

Internet Studies Lab, Department of Applied
Mathematics and Business Informatics



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

ПРЕПРОЦЕССИНГ РУССКОЯЗЫЧНЫХ ТЕКСТОВ В



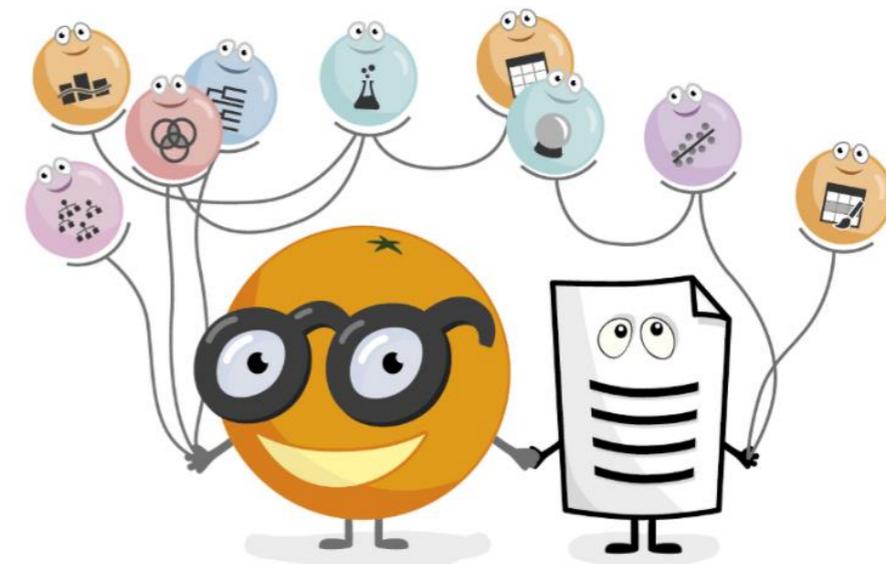
Анализ баз данных в публичном управлении
Кольцов С.Н.

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 токенизации 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

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 Concatenate...

Transpose Randomize Preproc... Impute

Outliers Edit Domain Python Script Image Viewer

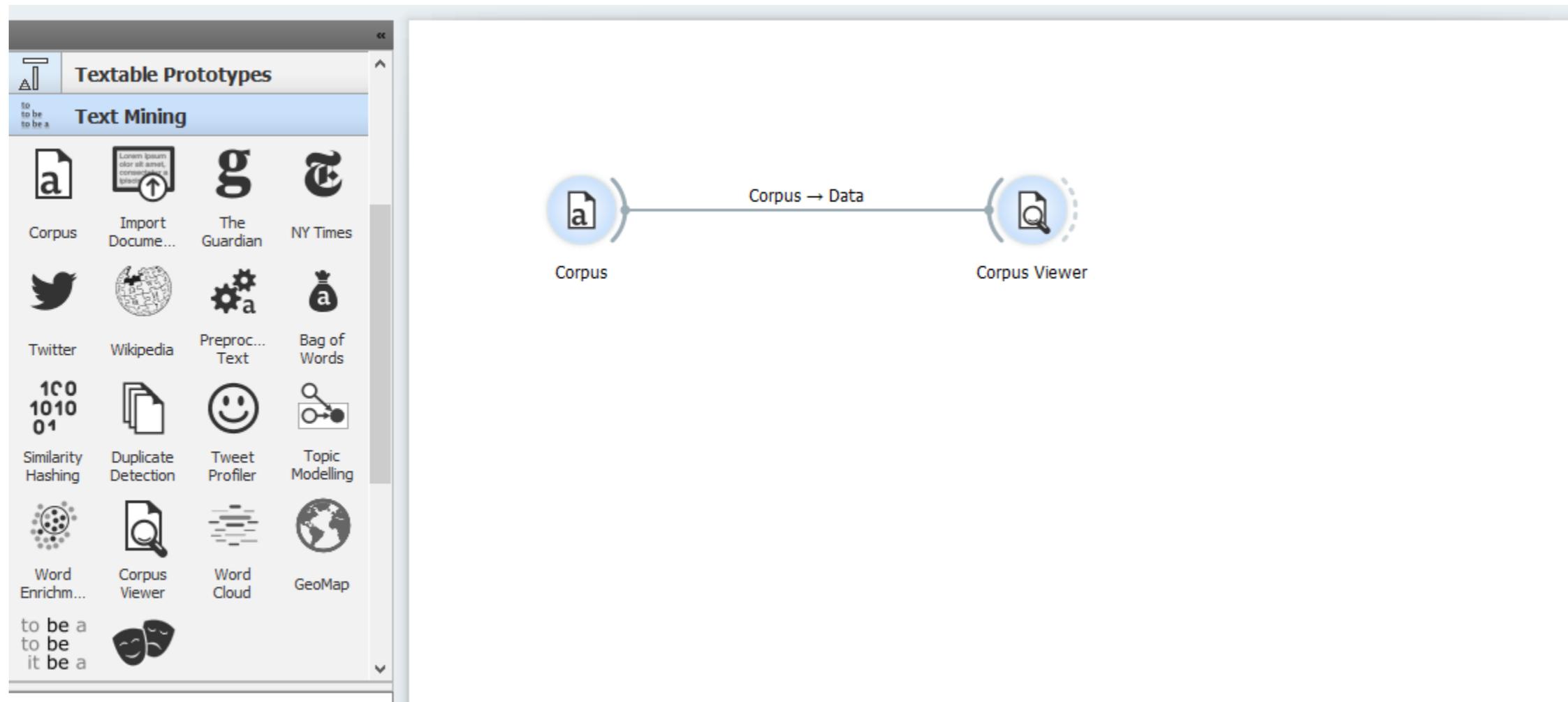
В данном пакете можно сделать алгоритм на основе визуального программирования.

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

Corpus → Data

Corpus Corpus Viewer

ORANGE : ПРЕПРОЦЕССИНГ



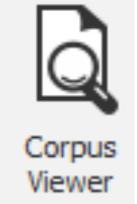
На вкладке ‘Text Mining’ нужно взять виджеты: ‘Corpus’  ‘Corpus Viewer’ 
Далее нужно соединить два виджета линией.

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.

The image shows two side-by-side screenshots of a web application interface. The left screenshot shows the 'Corpus' window. At the top, there is a 'Corpus file' section with a dropdown menu showing 'my_exel_test.xlsx' and buttons for 'Browse' and 'Reload'. Below this is the 'Corpus info' section, which displays '499 document(s), 1 text features(s), 0 other feature(s). Data has no target variable.' At the bottom, there are two panels: 'Used text features' and 'Ignored text features'. The 'Used text features' panel shows a single feature: 'S СНОУБОРД Алена Заварзина стала'. The 'Ignored text features' panel is empty. A blue arrow points from the 'Used text features' panel to the right screenshot. The right screenshot shows the 'Corpus Viewer' window. It has an 'Info' section with statistics: 'Documents: 500', 'Preprocessed: False', 'Tokens: n/a', 'Types: n/a', 'POS tagged: False', 'N-grams range: 1-1', and 'Matching: 500/500'. Below this is a 'Search features' section with a dropdown menu showing 'S Document'. There is also a 'Display features' section with a dropdown menu showing 'S Document' and a checkbox for 'Show Tokens & Tags'. At the bottom, there is a checkbox for 'Auto send is on' which is checked. The main part of the window is a list of 16 documents, numbered 1 to 16. Document 1 is selected and highlighted in blue. To the right of the list, the text of Document 1 is displayed: 'СНОУБОРД Алена Заварзина стала третьей на австрийском этапе Кубка мира в параллельном слаломе. Заварзина в малом финале опередила норвежку Хильде Катрин Энгели. В главном финале австрийка Сабина Шоффманн оказалась сильнее немки Амели Кобер. У мужчин первенстввал итальянец Роланд Фишналлер. Из россиян лучшим на этапе стал супруг Заварзиной Вик Уайлд, шедший четвертьфинала.'

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

ПРЕПРОЦЕССИНГ И ОБЛАКО ЛЕМАТИЗИРОВАННЫХ СЛОВ

The screenshot shows a software interface with a menu bar (File, Edit, View, Widget, Options, Help) and a toolbar. On the left is a 'Text Mining' widget palette containing various tools like 'Corpus', 'Import Document', 'The Guardian', 'NY Times', 'Twitter', 'Wikipedia', 'Preprocess Text', 'Bag of Words', 'Similarity Hashing', 'Duplicate Detection', 'Tweet Profiler', 'Topic Modelling', 'Word Enrichment', 'Corpus Viewer', 'Word Cloud', and 'GeoMap'. The main workspace displays a workflow diagram with four nodes: 'Corpus' (document icon), 'Corpus Viewer' (magnifying glass icon), 'Preprocess Text' (gears icon), and 'Word Cloud' (word cloud icon). Connections are as follows: 'Corpus' to 'Corpus Viewer' labeled 'Corpus → Data'; 'Corpus' to 'Preprocess Text' labeled 'Corpus'; and 'Preprocess Text' to 'Word Cloud' labeled 'Corpus'.

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.



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

<https://linis.hse.ru/>

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru