Corpus

Load a corpus of text documents, (optionally) tagged with categories, or change the data input signal to the corpus.

Inputs

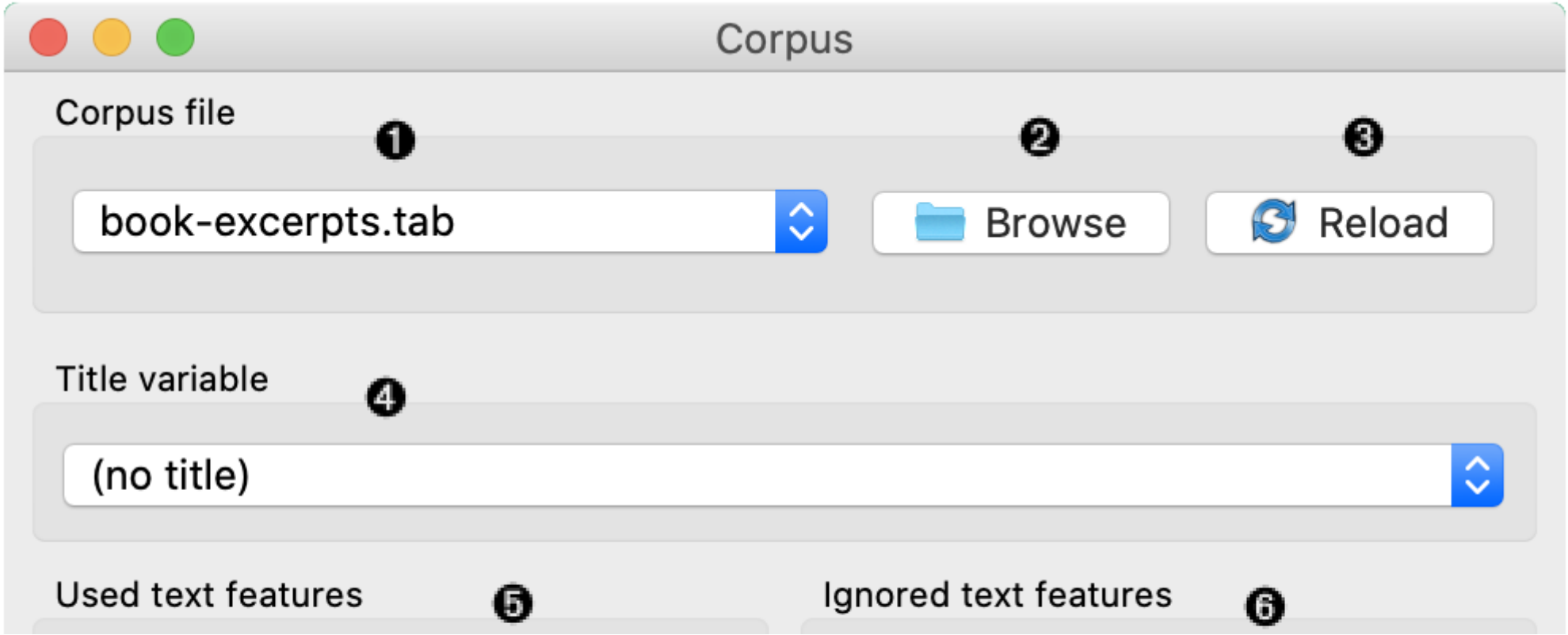
- Data: Input data (optional)

