НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Лаборатория интернет-исследований

ПРЕПРОЦЕССИНГ РУССКОЯЗЫЧНЫХ ТЕКСТОВ В

Кольцов С.Н

Санкт-Петербург, 2018

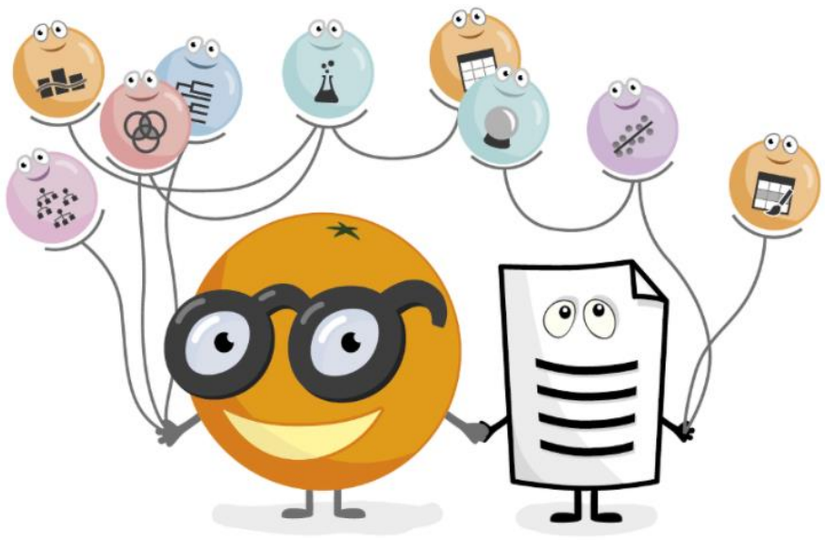
ПРЕПРОЦЕССИНГ

Препроцессинг состоит из следующих этапов

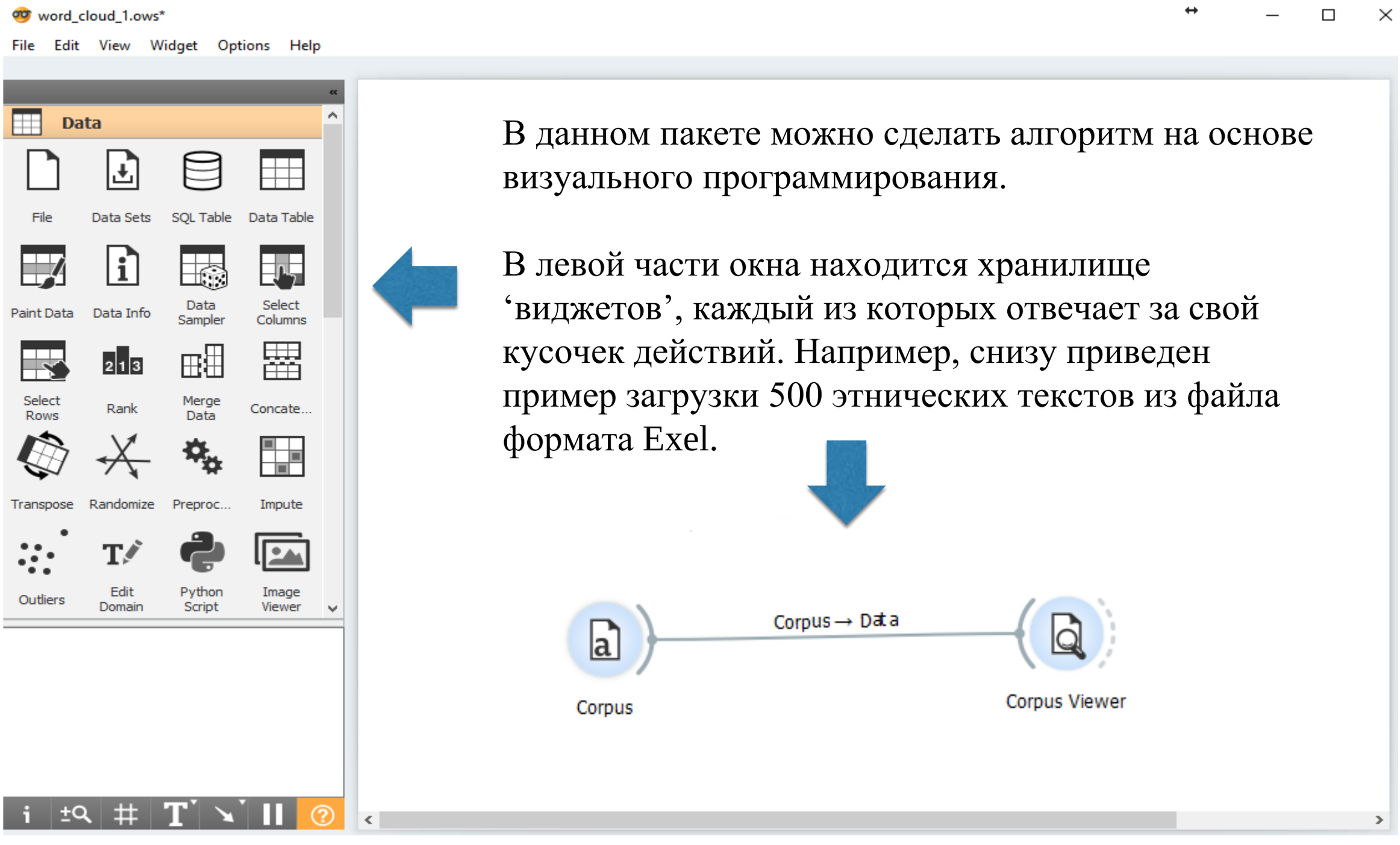
1. **Загрузка данных из источника данных**, например csv файл.
2. **Процедура лематизации**. Данная процедура заключается в процессе токинизации текстов.
3. **Удаление стоп слов**. Некоторые слова являются общими и при этом часто встречаются, поэтому их необходимо удалять.
4. **Визуализация текстовых данных в виде облака слов**. Наиболее частотные слова можно представить в виде облака слов, что упрощает некоторое понимание того какие темы внутри коллекции документов.

В дальнейшем использовать пакет Orange.

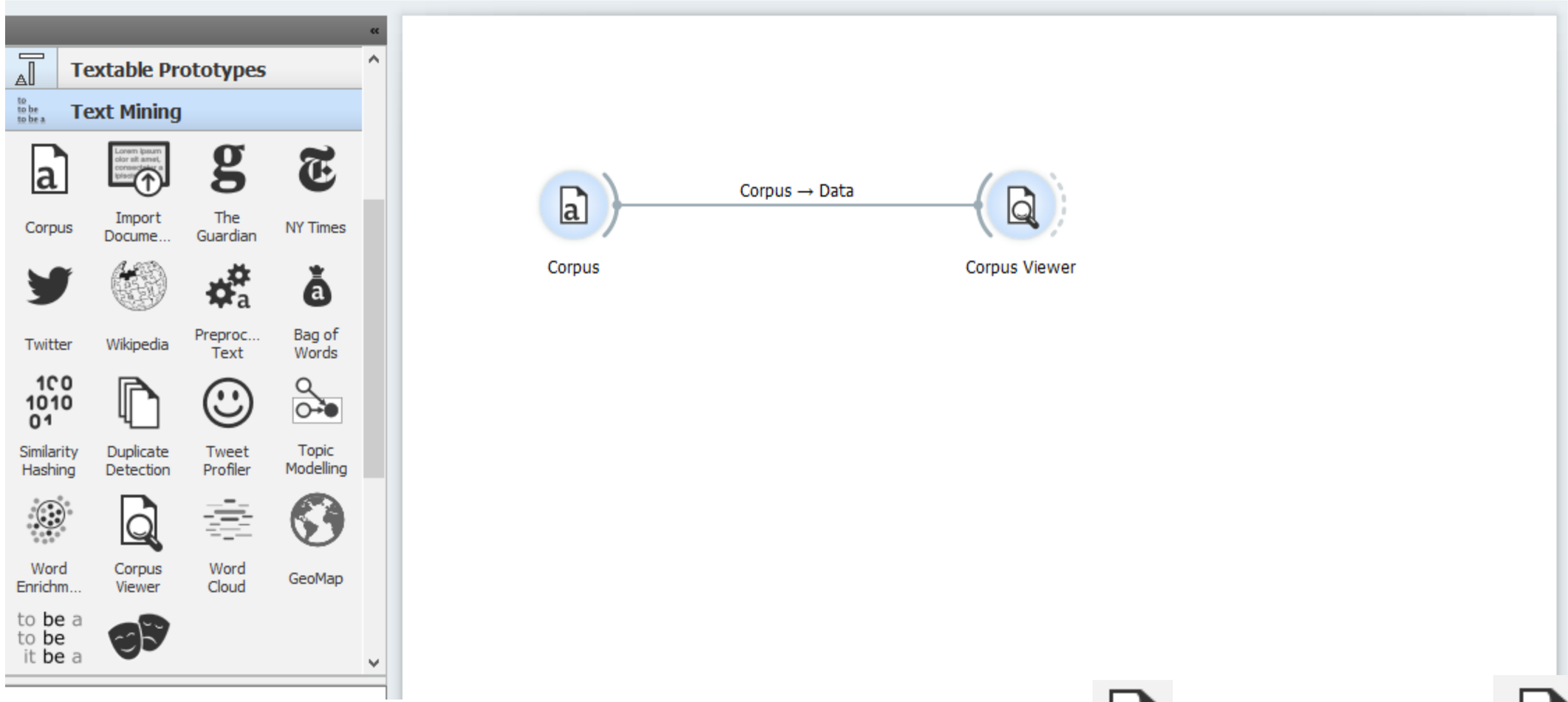
Данный пакет можно скачать по адресу:
https://orange.biolab.si/



