Statistics

Create new statistic variables for documents.

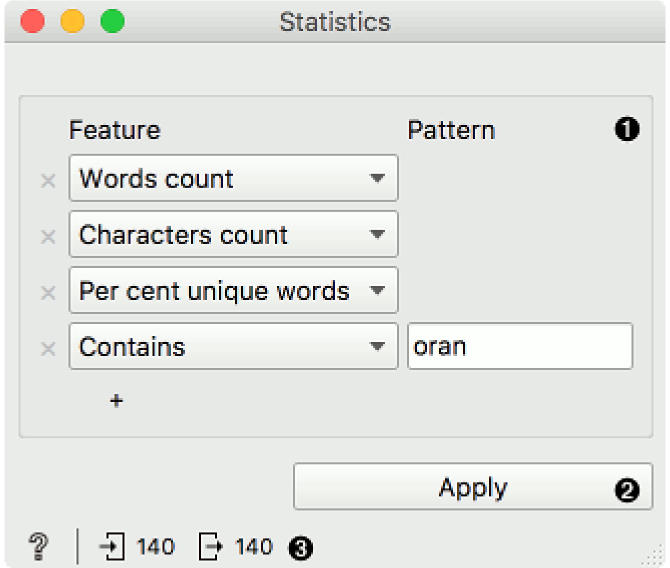
Inputs

- Corpus: A collection of documents.

Outputs

- Corpus: Corpus with additional attributes.

Statistics is a feature constructor widget that adds simple document statistics to a corpus. It supports both standard statistical measures and user-defined variables.



1. Add or remove features. Features can be added with the + sign below. They can be removed with the x sign on the left side. Feature options are:
   - Words count: number of words in the document.
   - Characters count: number of characters in the document.
   - N-grams count: number of n-grams. Define n-grams in [Preprocess Text], otherwise only unigrams will be reported.
   - Average word length: ratio between character count and the number of words
   - Punctuations count: number of punctuations
   - Capitals count: number of capital letters

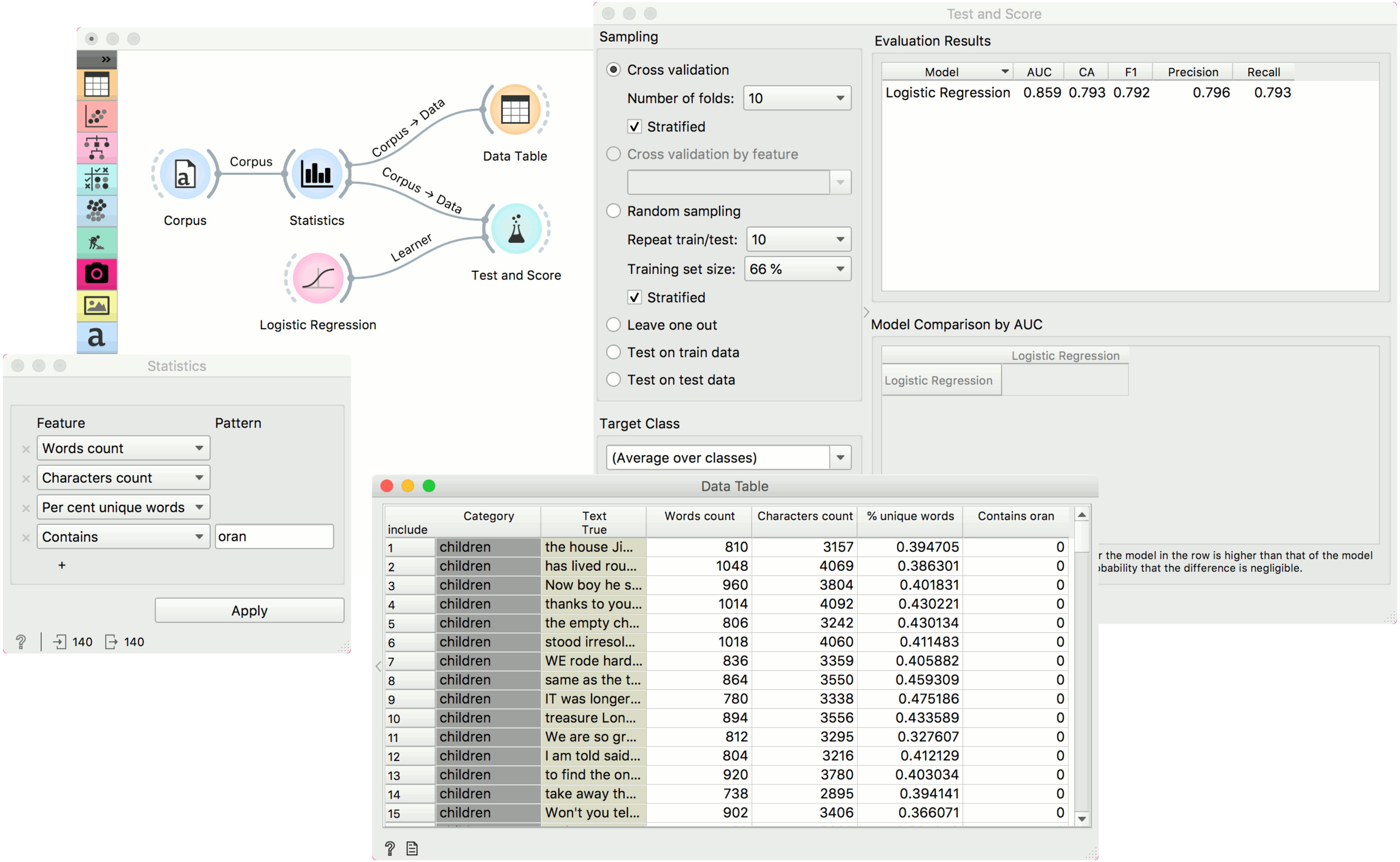
- Vowels count: number of vowels. The default is ‘a, e, i, o, u’, but the user can add her own.
- Consonants count: number of consonants. Default is given, but the user can adjust it.
- Per cent unique words: ratio of unique words to all the words (types/tokens).
- Starts with: number of times a token begins with the specified sequence.
- Ends with: number of times a token ends with the specified sequence.
- Contains: number of times a specified sequence is in the token.
- Regex: number of times the provided regular expression matches the token.
- POS tag: count specified POS tags. Requires POS tagged tokens from Preprocess Text. List of Tree POS tags for English can be found here.

2. Press Apply to output corpus with new features.
3. Status line with help on the left and input and output on the right.

Example

Here is a simple example how **Statistics** widget works. As it is a basic feature construction widget, it can be used directly after Corpus. We have added a couple of features, namely word count, character count, percent unique words and number of words containing ‘oran’. We can observe the table with additional columns in a Data Table.

We can also use the output of Statistics for predictive modeling with Test and Score. Normally, however, we would use Statistics only to enhance features from the Bag of Words widget. Some features require POS tagged tokens, which can be created with Preprocess Text widget.



Test and Score

Sampling

● Cross validation

Number of folds: 10

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Target Class

(Average over classes)

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.859	0.793	0.792	0.796	0.793

Model Comparison by AUC

	Logistic Regression
Logistic Regression	

r the model in the row is higher than that of the model
›bability that the difference is negligible.

Statistics

	Feature	Pattern
×	Words count	
×	Characters count	
×	Per cent unique words	
×	Contains	oran

+

Apply

140 140

Data Table

	Category	Text	Words count	Characters count	% unique words	Contains oran
include		True				
1	children	the house Ji...	810	3157	0.394705	0
2	children	has lived rou...	1048	4069	0.386301	0
3	children	Now boy he s...	960	3804	0.401831	0
4	children	thanks to you...	1014	4092	0.430221	0
5	children	the empty ch...	806	3242	0.430134	0
6	children	stood irresol...	1018	4060	0.411483	0
7	children	WE rode hard...	836	3359	0.405882	0
8	children	same as the t...	864	3550	0.459309	0
9	children	IT was longer...	780	3338	0.475186	0
10	children	treasure Lon...	894	3556	0.433589	0
11	children	We are so gr...	812	3295	0.327607	0
12	children	I am told said...	804	3216	0.412129	0
13	children	to find the on...	920	3780	0.403034	0
14	children	take away th...	738	2895	0.394141	0
15	children	Won't you tel...	902	3406	0.366071	0