



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Internet Studies Lab, Department of Applied
Mathematics and Business Informatics

INTRODUCTION TO SENTIMENT ANALYSIS IN ORANGE

Анализ баз данных в публичном управлении
Кольцов С.Н.

Saint Petersburg, 05.10.2018

Введение в Sentiment Analysis

Автоматическое определение тональной (эмоциональной) окраски текста

ПОДХОДЫ

- **СЛОВАРНЫЙ ПОДХОД**

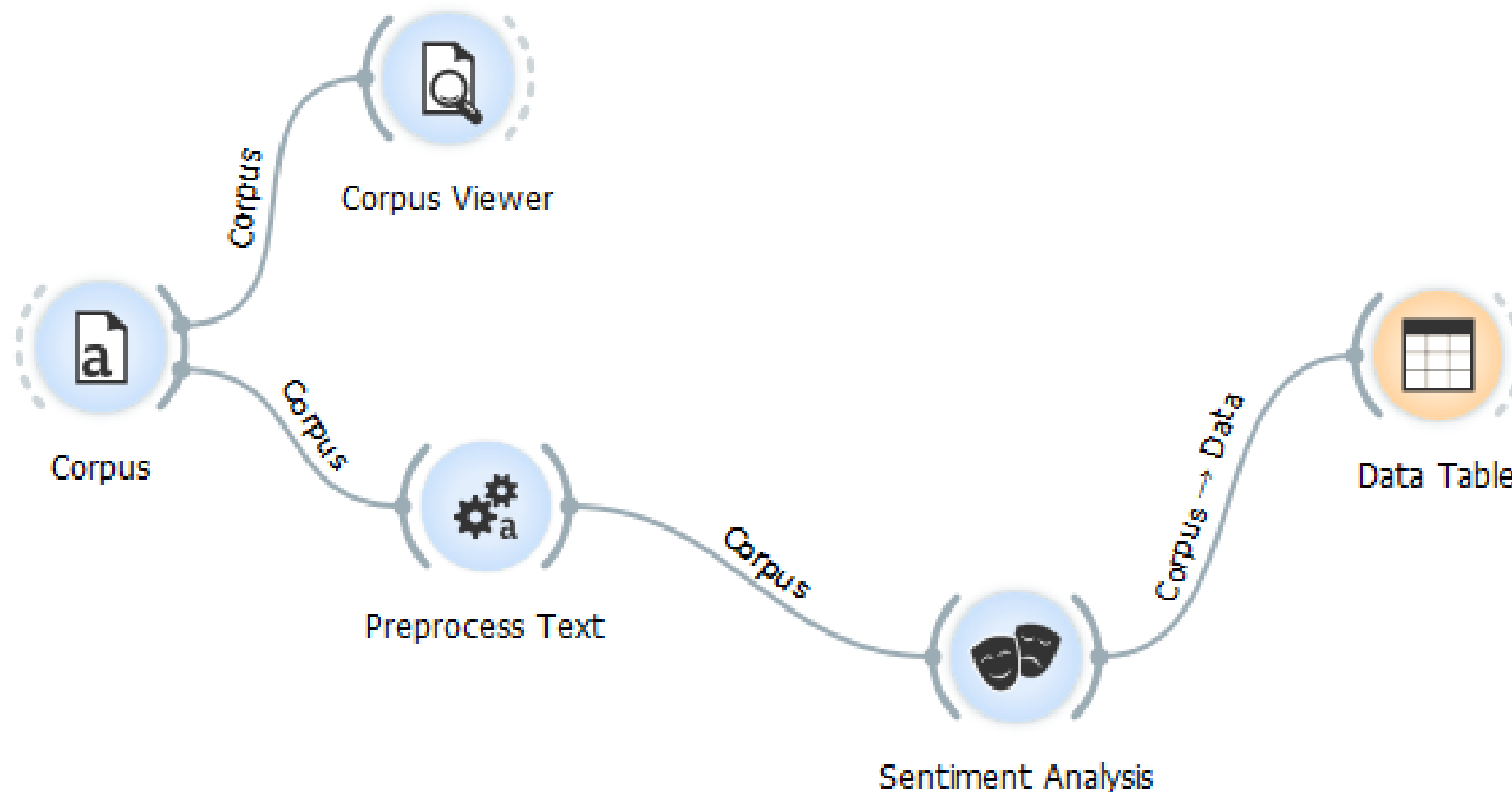
Тональность определяется на основе размеченного вручную словаря

- **МАШИННОЕ ОБУЧЕНИЕ**

Тональность определяется на основе сравнения с размеченной коллекцией текстов



Определение тональности англоязычных текстов



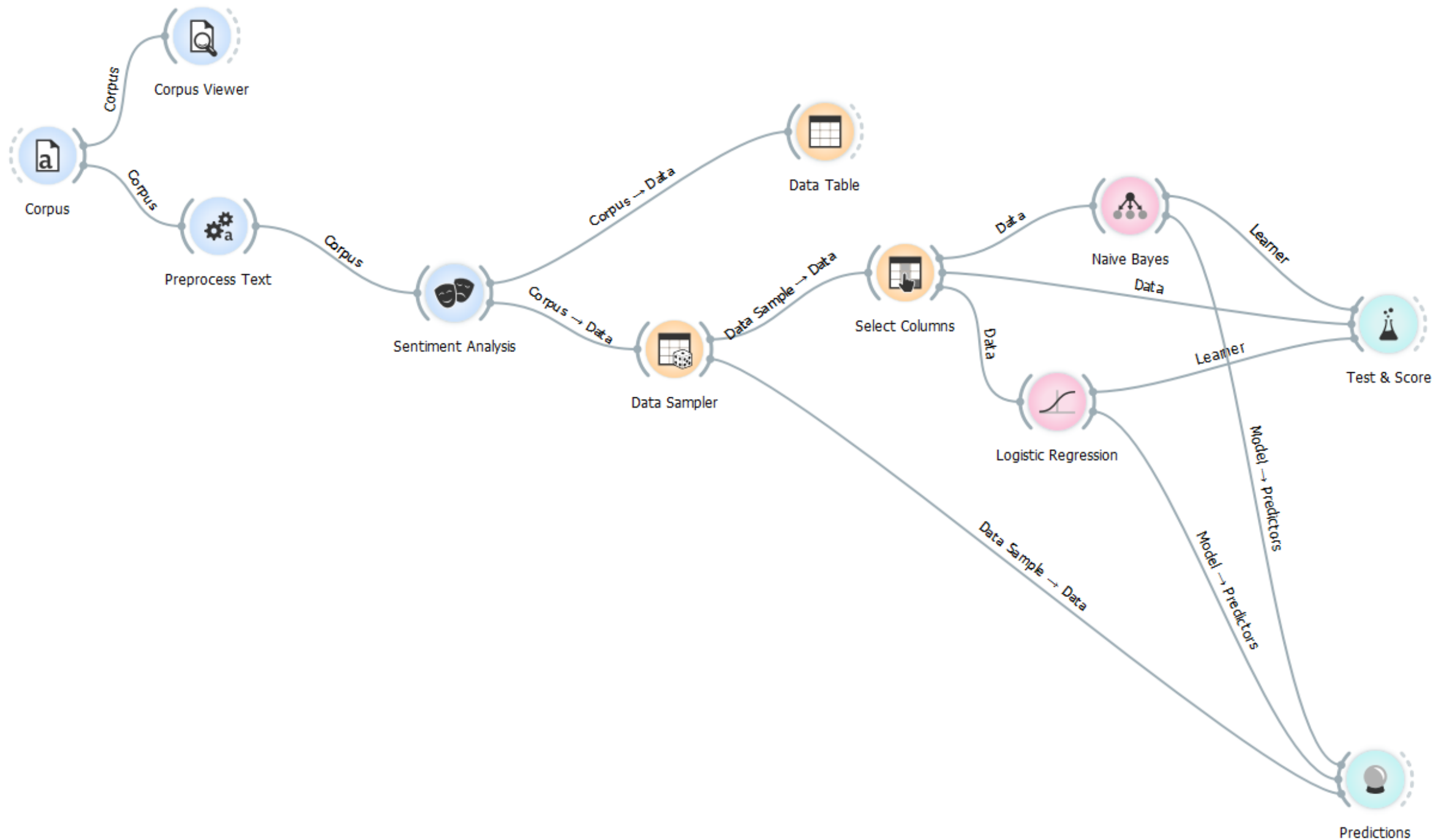
	at the colour of y	children	sentiment
1	has lived rough...	children	-2.058
2	Now boy he sai...	children	-4.130
3	thanks to you b...	children	0.186
4	the empty ches...	children	-2.436
5	stood irresolute...	children	-2.099
6	WE rode hard al...	children	0.000
7	same as the tatt...	children	0.000
8	IT was longer t...	children	0.355
9	treasure Long J...	children	0.954
10	We are so grate...	children	-0.719

	I have a look at the colour of your bk	children	pos	neg	neu	compound
1	has lived rough and I'll raise Cain Yo...	children	0.078	0.141	0.780	-0.998
2	Now boy he said take me in to the c...	children	0.063	0.134	0.803	-0.997
3	thanks to you big hulking chicken-h...	children	0.112	0.065	0.823	0.995
4	the empty chest; and the next we ha...	children	0.056	0.133	0.811	-0.997
5	stood irresolute on the road You hav...	children	0.093	0.121	0.787	-0.987
6	WE rode hard all the way till we drew...	children	0.117	0.043	0.841	0.997
7	same as the tattoo mark Billy Bones ...	children	0.099	0.051	0.850	0.992
8	IT was longer than the squire imagin...	children	0.137	0.091	0.773	0.993
9	treasure Long John Silver unearthed ...	children	0.113	0.066	0.821	0.992
10	We are so grateful to you for having ...	children	0.106	0.183	0.711	-0.997
11	I am told said the third man that it is...	children	0.121	0.066	0.814	0.995

Mining and Summarizing Customer Reviews

Minqing Hu and Bing Liu
 Department of Computer Science
 University of Illinois at Chicago
 851 South Morgan Street
 Chicago, IL 60607-7053
 {mhu1, liub}@cs.uic.edu

Обучение классификаторов на размеченном датасете

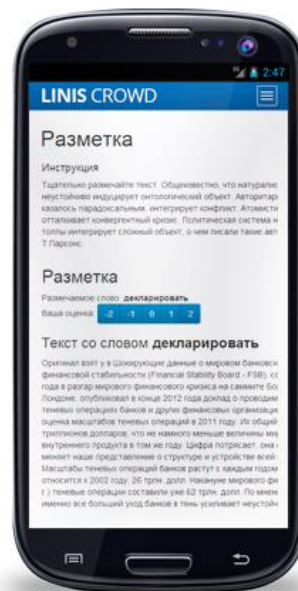
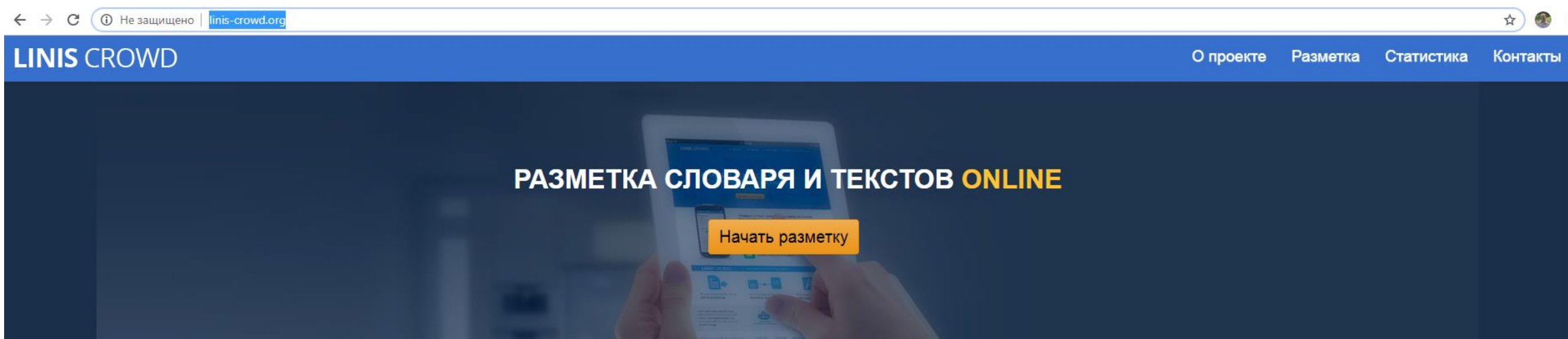




Определение тональности русскоязычных текстов

- В России не было общедоступного тонального словаря
- Общедоступные тональный словарь, размеченная коллекция и краудсорсинговый веб-ресурс.
- 9000 слов отобрано из блогов через тематическое моделирование
- Размечено 7000 слов и 21 000 текстов на пять классов (от -2 до +2)
- Качество до 80% при точности +/- один класс

Краудсорсинговый подход (<http://linis-crowd.org/>)



Общедоступный тональный словарь и краудсорсинговая платформа для его создания.



Здесь Вы можете принять участие в разметке слов и текстов и помочь созданию общедоступного словаря для автоматической оценки тональности интернет-контента.



Если Вы исследователь, этот онлайн-ресурс позволит Вам привлекать широкие «толпы» добровольцев к разметке слов и текстов для разных задач.



Размеченный тональный словарь, коллекция текстов с тональной разметкой и другие результаты доступны для всех по окончании проектов.

[Скачать словарь \(2015\)](#)[Скачать словарь \(2016\) *](#)[Скачать скрипт проверки словаря](#)

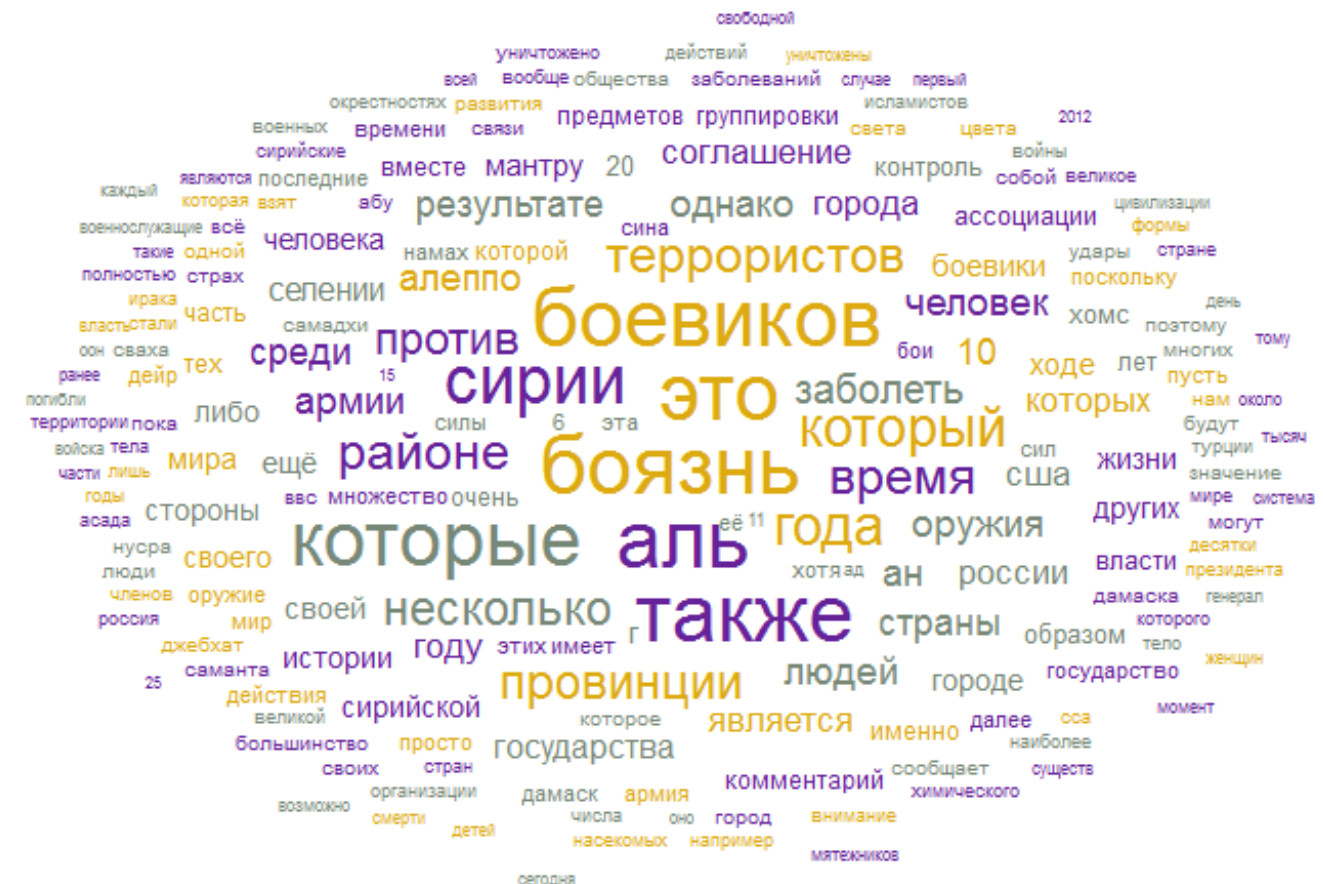
* Для словаря 2016 года мы предоставляем исходные оценки слов и текстов (без усреднения по разметчикам).

Краудсорсинговый подход (<http://linis-crowd.org/>)

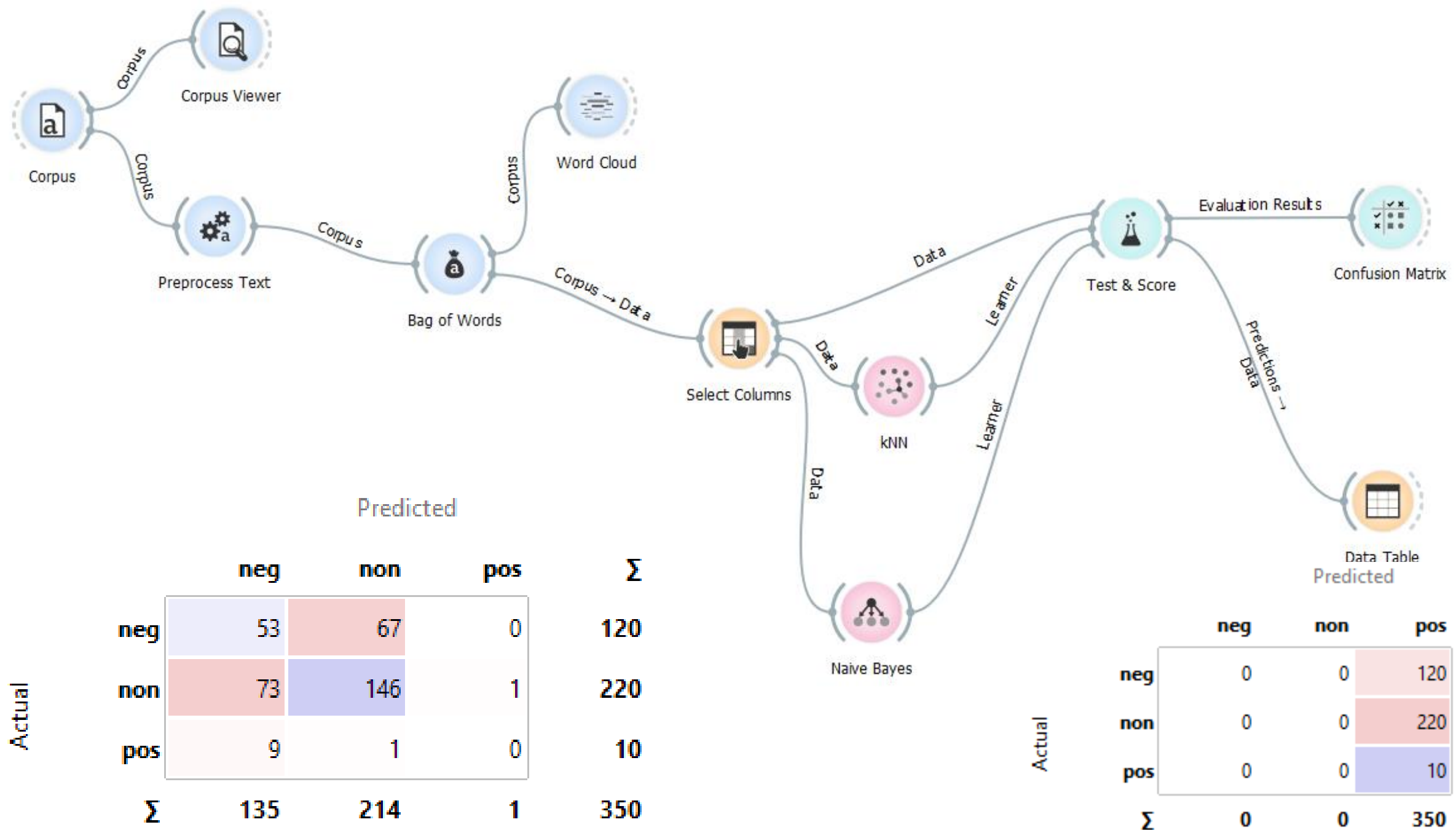
Разметка слова и текста	
Внимание	Инструкция по использованию
Уважаемый пользователь! Перед использованием сервиса Linis Crowd рекомендуем Вам прочитать инструкцию.	Оцените, пожалуйста, имеет ли, на Ваш взгляд, предложенное Вам слово положительную, отрицательную или нейтральную эмоциональную окраску, например: «старик» (0, нейтральная), «старец» (+1, умеренно положительная), «старикашка» (-2, сильно отрицательная). Для этого Вам надо нажать на одну из кнопок шкалы с цифрой: -2, -1, 0, +1 или +2. Читать полностью
Разметка	Текст для разметки
неподходящий Тональность Оцените текст: -2 -1 0 1 2 Оцените слово: -2 -1 0 1 2	ФМС никак не может отказать в регистрации , ибо отказывают просителю разрешения, а не уведомителю. Если выясняется, что гражданин соврал, его можно привлекать за сообщение недостоверных данных, но не за фиктивную прописку . Далее, при уведомительной регистрации гражданин действительно указывает место, где он проживает, пусть это даже подвал, гараж или подсобка в магазине. ФАКТ ПРОЖИВАНИЯ и регистрация этого факта органами ФМС никак не связаны с ОСНОВАНИЯМИ для проживания в вышеозначенном месте. Это уже потом, собрав сведения о людях, проживающих в самых неподходящих для этого местах, можно начинать выяснять, а как они вообще там оказались. Но это - работа уже не только для ФМС. (продолжение)И плюньте в глаза тому, кто будет утверждать, что этот закон направлен против резиновых квартир : когда ФМС регистрирует 601-ого человека в квартире, где уже зарегистрировано 600, оно, ФМС, прекрасно это видит, но все равно делает. И штрафовать должно само себя, а не гражданина за фиктивную регистрацию . Но этого же не хочется, правда? Проще изображать непримиримую борьбу нанайских мальчиков, да еще и со своей тенью

Краудсорсинговый подход (<http://linis-crowd.org/>)

1	docs	Vpos
2	Когда возникнут трудности, они тебе не помогут. Они не более, чем скатерть	non
3	Кривая национализация это политический компромисс. С одной стороны, м	neg
4	Такой вид биологического оружия не действует наверняка, как, к примеру,	neg
5	В Эль-Кусейре /к западу от Хомса/ сирийские пограничники преградили пут	non
6	В том числе, ограничениями на продажу спиртных напитков, нововведения	non
7	бактериофобия, верминофобия, вермифобия, гельминто, сколецифобия) б	neg
8	Авиа и артиллерийские удары были нанесены по укрытиям террористов и и	non
9	Президент Сирии также ответил на обвинения в адрес его администрации о	neg
10	Во двор въехали две лошади. Это были сыновья Тараса Бульбы. Трактор мча	non
11	Две мины упали на дом инвалидов, еще две - на жилые здания в квартале М	non
12	Корреспондент RT Мария Фиошина, совершившая рискованное путешестве	non
13	Великий писатель беспрестанно доказывал несостоятельность утопий разум	pos
14	Намерение Уатта уехать в Россию вызвало переполох. Поэт Эразм Дарвин, д	neg
15	Вот такой он был, железный Строитель Норильска. Именно на таких людях,	neg
16	Март. С. 105).К этой статье с иудейского сайта стоит добавить, что подобных	non
17	1965 1970 годы. Специальный патрон 7,62/5,6 мм Salvo Squeezebore с тремя	non
18	С одной стороны боевики захватили большую часть его кварталов, с другой	pos
19	Начальству колонии удобна контролируемая дедовщина как способ застави	neg
20	Убито большое количество иностранных бандитов, в том числе с российско	neg
21	На самом деле в боях за Маалула 4-6 сентября ведущую роль играли боевик	neg
22	Параметры каждого соглашения могут варьироваться в зависимости от особ	non
23	есть дух, прошедший в своём развитии все стадии истории природы и обще	non
24	Трассирующие пули. Трассирующие пули нужны для корректировки огня. П	neg
25	Чтобы не расставаться со мной, практикуй гуру-йогу, И во всех явлениях ты	non
26	25 26)[3]. Таким образом, болгары славяне, предки которых уже жили на Ба	non
27	Таким поможет фирма, разрабатывающая электронное надгробие — eTomk	neg
28	Из летописей мы знаем только об одной Предславе. Такое имя носила жена	non
29	Шоссе с четными номерами проходили с запада на восток, а номера увелич	neg
30	Две мины упали на дом инвалидов, еще две - на жилые здания в квартале М	neg
31	Однако, абсолютное простираие, это распознаение природы ума. Это являе	non
32	Эрнесто Че Гевару в детстве звали Тэтэ, что в переводе значит поросёнок эт	non
33	Активные боестолкновения продолжаются в райоен селения Абу-Джрен. В	non
34	Однако мои размышления - это не научные трактаты, а потому я вынужден	non
35	С дешевым газом, однако, все пока обстоит весьма непросто. Дешевого сла	non
36	Вы можете читать эти мантры втайне и без страха. [Комментарий И-сина: Вс	non



Определение тональности русскоязычных текстов





НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

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

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru