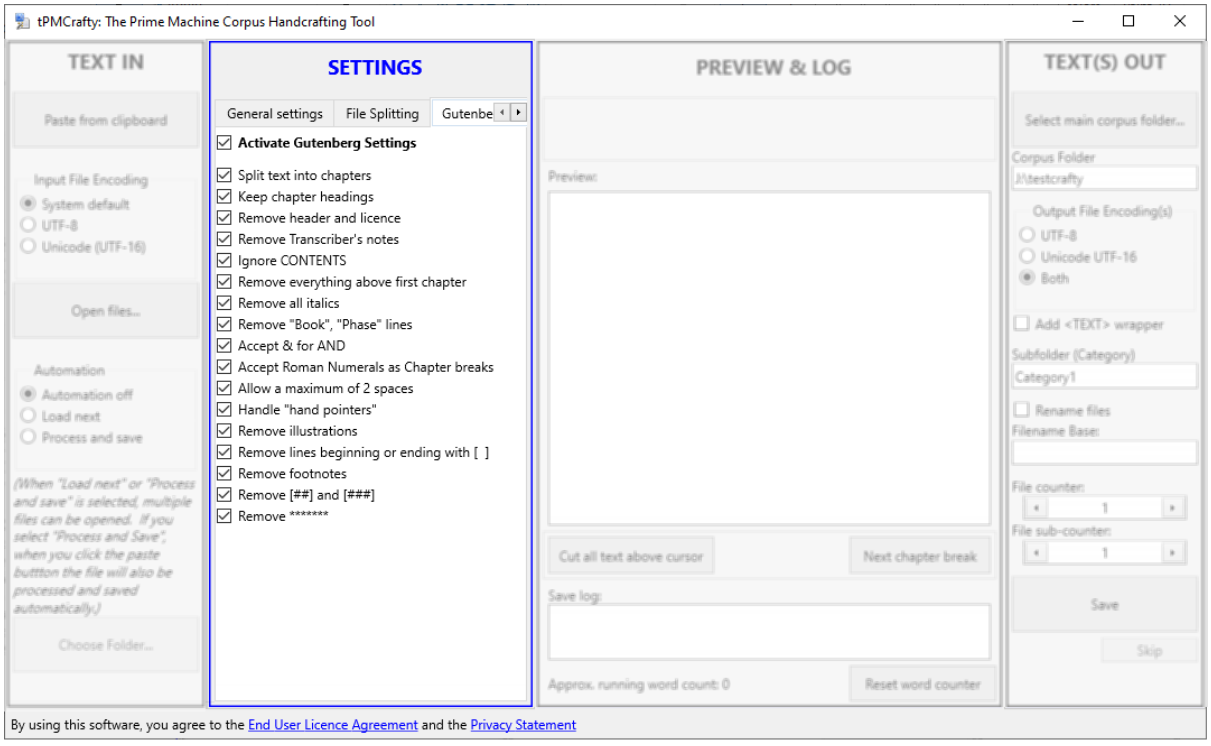
**Please respect the Project Gutenberg restrictions and do not share your processed files inappropriately!**

**For more information about Project Gutenberg see their website: [http://www.gutenberg.org/](http://www.gutenberg.org/)**

*tPMCrafty* will work well with many Gutenberg ebooks, correctly identifying divisions between the header and the first chapter, chapter breaks and the start of the Gutenberg ebook licence. However, due to the vast number of ebooks available, different formatting conventions in different ebooks may mean these are not correctly identified. See Section 4 for details on how to complete manual checks.

If you want to export a complete ebook (without splitting it into multiple text files with one text file for each chapter), you should untick the first option **Split text into chapters**.

As with the general settings, you can experiment with the different options and see a preview in the preview box in the third block.



### 3.4 Load/Save Settings

If you want to save all the settings to use again in the future, you can use the last tab until **SETTINGS**. This will allow you to save the status of the tick boxes and text boxes to load again.