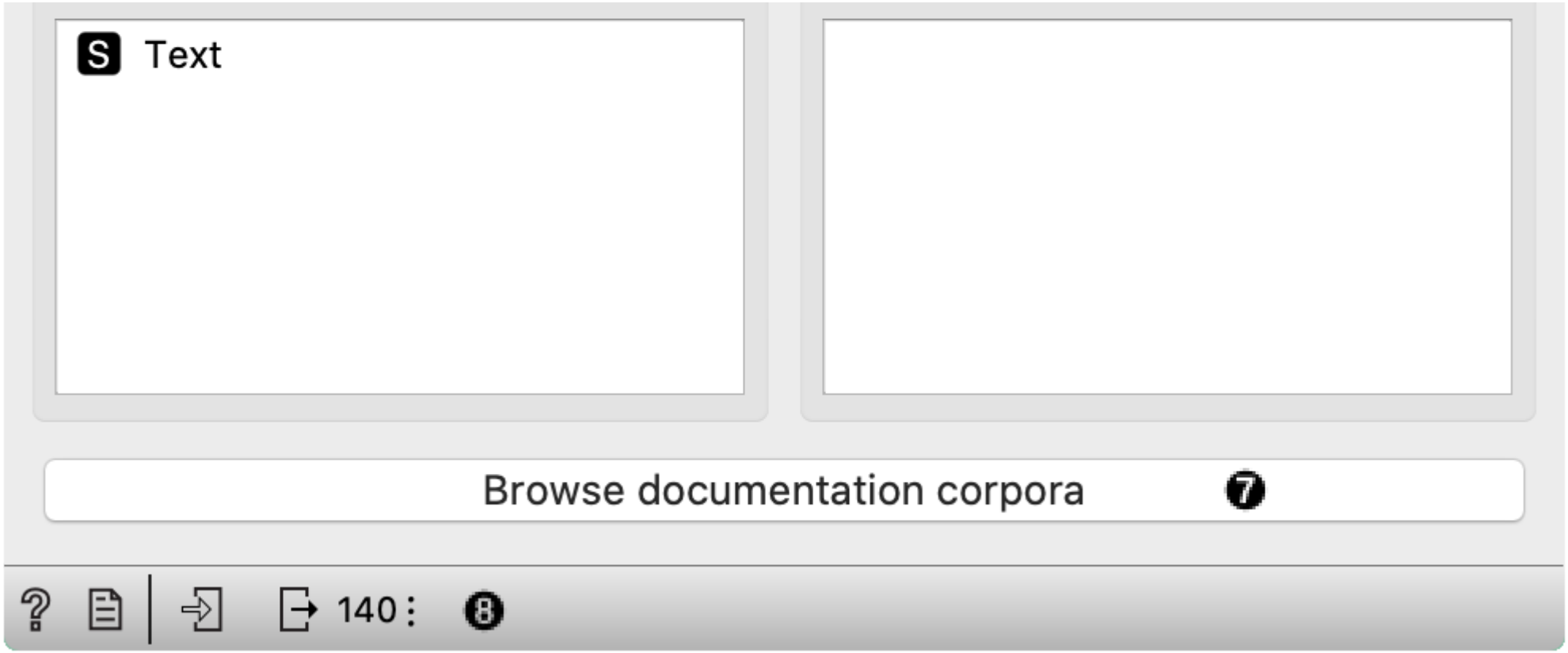
Outputs

- Corpus: A collection of documents.

Corpus widget can work in two modes:

- When no data on input, it reads text corpora from files and sends a corpus instance to its output channel. History of the most recently opened files is maintained in the widget. The widget also includes a directory with sample corpora that come pre-installed with the add-on. The widget reads data from Excel (**.xlsx**), comma-separated (**.csv**) and native tab-delimited (**.tab**) files.
- When the user provides data to the input, it transforms data into the corpus. Users can select which features are used as text features.



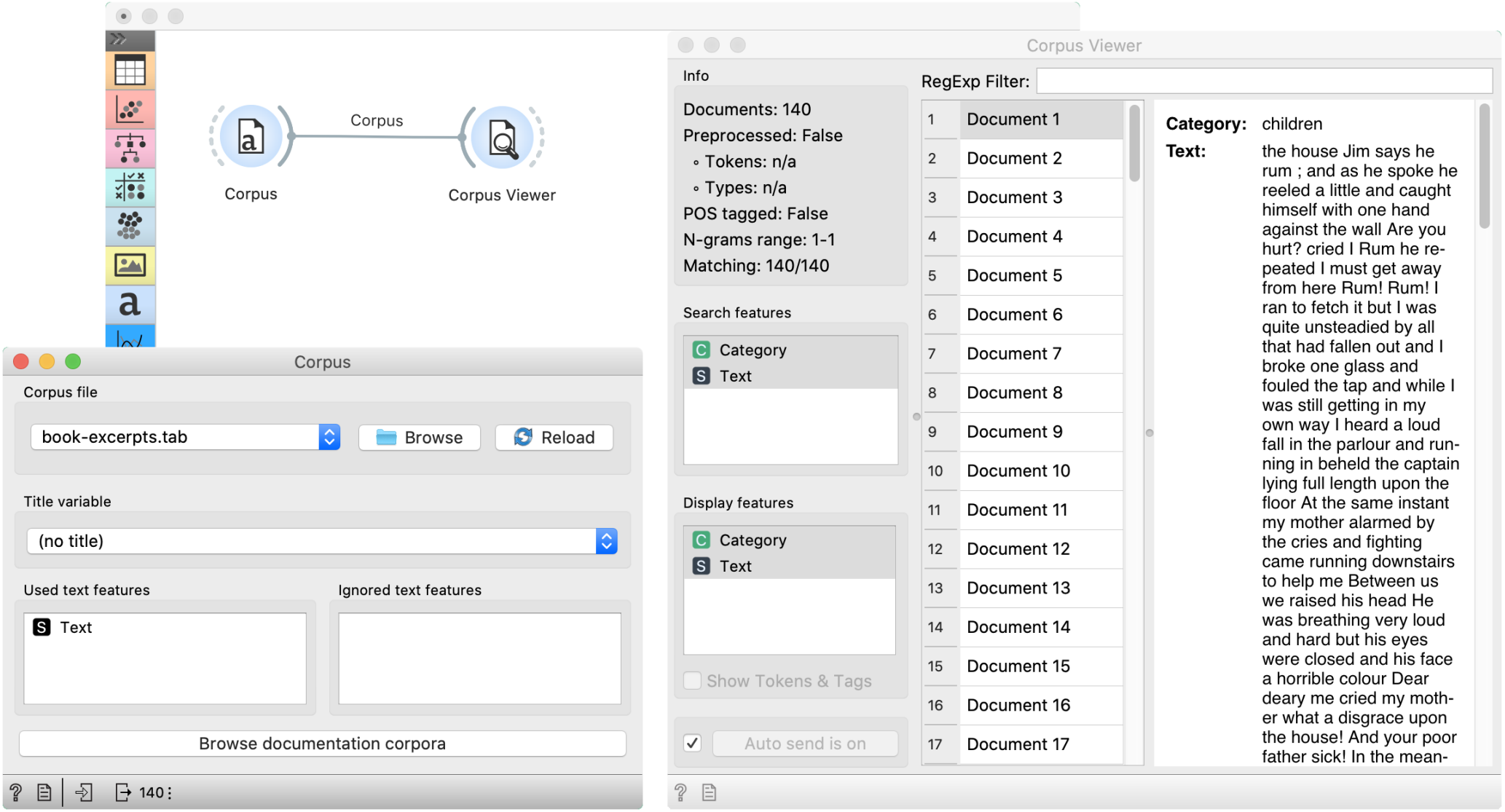


1. Browse through previously opened data files, or load any of the sample ones.
2. Browse for a data file.
3. Reloads currently selected data file.
4. Select the variable that is shown as a document title in Corpus Viewer.
5. Features that will be used in text analysis.
6. Features that won't be used in text analysis.
7. Browse through the datasets that come together with an add-on.
8. Access help, make a report and get information on the loaded data set.

