НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Internet Studies Lab, Department of Applied Mathematics and Business Informatics

ПРЕПРОЦЕССИНГ РУССКОЯЗЫЧНЫХ ТЕКСТОВ В

Анализ баз данных в публичном управлении

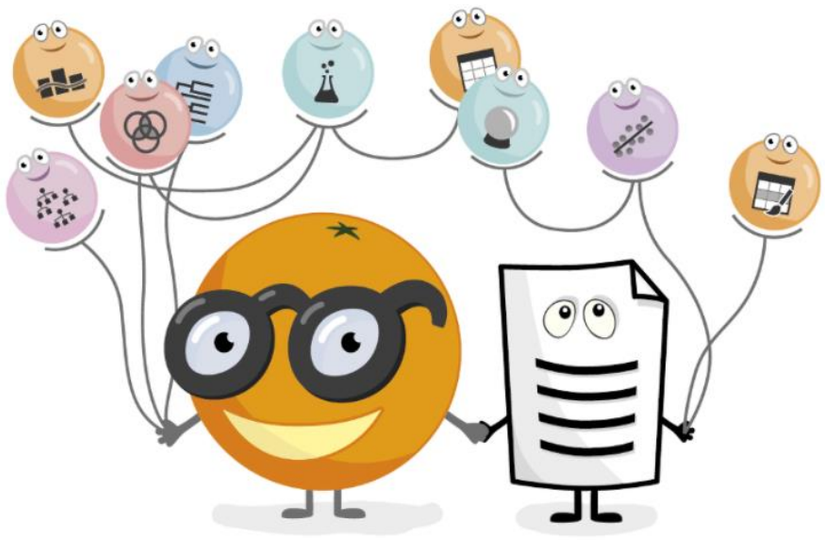
Кольцов С.Н.

Saint Petersburg, 07.09.2018

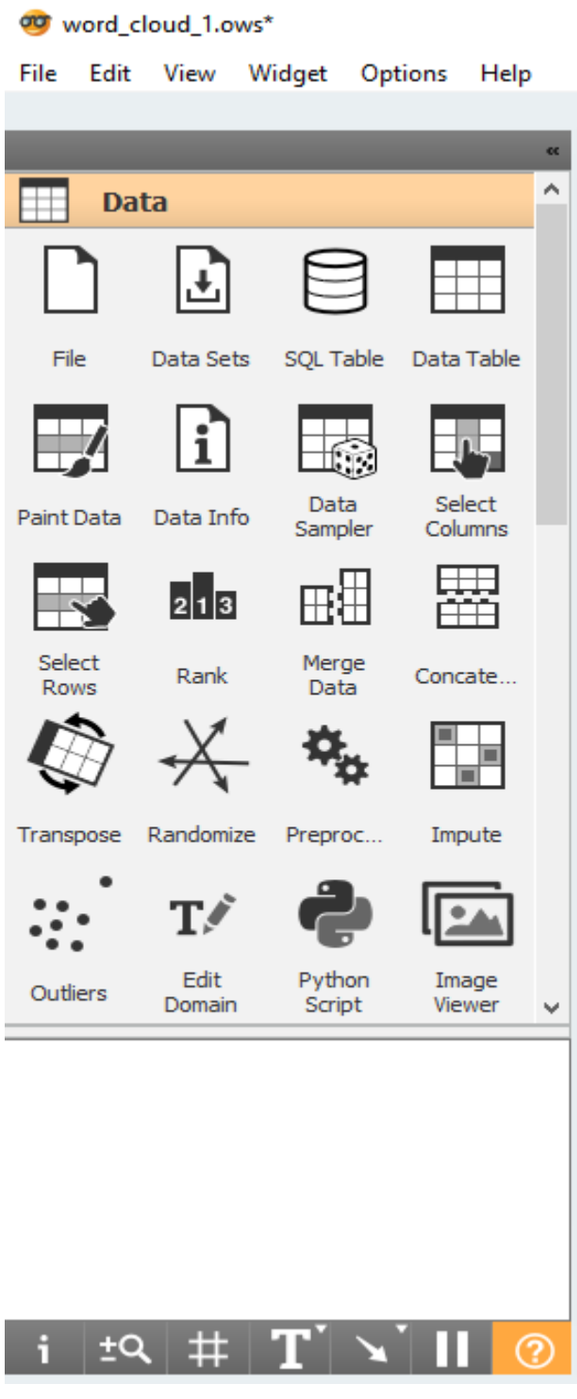
ORANGE : ПРЕПРОЦЕССИНГ

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.

