Corpus Slicer User Guide

Author: Hamish Croser
Version: 0.3.6

Contents

Overview

The Corpus Slicer is a tool that can filter a corpus by applying conditions on the textual data and/or metadata.
The Corpus Slicer incorporates the Corpus Loader, another Australian Text Analytics Platform tool that builds a corpus from a set of documents.

Instructions

Building and selecting a corpus

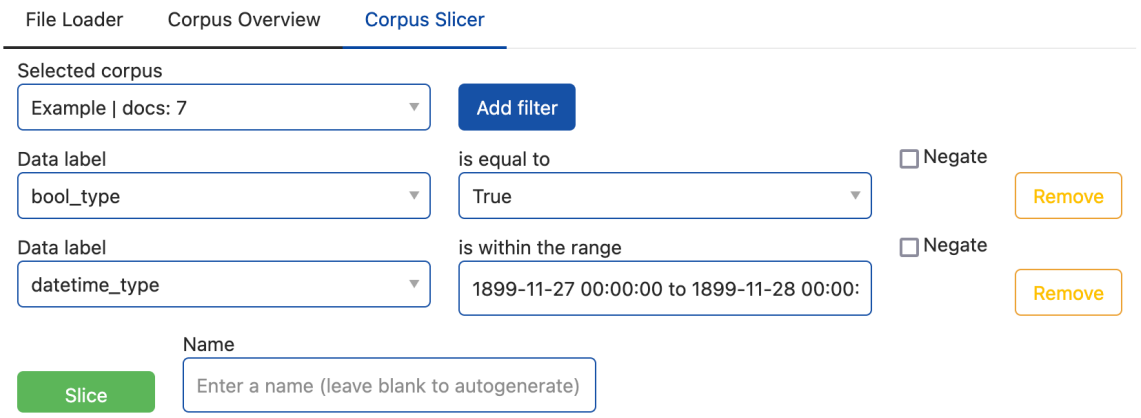
Before you can slice a corpus, you must build a corpus. To do this, upload your corpus documents and build your corpus using the Corpus Loader. Refer to the Corpus Loader User Guide for detailed instructions on how to do this.

Once you have built a corpus, navigate to the 'Corpus Slicer' tab. From the 'Selected Corpus' dropdown, select the corpus you wish to slice. If only one corpus has been built, it will be selected by default. Once you are sure the correct corpus is selected, move onto the next step.

Adding and removing filters

When you select a corpus, there will be a single filter line. If you wish to chain multiple filters together, click the 'Add filter' button next to the 'Selected corpus' dropdown. If you wish to remove a filter, click the 'Remove' button on the right side of the filter line.

When multiple filters are applied, they are chained together such that a document must pass all the filters in order to be included in the sliced corpus. This is known as an AND operation, distinct from an OR operation. Below is a screenshot showing multiple filters being applied to a corpus.



Filter operations

There are many operations that can be applied using a filter. The operations available are dependent on the data type the filter is being applied to. Because of this, it is important to understand the data you are working with and select appropriate data types in the Corpus Loader. If a metadata has the wrong data type for the filter condition you wish to apply, navigate back to the 'File Loader' tab and rebuild the corpus with the correct data types selected.

Below are explanations of the operations available for each data type.

Text operations

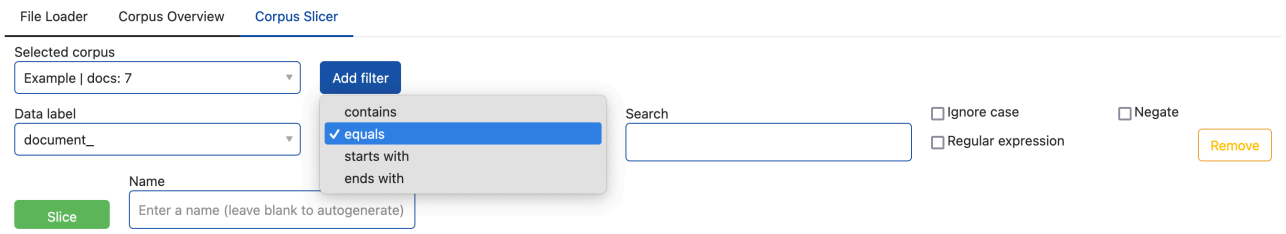
There are several operations available when filtering text data:

- Contains: Checks if the search appears anywhere in the text
  - When using the contains operation, you can additionally select 'at most' or 'at least' a specific number of occurrences
- Equals: Checks if the search matches the text in its entirety
- Starts with: Checks it the search matches the start of the text
- Ends with: Checks it the search matches the end of the text

Each text operation also includes the following checkboxes:

- Ignore case: the text will match regardless of the case of characters in the search

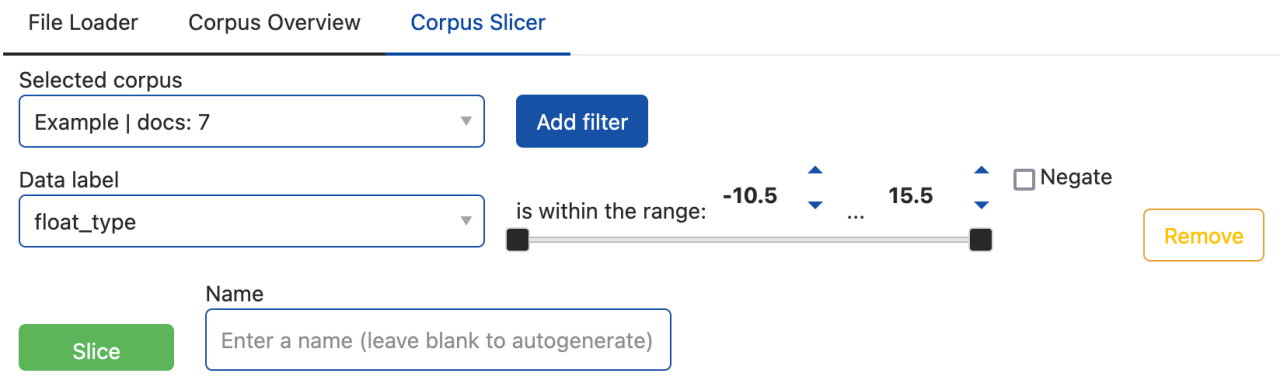
- Regular expression: your search will be interpreted as a regular expression. This allows a wide range of complex operations to be applied, but constructing regular expressions is equally as complex.



Numeric operations

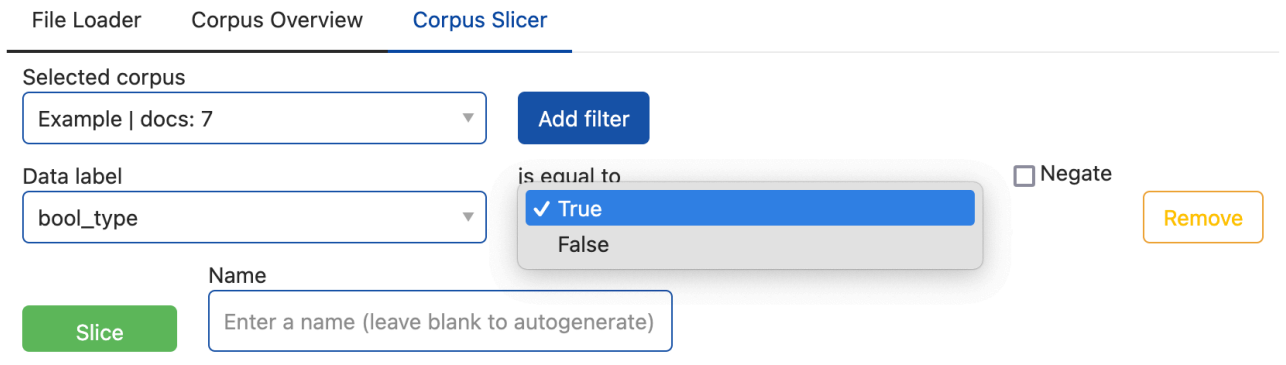
Numeric data (integer and decimal data types) has a range slider for filtering. Data inside the bounds of the range (inclusive of the bound edges) will pass the filter, while data outside the bounds will not pass the filter.

The slider can be controlled by clicking and dragging or by entering a number into the fields. The default value for the slider is the minimum and maximum values found within the data selected.



Boolean operations

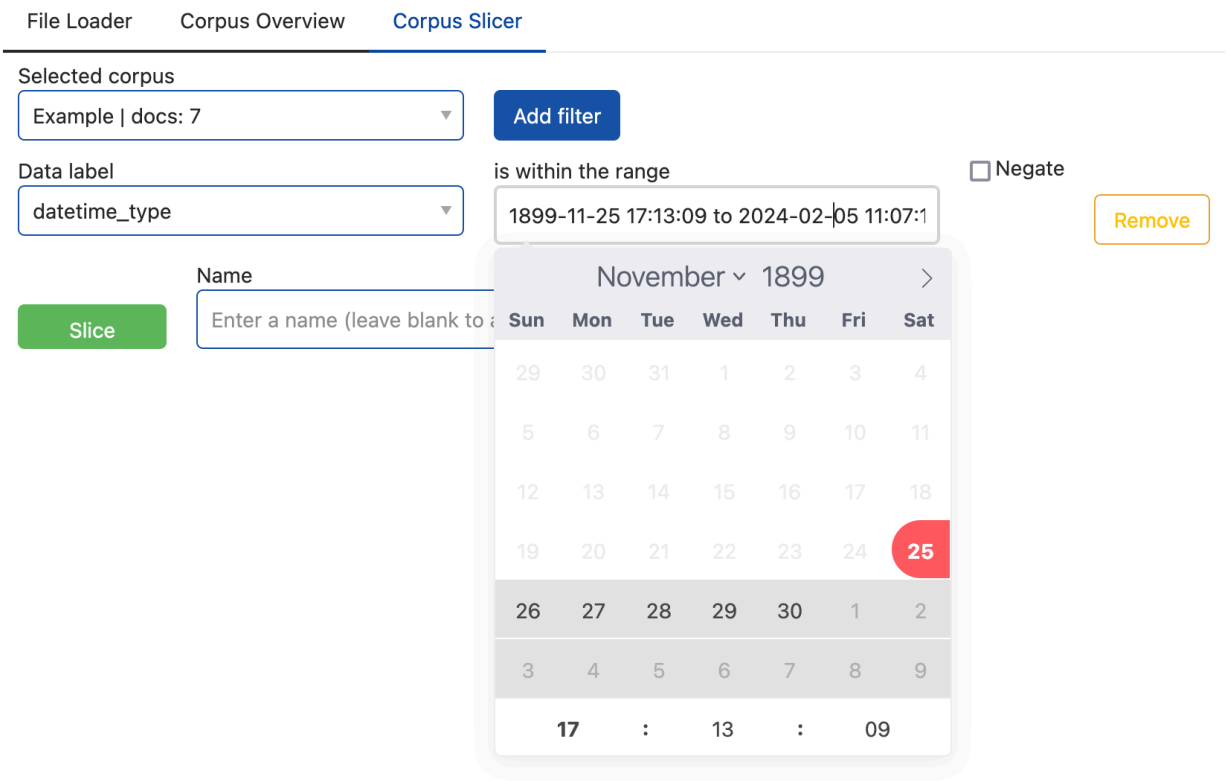
Boolean data must either be True or False, and so the boolean operation is a single dropdown with each of these values.



Datetime operations

Datetime data has a datetime picker allowing the selection of a range, similar to the number slider for numeric data. Data inside the bounds of the range (inclusive of the bound edges) will pass the filter, while data outside the bounds will not pass the filter.

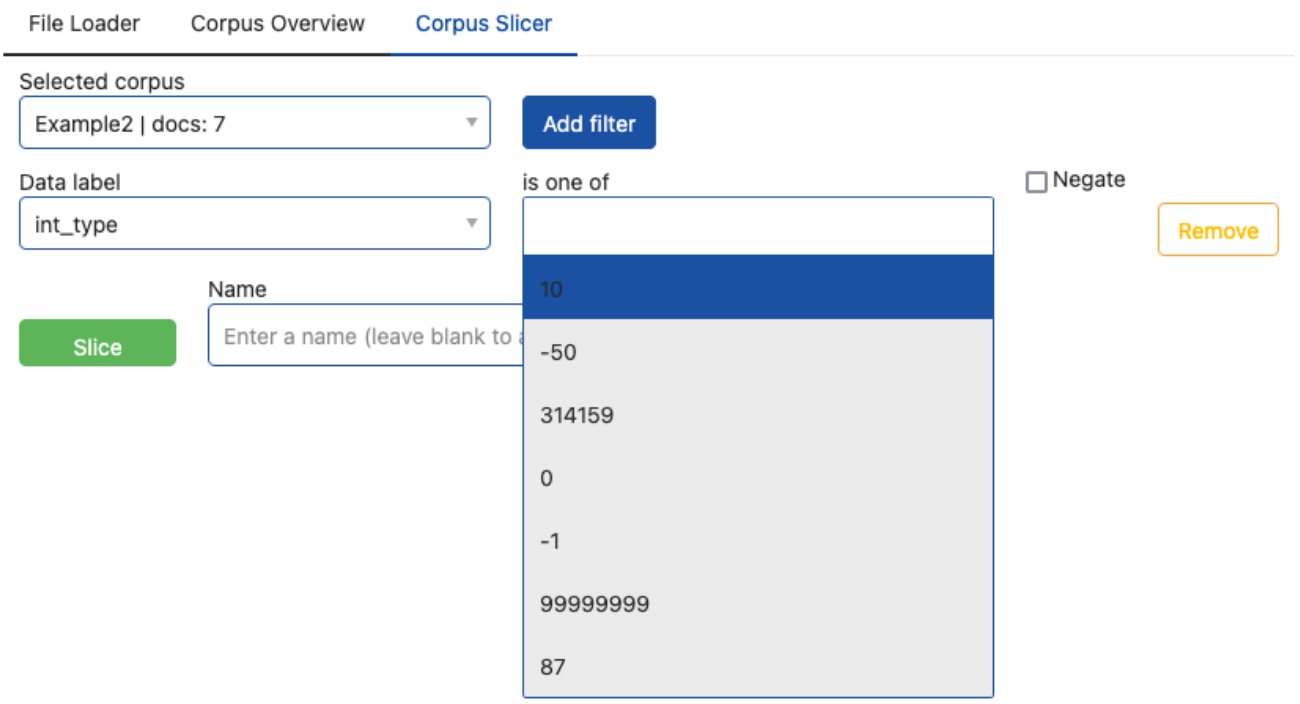
The picker allows selection of the date range in the upper section, and selection of a time range in the lower section. If your data only includes a date and no time or a time and no date, the missing component will be ignored when applying the filter.



Categorical operations

The categorical data type has a simple dropdown containing all possible values. As such it is advised to only select categorical for data that contains few unique values.

The operation allows selecting multiple values and the filter will pass if any of the values match.



Negations

Each data type filter provides a checkbox labelled 'Negate'. If selected, this inverts the filter conditions.

In the below screenshot a numeric data type is being filtered. With the 'Negate' checkbox unchecked (default), a value inside the range would pass this filter and a value outside the range would not pass this filter. With the 'Negate' checkbox checked, a value inside the range would not pass this filter and a value outside the range would pass this filter.

Slicing and exporting

Once the corpus has been selected and all filters have been configured, you can slice the corpus to produce a new filtered corpus by clicking the 'Slice' button.

There is an optional text field called 'Name' that allows setting the name of the new corpus. If left blank, the new corpus name will be autogenerated.

Note: once the corpus slicing is complete, the filtered corpus will become the selected corpus.