
Orange3 Text Mining Documentation

Release 0.1.1

Biolab

January 26, 2017

1	Widgets	1
2	Scripting Reference	3
3	Indices and tables	5

1.1 Preprocess

The description of the tokenizer goes here.

Scripting Reference

2.1 Preprocessor

```
class orangecontrib.text.preprocess.Preprocessor (incl_punct=False, lowercase=True,  
stop_words='english', trans=None,  
min_df=1)
```

Holds pre-processing flags and other information, about stop word removal, lowercasing, text morphing etc.(the options are set via the Preprocess widget).

```
Preprocessor.__init__ (incl_punct=False, lowercase=True, stop_words='english', trans=None,  
min_df=1)
```

Parameters

- **incl_punct** (*boolean*) – Determines whether the tokenizer should include punctuation in the tokens.
- **lowercase** (*boolean*) – If set, transform the tokens to lower case, before returning them.
- **stop_words** (*'english' or list or None*) – Determines whether stop words should(“english”), or should not(None) be removed. If this is list, it should contain stop-words.
- **trans** – An optional pre-processor object to perform the morphological transformation on the tokens before returning them.

Returns

class *orangecontrib.text.preprocess.Preprocessor*

Indices and tables

- `genindex`
- `modindex`
- `search`

Symbols

`__init__()` (orangecontrib.text.preprocess.Preprocessor
method), 3

P

Preprocessor (class in orangecontrib.text.preprocess), 3