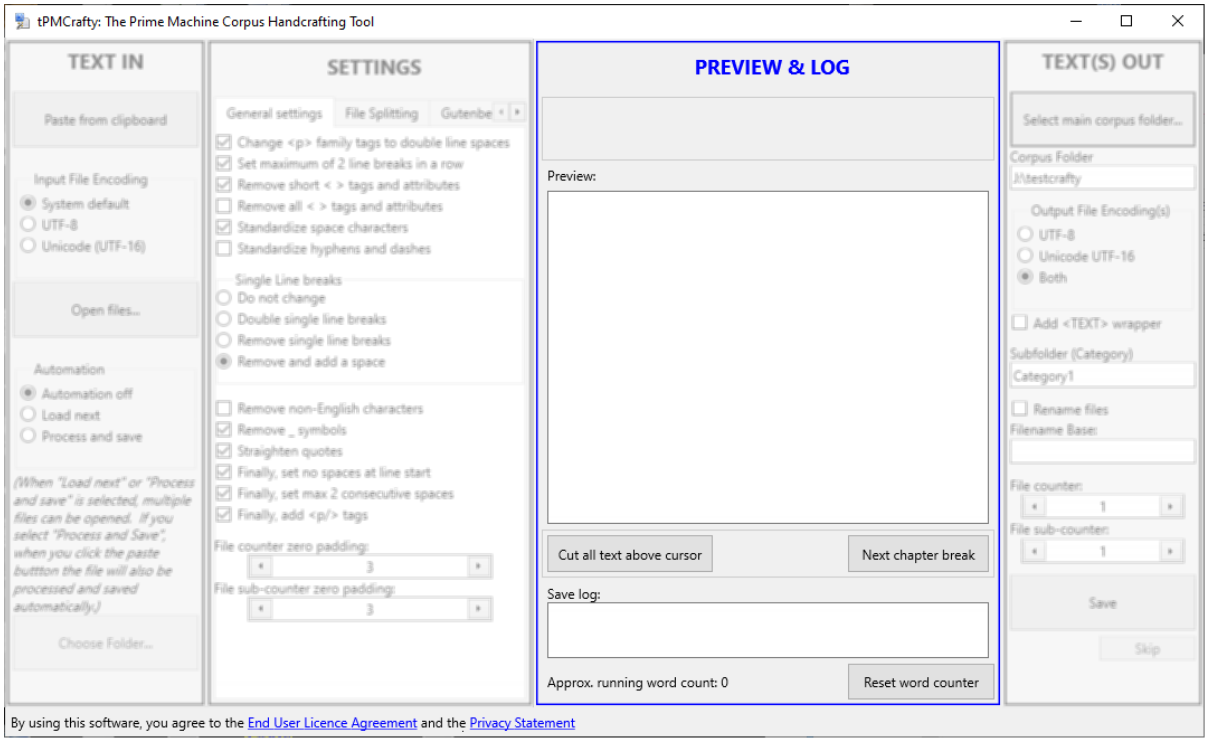
## Section 4: Manual editing and logs

The third block is headed **PREVIEW & LOG** and this allows you to see how the text has been transformed.

You can edit the text in this box before saving.

The **Cut all text above cursor** will delete everything before the cursor and can be useful if a text file has a header or some additional unwanted text at the beginning.

If you have used the **File Splitting** or **Gutenberg** settings, the **Next chapter break** will move the cursor to the next point in the file that has been identified a place to split the output file. The codes you see such as **#NEWTEXTFILE#** can be manually deleted if you do not want to split the file where indicated.




At the bottom of the screen, a **Save log** shows a date-time stamp along with filenames and estimated word counts.

**Reset word counter** simply resets the word counter to zero. This section is for information only and does not affect the output files themselves.

## Section 5: Saving texts

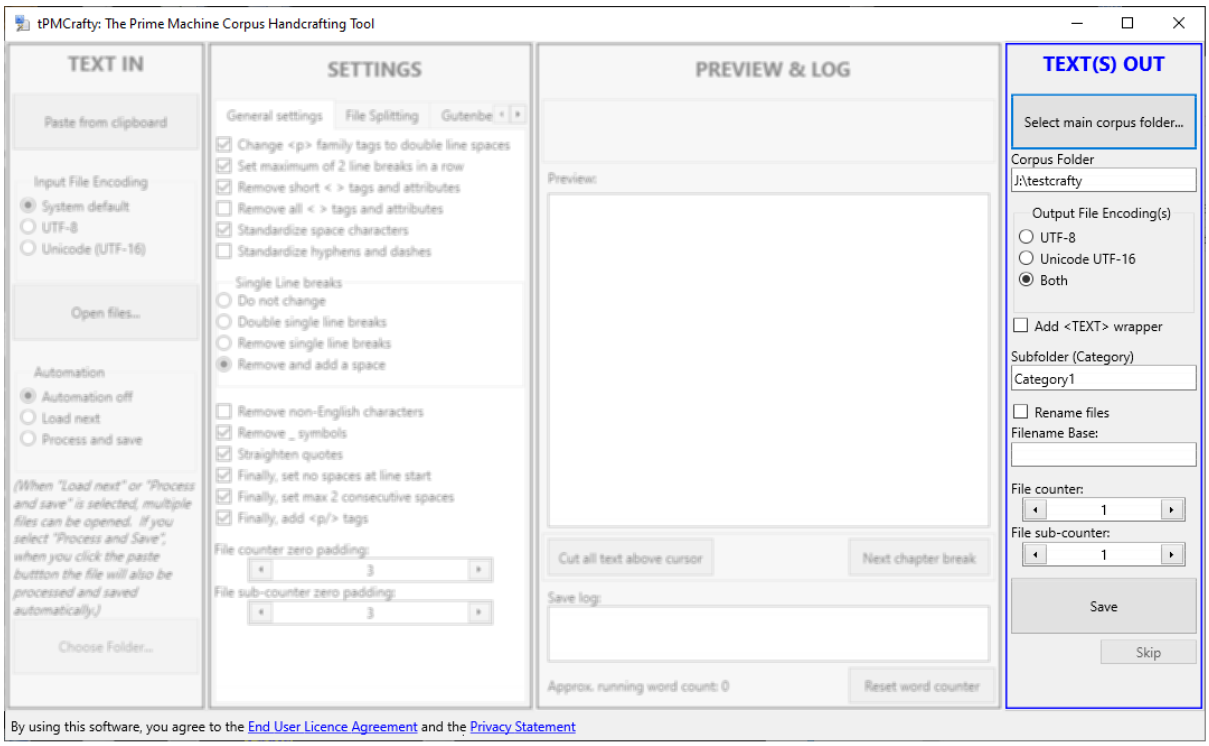
The last block contains settings related to the output of the text.

*tPMCrafty* has been designed to help prepare texts for *tPM* (which will accept UTF-8 or UTF-16) as well as some other popular corpus tools which may need (or prefer) either of these formats.

For some applications, you may need to add <TEXT> and </TEXT> around the file. This can be particularly useful for certain taggers.

The Subfolder (Category) is a text box where you can enter the name of a new folder in which to store your corpus files.

If you select the **Rename files** box, you can enter a new base for the filenames. The **File counter** number will be automatically added onto the end of the text files. If a text file is split into multiple files, the **File sub-counter** number will also be added onto the end of the new filename.




Finally, the **Save** button takes the text from the **Preview** box, splitting it into separate files if necessary and creates the output text files in the specified folder.

If you use **Automation,** clicking the **Save** button will save the text file(s) and move to the next item in the queue. If you click **Skip**, the currently displayed text will not be saved and the next file in the queue will be loaded and processed.


www.theprimemachine.net/tpmcrafty

Last updated: 21/04/2021 06:12