Document Embedding

Embeds documents from input corpus into vector space by using pretrained fastText models described in E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning Word Vectors for 157 Languages. Proceedings of the International Conference on Language Resources and Evaluation, 2018.

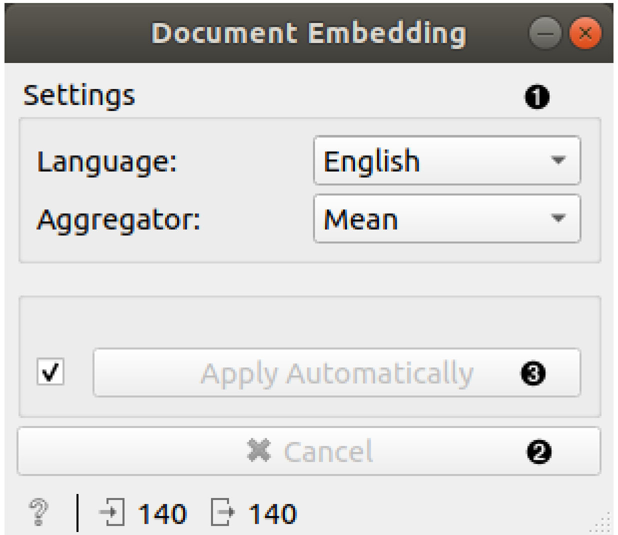
Inputs

- Corpus: A collection of documents.

Outputs

- Corpus: Corpus with new features appended.

Document Embedding parses ngrams of each document in corpus, obtains embedding for each ngram using pretrained model for chosen language and obtains one vector for each document by aggregating ngram embeddings using one of offered aggregators. Note that method will work on any ngrams but it will give best results if corpus is preprocessed such that ngrams are words (because model was trained to embed words).



Document Embedding

Settings ❶

Language: English

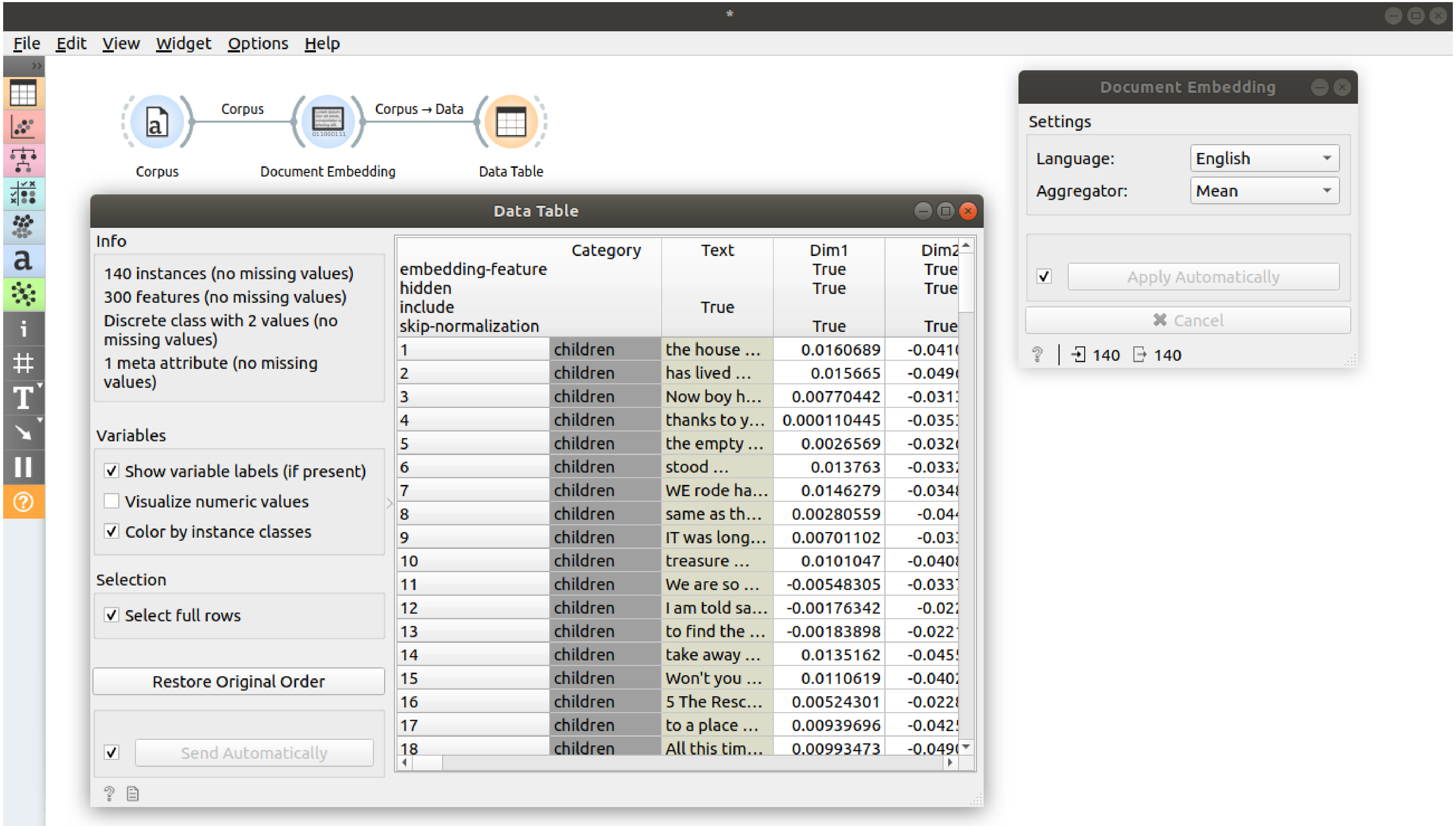
Aggregator: Mean

☑ Apply Automatically ❸

1. Widget parameters:
   - Language: widget will use a model trained on documents in chosen language.
   - Aggregator: operation to perform on ngram embeddings to aggregate them into a single document vector.
2. Cancel current execution.
3. If *Apply automatically* is checked, changes in parameters are sent automatically. Alternatively press *Apply*.

Examples

In first example, we will inspect how the widget works. Load *book-excerpts.tab* using Corpus widget and connect it to **Document Embedding**. Check the output data by connecting **Document Embedding** to **Data Table**. We see additional 300 features that we widget has appended.



Data Table

Info

140 instances (no missing values)
300 features (no missing values)
Discrete class with 2 values (no missing values)
1 meta attribute (no missing values)

Variables

☑ Show variable labels (if present)
□ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

☑ Send Automatically

		Category	Text	Dim1	Dim2
embedding-feature				True	True
hidden				True	True
include			True		
skip-normalization				True	True
1		children	the house ...	0.0160689	-0.041
2		children	has lived ...	0.015665	-0.049
3		children	Now boy h...	0.00770442	-0.031
4		children	thanks to y...	0.000110445	-0.035
5		children	the empty ...	0.0026569	-0.032
6		children	stood ...	0.013763	-0.033
7		children	WE rode ha...	0.0146279	-0.034
8		children	same as th...	0.00280559	-0.04
9		children	IT was long...	0.00701102	-0.03
10		children	treasure ...	0.0101047	-0.040
11		children	We are so ...	-0.00548305	-0.033
12		children	I am told sa...	-0.00176342	-0.02
13		children	to find the ...	-0.00183898	-0.022
14		children	take away ...	0.0135162	-0.045
15		children	Won't you ...	0.0110619	-0.040
16		children	5 The Resc...	0.00524301	-0.022
17		children	to a place ...	0.00939696	-0.042
18		children	All this tim...	0.00993473	-0.049

Document Embedding

Settings

Language: English
Aggregator: Mean

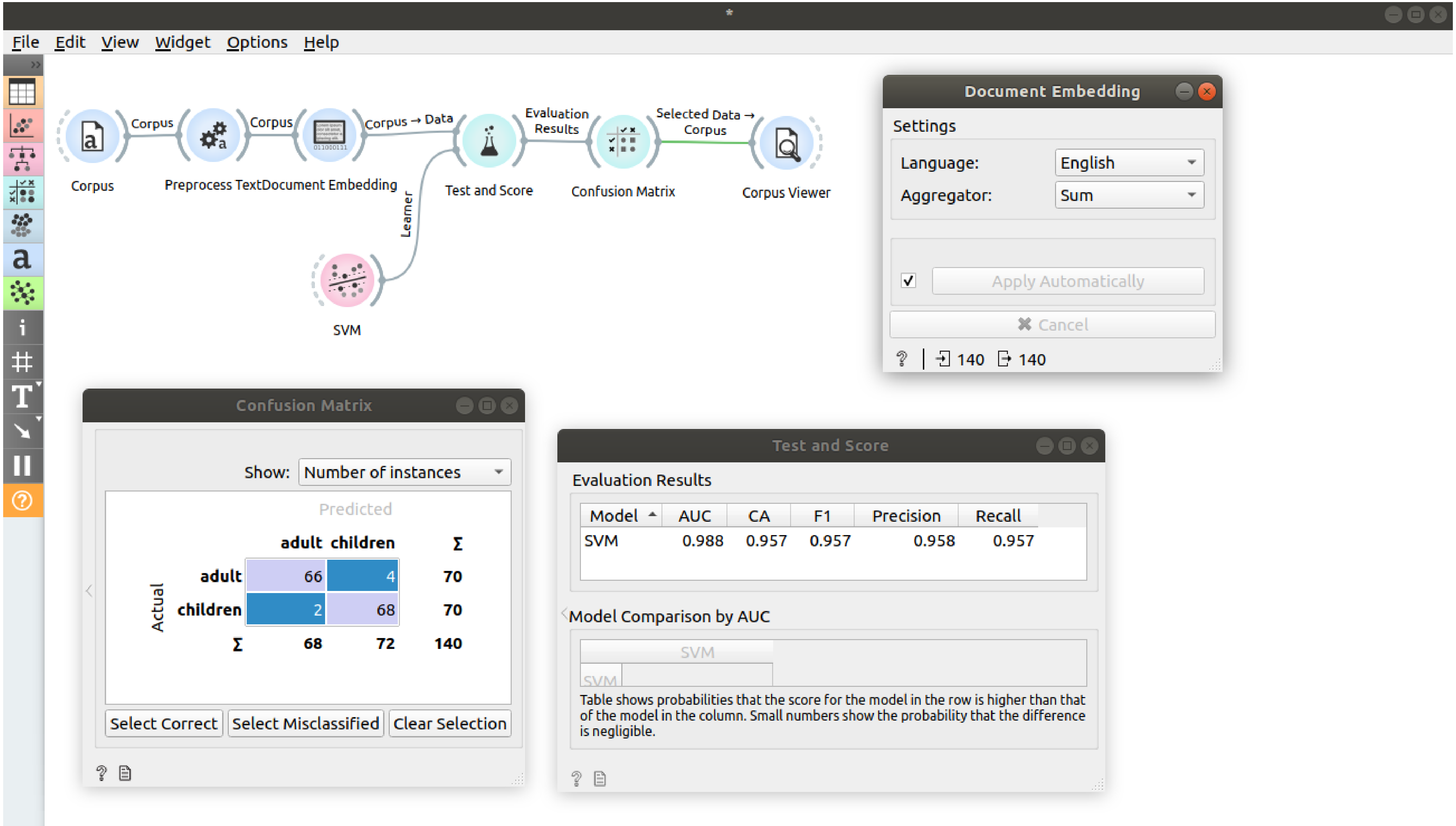
☑ Apply Automatically

140 140

In the second example we will try to predict document category. We will keep working on *book-excerpts.tab* loaded with Corpus widget and sent through Preprocess Text with default parameters. Connect **Preprocess Text**to **Document Embedding** to obtain features for predictive modelling. Here we set aggregator to Sum.

Connect **Document Embedding** to **Test and Score** and also connect learner of choice to the left side of **Test and Score**. We chose SVM and changed kernel to Linear. **Test and Score** will now compute performance of each learner on the input. We can see that we achieved great results.

Let’s now inspect confusion matrix. Connect **Test and Score** to **Confusion Matrix**. Clicking on *Select Misclassified* will output documents that were misclassified. We can further inspect them by connecting Corpus Viewer to **Confusion Matrix**.



Document Embedding

Settings

Language: English

Aggregator: Sum

☑ Apply Automatically

✖ Cancel

? 140 140

Confusion Matrix

Show: Number of instances

Predicted

Actual	adult	children	Σ
adult	66	4	70
children	2	68	70
Σ	68	72	140

Select Correct | Select Misclassified | Clear Selection

Test and Score

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
SVM	0.988	0.957	0.957	0.958	0.957

Model Comparison by AUC

	SVM
SVM	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.