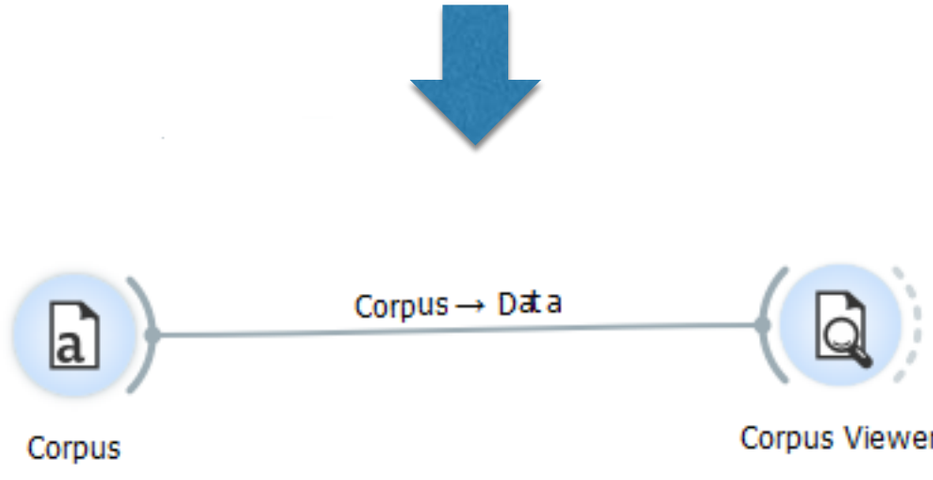


ORANGE : ПРЕПРОЦЕССИНГ

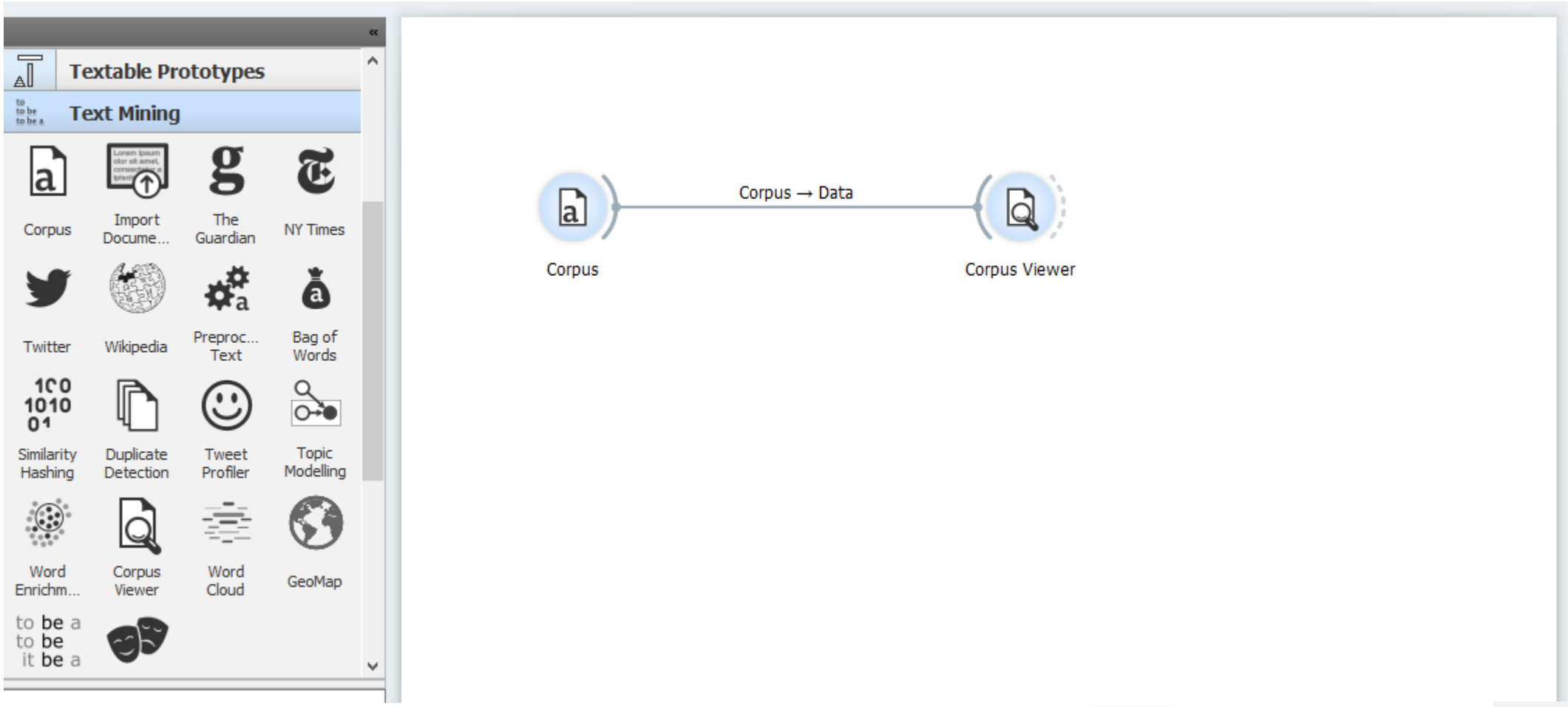


В данном пакете можно сделать алгоритм на основе визуального программирования.

В левой части окна находится хранилище ‘виджетов’, каждый из которых