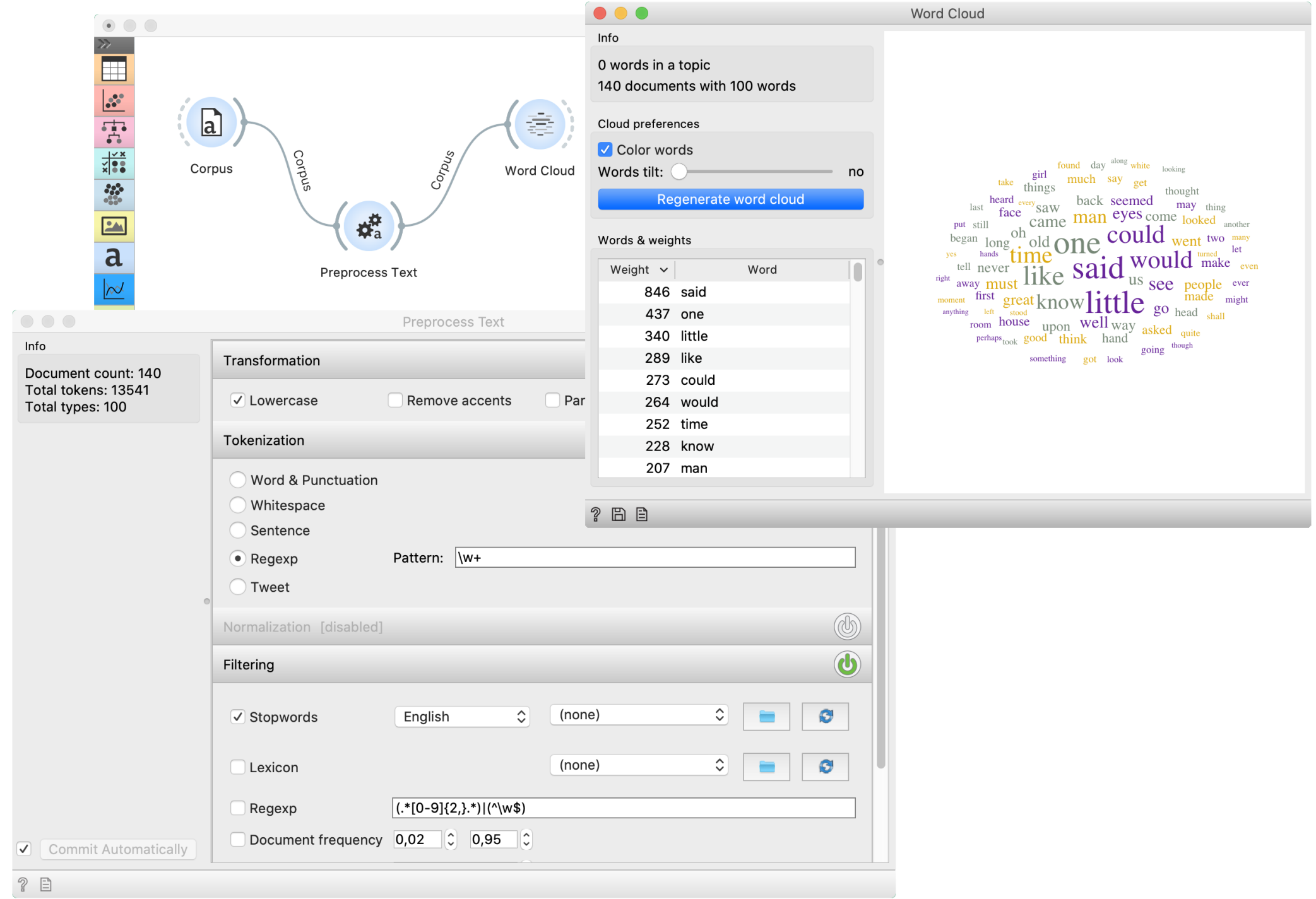
You can drag and drop features between the two boxes and also change the order in which they appear.

Example

The first example shows a very simple use of **Corpus** widget. Place **Corpus** onto canvas and connect it to Corpus Viewer. We've used *book-excerpts.tab* data set, which comes with the add-on, and inspected it in **Corpus Viewer**.



The second example demonstrates how to quickly visualize your corpus with Word Cloud. We could connect **Word Cloud** directly to **Corpus**, but instead, we decided to apply some preprocessing with Preprocess Text. We are again working with *book-excerpts.tab*. We've put all text to lowercase, tokenized (split) the text to words only, filtered out English stopwords and selected 100 most frequent tokens.


Corpus
Corpus
Preprocess Text
Corpus
Word Cloud
Word Cloud
Info
0 words in a topic
140 documents with 100 words
Cloud preferences
Color words
Words tilt:
no
Regenerate word cloud
Words & weights
Weight
Word
846 said
437 one
340 little
289 like
273 could
264 would
252 time
228 know
207 man
Preprocess Text
Info
Document count: 140
Total tokens: 13541
Total types: 100
Transformation
Lowercase
Remove accents
Par
Tokenization
Word & Punctuation
Whitespace
Sentence
Regexp
Pattern:
\w+
Tweet
Normalization [disabled]
Filtering
Stopwords
English
(none)
Lexicon
(none)
Regexp
(.*[0-9]{2,}.*)|(^\w$)
Document frequency
0,02
0,95
Commit Automatically