

Corpus Loader User Guide

Author: Hamish Croser

Contents

- Overview
- File loader
 - Preparing your data
 - Notes on file type specifics
 - Upload
 - Load a corpus
 - Build a corpus
 - View the corpus
- Oni Loader
 - Select a provider
 - Add a provider
 - Set the API key
 - Retrieve a collection
 - Building the corpus

Overview

The Corpus Loader is a tool used to build a corpus from a wide range of file types. The tool is designed to be used in conjunction with corpus analysis tools from the Australian Text Analytics Platform.

The Corpus Loader features a file loader, where users can upload files to the notebook environment and load them as a corpus. There is also an Oni loader feature, which enables loading a corpus directly from a platform running the data archiving service Oni.

File loader

Preparing your data

The Corpus Loader accepts the following file types:

TXT, DOCX, ODT, CSV, TSV, XLSX, ODS, XML

Files of the above file types can also be archived into a ZIP file and loaded.

When loading files, the Corpus Loader will load textual and tabular files differently.

- Textual files (TXT, DOCX, ODT, XML) will have their text content read in as a document with no consideration for their internal structure. This means that no metadata can be

extracted from the contents of the file, just the file name and path.

- Tabular files (CSV, TSV, XLSX, ODS) will have their table-like structure preserved when loading. Once loaded, you can include/exclude metadata columns from the final built corpus, and select a column to be used as the document column

The Corpus Loader allows you to load a set of textual files as a corpus with no metadata (e.g. a collection of TXT files) and separately a tabular file for metadata (e.g. a CSV).

Crucially, the metadata file must contain a linking column, which is a metadata column used to link each metadata row to a corpus document.

For textual files, the Corpus Loader constructs two metadata columns for textual files: filename and filepath. The filepath for a document is the path displayed in the file selector window of the Corpus Loader. The filename for a document is the name of the file without the file type extension, e.g. a file 'example.txt' will have a filename metadata of 'example'.

Notes on file type specifics

All files are expected to be UTF-8 encoded.

TXT

This file type has almost no pre-processing applied when being loaded into the corpus, i.e. the text is simply read into the corpus as-is.

This means that if you would like to load another file type verbatim (e.g. you wish to load a CSV as a text document rather than a tabular document), you should rename that file's extension to TXT.

DOCX, ODT

The Corpus Loader will not include tables, images, text boxes, and their contents in the loaded corpus. Only top-level text will be included in the corpus.

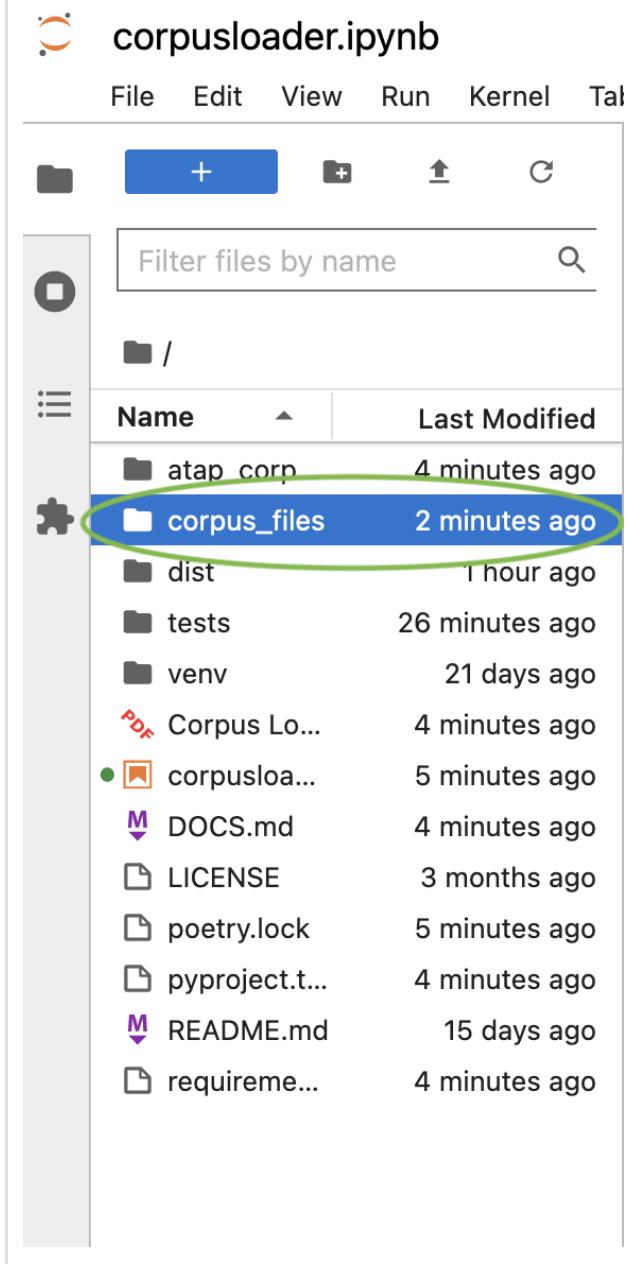
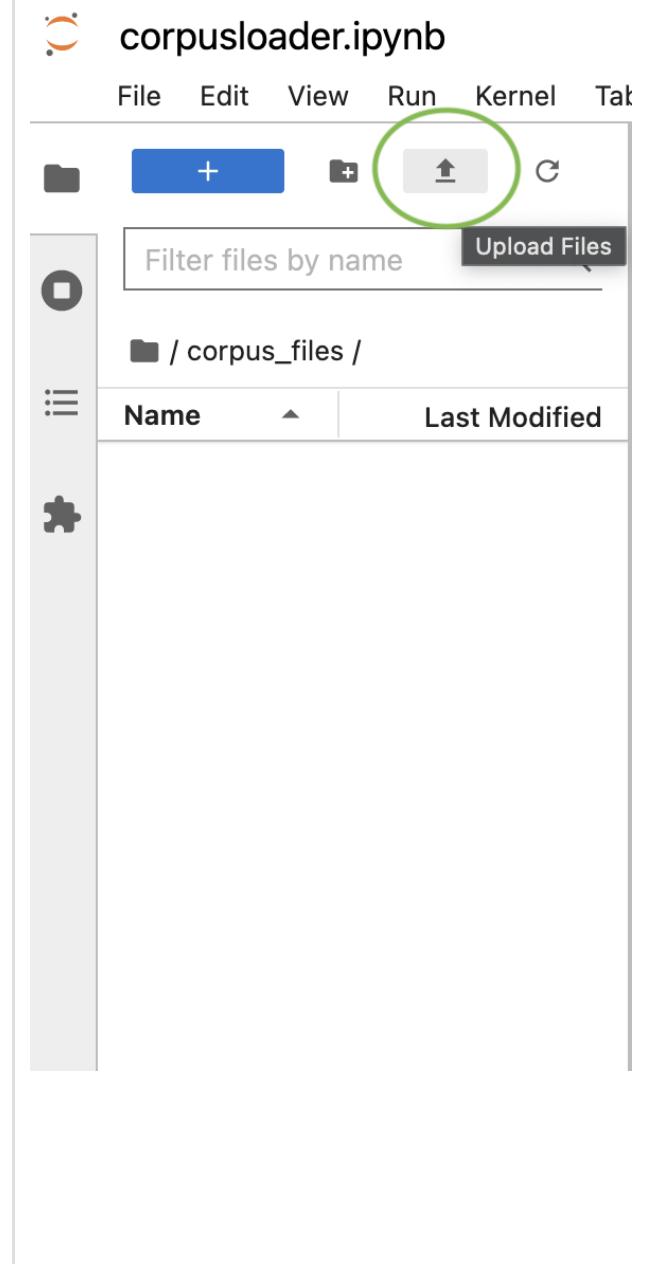
XML

XML tags are removed before loading into the corpus. Only the text between the tags is kept.

Upload

The Corpus Loader will display files from a specific folder, defined in the cell that starts the Corpus Loader in the notebook.

Use the notebook file browser to upload corpus and metadata files to the specified folder. This can be done by dragging the files from your computer's file browser into the notebook file browser. You can also use the 'Upload files' button near the top left of the notebook file browser.

Navigate to the target folder	Upload using the upload button																														
 <p>corpusloader.ipynb</p> <p>File Edit View Run Kernel Tab</p> <p>+ Filter files by name</p> <p>Filter files by name</p> <p>corpus_files /</p> <table border="1"> <thead> <tr> <th>Name</th> <th>Last Modified</th> </tr> </thead> <tbody> <tr> <td>atap_corp</td> <td>4 minutes ago</td> </tr> <tr> <td>corpus_files</td> <td>2 minutes ago</td> </tr> <tr> <td>dist</td> <td>1 hour ago</td> </tr> <tr> <td>tests</td> <td>26 minutes ago</td> </tr> <tr> <td>venv</td> <td>21 days ago</td> </tr> <tr> <td>Corpus Lo...</td> <td>4 minutes ago</td> </tr> <tr> <td>corpusloa...</td> <td>5 minutes ago</td> </tr> <tr> <td>DOCS.md</td> <td>4 minutes ago</td> </tr> <tr> <td>LICENSE</td> <td>3 months ago</td> </tr> <tr> <td>poetry.lock</td> <td>5 minutes ago</td> </tr> <tr> <td>pyproject.t...</td> <td>4 minutes ago</td> </tr> <tr> <td>README.md</td> <td>15 days ago</td> </tr> <tr> <td>requireme...</td> <td>4 minutes ago</td> </tr> </tbody> </table>	Name	Last Modified	atap_corp	4 minutes ago	corpus_files	2 minutes ago	dist	1 hour ago	tests	26 minutes ago	venv	21 days ago	Corpus Lo...	4 minutes ago	corpusloa...	5 minutes ago	DOCS.md	4 minutes ago	LICENSE	3 months ago	poetry.lock	5 minutes ago	pyproject.t...	4 minutes ago	README.md	15 days ago	requireme...	4 minutes ago	 <p>corpusloader.ipynb</p> <p>File Edit View Run Kernel Tab</p> <p>+ Filter files by name</p> <p>Filter files by name</p> <p>corpus_files /</p> <table border="1"> <thead> <tr> <th>Name</th> <th>Last Modified</th> </tr> </thead> </table>	Name	Last Modified
Name	Last Modified																														
atap_corp	4 minutes ago																														
corpus_files	2 minutes ago																														
dist	1 hour ago																														
tests	26 minutes ago																														
venv	21 days ago																														
Corpus Lo...	4 minutes ago																														
corpusloa...	5 minutes ago																														
DOCS.md	4 minutes ago																														
LICENSE	3 months ago																														
poetry.lock	5 minutes ago																														
pyproject.t...	4 minutes ago																														
README.md	15 days ago																														
requireme...	4 minutes ago																														
Name	Last Modified																														

Load a corpus

Use the file selector to select one or more files, and then load these files as corpus or metadata.

- Loading a file as corpus indicates that it contains 'document' data (the primary data to be analysed). 'Corpus' files can also contain metadata, but may not be exclusively metadata
- Loading a file as metadata indicates that it does not contain 'document' data, and only contains metadata

The file selector contains several features to filter and display the files to be loaded:

- The 'Select all' button: this will select all files displayed in the selector (including those files that are off-screen and must be scrolled to view). Using 'Select all' can be slow if there are many files (~ >1000) in the selector.

- The filter text field: typing text in this field will filter files whose file path matches the text, e.g. typing 'example' in the field will display the files 'example.txt' and 'example_file.txt' but not 'another_file.txt'.

This can be used to filter for subfolders. The '*' character is a 'wildcard', and represents any text, e.g. typing '197*record' in the field will display the files '1971-record.txt' and '1975-record.txt' but not '1980-record.txt' or '1907-record.txt'

The filter will be applied when you press 'enter' or if you deselect the field. To clear the filter, delete all text within the field.

- The 'Show hidden' checkbox: when checked, this will display files whose filename begins with '.', e.g. '.example.txt'. These files are unlikely to appear and so can be ignored by most users
- The 'Expand archives' checkbox: when checked, all ZIP files displayed in the file selector will have every internal file displayed in the selector, allowing files to be selected individually. When unchecked, selecting and loading a ZIP file will load all internal files.
- The file type dropdown filter: this allows filtering of a specific file type. It defaults to 'All valid filetypes'. This dropdown will filter internal files in a ZIP file.
The file selector will not display files that are not one of the valid file types.
- Keyboard keys:
 - Click and drag will select/deselect multiple files
 - Cmd+click (Mac) / Ctrl+click (other) will select/deselect multiple files that are not adjacent
 - Up/down arrow keys can be used to navigate the file selector. Holding shift while navigating will select all files visited

Select the files

The screenshot shows the 'File Loader' interface. At the top, there are two tabs: 'File Loader' (which is active) and 'Corpus Overview'. Below the tabs are several controls:

- A 'Filter displayed files' input field with a clear button (indicated by a left arrow).
- Two checkboxes: 'Show hidden' and 'Expand archives'.
- A 'Filter by filetype' dropdown set to 'All valid filetypes'.
- A blue 'Select all' button.
- A list area containing a single item: 'corpus_files/txt_corpus.zip'. This item is highlighted with a green oval.
- At the bottom, there are three buttons: 'Load as corpus' (green), 'Total files: 0' (with a question mark icon), and 'Unload selected' (red). To the right of these is a brown 'Unload all' button.

Load the files

The screenshot shows the 'File Loader' and 'Corpus editor' sections of a software interface. In the 'File Loader' section, a file named 'corpus_files/txt_corpus.zip' is selected. Below it are buttons for 'Load as corpus' (circled in green), 'Corpus name' (text input), 'Build corpus' (green button), 'Total files: 5' (info), 'TXT: 5' (info), 'Unload selected' (button), and 'Unload all' (button). In the 'Corpus editor' section, a dropdown menu 'Select document label' is set to 'document'. A dashed box highlights the 'Data label' and 'Datatype' columns for three rows: 'document' (datatype TEXT, include checked), 'filename' (datatype TEXT, include unchecked), and 'filepath' (datatype TEXT, include unchecked).

Build a corpus

If both corpus and separate metadata files are loaded, then the metadata and corpus files must be linked. To link the two, use the linking dropdown selectors between the corpus editor and the metadata editor choose metadata labels to join on. Once the metadata label is chosen, a 'link' symbol will be seen next to the selected label.

Once metadata linking has been done, or if not required, prepare the metadata for loading as follows:

1. A document label should be selected in the corpus editor. This is the data to be used as the primary data to be analysed. For textual file types, this will default to the content of the file. For tabular file types, this will default to the left-most column.
2. All metadata is included by default, but can be excluded from the final corpus by unchecking the checkbox under the 'Include' heading. Document data and linking metadata cannot be excluded from the corpus.
3. Data types for metadata are inferred but can be modified using the dropdown selectors under the 'Datatype' heading. The TEXT data type is unconstrained and allows the most freedom for the metadata (inconsistencies, missing values, etc.). The other data types have the following constraints:
 - INTEGER - each entry must be a whole number
 - DECIMAL - each entry must be a number
 - BOOLEAN - each entry must be either 'True' or 'False'
 - DATETIME - each entry must be of the format 'yyyy-mm-dd hh:mm:ss'
 - CATEGORY - similar to text but some analyses will treat equal values as part of the same category

Once the metadata has been prepared, the corpus is ready to be built. In the bottom left of the tool, type in a name for the corpus and then click the 'Build corpus' button. A progress bar will appear while the corpus is being built, and a green notification in the bottom right of the window will appear when the corpus is complete.

If there is an error during a corpus build, a red notification in the bottom right of the window will appear with information on why the build failed.

Select the document label

Corpus editor ?

Select document label

document
filename
filepath

Data label **Datatype**

Include

document	TEXT	<input checked="" type="checkbox"/>
filename	TEXT	<input type="checkbox"/>
filepath	TEXT	<input type="checkbox"/>

The screenshot shows the 'Corpus editor' interface. At the top, a green oval highlights the 'Select document label' dropdown menu, which contains three options: 'document' (selected), 'filename', and 'filepath'. Below this, a dashed box encloses the 'Data label' and 'Datatype' columns. An arrow points from the word 'Include' to the first row of the table below. The table lists three items: 'document' with datatype 'TEXT' and checked 'Include' status; 'filename' with datatype 'TEXT' and unchecked 'Include' status; and 'filepath' with datatype 'TEXT' and unchecked 'Include' status.

Link metadata labels (optional)

File Loader Corpus Overview

Filter displayed files ↲ Show hidden Filter by filetype
Expand archives All valid filetypes

Select all

corpus_files/txt_corpus.zip
corpus_files/xlsx_meta.zip

Load as corpus Load as metadata Total files: 6
Corpus name Build corpus TXT: 5
XLSX: 1 Unload selected Unload all

Corpus editor

Select document label

document

Data label	Datatype	Include	Link
document	TEXT	<input checked="" type="checkbox"/>	
filename	TEXT	<input checked="" type="checkbox"/>	
filepath	TEXT	<input type="checkbox"/>	

Select corpus linking label

filename

Select metadata linking label

- filename
- philosopher_name
- birth_year
- teacher

Metadata editor

Data label	Datatype	Include	Link
filename	TEXT	<input type="checkbox"/>	
philosopher_name	TEXT	<input type="checkbox"/>	
birth_year	INTEGER	<input type="checkbox"/>	
teacher	TEXT	<input type="checkbox"/>	

Name and build the corpus

File Loader Corpus Overview

Filter displayed files ↲ Show hidden Filter by filetype
Expand archives All valid filetypes

Select all

corpus_files/txt_corpus.zip

Load as corpus my_corpus Build corpus Total files: 5
TXT: 5 Unload selected Unload all

Corpus editor

Select document label

document

Data label	Datatype	Include
document	TEXT	<input checked="" type="checkbox"/>
filename	TEXT	<input type="checkbox"/>
filepath	TEXT	<input type="checkbox"/>

View the corpus

The 'Corpus Overview' tab lists all corpuses that have been built in the current session. It contains information about each corpus (name, number of documents, data labels, and data

types).

To view information about a corpus, click on the corpus name to expand the panel. From here, the corpus can be renamed by typing in the 'Rename corpus' text field and pressing enter.

The corpus can be exported by selecting an export filetype using the 'Export filetype' dropdown, and then clicking the 'Export' button. This will trigger a download of the corpus as a single file.

The corpus can be deleted by pressing the 'Delete' button.

View the corpus summary

The screenshot shows the 'Corpus Overview' tab selected. A blue header bar displays the title '▼ my_corpus - 5 documents'. Below it is a control panel with three buttons: 'Rename corpus' (containing 'my_corpus'), 'Export filetype' (set to 'csv'), and 'Export' (blue button). A red 'Delete corpus' button is also present. The main content area shows a table for the first document:

Data label	document_
Datatype	TEXT
First document	A pupil of Plato, Aristotle became one of history'

Oni Loader

The Oni loader enables loading a corpus directly from a platform running the data archiving service Oni. To use the Oni loader, click the tab at the top of the Corpus Loader labelled "Oni Loader".

The screenshot shows the 'Oni Loader' tab selected. It includes fields for 'Provider selector' (set to 'LDaCA'), 'API Key' (containing 'af6391e0-f873-11ee-8355-bae397411a92'), and 'Collection ID' (containing 'arcp://name,doi10.26180%2F23961609'). A green 'Retrieve collection information' button is located next to the collection ID field.

Select a provider

Select an Oni provider using the dropdown to set the current Oni provider. The link next to the selector displays the current provider address.

Add a provider

You can add an Oni provider if your data is stored somewhere running an Oni implementation that isn't listed in the dropdown.

To add a provider, click the "Add new provider" button and fill out the fields. The provider name can be anything but not cannot be left empty. The Provider address should be the base URL of the provider, e.g. <https://data.idaca.edu.au>

Set the API key

An API key from the selected Oni platform must be provided to access a collection.

To obtain your API key, do the following:

1. Visit the provider portal by clicking the "Visit provider portal" button
2. If not logged in, login using the button in the top right
3. If logged in, navigate to the 'User Information' page
4. Under 'User Details' > 'API Key', click "Generate"
5. Copy the API Key shown. Navigate back to this tool and paste in the field provided

Retrieve a collection

To access a collection at the Oni provider, enter the collection ID in the provided field and click the "Retrieve collection information" button.

To find the collection ID visit the provider portal, view the page of the collection you want to access, and the collection ID will be the text labelled "@id". Ensure you copy the *text* of the ID and not the link of the ID.

Building the corpus

Once the collection is retrieved, the corpus can be built by following the steps outlined in the File loader instructions above.