отвечает за свой кусочек действий. Например, снизу приведен пример загрузки текстов из файла формата Exel.



ORANGE : ПРЕПРОЦЕССИНГ



На вкладке ‘Text Mining’ нужно взять виджеты: ‘Corpus’ 

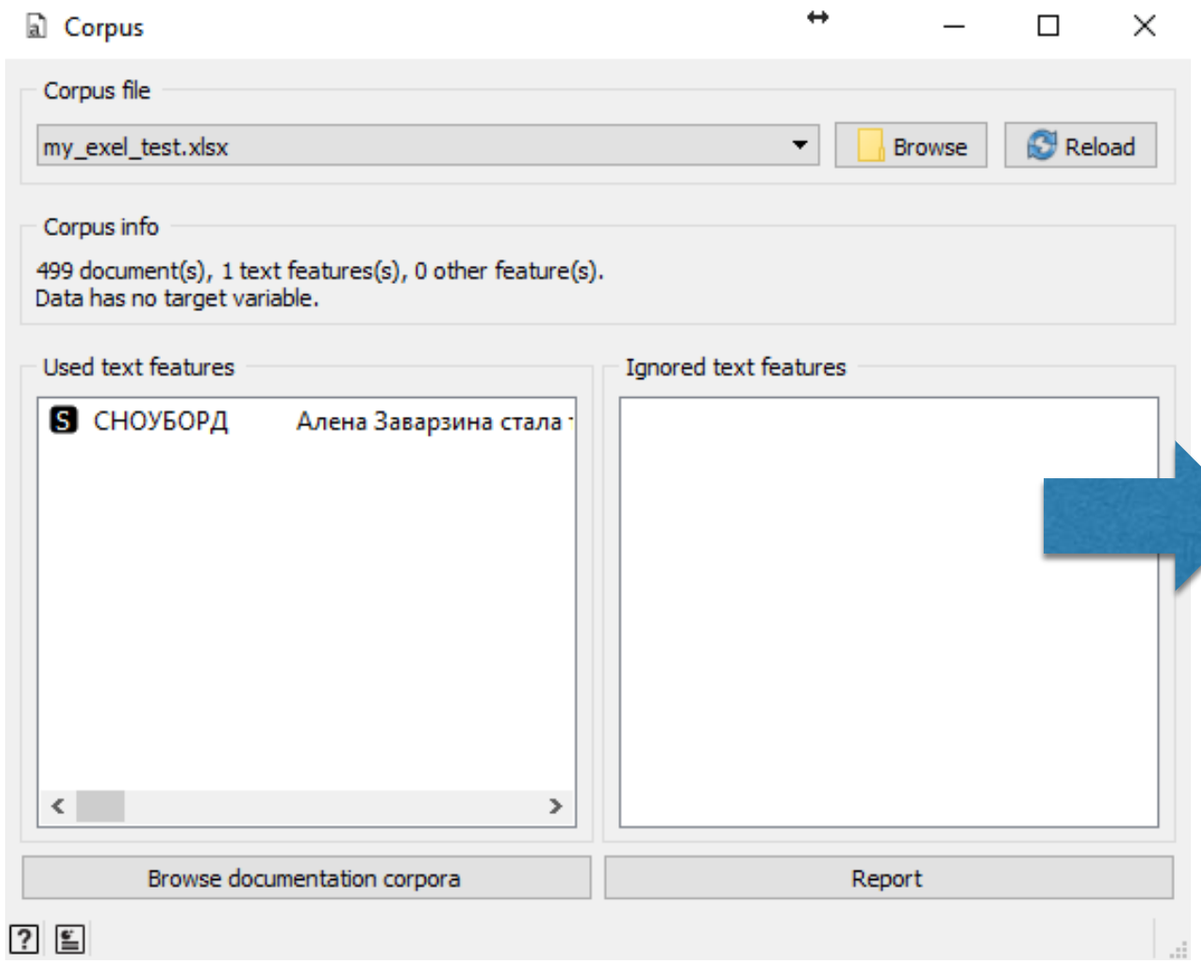
‘Corpus Viewer’ 

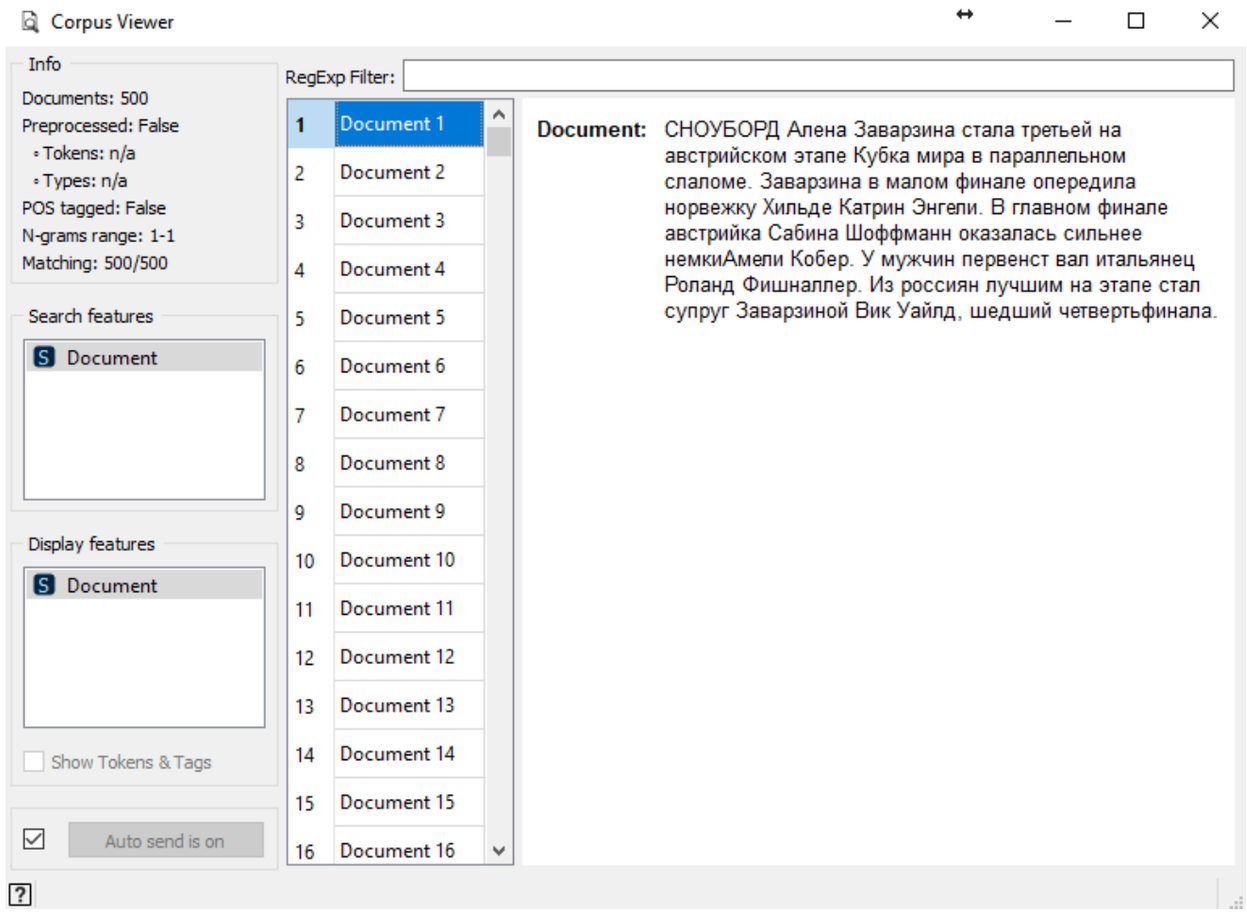
Далее нужно соединить два виджета линей.

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

ЗАГРУЗКА И ПРОСМОТР РУССКОЯЗЫЧНЫХ ТЕКСТОВ

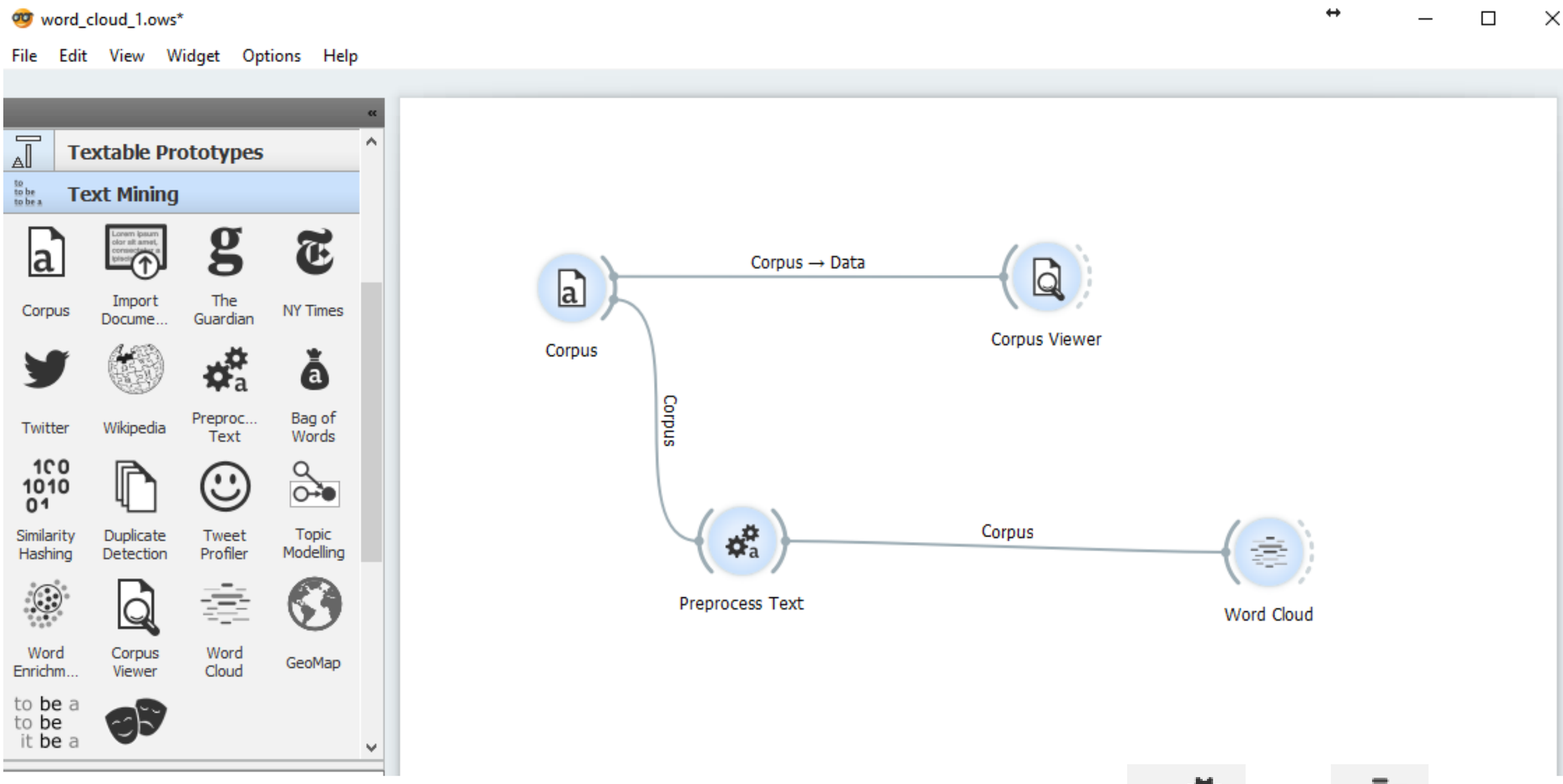
Для того что бы загрузить данные нужно кликнуть на виджете ‘Corpus’ и указать имя файла в формате книги Excel.



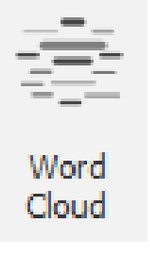


Для того что бы просмотреть результаты загрузки нужно кликнуть на виджете ‘Corpus Viewer’.

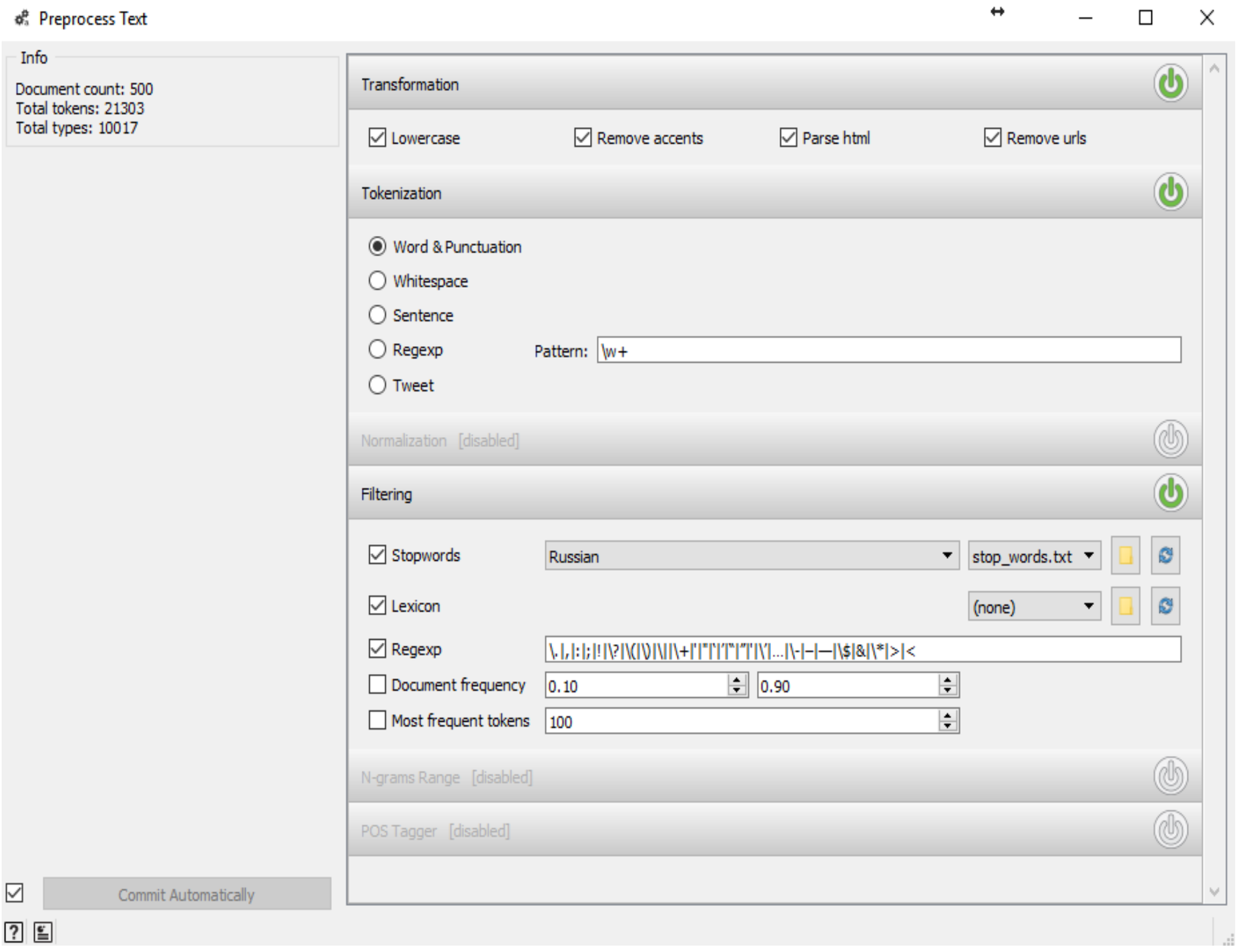
ПРЕПРОЦЕССИНГ И ОБЛАКО ЛЕМАТИЗИРОВАННЫХ СЛОВ



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

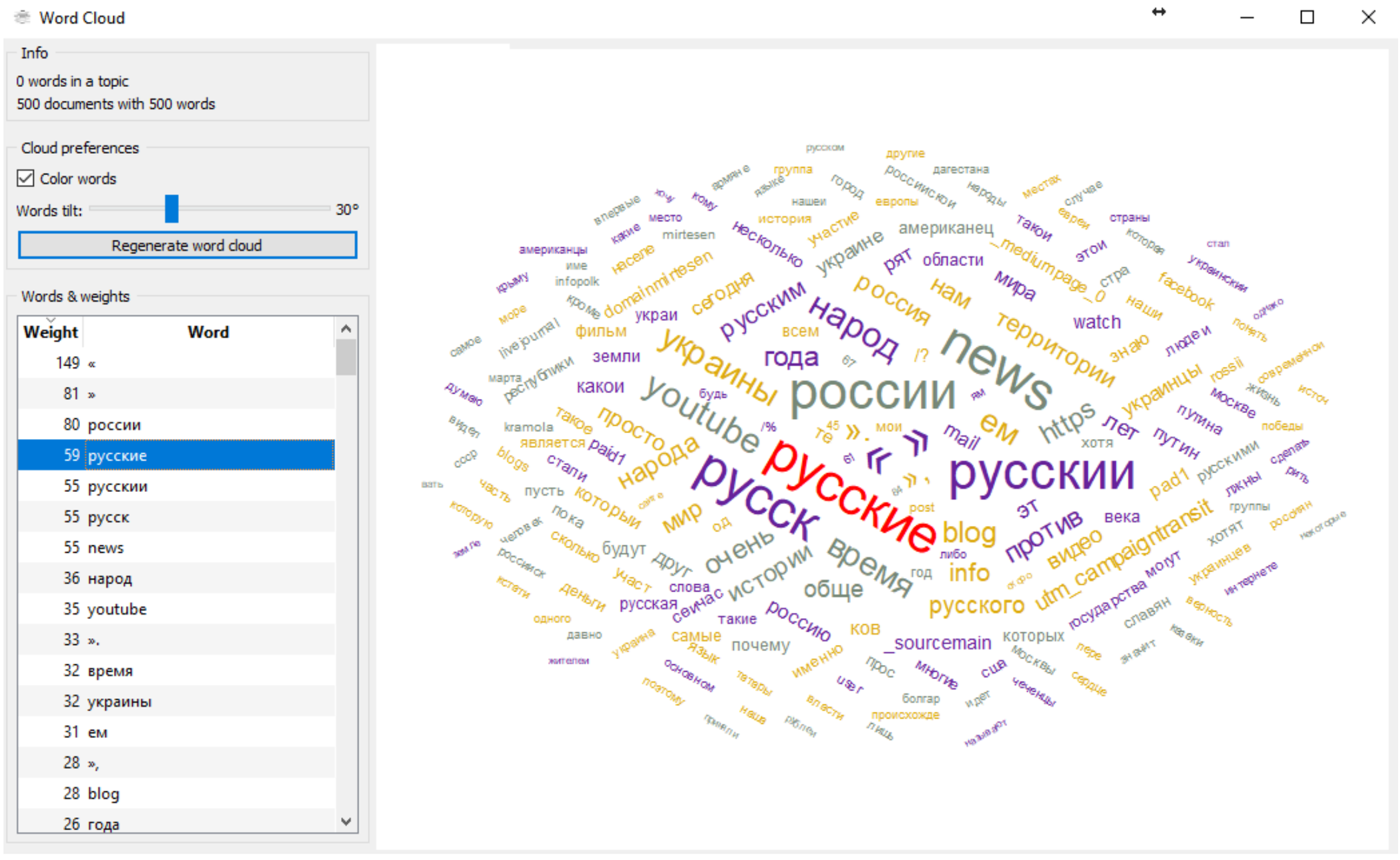


ПРЕПРОЦЕССИНГ В ORANGE



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

ОБЛАКО НЕЛЕМАТИЗИРОВАННЫХ СЛОВ



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

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

https://linis.hse.ru/

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru