tPMCrafty V2

Getting Started Guide

Thank you for your interest in tPMCrafty – the corpus-building tool designed to help you create handcrafted DIY corpora.

Table of Contents

App overview

The basic idea of tPMCrafty is to allow you to gather text for processing using the lefthand side of the interface (Text in), and then to save plain text in an organized way using the righthand side of the interface (Text out).

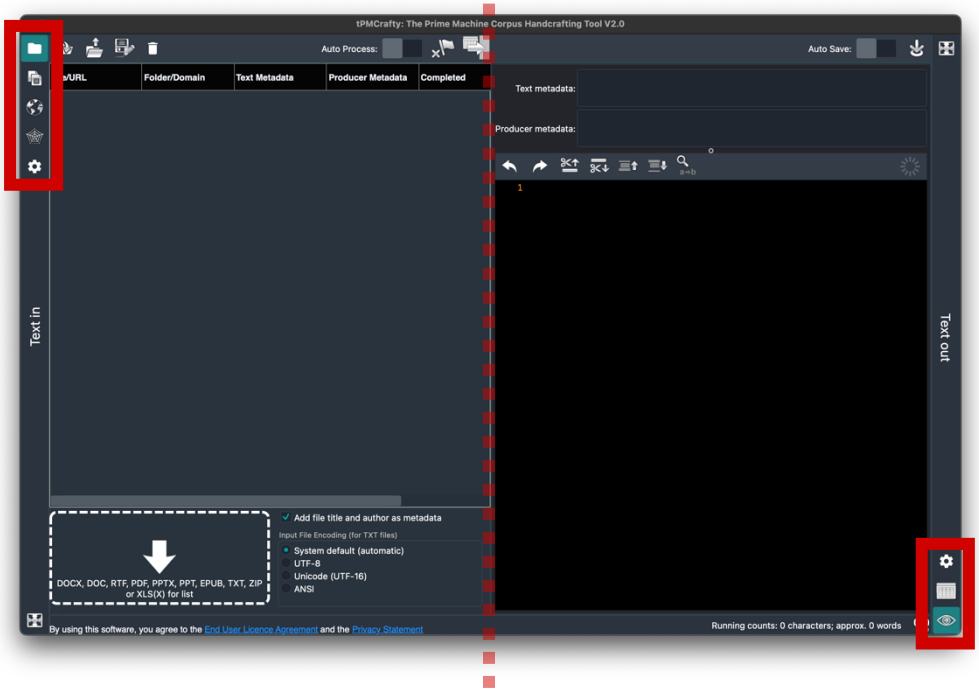
The **Text in** half of the interface has 5 main tabs (located in the top-left): • The Filelist Tab – for loading various kinds of text documents and for processing multiple webpage addresses. • The Copy & Paste Tab – for processing text from the clipboard (also allowing you to automatically collect clipboard clippings as you use the copy to clipboard function in other apps). • The Web Browser Tab – for exploring the web and using a single web page as the input source for processing. • The Sitemap Tab – for exploring website sitemaps and selecting web pages from these to add to the Filelist. • The Input Settings Tab – for controlling a wide range of settings related to text processing process (from paragraphing, to book chapters, to splitting files).	The **Text out** half of the interface has 3 main tabs (located in the bottom-right): • The Output Settings Tab – for selecting the destination folder (where the new corpus texts will be saved). • The Log Tab – for seeing a list of processed files and their metadata (information about the texts). • The Preview Tab – for seeing how the extracted text will be saved, and for manually editing text before it is saved.

Typically, you'll want to work with both panels simultaneously, but if you want to zoom the left or right side of the interface, you can use the zoom buttons located in the bottom-left (Text in) and top-right (Text out).

Formatted text comes in…

… and plain text comes out

Tex
t In
Tab



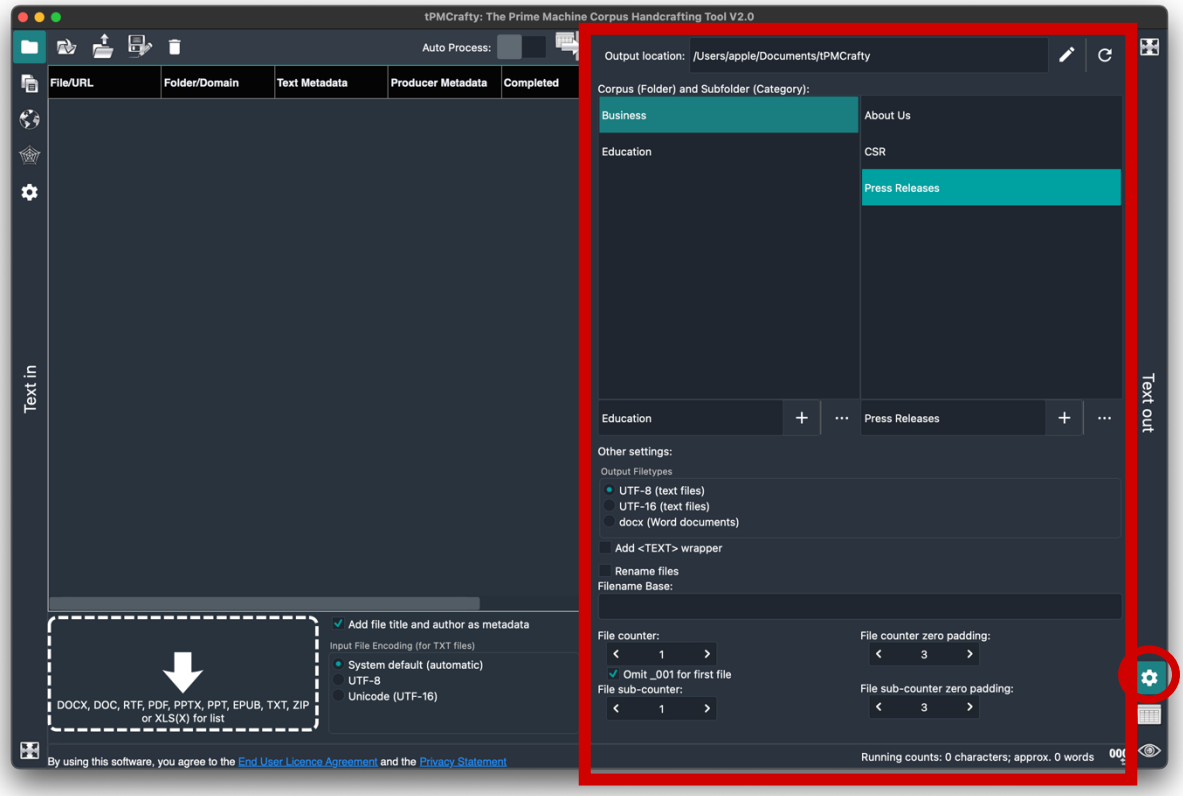
Tex
t
Out
Tab

There are many things you can do with tPMCrafty to quickly process many texts. But it is best to take things slowly; paste in, load or grab text from a website; check everything looks good in the Preview panel; and finally save your handcrafted text file. A corpus of clean texts that is carefully divided into categories will give you many more ways to analyze your topic than a jumble of junk collected from anywhere!

Above the Preview panel, there are two boxes you can edit to add information about the text (Text metadata) and information about the author(s) or speaker(s) (Producer metadata). Some files and websites may contain information about these that tPMCrafty can collect automatically. The information in these boxes will appear in the spreadsheet log file which is automatically created each time you save a text or collection of texts.

Deciding where to store your new corpus

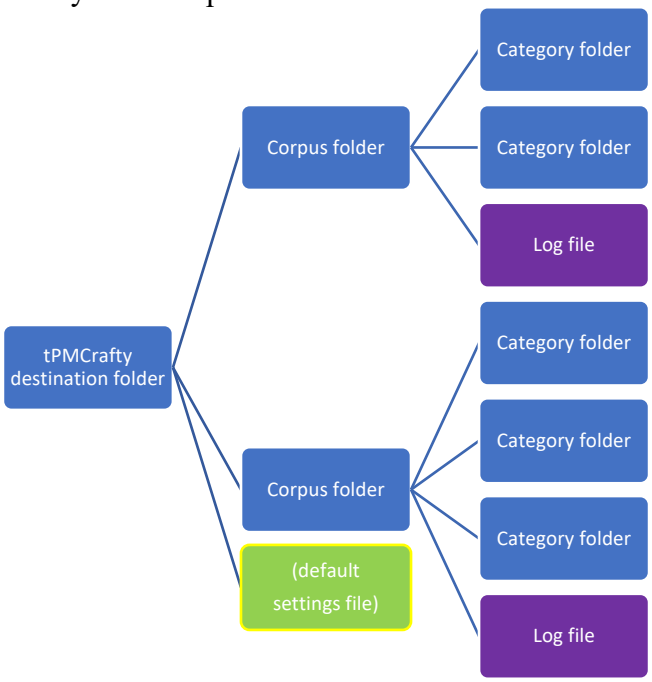
When you first open tPMCrafty, it will show the Filelist Tab on the left and the Preview on the right. It is possible to view texts on the Preview Tab without saving them, but generally the first thing to do when opening tPMCrafty is to choose a destination folder – the folder on your device where you want the corpus texts to be saved. Click on the cog icon on the Output Tab to choose Output Settings.



The default output location is your computer's Documents folder (Mac/Windows) or tPMCrafty's Documents folder (iPad[1]). On Mac & Windows you can change this location using the pencil button in the top-right corner.

[1] The iPad version is still currenty under development

The structure of a tPMCrafty DIY Corpus is as follows:



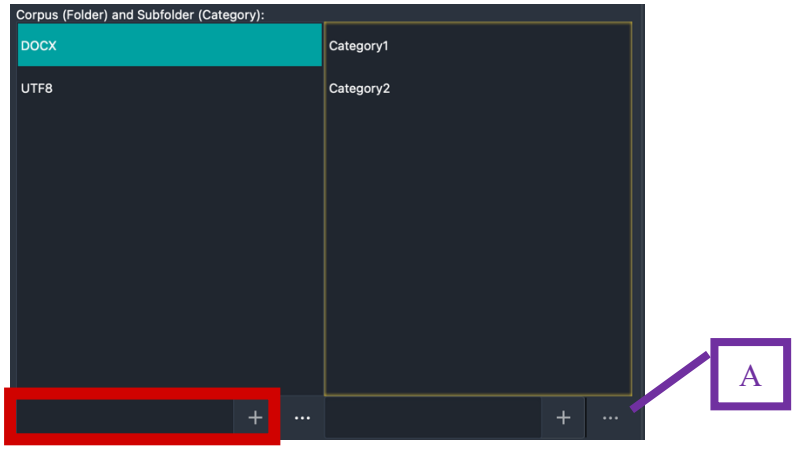
Inside the tPMCrafty destination folder each DIY Corpus will have its own subfolder. The main output destination folder is also the location tPMCrafty will look for default settings when it is launched.

Under each DIY Corpus subfolder, there will be one subfolder for each DIY Corpus Category. The DIY Corpus subfolder will also hold the spreadsheet log file for all texts in the DIY Corpus.

Inside each DIY Corpus Category folder, there will be the set of TXT or DOCX files you create.

Note: If you change the default destination, you will not be able to delete folders from within tPMCrafty. It is not recommended to move or delete files or folders in tPMCrafty's destination folder using other apps while tPMCrafty is running.

Remember: It is easy to merge files from different folders, but much harder to re-organize files by text type, topic or source. Saving in separate Corpus Categories means you can keep similar texts together and then re-organize your corpus if you later discover you want to compare and contrast sub-corpora.



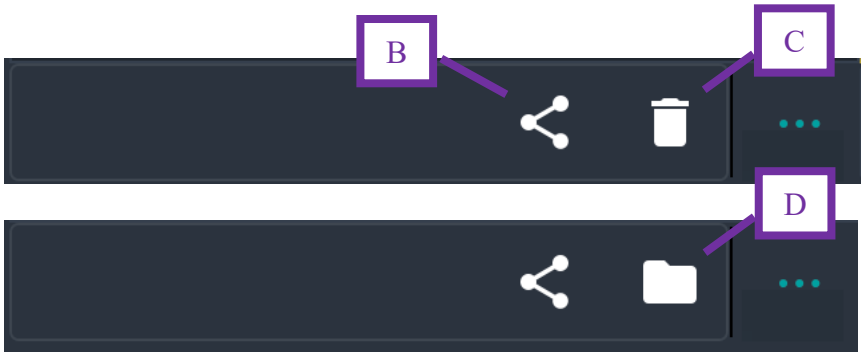
To create a new corpus, enter a name in the box at the bottom of the listbox and press the plus button. The name will become a subfolder, so remember to avoid using symbols not permitted in folders or filenames on your system.

When you click on a DIY Corpus name in the list, it will update the Category list to the right and also load the corpus log into the grid on the Log Tab.

Every DIY Corpus for tPMCrafty must have both a DIY Corpus name and a DIY Corpus Category.

For additional actions (A), you can select and item from either of the lists and click the button showing three dots. This will reveal a menu where you can share (B) or delete (C) the DIY Corpus or DIY Corpus Category. On Mac and Windows, the share button will ZIP the selected corpus or corpus category and show the folder containing the new zip file.

If the main destination folder has been changed from its default, the delete button will change (D) to allow you to view the files in your device's file manager.



The other elements on the Output Settings Tab allow you to control the output file naming system and to choose between three text output modes:

- UTF-8 (text files) – recommended for most applications;
- UTF-16 (text files) – for compatibility with Windows programs developed some years ago;
- docx (Word documents) – to save the files using the Open XML formatted Microsoft Word document file format. Note: the contents will all be plain text except for headings and sub-headings. The whole purpose of tPMCrafty is to remove other formatting from documents or websites, so DIY Corpora saved in docx format will not look the same as they did originally!

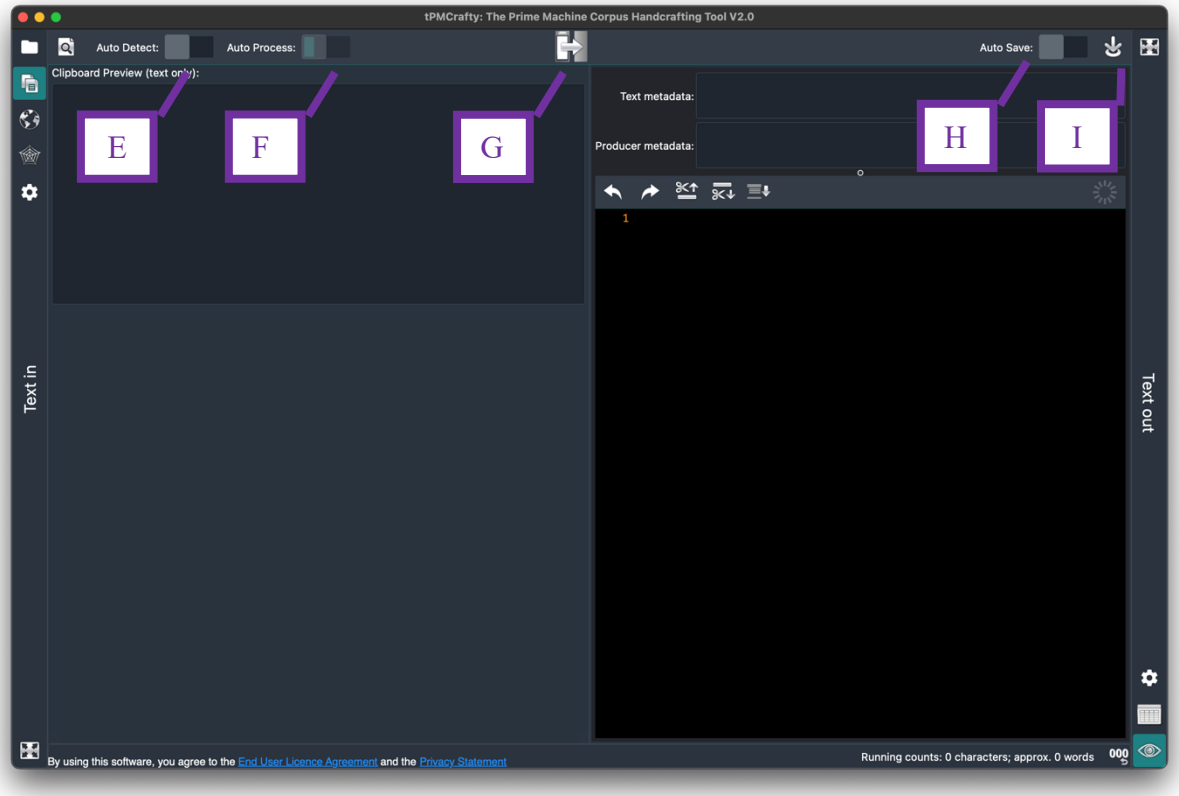
The **Add <TEXT> wrapper** option is another important option to consider using if you are planning to load tPMCrafty's DIY Corpora into Natural Language Processing apps or other taggers. This simply puts <TEXT> at the beginning and </TEXT> at the end, which may be required for some scripts and tagging services.

Once you have set up your Text out settings, you are ready to start importing text and saving it to your new corpus!

When you want to save files to a different subfolder simply return to the Output Settings Tab and add or choose the name of the folders you want to use as the DIY Corpus and the DIY Corpus Category. The next file you save will be saved in the new destination.

Text Input Method#1: Copy and (automatically) Paste (and save)

The easiest way to use tPMCrafty is to import text from the clipboard. Just like other apps, tPMCrafty allows you to paste in the text you've copied from another app using the paste button (G). However, what makes tPMCrafty a bit more special is its ability to automatically detect (E), automatically process (F) and automatically save (H) each time you copy new text in another app.



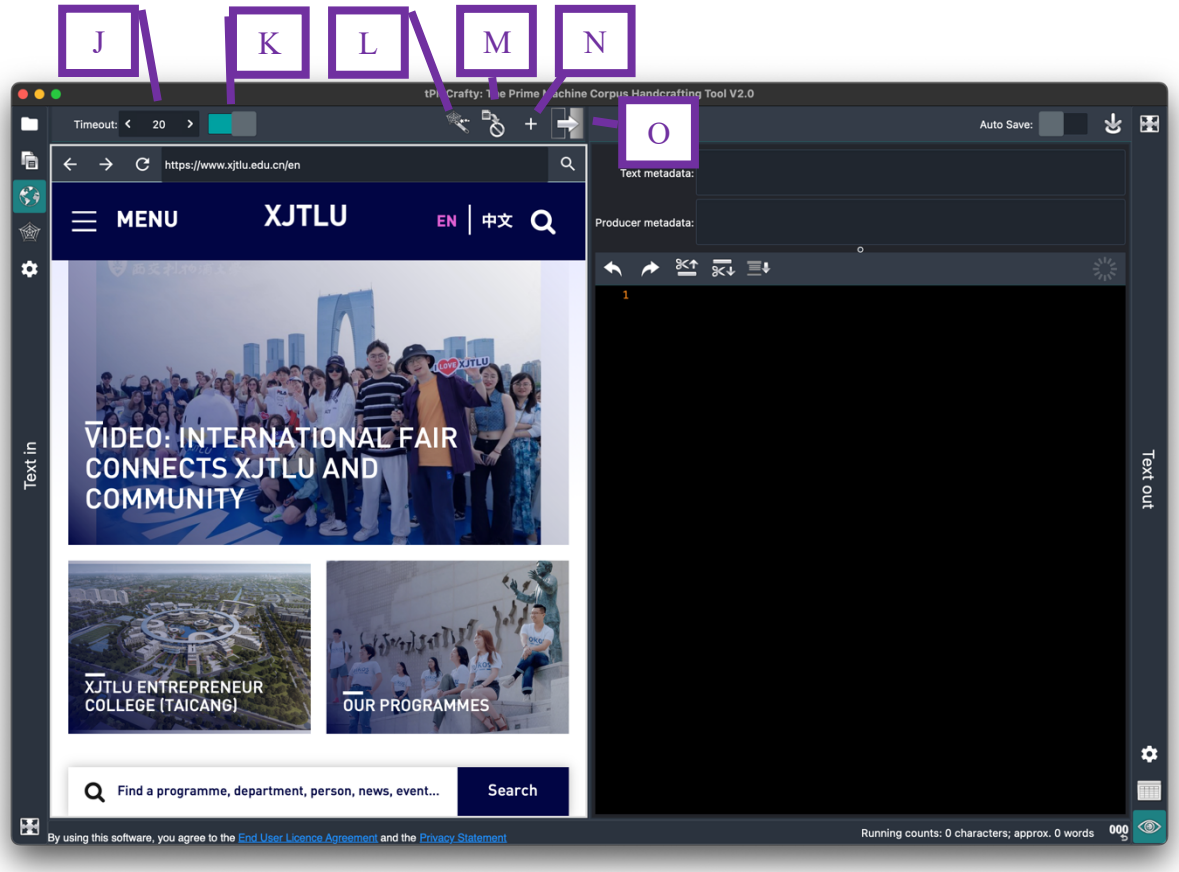
You can use this tab in conjunction with the Auto Save switch (located on the Preview Tab), allowing you to save new corpus files as soon as pasted text has been processed.

You can use this tab in conjunction with the Auto Save switch (located on the Preview Tab), allowing you to save new corpus files as soon as pasted text has been processed.

Mode	Switch settings	Purpose
Normal copy & paste behaviour (default)	Auto Detect: Off Auto Process: Off Auto Save: Off	When you flick to the Copy & Paste Tab, the defaults are restored and it will work like other apps; The clipboard will only be accessed when you click the paste preview or the paste button (G).
Paste and then save	Auto Detect: Off Auto Process: Off Auto Save: On (H only)	When you click the paste button (G), the contents of the clipboard will be processed and immediately saved to the corpus.
Automatically Detect	Auto Detect: On Auto Process: Off Auto Save: Off (E only)	PMCrafty will check the clipboard every second to see if there have been any changes. It will display the first 500 characters in the preview box. Text will not be processed until you click the paste button (G).
Automatically Detect and Process	Auto Detect: On Auto Process: On Auto Save: Off (E + F)	tPMCrafty will automatically process the contents of the clipboard when it detects changes (every second). The processed text will not be saved until you click the save button (I).
Automatically Detect, Process and Save	Auto Detect: On Auto Process: On Auto Save: On (E + F + H)	With this mode, you can switch to any other app that supports copying to the clipboard, and every time tPMCrafty detects a change it will automatically process and save the text to your corpus. Notifications will appear to confirm each save (if you allow tPMCrafty to show notifications); these appear as number badges on Mac or pop-up messages on Windows.

Windows users who are unable to use tPMCrafty's built-in web-browsing capabilities are encouraged to use the Copy & Paste Tab for their web-page collection needs.

Text Input Method#2: Basic Web Browsing



tPMCrafty has your operating system's browser running on the Web Browser Tab. You can navigate to different websites much as you would in your usual browser. However, no windows or tabs will open; all content will be displayed in the single view.

At the top you can set a timeout for web page loading (J). Typically, the slower content on web pages tends to be videos and images which won't be imported into the corpus anyway. The default is for each page to load for a maximum of 20 seconds. The switch to the right of the Timeout spinner (K) allows you to disable this timeout feature if you wish. At the top on the right, there are four buttons. The first two of these will be described in the following section. The Plus button (N) will add the currently viewed web address to the Filelist Tab's table. The right-most button (O) will take the text content from the currently viewed website, process it and display it in the preview.

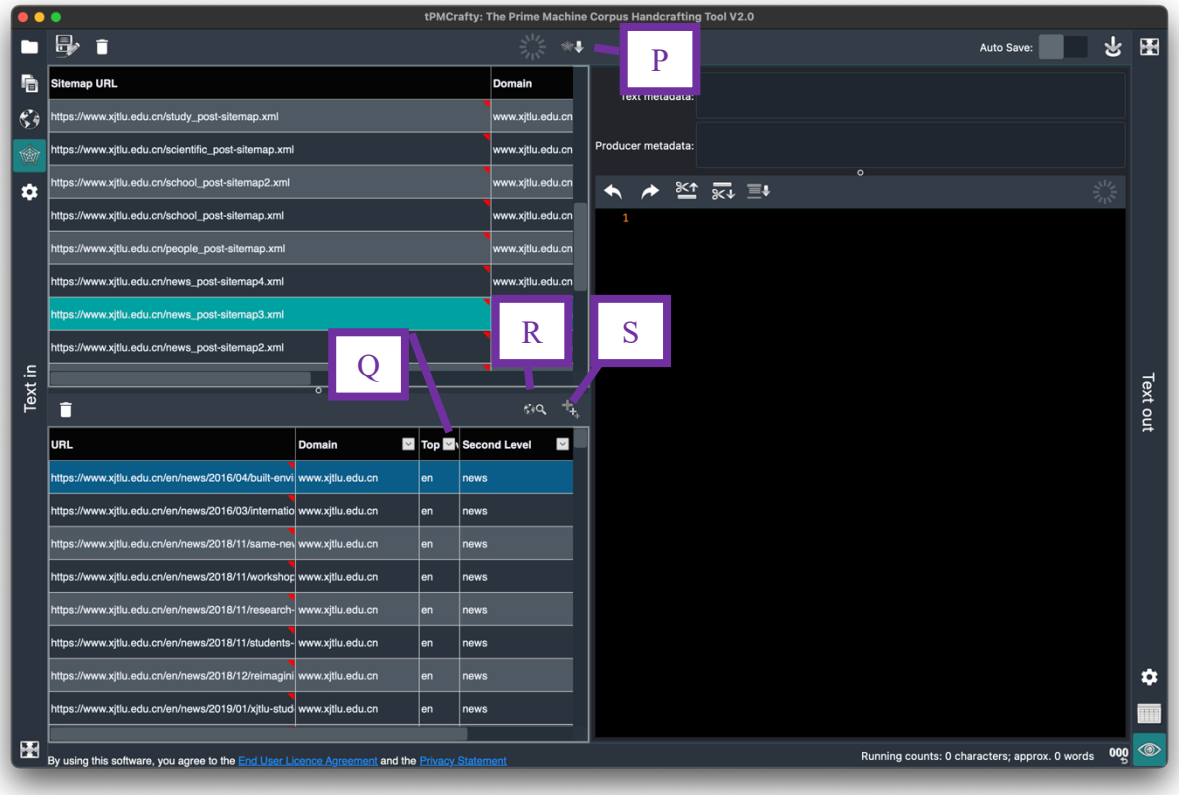
More advanced Web Browser functions

Finding and using Sitemaps

Most websites have a sitemap file which contains a list of all the webpages the web designers think might be useful for search engines and web crawling. When you click the spider's web button (L), tPMCrafty will try to look for sitemap files using commonly used filenames. If it is successful, it will show a list of sitemaps, a list of web pages, or a combination of both sitemaps and web pages. The list of sitemaps appears in the top table. Clicking on one of these and then selecting the right-most button in the top toolbar (P) will download that sitemap and fill the bottom table with the list of its web page addresses.

If you know the actual address of a sitemap.xml file, you can paste the address into the browser and then click the sitemap wizard. tPMCrafty will detect that you are already looking at an XML file and try to use this as the sitemap..

In the bottom half of the screen, the list of web pages will be displayed. These can be filtered by any of the columns (Q), with sections of the URL divided up automatically. There are two function buttons in the righthand side of the toolbar for the bottom table: the first will open the selected web address in the browser (R); the second will add all the filter results to the Filelist table for processing (S).



tPMCrafty checks each sitemap file address very slowly. This is because it is rather unfair to a website if you try to open dozens of pages at once!

Once in the Filelist, only one web page from each domain will be processed every 20 seconds. This means you are using a website more like a real person (conveniently being given a tour by the sitemap).

You can collect more web texts more quickly if you include several different domains in your Filelist. tPMCrafty will hit each domain once every 20 seconds, but if you have multiple domains listed in your Filelist, it won't have as long to wait between loading new pages..

Blocklist

Modern websites include a host of labels, information and footnotes which appear on multiple pages. For corpus analysis, it usually isn't ideal to include multiple copies of text which is essentially 'boilerplate' or from site template.

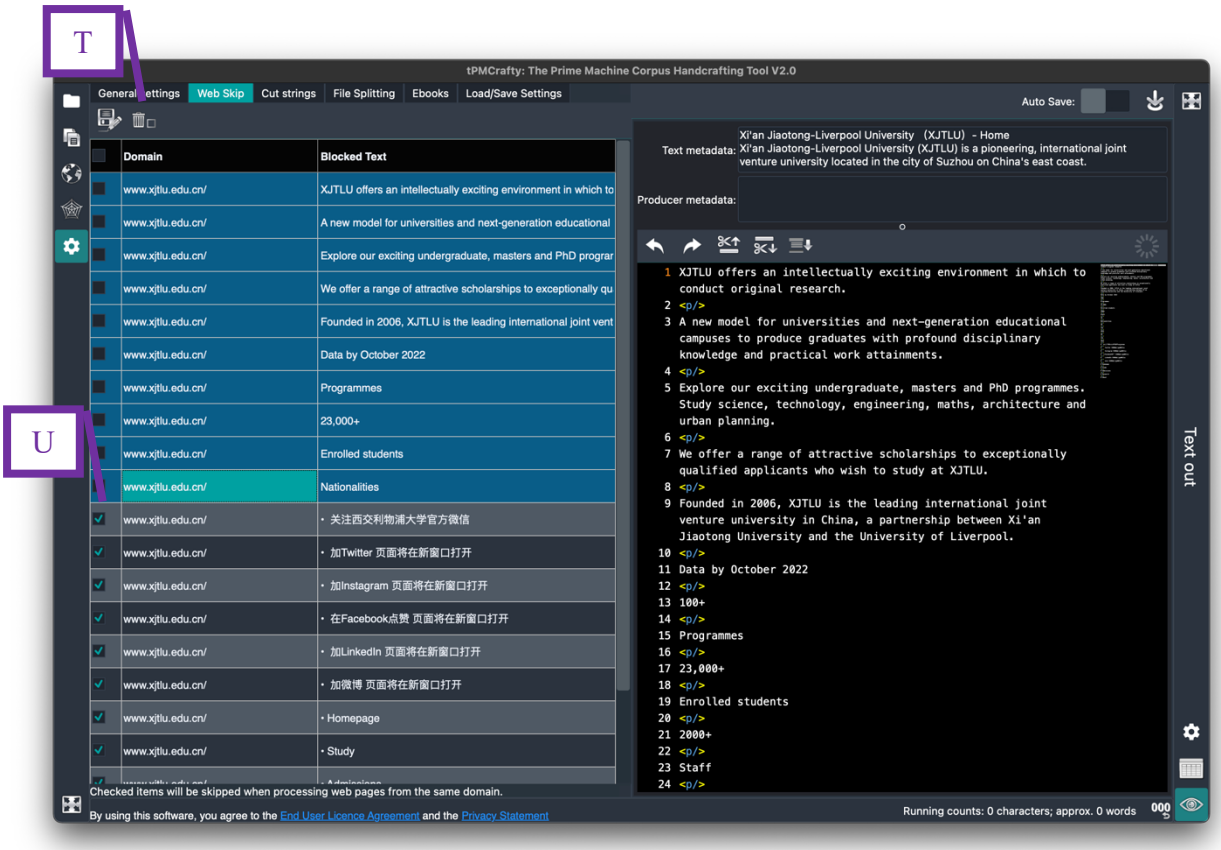
Of course, you can check each web page one by one and use the "Cut all above / below" buttons to manually remove text from the preview box before it is saved.

But for your convenience, you can also use the Blocklist generator button (M) to capture all the text elements in the currently viewed web page, and you can select which items you want to skip over or block when tPMCrafty extracts text from any page on the same website.

For many websites, it is best to try the Block list generator function from the website's home page. But you can use the function on any web-page.

The items which are selected (U) will be skipped over when you process text from other web-pages on the same domain, so they won't appear in your corpus texts.

You can delete unselected items using the rubbish bin button in the top-left corner.



Text Input Method#3: Loading files and processing lists of web site addresses

As well as using the clipboard or the built-in web browser, tPMCrafty can work with lists of URLs and files.

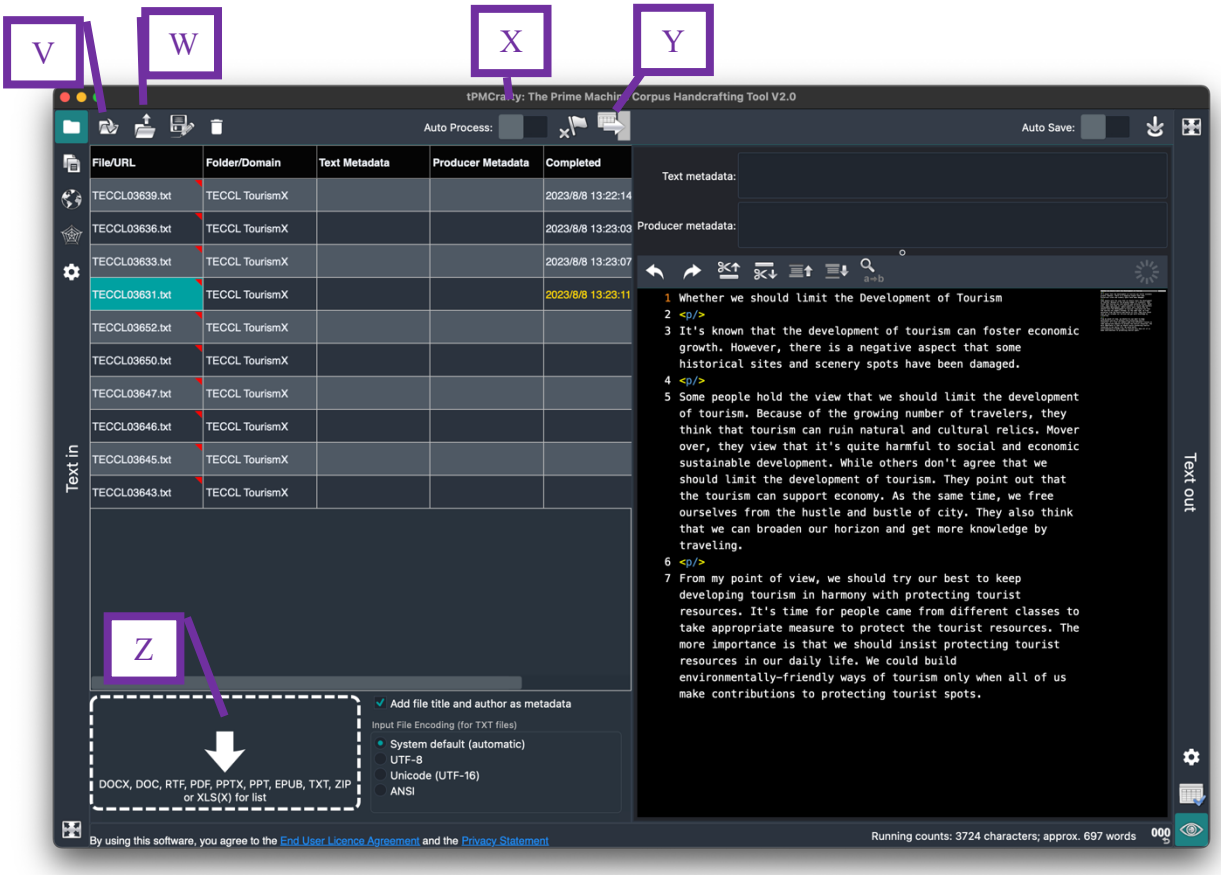
Importing files

On Mac/Windows, there are four ways to add files to the Filelist Tab for processing:

1. Drag and drop files into the target area in the bottom-left corner of the screen (Z);
2. Use the Open files button (V) to add one or more files from inside a folder.
3. Use the Open folder button (W) to add all the files from one folder.
4. Update a saved XLSX spreadsheet (or use the same headings in the first row), and drag and drop or open the XLSX file.

On iPad, to protect your device tPMCrafty will only open files which you explicitly share with the app. Therefore, on iPad you should use the iPad File Manager App or another app which has access to the files you want to process and then use the Share menu in that app to share with tPMCrafty.

*To add a list of web pages from the sitemap of a website, you can use the Sitemap wizard function, which is explained under "*Finding and using Sitemaps*".*



Processing one URL or file at a time

The button in the top-right corner of the Filelist Tab (Y) will find the next item on the list and process it. If the next item is a website, the page will be loaded in the web browser. If the next item is a file, it will be loaded from disk. The button to the left of this (next to Y) has an x and a picture of a flag. This button will clear the Completed column for the currently selected line if you want to reset the file. This will not delete any files that have already been saved.

Processing all

By turning on the Auto Process switch (X), tPMCrafty will move through the entire list of items one by one.

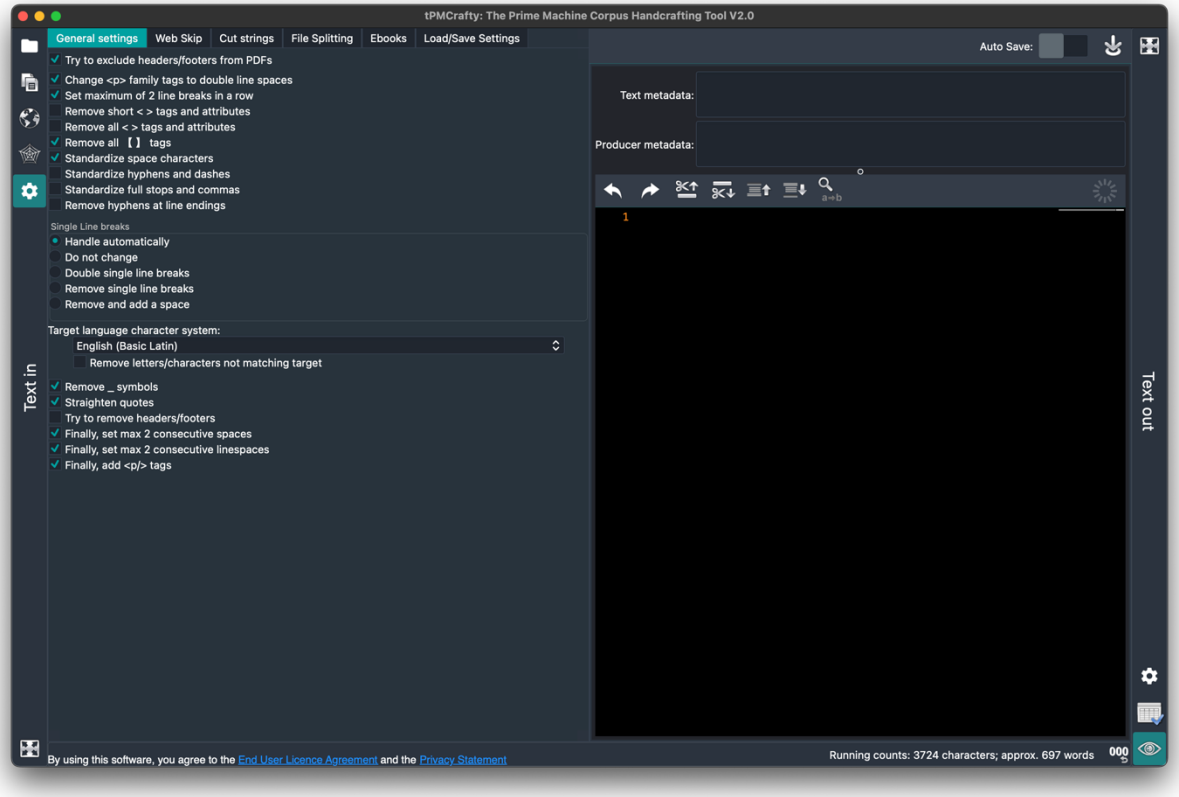
To prevent overloading websites, Filelist, only one web page from each domain will be processed every 20 seconds. This means you are using a website more like a real person (conveniently being given a tour by the sitemap).

If all file items have been processed and there are multiple pages from the same domain listed, tPMCrafty will display a clock icon and wait 20 seconds before opening the next web page.

- Items which are currently being processed appear in yellow.
- Items which have been processed and saved appear in white.
- Items which have been skipped appear in red. After all items have been processed, you will be given the option to retry skipped items.

Dozens of options for processing text

tPMCrafty has many different options for processing text in different ways. For many of these, the best way to understand what a tickbox is for is to just try it and see. However, there are a few advanced features which may need more explanation.



Splitting text by using strings of characters

The File Splitting tab allows you to enter a list of strings you want to use to mark the beginning or end of separate texts.

For example, if you are processing customer review and you want to split all the reviews into separate text files, you may find that each review ends with “Report” or “Rate this comment”.

There are options to keep the strings you are using to identify breaks in the text (adding a text break just before or just after the string). There is also the option to remove the string used for identification.

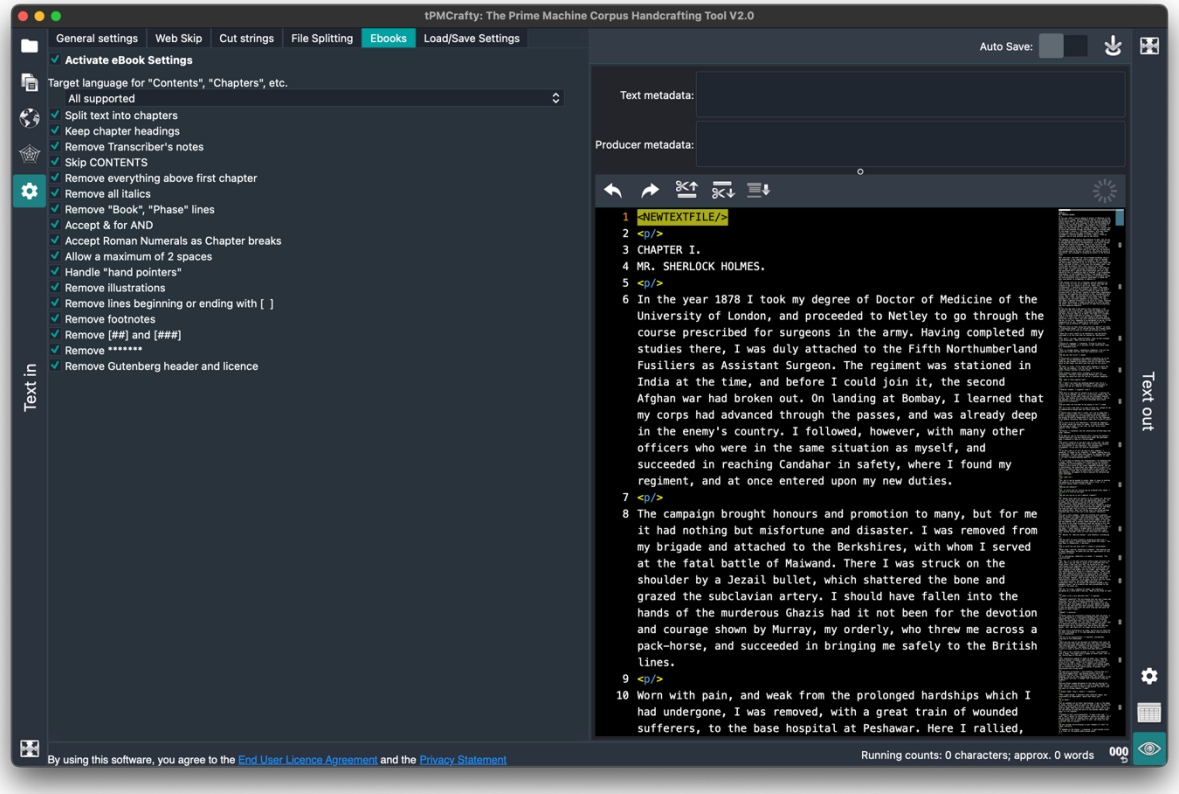
Want to compare the speeches of Romeo with those of Juliet? Or compare candidates in a televised presidential debate? Try "group matches together" on the File Splitting tab.

Another way this feature can be used is to divide up a text such as a transcript from a debate or the script of a play. If the names of the speakers are known, and the names appear in capitals at the beginning of each turn, you could enter the names of the speakers on separate lines in the box. Then you could use the "Group matches together" tickbox to save each speaker's turns separately.

Processing e-books

When working with e-books – and novels in particular – there are many different things to consider about how you organize your texts. tPMCrafty has a large number of features designed to assist in analyzing novels such as those freely available from Project Gutenberg.

Remember to respect copyright and the terms of use of websites and e-books. tPMCrafty has these features to allow you to do your own personal research using texts to which you already have access.



If you want to explore development of themes through the chapters of a novel, one of the really useful functions is "Split text into chapters". tPMCrafty can't always locate chapter breaks (as a wide variety of words or symbols might be used to indicate these in an e-boook). However, it will try to find common patterns in several languages and it will add a marker in

the preview box to show where each text break will occur. The marker "NEWTEXTFILE" won't appear in the files that are saved; this marks the point where a new file will be started.

Additional features of the Preview screen

The Preview Tab also has a number of additional features.

Undo / Redo

Just like in other text editors, you can undo or redo changes you make directly in the Preview Editing box. Unfortunately, this is not available on Windows systems where the Edge Chromium browser can't be started. These buttons only work on changes you make directly in the preview box.

Cut all text above / all below

The two buttons with scissors make it convenient to remove all the text above or below the cursor. For example, if you want to remove the title page and contents from a document, you can scroll down to the beginning of the text proper and click the Cut all text above button.

Go to previous or next chapter or section

If you have use the File Splitting or e-book chapter identification functions, you can use these buttons to jump to the previous or next section break.

Find and Replace (Command/CTRL F)

The editor also has built-in find and replace. To use this click this button or press Command + F (Mac) or Ctrl + F (Windows).

Notes on Windows Edge Chromium

If you have updates installed on Windows 10 / 11, you should automatically have the Edge Chromium browser on your computer, ready for apps like tPMCrafty to use for web surfing. If you want to install manually, you can try this link: https://www.microsoft.com/en-us/edge

Another option on Windows systems is to install WebView2 Runtime. This is a Microsoft runtime library to allow apps like tPMCrafty to have web surfing capabilities. You can download it here: https://developer.microsoft.com/en-us/microsoft-edge/webview2/

If you can't get it to work, don't panic. The Preview Tab will work almost as well (although some features will be missing), and you can always use the Copy and Paste Tab to make it really easy to grab text from other apps simply by copying text to the clipboard.

Other known issues

There are a number of other known limitations with the current version:

- The Undo/Redo buttons do not capture all edit changes on macOS.
- Some websites use JavaScript to handle downloads and these will not always work on macOS. If nothing happens when you click a link, try opening it in your normal browser and copying the address of the download and pasting it into the tPMCrafty browser.

Something not working or got a new idea?

Please get in touch with the developer: Dr. Stephen Jeaco (stevejeaco @ theprimemachine.net). Text comes in so many shapes and sizes! And some requests are easy to implement; others might be more challenging. If you email with your suggestions I'll try to do what I can!

https://www.theprimemachine.net/tpmcrafty/index.html