ORANGE


word_cloud_1.ows*
File Edit View Widget Options Help
Data
File
Data Sets
SQL Table
Data Table
Paint Data
Data Info
Data Sampler
Select Columns
Select Rows
Rank
Merge Data
Concate...
Transpose
Randomize
Preproc...
Impute
Outliers
Edit Domain
Python Script
Image Viewer
В данном пакете можно сделать алгоритм на основе визуального программирования.
В левой части окна находится хранилище ‘виджетов’, каждый из которых отвечает за свой кусочек действий. Например, снизу приведен пример загрузки 500 этнических текстов из файла формата Exel.
Corpus → Data
Corpus
Corpus Viewer

ЗАГРУЗКА РУССКОЯЗЫЧНЫХ ДАННЫХ В ORANGE



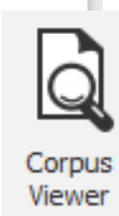
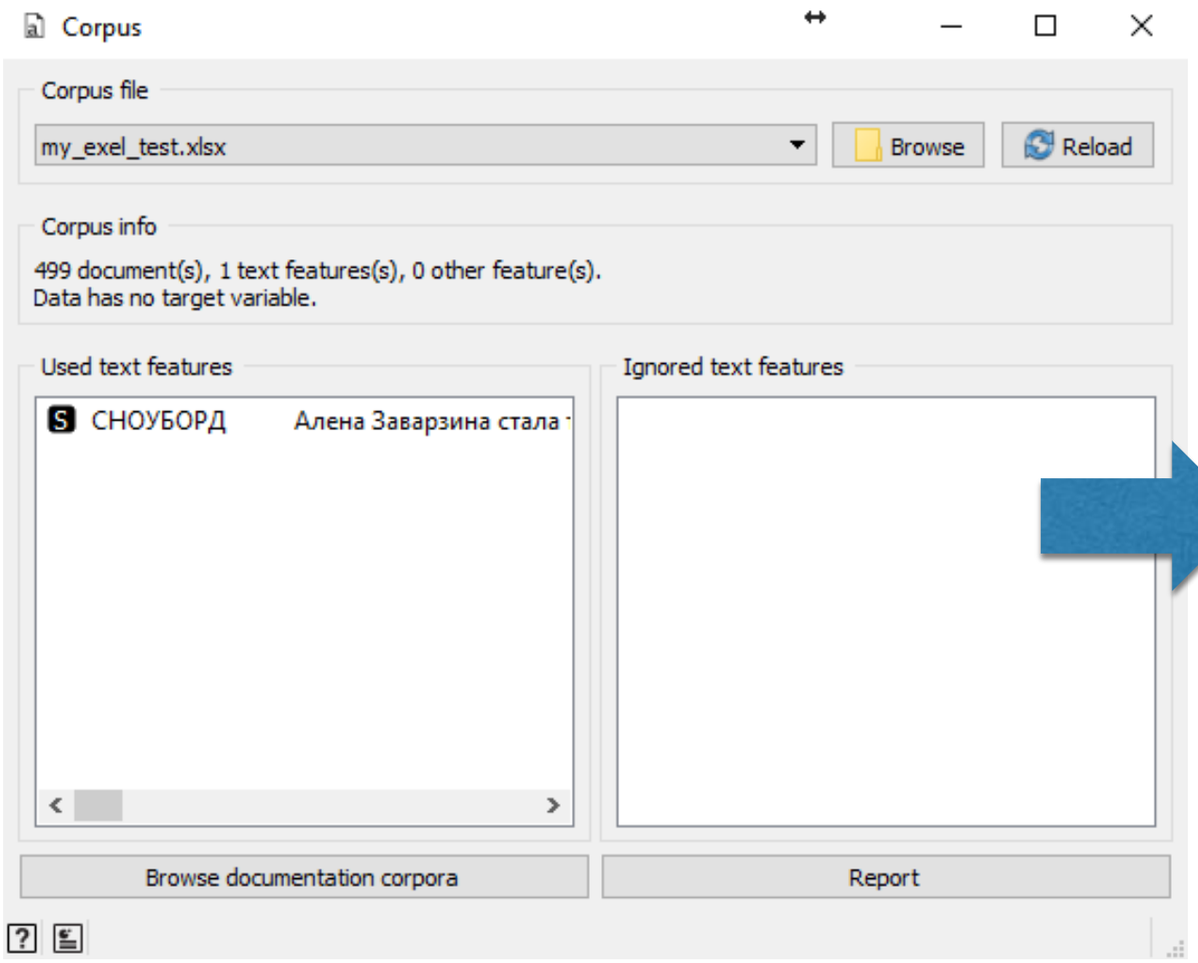
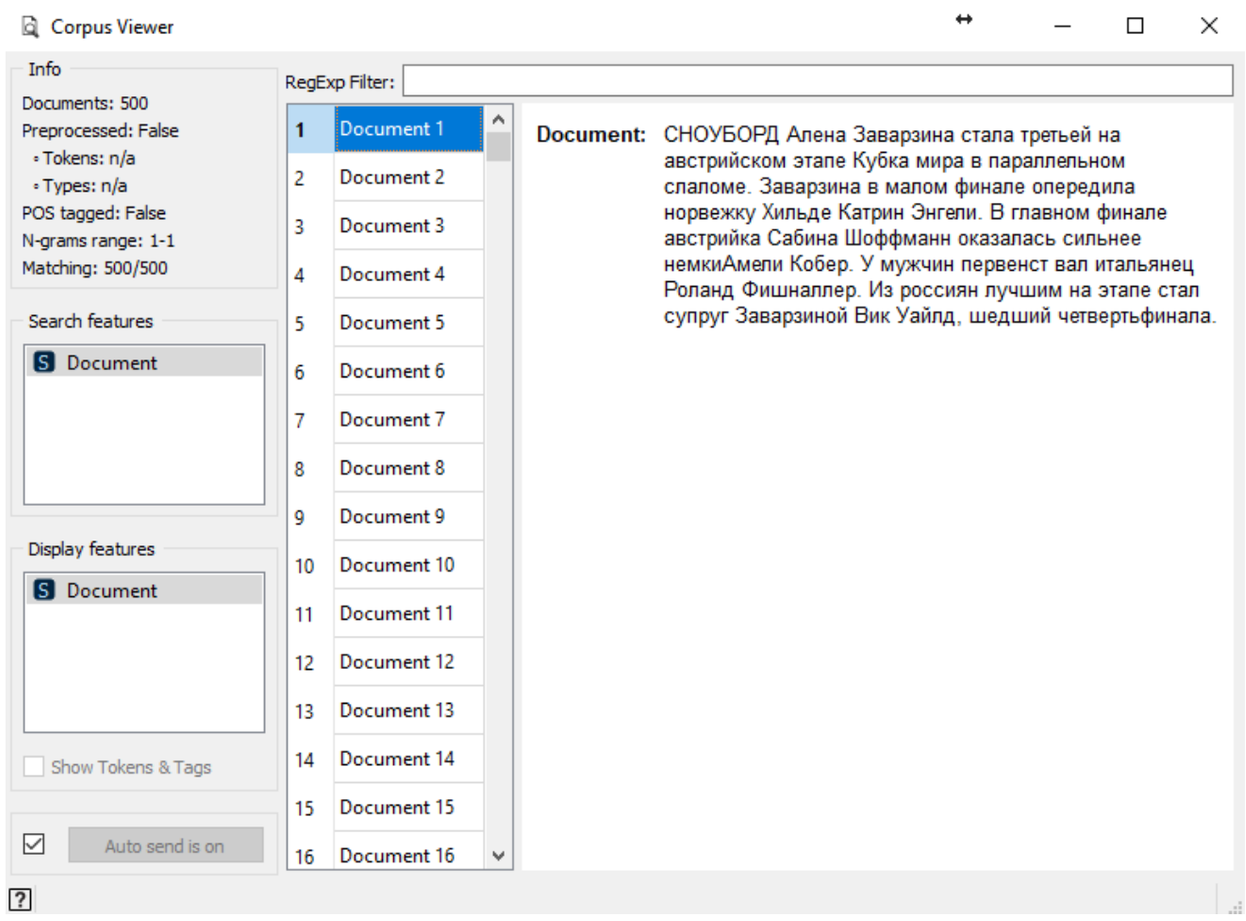
На вкладке ‘Text Mining’ нужно взять виджеты: ‘Corpus’  ‘Corpus Viewer’ 
Далее нужно соединить два виджета линей.

Схема загрузки готова, теперь нужно лишь указать имя файла для загрузки и просмотра.

ЗАГРУЗКА И ПРОСМОТР РУССКОЯЗЫЧНЫХ ТЕКСТОВ

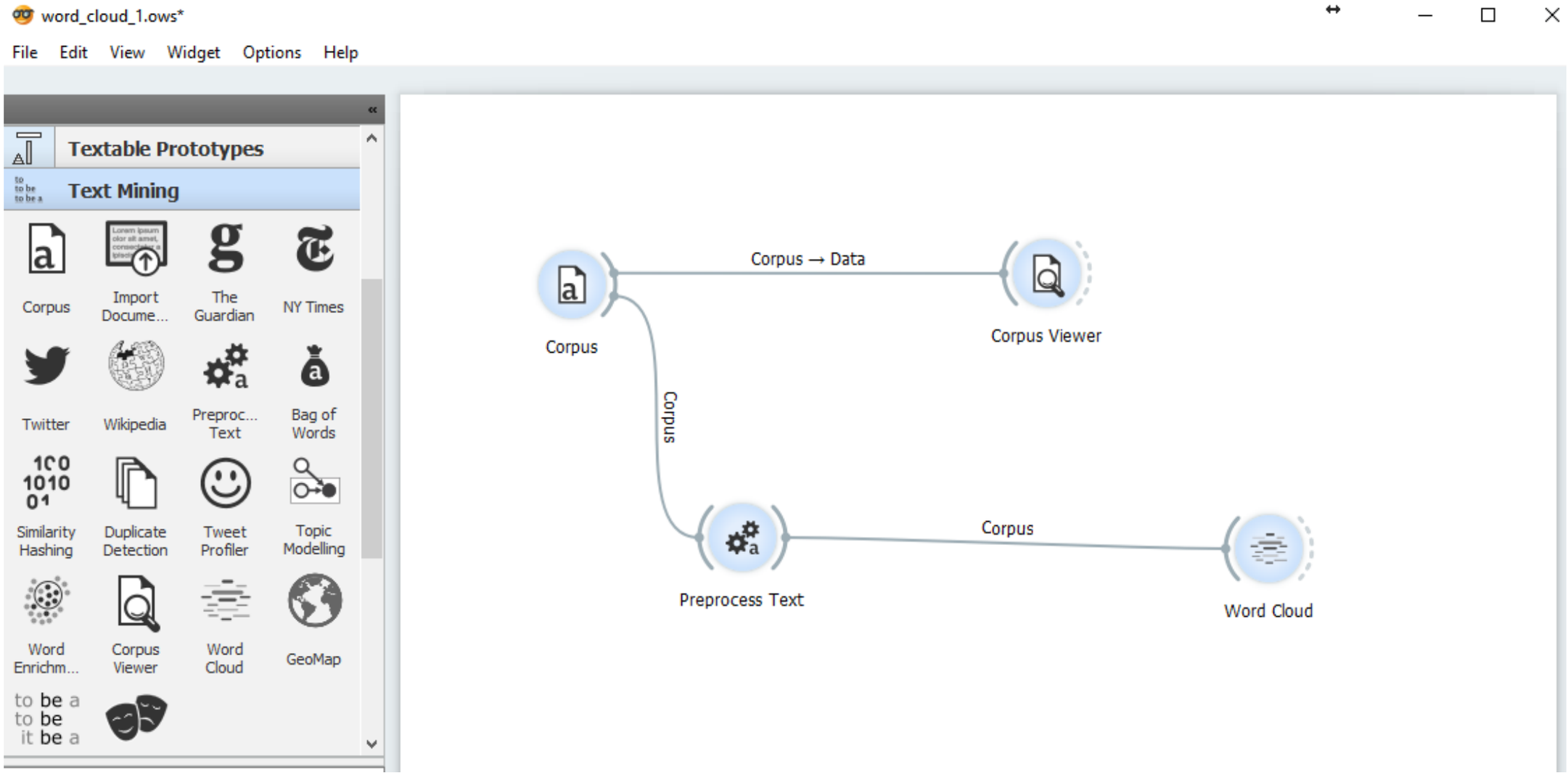
Для того что бы загрузить данные нужно кликнуть на виджете ‘Corpus’ и указать имя файла в формате книги Excel.



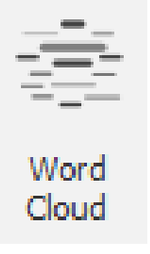


Для того что бы просмотреть результаты загрузки нужно кликнуть на виджете ‘Corpus Viewer’.

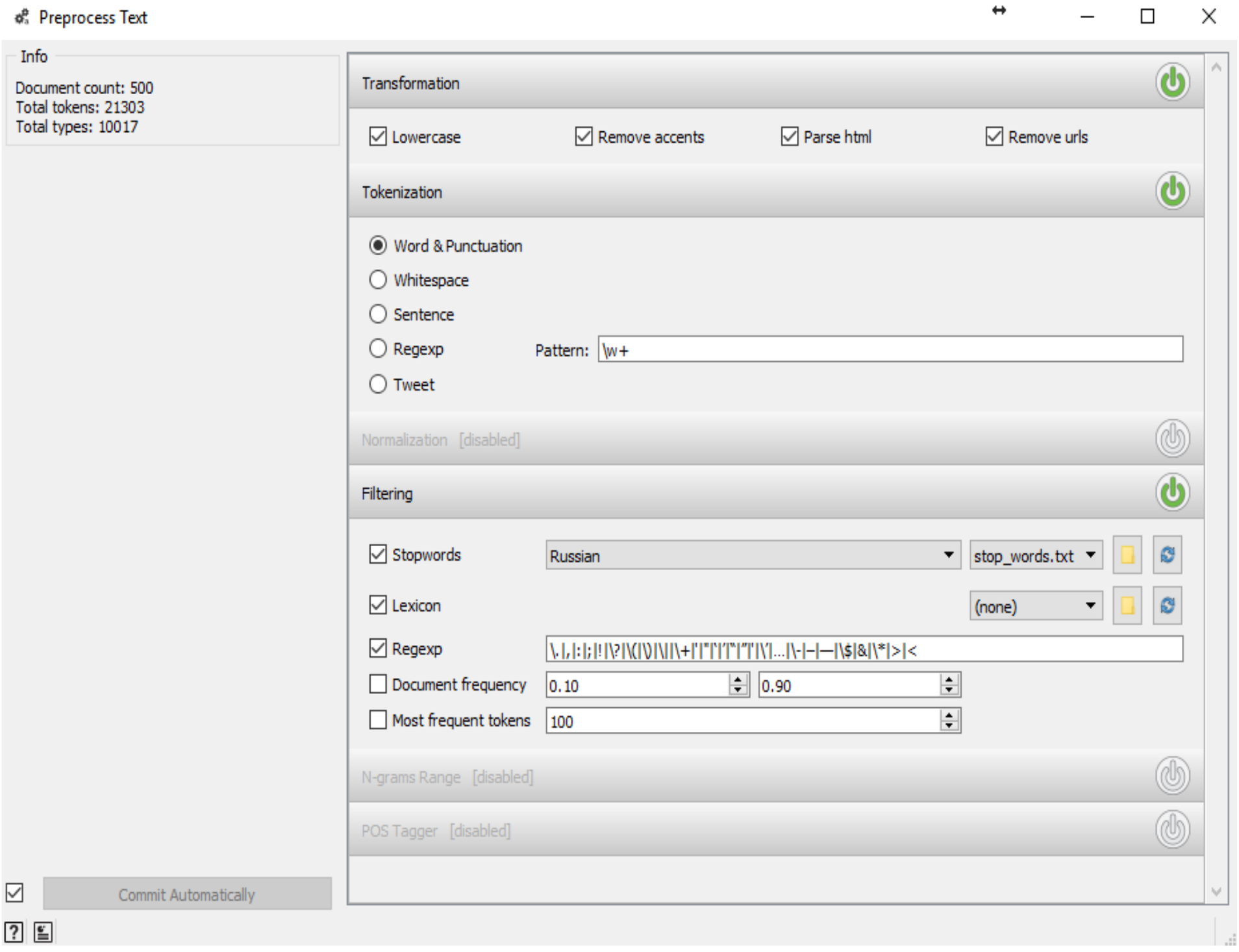
ПРЕПРОЦЕССИНГ И ОБЛАКО ЛЕМАТИЗИРОВАННЫХ СЛОВ



Для удаления стоп слов, пунктуации (и так далее), а также для построения облака слов, достаточно добавить еще два виджета:

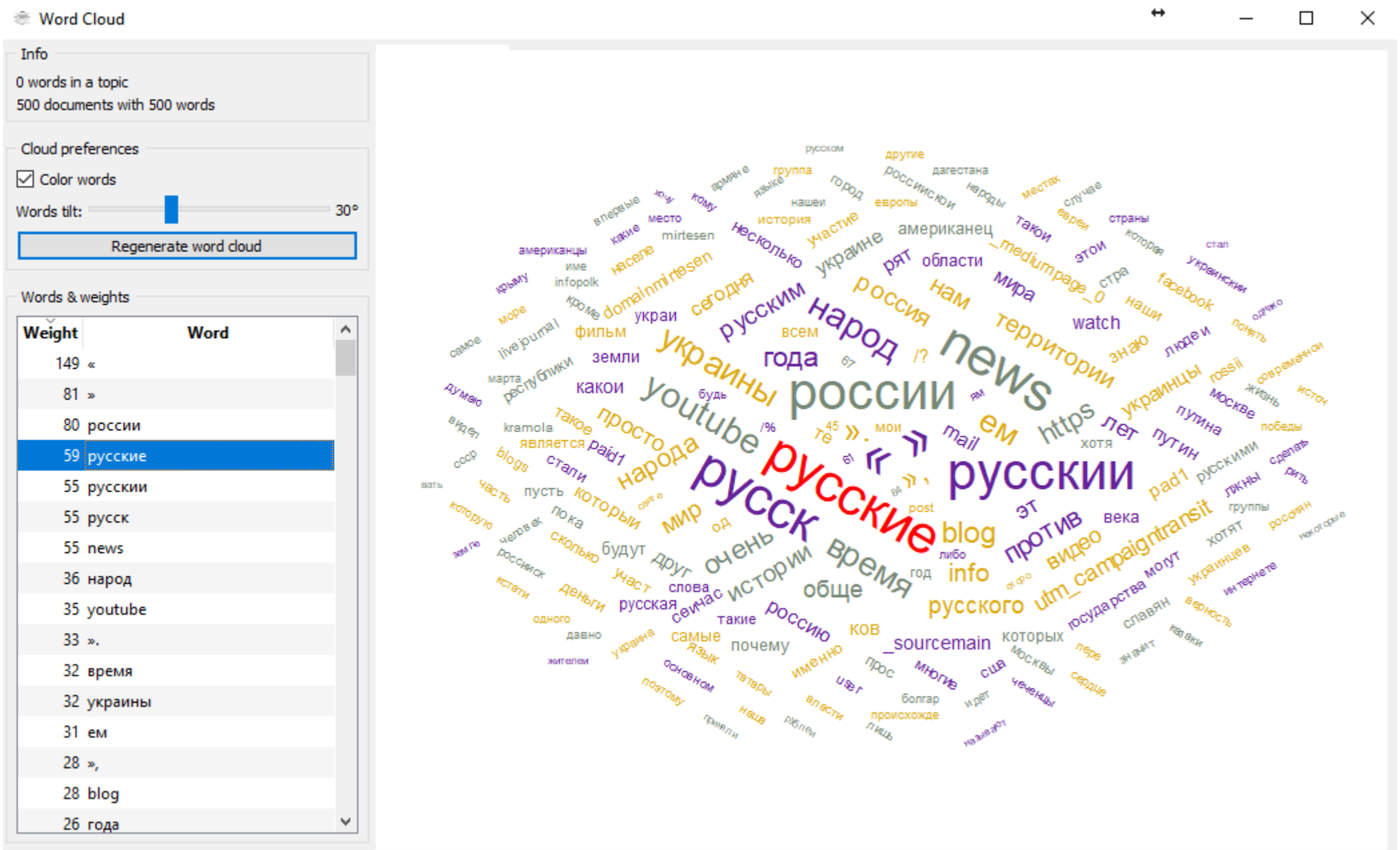


ПРЕПРОЦЕССИНГ В ORANGE



1. Можно указать параметры препроцессинга
2. Указать файл со списком стоп слов.
3. Указать язык.

ОБЛАКО НЕЛЕМАТИЗИРОВАННЫХ СЛОВ



Info

0 words in a topic
500 documents with 500 words

Cloud preferences

☑ Color words

Words tilt: 30°

Words & weights

Weight	Word
149	«
81	»
80	россии
59	русские
55	русскии
55	русск
55	news
36	народ
35	youtube
33	».
32	время
32	украины
31	ем
28	»,
28	blog
26	года

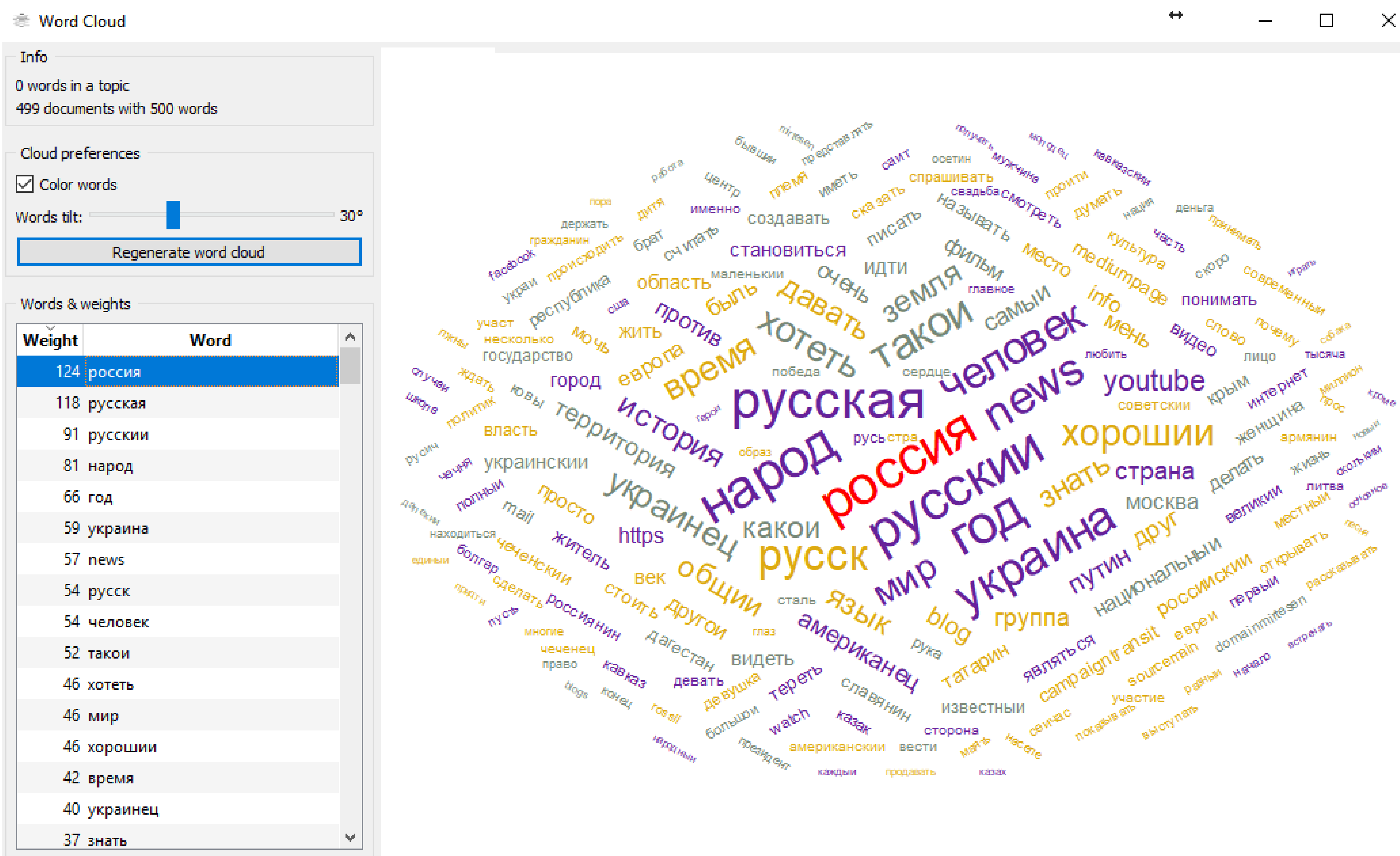
Таким образом, за счет списка стоп слов можно построит нормальное облако слов.

SNOWBALL STEMMER

Данный стеммер просто обрезает концы слов, то есть не годится для русского языка

TOPICMINER - MYSTEM

Если сделать препроцессинг в TopicMiner, и выгрузить лематизированные данные в csv формате, и загрузить в Orange, то получится наиболее качественный вариант облака.



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

https://linis.hse.ru/

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru