

Что такое интеллектуальный анализ данных и как можно использовать его методы для достижения прикладной цели выбора Вуза?

Sergei Koltcov



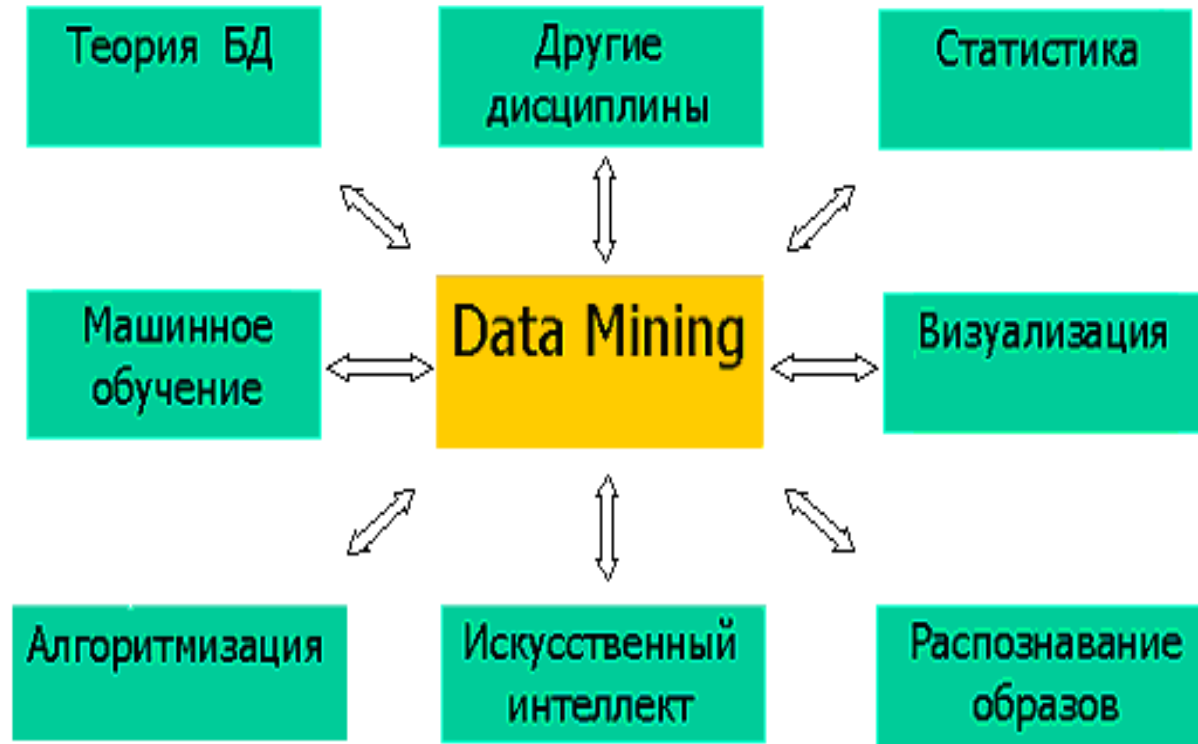
Интеллектуальный анализ данных

Интеллектуальный анализ данных (ИАД)— это сочетание широкого математического инструментария и последних достижений в сфере информационных технологий.

В технологии **ИАД** гармонично объединились строго формализованные методы и методы неформального анализа, т.е. количественный и качественный анализ данных.

Интеллектуальный анализ данных (**добыча данных**, глубинный анализ данных) — собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

ЧТО ТАКОЕ DATA MINING?



1. Data Mining технология не может заменить аналитика.
 2. Технология не может дать ответы на те вопросы, которые не были заданы.
 3. Извлечение полезных сведений невозможно без хорошего понимания сути данных.
- Необходим тщательный выбор модели и интерпретация зависимостей или шаблонов, которые обнаружены в данных.

Программные средства

Языки программирования с большим количеством библиотек.

1. Python, c++



Скриптовые программные средства

2. R studio



Среды с визуальным вариантом программирования

3. Orange или Knime



Методы кластеризации

Кластерный анализ — задача разбиения заданной выборки объектов на непересекающиеся подмножества, называемые кластерами так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

1. Цель исследования. Это может быть или определение кластерной структуры данных (проверка наличия кластеров в данных) или разбиение объектов на группы по заранее определенному принципу.
2. Преобразование имеющихся данных в объекты для их кластеризации, например, определение набора слов, которые характеризуют тексты, определение переменных, характеризующих объекты.
3. Преобразование выбранных объектов в “компьютерный вид”, т.е. сопоставление значениям переменных числовые значения.
4. Задание метрики пространства. (Формула расчета расстояния между кластерами)
5. Выбор алгоритма кластеризации.
6. Расчет и анализ результатов.

Цели кластерного анализа

Кластерный подход может быть использован для различных целей например:

1. Различные школы могут быть кластеризованы на основании множества различных параметров, включающие число школьников, соотношения мальчиков/девочек. Зарплата преподавателей, перечень дополнительных образовательных услуг и так далее.
2. Университеты также могут быть сгруппированы по таким параметрам как: Доля студентов получивших работу после окончания, доля средств от реализации интеллектуальной собственности, индекс научного цитирования, что отражает научный уровень университета и так далее.
3. Студенты из разных университетов, также могут сгруппированы по различным параметрам. Если производит кластерный анализ внутри одного университета то можно выявить группы, характеризующиеся определенными паттернами поведения.

Проблемы определения сходства

Процедура кластеризации — зависит от меры сходства или не сходства. Такие меры выражаются виде функций расстояний, выраженных в виде той или иной функции.



Задача определения сходства является задачей Machine learning.



Как измерять сходства между объектами

Одной из основных мер сходства/ различия может быть использовано Евклидово расстояние. А именно, каждый объект характеризуется набором параметров (фичей). Можно рассчитать расстояние между объектами в многомерном пространстве фичей при помощи теоремы Пифагора, только с учетом поправки на многомерность пространства.

$P=(p_1, p_2, \dots, p_n)$, $Q=(q_1, q_2, \dots, q_n)$

$$d(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

Перебирая последовательно все объекты, мы можем рассчитать расстояния между всеми объектами. Соответственно, два объекта похожи, если эти объекты находятся рядом друг с другом в многомерном пространстве. Также используются немного другие меры:

Квадрат расстояния

$$d(p, q) = \sum_{i=1}^n (p_i - q_i)^2$$

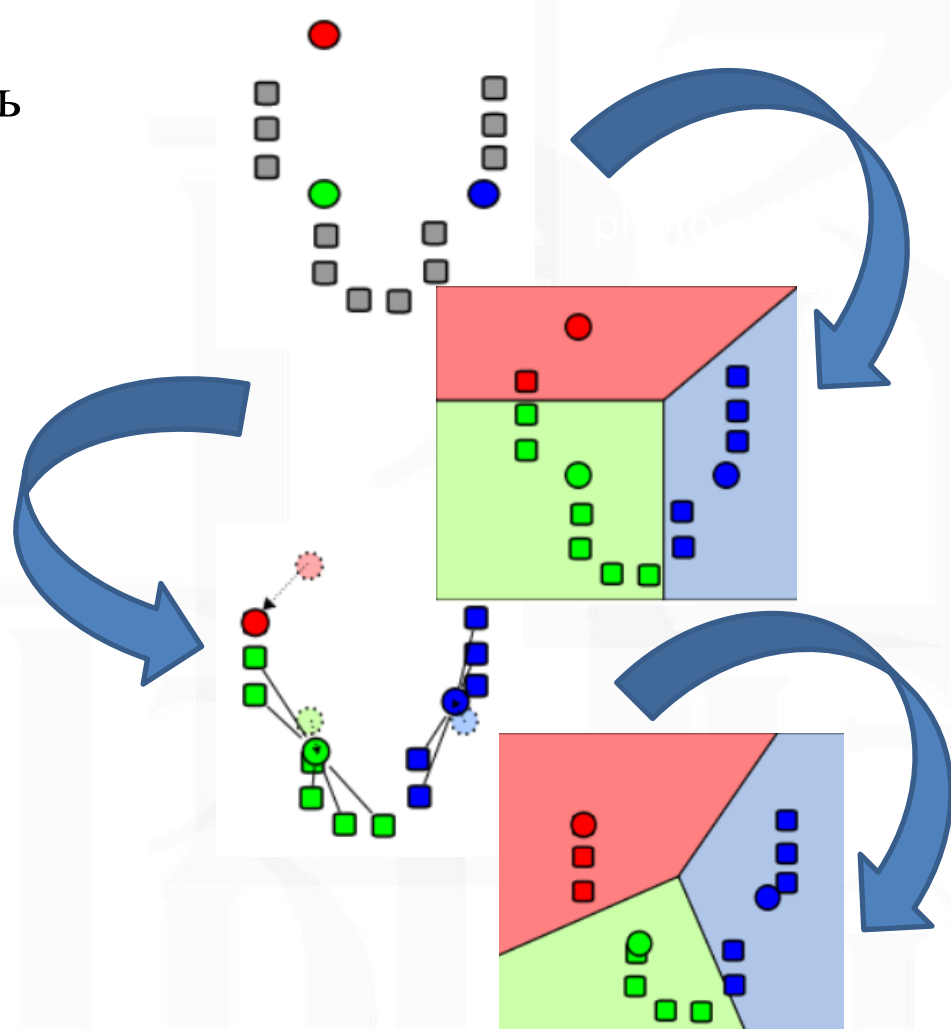
City block расстояние

$$d(p, q) = \sum_{i=1}^n |p_i - q_i|$$

Алгоритм K means

Основная суть кластеризации заключается в следующем: Пусть у нас есть совокупность объектов.

1. Выбираем начальные точки для кластеров.
2. Привязать ближайшие точки к центрам кластеров.
3. Пересчитать центры кластеров, исходя из того, что в кластер были добавлены новые объекты.
4. После того как нашли новые центры кластеризации, снова перераспределяем ближайшие точки по кластерам.



Orange – Open source software for data visualization and data analysis



Data Mining Fruitful and Fun

Open source data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

[Download Orange](#)

The **old version**, Orange 2.7, is still available.



Interactive workflows

Create your own interactive workflows to analyse and visualize your data.



Visualization

Orange is packed with different visualizations, from scatter plots, bar charts, trees, to dendrograms, networks and heat maps.



Large Toolbox

Over 100 widgets and growing. We cover most of standard data analysis tasks. Specialized add-ons available, like Orange Bioinformatics.

Можно скачать по
адресу:
<http://orange.biolab.si/>

Faculty of
Computer and
Information Science
University of
Ljubljana

Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research 14(Aug):2349–2353.

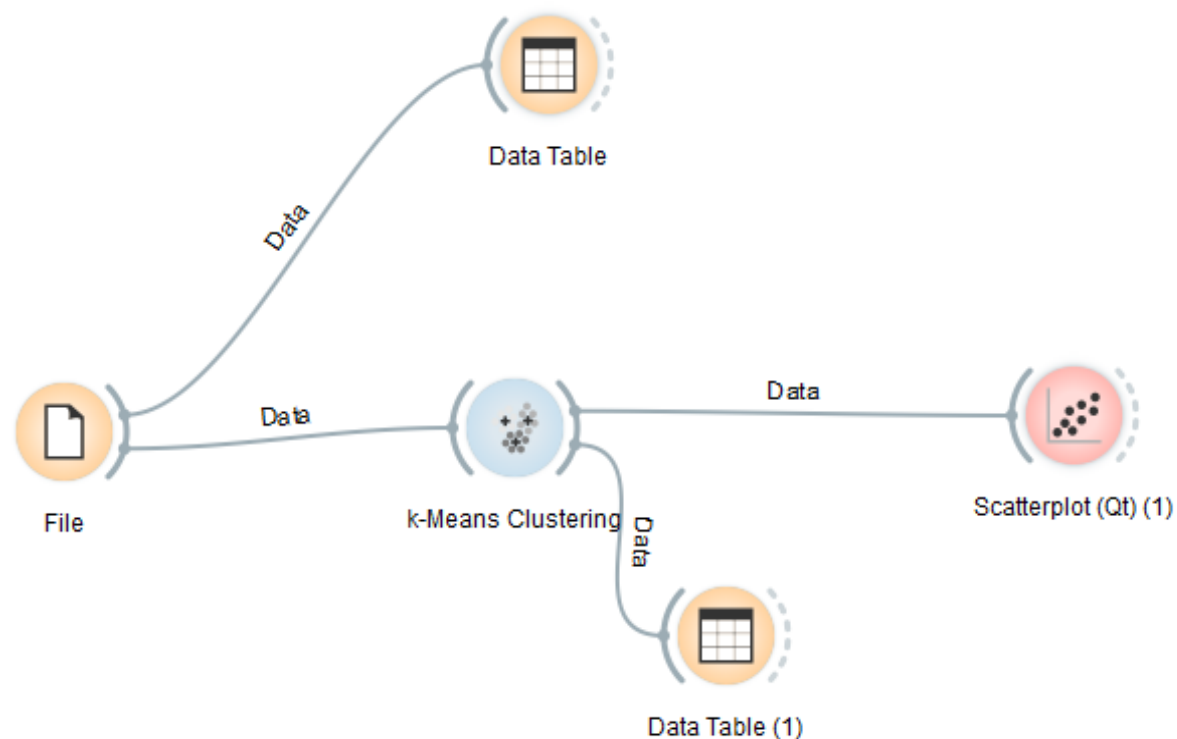
Orange consists of a canvas interface onto which the user places widgets and creates a data analysis workflow.

cluster_analysis*

File Edit View Widget Options Help

Data

			
File	Paint Data	Info	Save
			
Data Table	Multi-Key Merge D...	Select Attributes	Rank
			
Merge Data	Concatenate	Data Sampler	Purge Domain
			
Select Data	Quick Select	Discretize	Preproc...
			
Continuize	Impute	Outliers	Feature Constru...
			



Исходные данные

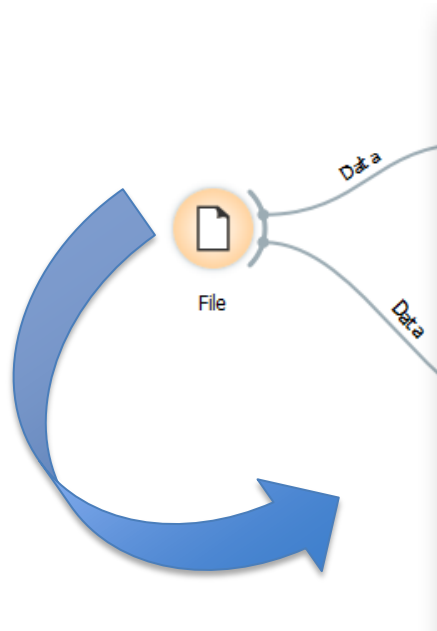
Таблица 1. 100 лучших вузов России, 2017 год

Место, 2017 год	Место, 2016 год	Название	Рейтинговый функционал	Условия для получения качественного образования, ранг	Уровень востребованности выпускников работодателями, ранг	Уровень научно-исследовательской деятельности, ранг
1	1	Московский государственный университет им. М.В. Ломоносова	4.729	1	2	1
2	2	Московский физико-технический институт (государственный университет)	4.603	3	5	3
3	3	Национальный исследовательский ядерный университет «МИФИ»	4.424	7	4	2
4	5	Санкт-Петербургский государственный университет	4.244	4	12	5
5	7	Московский государственный институт международных отношений (университет) МИД РФ	4.223	2	7	34
6	6	Национальный исследовательский университет «Высшая школа экономики»	4.216	5	6	15
7	4	Московский государственный технический университет имени Н.Э. Баумана	4.136	10	1	11
8	8	Национальный исследовательский Томский политехнический университет	4.100	6	15	7
9	9	Новосибирский национальный исследовательский государственный университет	3.957	8	16	6
10	11	Санкт-Петербургский политехнический университет Петра Великого	3.890	12	13	8
11	12	Российская академия народного хозяйства и государственной службы при Президенте РФ	3.724	9	11	30
12	10	Уральский федеральный университет имени первого Президента России Б.Н. Ельцина	3.704	25	9	9
13	14	Финансовый университет при Правительстве РФ	3.701	11	8	35
14	13	Национальный исследовательский Томский государственный университет	3.647	19	30	4
15	17	Казанский (Приволжский) федеральный университет	3.589	16	32	10
16	15	Сибирский федеральный университет	3.565	26	10	14
17	18	Национальный исследовательский технологический университет «МИСиС»	3.560	13	31	16
18	16	Российский государственный университет нефти и газа (национальный исследовательский университет) имени И.М. Губкина	3.553	24	3	37
19	19	Университет ИТМО	3.461	15	47	12
20	21	Российский университет дружбы народов	3.412	14	29	23
21	22	Первый Московский государственный медицинский университет имени И.М. Сеченова Министерства здравоохранения РФ	3.338	21	20	21
22	23	Первый Санкт-Петербургский государственный медицинский университет имени академика И.П. Павлова Министерства здравоохранения РФ	3.320	18	25	33
23	25	Российский экономический университет имени Г.В. Плеханова	3.253	23	14	41
24	27	Российский национальный исследовательский медицинский университет имени Н.И. Пирогова Министерства Здравоохранения РФ	3.175	22	23	31
25	20	Национальный исследовательский университет «МЭИ»	3.088	28	26	27
26	24	Новосибирский государственный технический университет	3.060	32	21	19
27	-	Московский государственный лингвистический университет	3.018	17	52	75
28	28	Национальный исследовательский Нижегородский государственный университет имени Н.И. Лобачевского	2.957	29	58	18

Исходные данные

	A	B	C	D	E
1	Название	Рейтинг	Качество образования	Уровень востребованности выпускников	Уровень научно-исследовательской деятельности
2	Московский государственный университет им. М.В. Ломоносова	4.729	1	2	1
3	Московский физико-технический институт (государственный университет)	4.603	3	5	3
4	Национальный исследовательский ядерный университет «МИФИ»	4.424	7	4	2
5	Санкт-Петербургский государственный университет	4.244	4	12	5
6	Московский государственный институт международных отношений (университет) МИД РФ	4.223	2	7	34
7	Национальный исследовательский университет "Высшая школа экономики"	4.216	5	6	15
8	Московский государственный технический университет имени Н.Э. Баумана	4.136	10	1	11
9	Национальный исследовательский Томский политехнический университет	4.1	6	15	7
10	Новосибирский национальный исследовательский государственный университет	3.957	8	16	6
11	Санкт-Петербургский политехнический университет Петра Великого	3.89	12	13	8
12	Российская академия народного хозяйства и государственной службы при Президенте РФ	3.724	9	11	30
13	Уральский федеральный университет имени первого Президента России Б.Н. Ельцина	3.704	25	9	9
14	Финансовый университет при Правительстве РФ	3.701	11	8	35
15	Национальный исследовательский Томский государственный университет	3.647	19	30	4
16	Казанский (Приволжский) федеральный университет	3.589	16	32	10
17	Сибирский федеральный университет	3.565	26	10	14
18	Национальный исследовательский технологический университет «МИСиС»	3.56	13	31	16
19	Российский государственный университет нефти и газа (национальный исследовательский университет)	3.553	24	3	37
20	Университет ИТМО	3.461	15	47	12
21	Российский университет дружбы народов	3.412	14	29	23
22	Первый Московский государственный медицинский университет имени И.М. Сеченова Министерства здравоохранения РФ	3.338	21	20	21
23	Первый Санкт-Петербургский государственный медицинский университет имени академика И.П. Павлова	3.32	18	25	33
24	Российский экономический университет имени Г.В. Плеханова	3.253	23	14	41
25	Российский национальный исследовательский медицинский университет имени Н.И. Пирогова Министерства здравоохранения РФ	3.175	22	23	31
26	Национальный исследовательский университет "МЭИ"	3.088	28	26	27
27	Новосибирский государственный технический университет	3.06	32	21	19
28	Московский государственный лингвистический университет	3.018	17	52	75
29	Национальный исследовательский Нижегородский государственный университет имени Н.И. Лобачевского	2.957	29	58	18
30	Всероссийская академия внешней торговли Министерства экономического развития РФ	2.865	20	39	127

Загрузка данных в Orange



File

Data

Data

File

File

File: рейтинг_2017.csv

URL:

Info

100 instance(s), 4 feature(s), 1 meta attribute(s)
Data has no target variable.

Columns (Double click to edit)

	Name	Type	Role	Values
1	Рейтинг	N numeric	feature	
2	Качество образования	N numeric	feature	
3	Уровень востребованно...	N numeric	feature	
4	Уровень научно-исследователь...	N numeric	feature	
5	Название	S string	meta	

Data Table

Info

100 instances (no missing values)
4 features (no missing values)
No target variable.
1 meta attribute (no missing values)

Variables

☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection

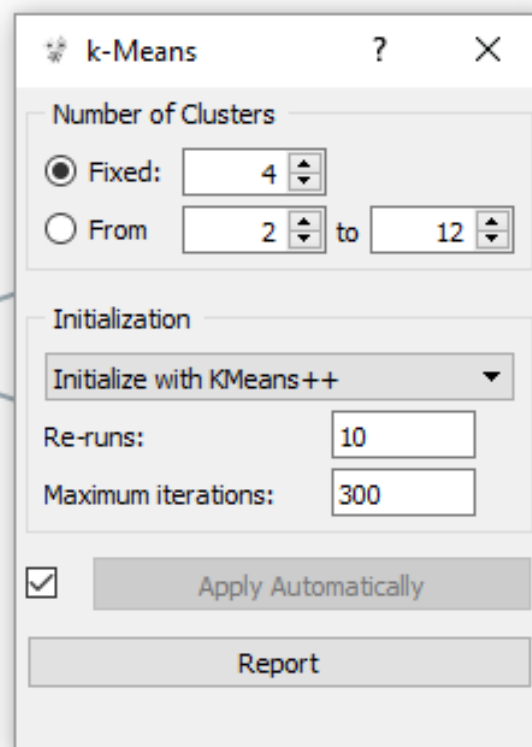
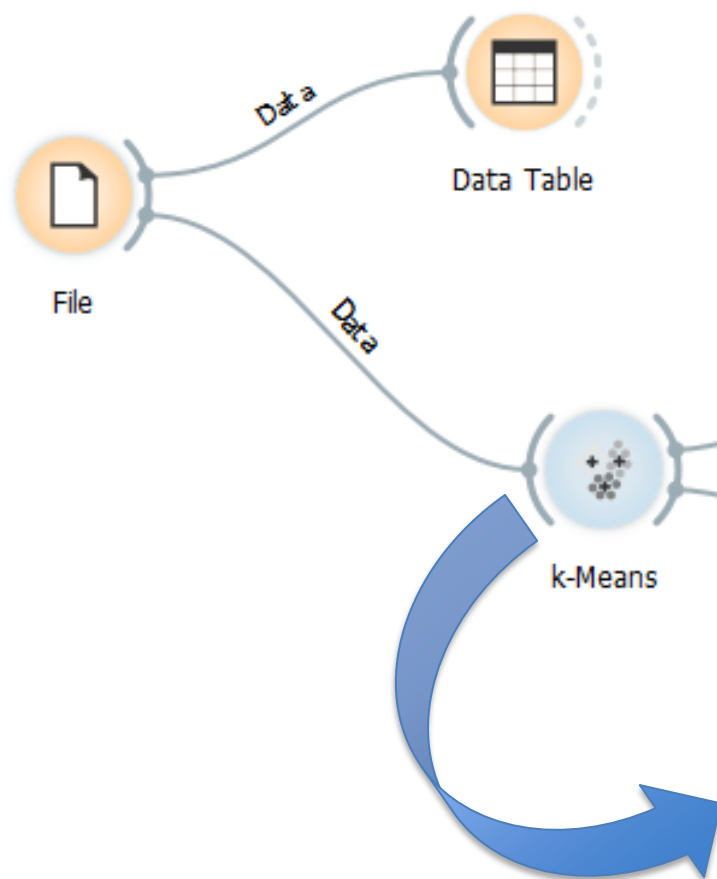
☒ Select full rows

	Название	Рейтинг	чество образова	ности выпускни	исследовательск
1	Московский г...	4.729	1	2	1
2	Московский ф...	4.603	3	5	3
3	Национальны...	4.424	7	4	2
4	Санкт-Петербу...	4.244	4	12	5
5	Московский г...	4.223	2	7	34
6	Национальны...	4.216	5	6	15
7	Московский г...	4.136	10	1	11
8	Национальны...	4.100	6	15	7
9	Новосибирски...	3.957	8	16	6
10	Санкт-Петербу...	3.890	12	13	8
11	Российская ак...	3.724	9	11	30
12	Уральский фед...	3.704	25	9	9
13	Финансовый у...	3.701	11	8	35
14	Национальны...	3.647	19	30	4
15	Казанский (Пр...	3.589	16	32	10

Report

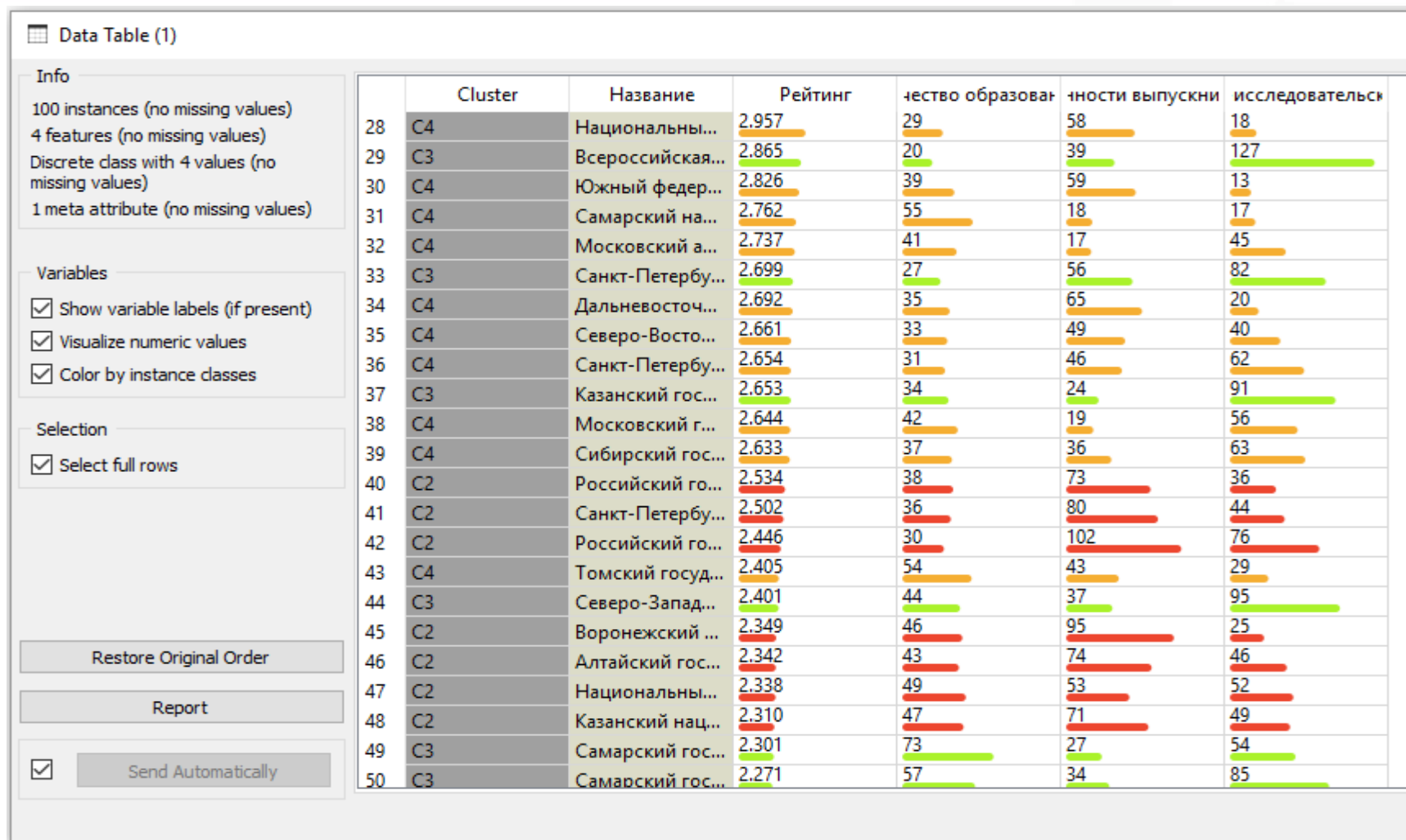
Apply

Задание числа кластеров

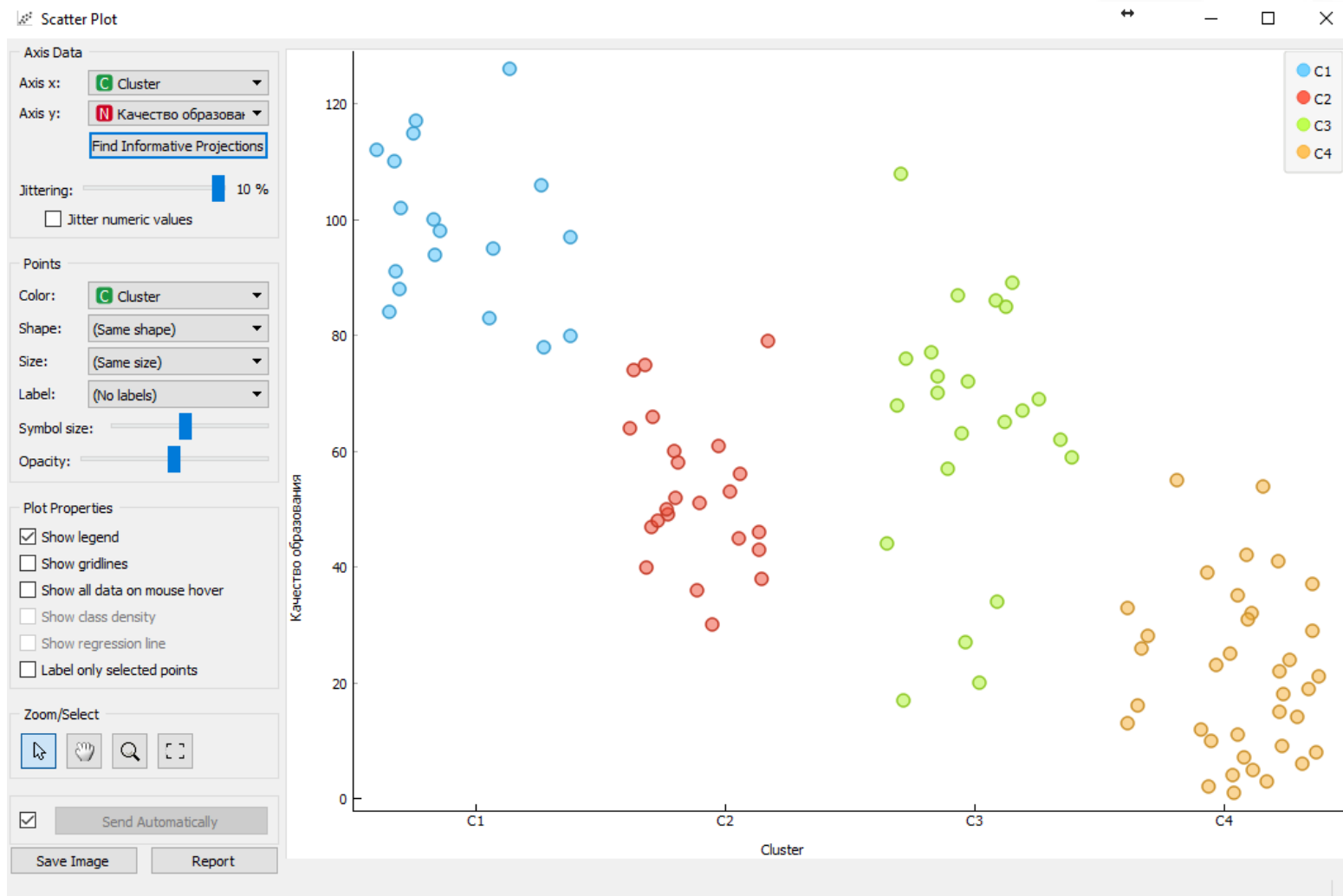


The screenshot shows the 'k-Means' configuration dialog box. It has a title bar with a question mark and a close button. The 'Number of Clusters' section has two options: 'Fixed' (selected) with a value of 4, and 'From' (unselected) with a range from 2 to 12. The 'Initialization' section has a dropdown menu set to 'Initialize with KMeans++'. Below this, there are input fields for 'Re-runs' (10) and 'Maximum iterations' (300). At the bottom, there is a checked checkbox for 'Apply Automatically' and a 'Report' button.

Результат кластеризации



Распределение университетов по качеству образования

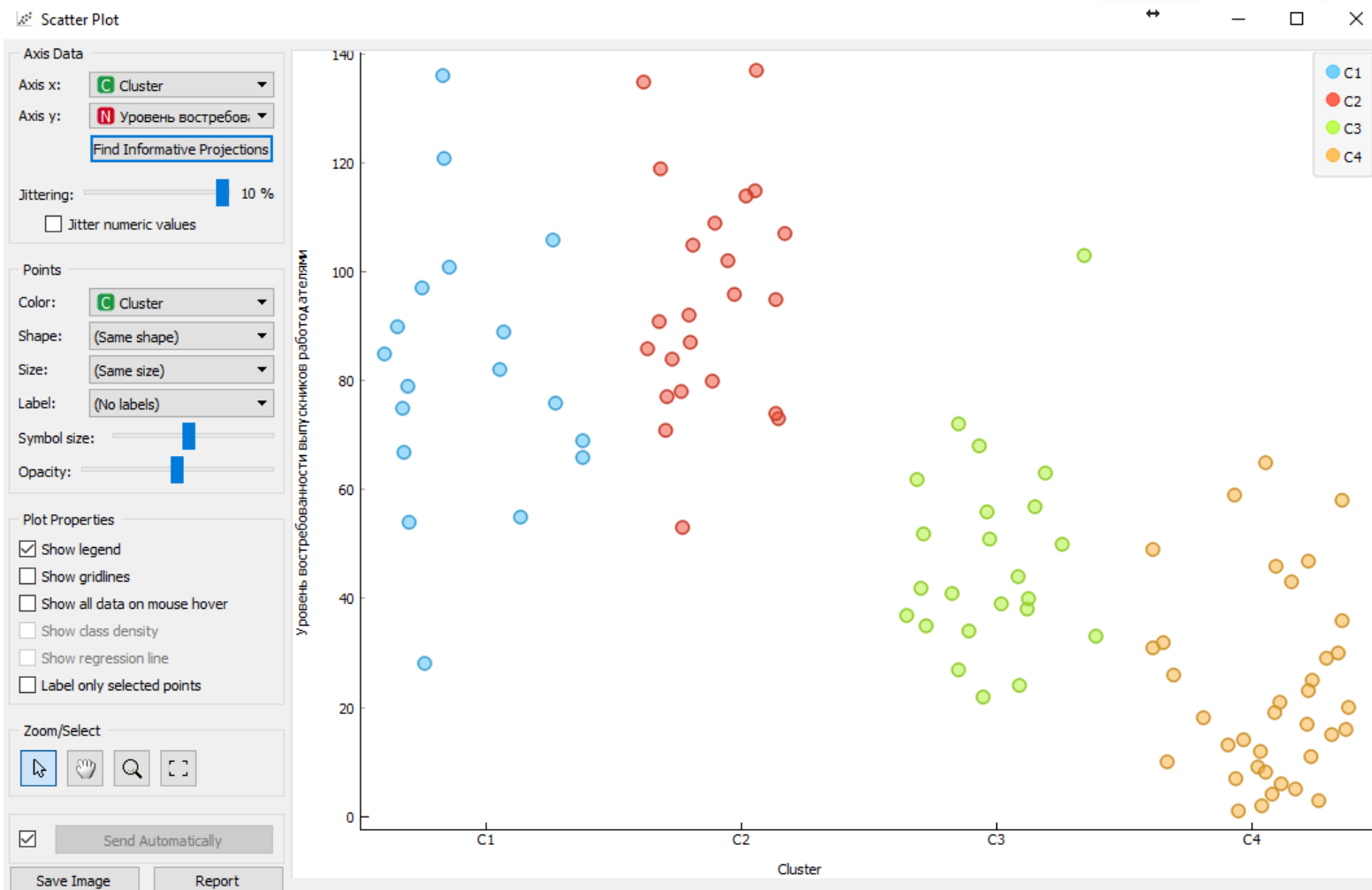


Содержимое кластера С4

	Cluster	Название	Рейтинг	Число образовательных программ
1	C4	Московский государственный университет им. М.В. Ломоносова	4.729	1
5	C4	Московский государственный институт международных отношений (университет) МИД РФ	4.223	2
2	C4	Московский физико-технический институт (государственный университет)	4.603	3
4	C4	Санкт-Петербургский государственный университет	4.244	4
6	C4	Национальный исследовательский университет "Высшая школа экономики"	4.216	5
8	C4	Национальный исследовательский Томский политехнический университет	4.100	6
3	C4	Национальный исследовательский ядерный университет «МИФИ»	4.424	7
9	C4	Новосибирский национальный исследовательский государственный университет	3.957	8
11	C4	Российская академия народного хозяйства и государственной службы при Президенте РФ	3.724	9
7	C4	Московский государственный технический университет имени Н.Э. Баумана	4.136	10
13	C4	Финансовый университет при Правительстве РФ	3.701	11
10	C4	Санкт-Петербургский политехнический университет Петра Великого	3.890	12
17	C4	Национальный исследовательский технологический университет «МИСиС»	3.560	13
20	C4	Российский университет дружбы народов	3.412	14
19	C4	Университет ИТМО	3.461	15
15	C4	Казанский (Приволжский) федеральный университет	3.589	16

Кластер университетов с наилучшим уровнем образования.

Распределение университетов по востребованности среди работодателей



Распределение университетов по востребованности среди работодателей

Data Table (1)

Info					
100 instances (no missing values)					
4 features (no missing values)					
Discrete class with 4 values (no missing values)					
1 meta attribute (no missing values)					
Variables					
☒ Show variable labels (if present)					
☒ Visualize numeric values					
☒ Color by instance classes					
Selection					
☒ Select full rows					
Restore Original Order					
Report					
☒ Send Automatically					
	Cluster	Название	Рейтинг	ество образова	ности ^
52	3	Тюменский индустриальный университет	2.262	63	22
24	4	Российский национальный исследовательский медицинский университет имени Н.И. Пирогова ...	3.175	22	23
37	3	Казанский государственный медицинский университет Министерства здравоохранения РФ	2.653	34	24
22	4	Первый Санкт-Петербургский государственный медицинский университет имени академика И.П. ...	3.320	18	25
25	4	Национальный исследовательский университет "МЭИ"	3.088	28	26
49	3	Самарский государственный технический университет	2.301	73	27
65	1	Нижегородский государственный технический университет имени Р.Е. Алексеева	2.053	117	28
20	4	Российский университет дружбы народов	3.412	14	29
14	4	Национальный исследовательский Томский государственный университет	3.647	19	30
17	4	Национальный исследовательский технологический университет «МИСиС»	3.560	13	31
15	4	Казанский (Приволжский) федеральный университет	3.589	16	32
53	3	Воронежский государственный медицинский университет имени Н.Н. Бурденко Министерства зд...	2.235	59	33
50	3	Самарский государственный медицинский университет Министерства здравоохранения РФ	2.271	57	34
61	3	Московский государственный университет путей сообщения Императора Николая II	2.071	76	35
39	4	Сибирский государственный медицинский университет Министерства здравоохранения РФ	2.633	37	36
44	3	Северо-Западный государственный медицинский университет имени И.И. Мечникова Министерс...	2.401	44	37
57	3	Уфимский государственный нефтяной технический университет	2.184	65	38
29	3	Всероссийская академия внешней торговли Министерства экономического развития РФ	2.865	20	39
63	3	Санкт-Петербургский государственный архитектурно-строительный университет	2.063	85	40
77	3	Омский государственный медицинский университет Министерства здравоохранения РФ	1.919	77	41
93	3	Тюменский государственный медицинский университет Министерства здравоохранения РФ	1.776	108	42
43	4	Томский государственный университет систем управления и радиоэлектроники	2.405	54	43

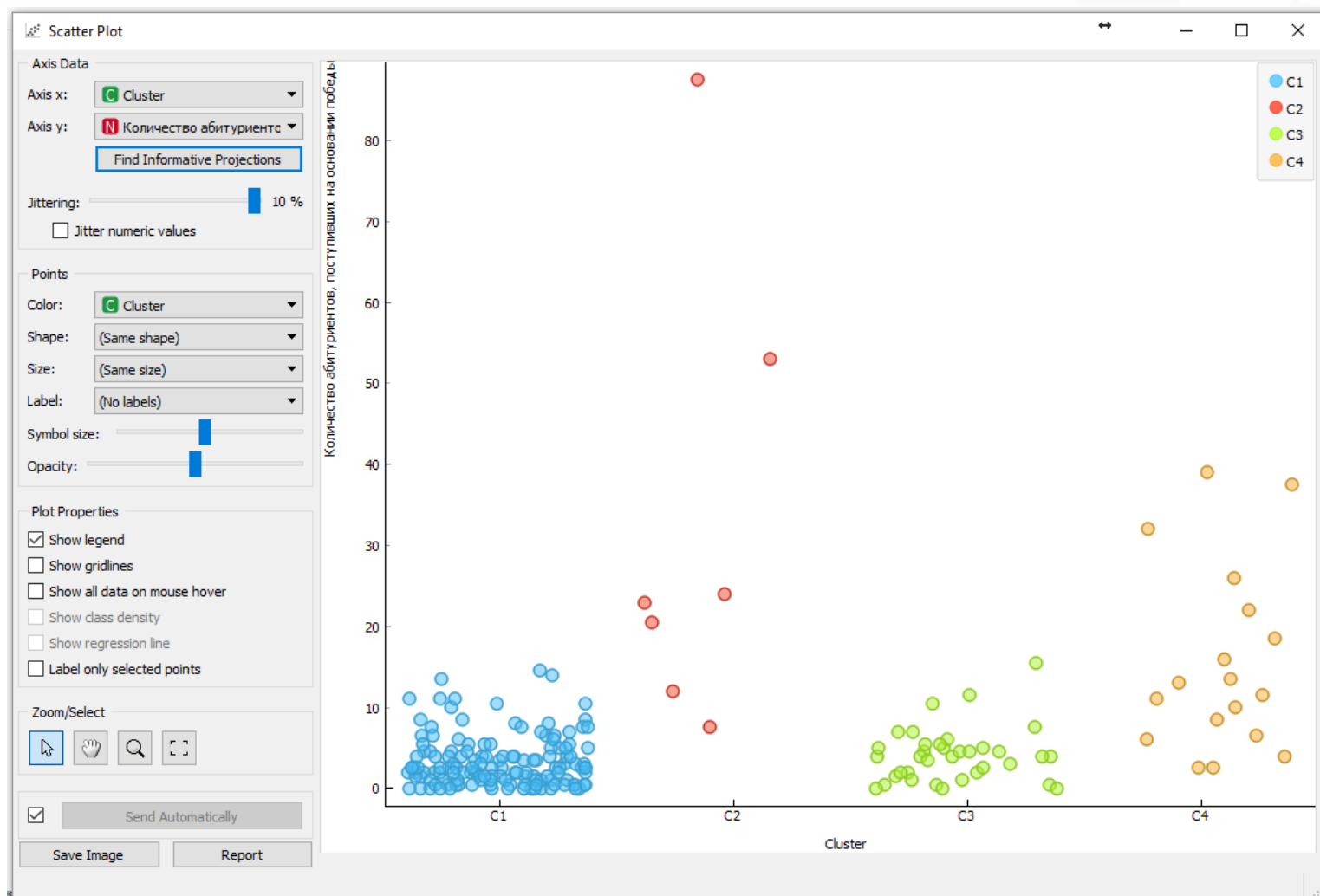
Данные по школам РФ

	A	B	C	D	E	F	G	H
1	Название	Город	Регион	Балл	Количество абитуриентов, поступивших	Количество абитуриентов, поступивших	Количество абитуриентов, поступивших	
2	Лицей № 1580 при МГТУ имени Н.Э. Баумана	Москва	Москва	166.8	168	16	87.5	
3	СУНЦ школа-интернат им. А. Н. Колмогорова МГУ им. М. В. Ломоносова	Москва	Москва	123.5	116	3	53	
4	Предуниверситетский НИЯУ МИФИ	Москва	Москва	110.4	163.5	38.5	7.5	
5	Лицей № 1502 при МЭИ	Москва	Москва	91.9	158.5	16.5	12	
6	Лицей №1535	Москва	Москва	87.8	84.5	43	23	
7	СУНЦ УрФУ им. первого Президента России Б.Н.Ельцина	Екатеринбург	Свердловская област	85.3	119	30	20.5	
8	Президентский физико-математический лицей №239	Санкт-Петербург	Санкт-Петербург	70.3	64.5	5.5	26	
9	Лицей при ТПУ	Томск	Томская область	66.6	96.5	3	24	
10	Школа 179 МИОО	Москва	Москва	65.1	34	18	39	
11	Лицей "Вторая школа"	Москва	Москва	57.9	60	13	18.5	
12	Лицей № 1581	Москва	Москва	52.6	35.5	7	37.5	
13	Центр образования № 57 "Пятьдесят седьмая школа"	Москва	Москва	52.2	25.5	18	32	
14	Санкт-Петербургский губернаторский физико-математический лицей	Санкт-Петербург	Санкт-Петербург	49.3	53.5	2	13.5	
15	Академическая гимназия имени Д.К. Фаддеева Санкт-Петербург	Санкт-Петербург	Санкт-Петербург	47.4	36.5	5	11	
16	Многопрофильный лицей № 1799	Москва	Москва	47	43	51.5	4	
17	Гимназия №1514	Москва	Москва	45.7	34.5	12	22	
18	Школа № 2086	Москва	Москва	45.1	26	51.5	10.5	
19	Лицей № 1568 имени Пабло Неруды	Москва	Москва	44.3	50.5	10.5	13	
20	Гимназия № 1567	Москва	Москва	41.3	42	23.5	8.5	
21	Гимназия № 9	Екатеринбург	Свердловская област	40.6	52.5	14	11.5	
22	Московская гимназия на Юго-Западе № 1543	Москва	Москва	40.5	36	10.5	16	
23	Школа № 2095 "Покровский квартал"	Москва	Москва	37.8	19.5	33.5	15.5	
24	Лицей № 1533 (информационных технологий)	Москва	Москва	37.5	36	33.5	4.5	
25	Лицей научно-инженерного профиля города Королева	Королев	Московская область	35.4	50	3	6.5	
26	Красноярская университетская гимназия №1 Универс	Красноярск	Красноярский край	35.2	49.5	40	2	
27	Лицей №130	Екатеринбург	Свердловская област	35.2	63.5	10.5	4	
28	СУНЦ НГУ	Новосибирск	Новосибирская облас	35.2	40.5	9.5	10	
29	Средняя общеобразовательная школа № 654 имени А.Д. Фри	Москва	Москва	33	32.5	20	7	
30	Гимназия №13 "Академ"	Красноярск	Красноярский край	32.6	52	25	4	

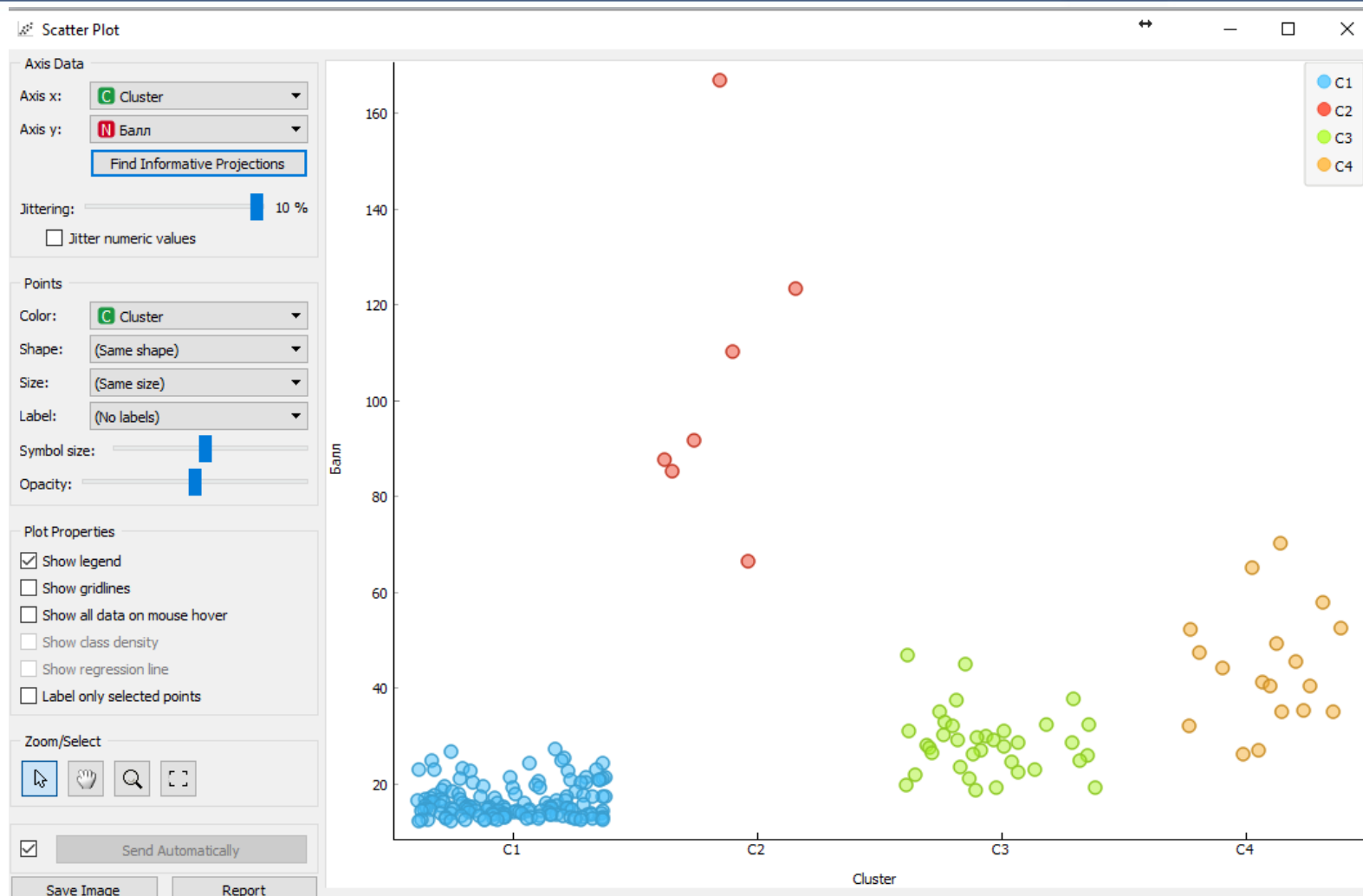
Результаты кластеризации по школам РФ

	Cluster	Название	Город	Регион	Балл	вших по конкур	поступивши
1	C2	Лицей № 1580 при МГТУ имени Н.Э. Баумана	Москва	Москва	166.8	168.0	16.0
2	C2	СУНЦ школа-интернат им. А. Н. Колмогорова МГУ им. М. В. Ломоно...	Москва	Москва	123.5	116.0	3.0
3	C2	Предуниверситарий НИЯУ МИФИ	Москва	Москва	110.4	163.5	38.5
4	C2	Лицей № 1502 при МЭИ	Москва	Москва	91.9	158.5	16.5
5	C2	Лицей №1535	Москва	Москва	87.8	84.5	43.0
6	C2	СУНЦ УрФУ им. первого Президента России Б.Н.Ельцина	Екатеринбург	Свердловская область	85.3	119.0	30.0
7	C4	Президентский физико-математический лицей №239	Санкт-Петербу...	Санкт-Петербург	70.3	64.5	5.5
8	C2	Лицей при ТПУ	Томск	Томская область	66.6	96.5	3.0
9	C4	Школа 179 МИОО	Москва	Москва	65.1	34.0	18.0
10	C4	Лицей "Вторая школа"	Москва	Москва	57.9	60.0	13.0
11	C4	Лицей № 1581	Москва	Москва	52.6	35.5	7.0
12	C4	Центр образования № 57 "Пятьдесят седьмая школа"	Москва	Москва	52.2	25.5	18.0
13	C4	Санкт-Петербургский губернаторский физико-математический лице...	Санкт-Петербу...	Санкт-Петербург	49.3	53.5	2.0
14	C4	Академическая гимназия имени Д.К. Фаддеева Санкт-Петербургског...	Санкт-Петербу...	Санкт-Петербург	47.4	36.5	5.0
15	C3	Многопрофильный лицей № 1799	Москва	Москва	47.0	43.0	51.5
16	C4	Гимназия №1514	Москва	Москва	45.7	34.5	12.0
17	C3	Школа № 2086	Москва	Москва	45.1	26.0	51.5
18	C4	Лицей № 1568 имени Пабло Неруды	Москва	Москва	44.3	50.5	10.5
19	C4	Гимназия № 1567	Москва	Москва	41.3	42.0	23.5
20	C4	Гимназия № 9	Екатеринбург	Свердловская область	40.6	52.5	14.0
21	C4	Московская гимназия на Юго-Западе № 1543	Москва	Москва	40.5	36.0	10.5
22	C3	Школа № 2095 "Покровский квартал"	Москва	Москва	37.8	19.5	33.5

Распределение школ по кластерам и количеству студентов, поступивших на основании победы в олимпиаде



Распределение школ по кластерам и параметру 'балл'



Кластер школ с низкими баллами

Cluster	Название	Город	Регион	Балл
C1	Лицей № 1550	Москва	Москва	13.7
C1	Лицей № 8	Красноярск	Красноярский край	13.7
C1	Гимназия № 9	Красноярск	Красноярский край	13.7
C1	Гимназия № 161	Екатеринбург	Свердловская область	13.6
C1	Гимназия № 26	Набережные Челны	Татарстан республика	13.4
C1	Лицей № 1367	Москва	Москва	13.4
C1	Гимназия № 1569 "Созвездие"	Москва	Москва	13.4
C1	Лицей г. Лесосибирска	Лесосибирск	Красноярский край	13.3
C1	Гимназия № 16	Красноярск	Красноярский край	13.2
C1	Гимназия № 37	Екатеринбург	Свердловская область	13.2
C1	Лицей № 88	Екатеринбург	Свердловская область	13.2
C1	Школа с углубленным изучением иностранных языков № 1205	Москва	Москва	13.2
C1	Гимназия № 1811 "Восточное Измайлово"	Москва	Москва	13.2
C1	Школа № 2006	Москва	Москва	13.1
C1	Школа № 18	Казань	Татарстан республика	13.1
C1	Школа № 1106	Москва	Москва	13.0
C1	Школа с углубленным изучением английского языка № 1208 имени ...	Москва	Москва	13.0
C1	Лицей-интернат "Республиканский лицей"	Якутск	Саха (Якутия) республика	13.0
C1	Школа № 1474	Москва	Москва	13.0
C1	Пушкинский лицей № 1500	Москва	Москва	12.9
C1	Общеобразовательный лицей "АМТЭК"	Череповец	Вологодская область	12.9
C1	Школа № 151	Красноярск	Красноярский край	12.8

Технология автоматического анализа тональности отзывов.

Сентимент-анализ — это раздел тексто - майнинга (text mining), система автоматического извлечения субъективных мнений из текста, дисциплина на стыке поиска информации и вычислительной лингвистики, которая исследует не столько содержание текста, сколько его тональность.



Что такое тональность текста и зачем она нужна?

Под определением тональности текста подразумевается выделение фрагментов текста, выражающих позитивную или негативную эмоциональность по отношению к объекту этой эмоциональной оценки (объекту тональности).

Объектом может выступать как имя собственное, так название продукта (например наименование бренда), организации (например Газпром), услуги, профессии и т. п. по отношению к которому анализируется текст.

Таким образом, тональность текста определяется такими тремя факторами:

1. **Субъект тональности** – это автор текста, рецензии, комментария и т.п.
 2. **Тональная оценка** (позитив/нейтрально/негатив) – это эмоциональное отношение автора к объекту тональности, выраженное в использовании специфической лексики при написании текста.
 3. **Объект тональности** — продукт, событие, персона или явления, о которых высказывается автор
- Оценка тональности текстов предназначена для получения обратной связи между потребителями услуги и производителями услуги. Например:
1. Оценка деятельности госучреждения в СМИ и в социальных сетях...
 2. Возможность предсказания котировок (в качестве дополнительного индекса) на фондовых рынках.
 3. Мониторинг экстремистских высказываний в СМИ, интернете.

Анализ тональности текста в Twitter



Модели в анализе тональности



подход
на правилах

Подход, основанный на правилах состоит из набора правил, применяя которые система делает заключение о тональности текста. Например, для предложения «**Я люблю кофе**», можно применить следующее правило:

если сказуемое ("люблю") входит в положительный набор глаголов ("люблю", "обожаю", "одобряю" ...) и в предложении не имеется отрицаний, то классифицировать тональность как "положительная"

Многие коммерческие системы (например **I-Teco**) используют данный подход, несмотря на то что он требует больших затрат, т.к. для хорошей работы системы необходимо составить большое количество правил. Зачастую правила привязаны к определенному домену (например, «ресторанная тематика») и при смене домена («обзор фотоаппаратов») требуется заново составлять правила.

Модели в анализе тональности



подход
со словарем

Подходы, основанные на словарях, используют так называемые тональные словари (affective lexicons) для анализа текста. В простом виде тональный словарь представляет из себя список слов со значением тональности для каждого слова.

слово	валентность (1-9)
счастливый	8.21
хороший	7.47
скучный	2.95
сердитый	2.85
грустный	1.61

Чтобы проанализировать текст, можно воспользоваться следующим алгоритмом: сначала каждому слову в тексте присвоить его значением тональности из словаря (если оно присутствует в словаре), а затем вычислить общую тональность всего текста.

Препроцессинг текста

Для того, чтобы векторная модель документа наиболее точно отображала тематику текста, необходимо каждое слово в тексте преобразовывать в уникальный терм и только потом считать статистику употребления термина.

Терм – это значимая часть слова, очищенная от незначащих аффиксов (т.е. словообразовательных морфем, которые присоединяются к корню слова). Для этого к текстам применяются процессы стемминга и лемматизации.

Лемматизация — процесс привода словоформы к лемме – её словарной форме.

Стемминг – это процесс нахождения основы слова для заданного исходного слова, то есть отсечение от слова окончаний и суффиксов. При преобразовании текста в векторную модель может также происходить удаление незначимых символов (например, знаков препинания).

После составления списка термов документа производится подсчет статистики употребления слов. Также для экономии времени вычислений в цикле подсчета статистики происходит проверка, входит ли терм в список стоп-слов.

Стоп-слова – слова, не несущие в себе смысла и не влияющие на его тематику. Такими словами чаще всего являются предлоги, союзы и другие части речи, предназначенные для связи слов в предложениях.

Процедура лемматизации

Все технологии /

MyStem

Программа MyStem производит морфологический анализ текста на русском языке. Она умеет строить гипотетические разборы для слов, не входящих в словарь. Первую версию программы написали [Илья Сегалович](#) и Виталий Титов.

[Документация](#) >

[Принципы работы MyStem](#)

[Условия использования](#) >

Все вопросы, замечания и предложения отправляйте на mystem@yandex-team.ru.

Файлы

Версия 3.0

[mystem-3.0-win7-32bit.zip](#)

15,51 МБ

[mystem-3.0-win7-64bit.zip](#)

15,80 МБ

[mystem-3.0-linux3.5-32bit.tar.gz](#)

15,76 МБ

[mystem-3.0-linux3.1-64bit.tar.gz](#)

15,70 МБ

Версия 2.1

[mystem-2.1-win7-32bit.zip](#)

4,63 МБ

[mystem-2.1-win7-64bit.zip](#)

4,92 МБ

[mystem-2.1-linux2.6-64bit.tar.gz](#)

4,72 МБ

[mystem-2.1-linux3.0-32bit.tar.gz](#)

4,76 МБ

Программа MyStem производит морфологический анализ текста на русском языке. Она умеет строить гипотетические разборы для слов, не входящих в словарь.

Результат лемматизации

{отзыв} : {санкт-петербургский} {государственный} {университет}
{http??}://{vuz??}.{edunetwork??}.{ru??}/78/v570/{opinions??}/{ирин|ирина}
{муравьев|муравьева} 16:43 23.06.2016{я} {обучаться} ({читать}: {терять}
{время}) {на} {филфак} {СПбГУ??}, {на} {отделение} {новогреческий} {язык}.
{впечатление} {один}: {отстой|отстоять}. {преподавательский} {состав}
{язык} {категорически|категорический} {не} {владеть}



отзыв : санкт-петербургский государственный
университет http?? :// vuz?? . edunetwork?? . ru?? /78/v570/
opinions?? / ирин|ирина муравьев|муравьева 16:43
23.06.2016 я обучаться (читать : терять время) на
филфак СПбГУ?? , на отделение новогреческий язык .
впечатление один : отстой|отстоять . преподавательский
состав язык категорически|категорический не владеть

Sentistrength

[Test](#) - [Download](#) - [Java Version](#) - [Non-English](#) - [Buy!](#) - [About](#)



SentiStrength

SentiStrength estimates the *strength* of positive and negative sentiment in *short texts*, even for informal language. It has [human-level accuracy](#) for short social web texts in English, except political texts. SentiStrength reports *two* sentiment strengths:

-1 (not negative) to -5 (extremely negative)

1 (not positive) to 5 (extremely positive)

It can also report binary (positive/negative), trinary (positive/negative/neutral) and single scale (-4 to +4) results. SentiStrength was originally developed for English and optimised for general short social web texts but can be configured for other languages and contexts by changing its input files - some variants are demonstrated below.

Quick Tests (English version):

Enter text:

Output: ☒ Dual, ☐ binary, ☐ trinary, ☐ scale

Keyword test:

Enter keywords (comma-separated list, no spaces):

Topic test:

Select domain (broad topic):

Other languages: [Finnish](#), [German](#), [Dutch](#), [Spanish](#), [Russian](#), [Portuguese](#), [French](#), [Arabic](#), [Polish](#), [Persian](#), [Swedish](#), [Greek](#), [Welsh](#), [Italian](#), [Turkish](#).

Работа с SentiStrength

Главная Вставка Дизайн Анимация Показ слайдов Рецензирование Вид Разработчик Acrobat

SentiStrength v2.2

File Sentiment Strength Analysis Sentiment Analysis Options Reporting Options Data Mining Help

Analyse ONE text -Reporting tagged and standard English versions, plus emotion categories

Analyse ALL Texts in File [each line separately] **Reporting all information selected in the Reporting Option menu

-As above for ALL files in folder

Classify column n in file, save column n and column m to new file

-As above, all files in folder

1 отзыв :

[old] Try adding and subtracting

Optimise the emotion dictionary

Run 10-fold cross-validation test

Run a 10-fold cross-validation

[Change number of folds in cross-validation]

[Speed algorithm by pre-loading dictionary]

Report frequency of use of sentiment words

Save a version of sentiment strength analysis

YouTube Networks - Calculate sentiment

Detect Sentiment in all Mozilla

меню

14 второе второй высокий на филологический факультет СпбГУ?? .полн

15 поступать не рекомендовать . качество образование не соответство

16 если вы хотите усовершенствовать /изучать /подтягивать английский

17 но не практический его он оно применение и отработка навык в

18 бьем?? информация приходится прорабатывать самостоятельно сам

19 жизнь и работа . весь все свое свой свободный время вы быть пос

20 выполнять в полный мера мера). курс рассчитывать на человек за

21 быть совсем иной уровень подготовка и вы придтись приходится

22 уничтожать настраивать настрой на продуктивный работа . требовани

23 и пр?? . во время сессия стресс зашкаливать по весь все фронт, ст

24 истощение и подрывать здоровье). на мой мыть взгляд, предлагат

25 взрослый человек достойный более много уважительный и равный

26 анонимный отзыв 21:57 03.08.2015

27 отзыв второе второй высокий филологический СпбГУ?? . отзыв . по

A	B	C	D	E	F	G	H
Positive	Negative	Emotion	Rationale				
1	-1	санкт[0] петербургский[0] государственный[0]					
1	-1	http[0] [[Sentence=-1,1=word max, 1-5]] ://[0]					
1	-1	ирин[0] ирина[0] муравьев[0] муравьева[0] 16:43 23.06.2016					
1	-2	я[0] обучаться[0] читать[0] терять[0] время[0]					
1	-2	без[0] преувеличение[-1] на[0] уровень[0] [[Sentence=-1,1=word max, 1-5]]					
2	-2	до[0] откровенно[0] откровенный[1] жалкий[0]					
2	-1	этот[0] греческий[0] стишок[0] [[Sentence=-1,1=word max, 1-5]]					
2	-1	не[0] мочь[0] внятно[0] внятный[1] объяснять[0]					
2	-2	плохо[0] плохой[-1] худо[0] худой[-1] точно[0]					
2	-2	год[0] лет[0] лета[0] выслушивать[0] чей[0] то					
2	-1	P[0] [[Sentence=-1,1=word max, 1-5]] S[0] [[Sentence=-1,1=word max, 1-5]]					
2	-1	анонимный[0] отзыв[1] 17:40 21.05.2016					
2	-2	второе[0] второй[0] высокий[0] на[0] филоло					
1	-1	поступать[0] не[0] рекомендовать[0] [[Sentence=-1,1=word max, 1-5]]					
2	-1	если[0] вы[0] хотите[0] усовершенствовать[0]					
1	-2	но[0] не[0] практический[0] его[0] он[0] оно[0]					
2	-2	бьем[0] [[Sentence=-1,1=word max, 1-5]] инфо					
2	-1	жизнь[0] и[0] работа[0] [[Sentence=-1,1=word max, 1-5]]					
2	-1	выполнять[0] в[0] полный[0] мера[0] мера[0]					
2	-4	быть[0] совсем[0] иной[0] уровень[0] подгото					
2	-2	уничтожать[-1] настраивать[0] настрой[0] на[0]					
2	-2	и[0] пр[0] [[Sentence=-1,1=word max, 1-5]] во[0]					
1	-2	истощение[-1] и[0] подрывать[-1] здоровье[0]					
2	-1	взрослый[0] человек[0] достойный[1] более[0]					
2	-1	анонимный[0] отзыв[1] 21:57 03.08.2015					
2	-1	отзыв[1] второе[0] второй[0] высокий[0] фи					



NATIONAL RESEARCH
UNIVERSITY

Thank you for your attention!

Room 216, building 2, 55 Sedova St., St.Petersburg, Russia
Laboratory for Internet Studies
www.linis.hse.ru