Московский институт электроники и математики им. А.Н. Тихонова

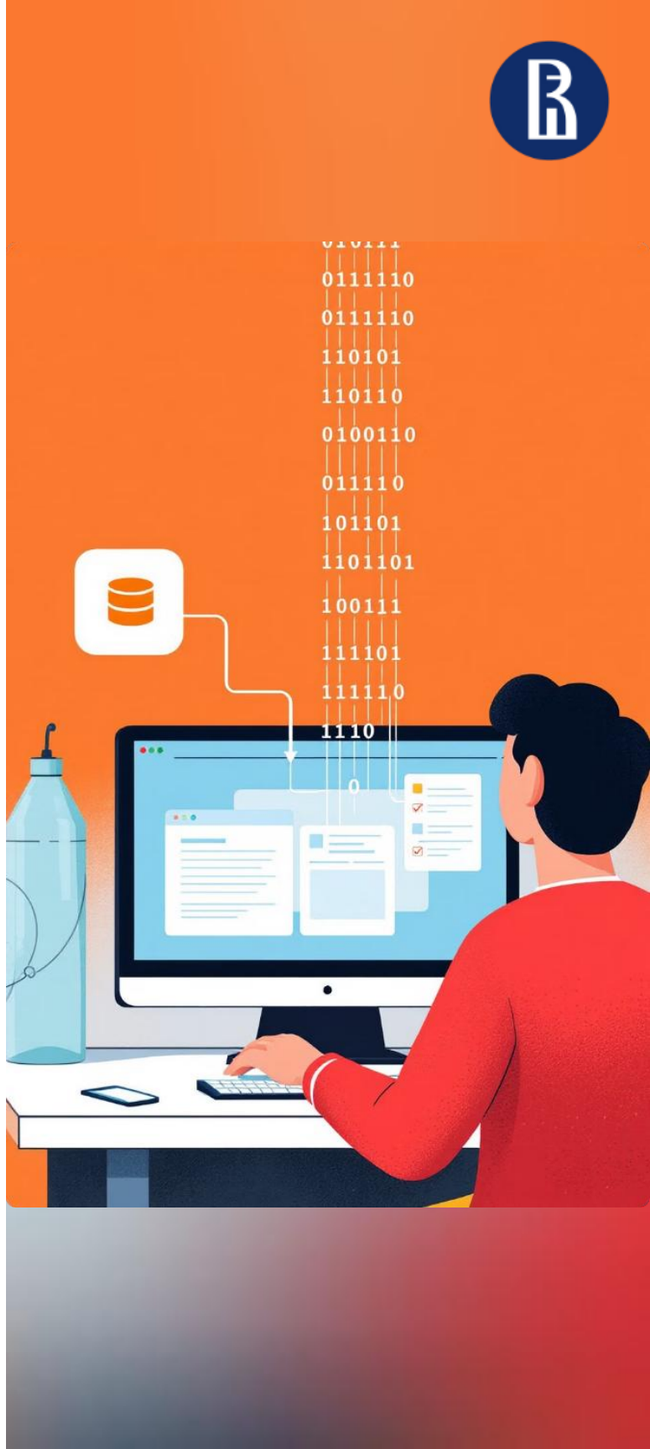
Кафедра информационной безопасности киберфизических систем

Москва 2025

Web Scraping и анализ данных

Web Scraping и анализ данных в RapidMiner

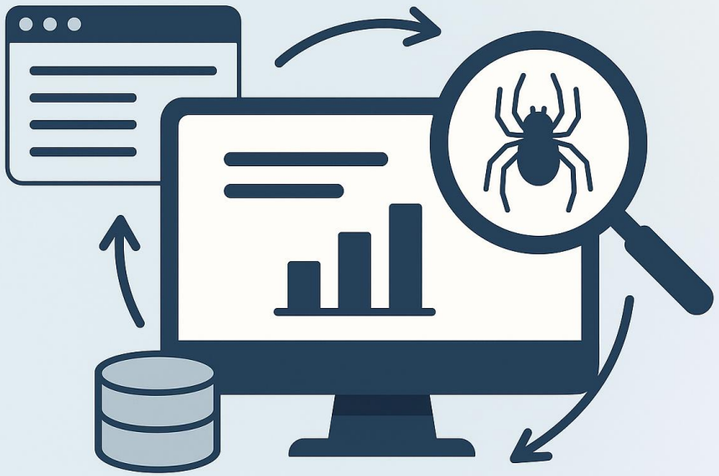
Технология web scraping позволяет извлекать и обрабатывать данные с веб-страниц. RapidMiner предлагает инструменты для автоматического сбора, очистки и анализа веб-данных без необходимости программирования.



Понятие Web Scraping

Web Scraping — метод автоматического извлечения больших объёмов данных с веб-сайтов.

Основная цель – преобразовать неструктурированную веб-информацию в структурированный и пригодный для анализа формат.



Web Scraping

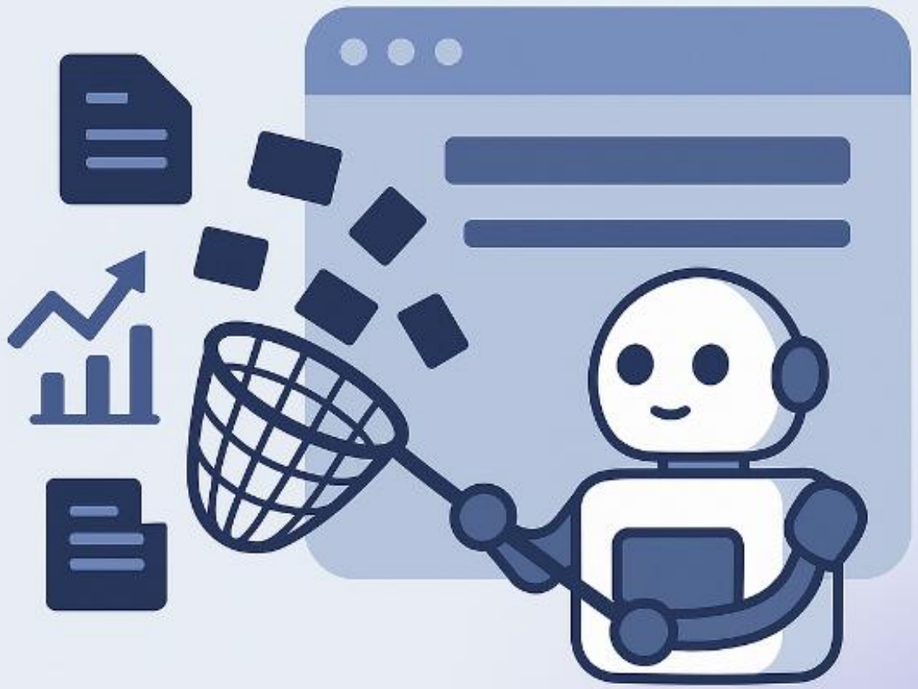
Зачем нужен Web Scraping

- Сбор данных о конкурентах и рынке.
- Мониторинг цен на товары и услуги.
- Извлечение данных для исследований и аналитических проектов в различных сферах.

Правовые аспекты Web Scraping

Использование технологии Web Scraping требует соблюдения законодательства о персональных данных, авторском праве и условий использования веб-сайтов. Важно проверять юридическую допустимость извлечения конкретной информации.

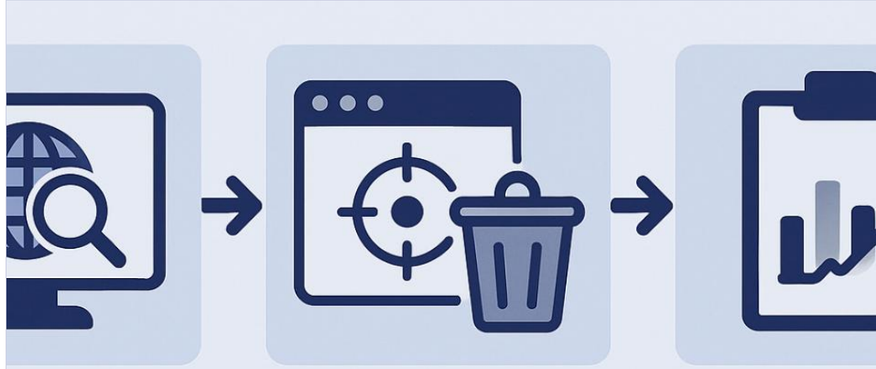
Этика и ограничения Web Scraping



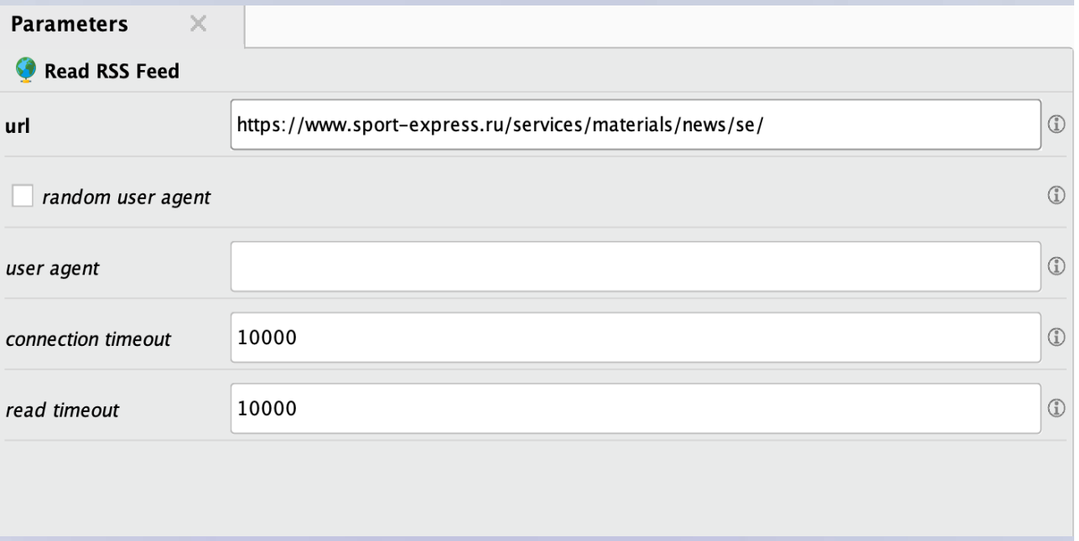
- Robots.txt и условий сайта.
- Запрет на массовое копирование защищённых данных.
- Рекомендуется избегать чрезмерной нагрузки на серверы сайтов.

Этапы Web Scraping в RapidMiner

Работа в RapidMiner включает последовательные этапы: подключение к сайту, извлечение информации, очистка данных от ненужных элементов и дальнейший структурированный анализ полученных результатов.



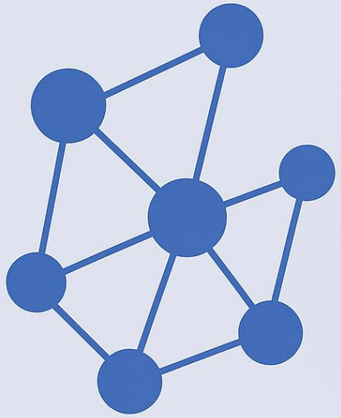
Подключение к веб-ресурсам

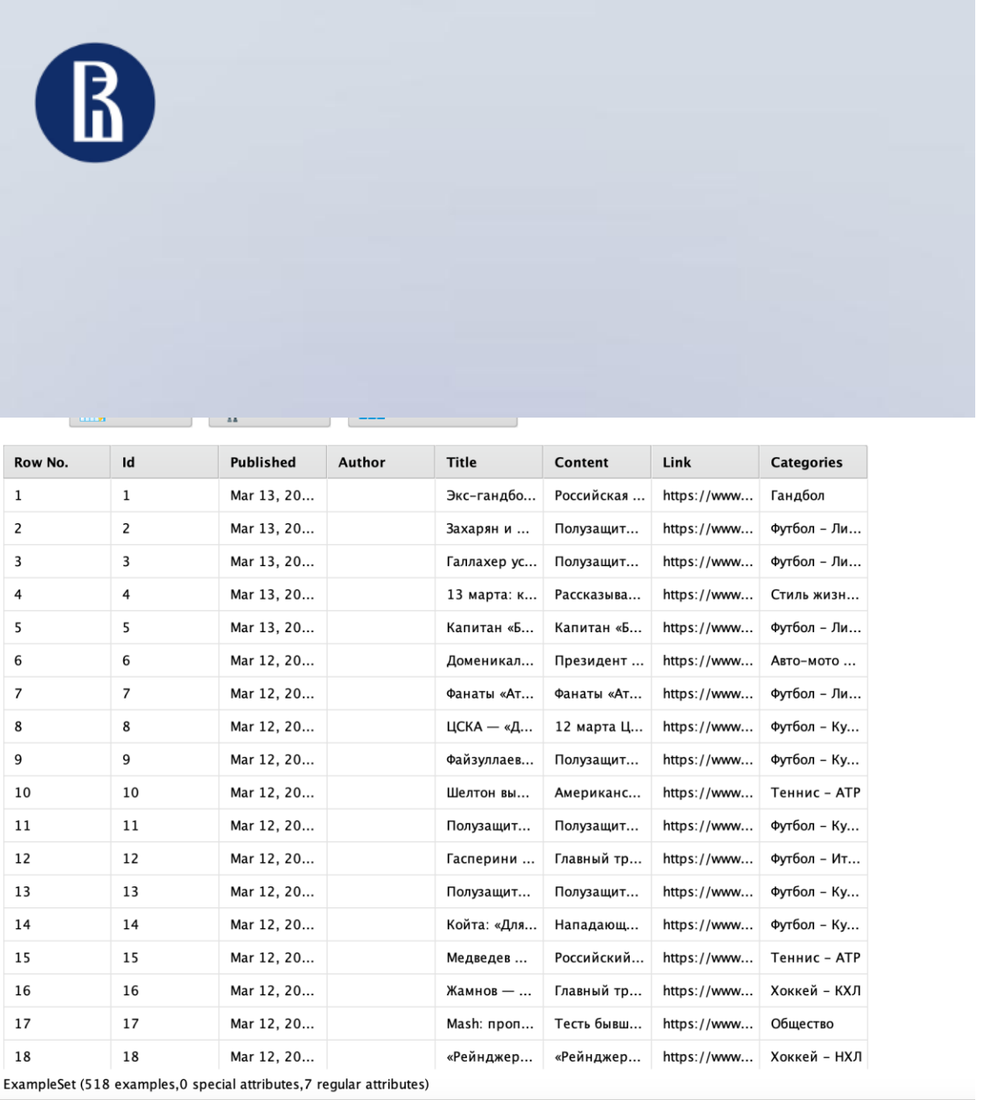


RapidMiner использует операторы Web Mining (Web Crawl, Get Page). Эти инструменты позволяют загружать HTML-страницы и получать контент для дальнейшей обработки и анализа.

Особенности оператора Web Crawl

Оператор Web Crawl предназначен для автоматического обхода множества страниц сайта по ссылкам. Удобен для извлечения данных с больших сайтов, имеющих многоуровневую структуру.





Row No.	Id	Published	Author	Title	Content	Link	Categories
1	1	Mar 13, 20...		Экс-гандбо...	Российская ...	https://www...	Гандбол
2	2	Mar 13, 20...		Захарян и ...	Полузащит...	https://www...	Футбол - Ли...
3	3	Mar 13, 20...		Галлахер ус...	Полузащит...	https://www...	Футбол - Ли...
4	4	Mar 13, 20...		13 марта: к...	Рассказыва...	https://www...	Стиль жизн...
5	5	Mar 13, 20...		Капитан «Б...	Капитан «Б...	https://www...	Футбол - Ли...
6	6	Mar 12, 20...		Доменикал...	Президент ...	https://www...	Авто-мото ...
7	7	Mar 12, 20...		Фанаты «Ат...	Фанаты «Ат...	https://www...	Футбол - Ли...
8	8	Mar 12, 20...		ЦСКА — «Д...	12 марта Ц...	https://www...	Футбол - Ку...
9	9	Mar 12, 20...		Файзуллаев...	Полузащит...	https://www...	Футбол - Ку...
10	10	Mar 12, 20...		Шелтон вы...	Американс...	https://www...	Теннис - ATP
11	11	Mar 12, 20...		Полузащит...	Полузащит...	https://www...	Футбол - Ку...
12	12	Mar 12, 20...		Гасперини ...	Главный тр...	https://www...	Футбол - Ит...
13	13	Mar 12, 20...		Полузащит...	Полузащит...	https://www...	Футбол - Ку...
14	14	Mar 12, 20...		Койта: «Для...	Нападающ...	https://www...	Футбол - Ку...
15	15	Mar 12, 20...		Медведев ...	Российский...	https://www...	Теннис - ATP
16	16	Mar 12, 20...		Жамнов — ...	Главный тр...	https://www...	Хоккей - КХЛ
17	17	Mar 12, 20...		Mash: проп...	Тесть бывш...	https://www...	Общество
18	18	Mar 12, 20...		«Рейнджер...	«Рейнджер...	https://www...	Хоккей - НХЛ

ExampleSet (518 examples,0 special attributes,7 regular attributes)

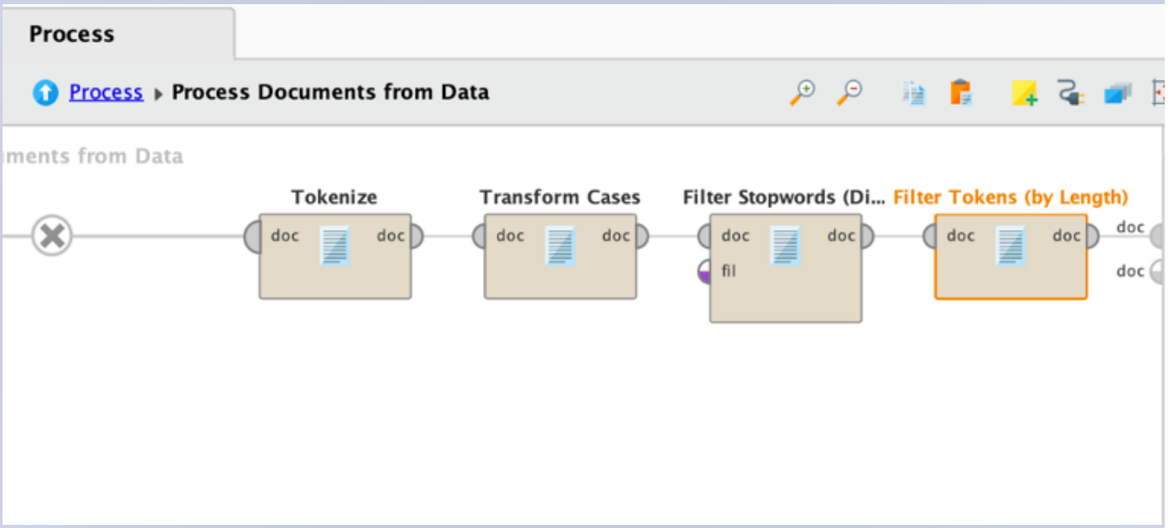
Извлечение элементов HTML

Оператор Extract Information использует XPath-запросы. XPath позволяет выбирать конкретные элементы (теги, классы, идентификаторы) из HTML-страницы и преобразовывать их в структурированный формат.

XPath-запросы: краткий обзор

XPath — язык запросов, используемый для навигации по XML и HTML-документам. Например, запрос "//div[@class='price']" извлечёт все элементы div с указанным классом.

Очистка и обработка данных

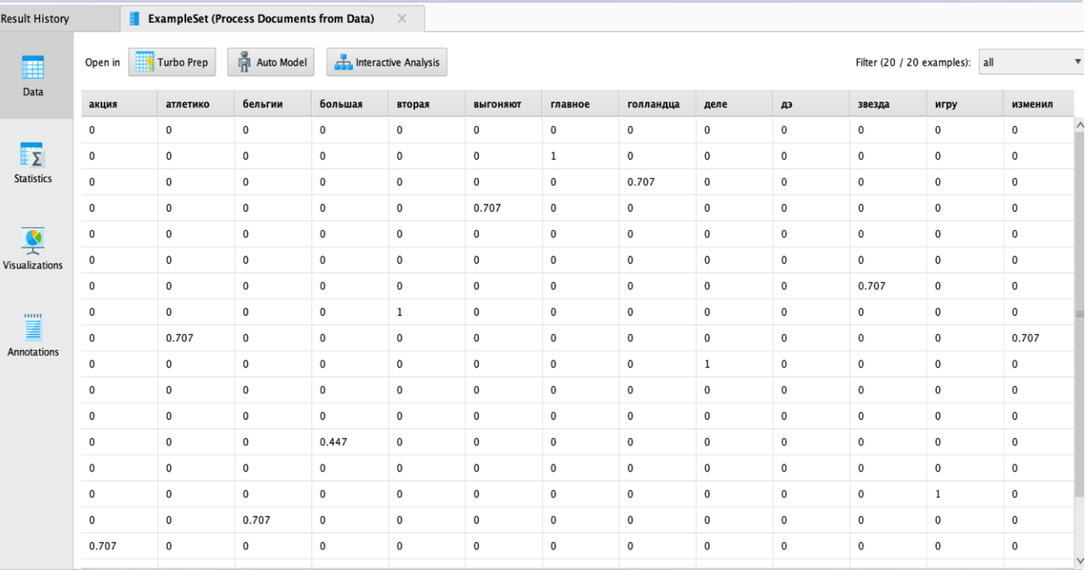


Полученные данные часто содержат лишние символы, HTML-теги или пробелы. Операторы Replace, Filter Tokens и Trim в RapidMiner помогают быстро привести данные к пригодному для анализа виду.

Типичные проблемы извлечённых данных

- Лишние пробелы и знаки препинания.
- HTML-теги внутри текста.
- Некорректные типы данных (текст вместо чисел).

Преобразование типов данных

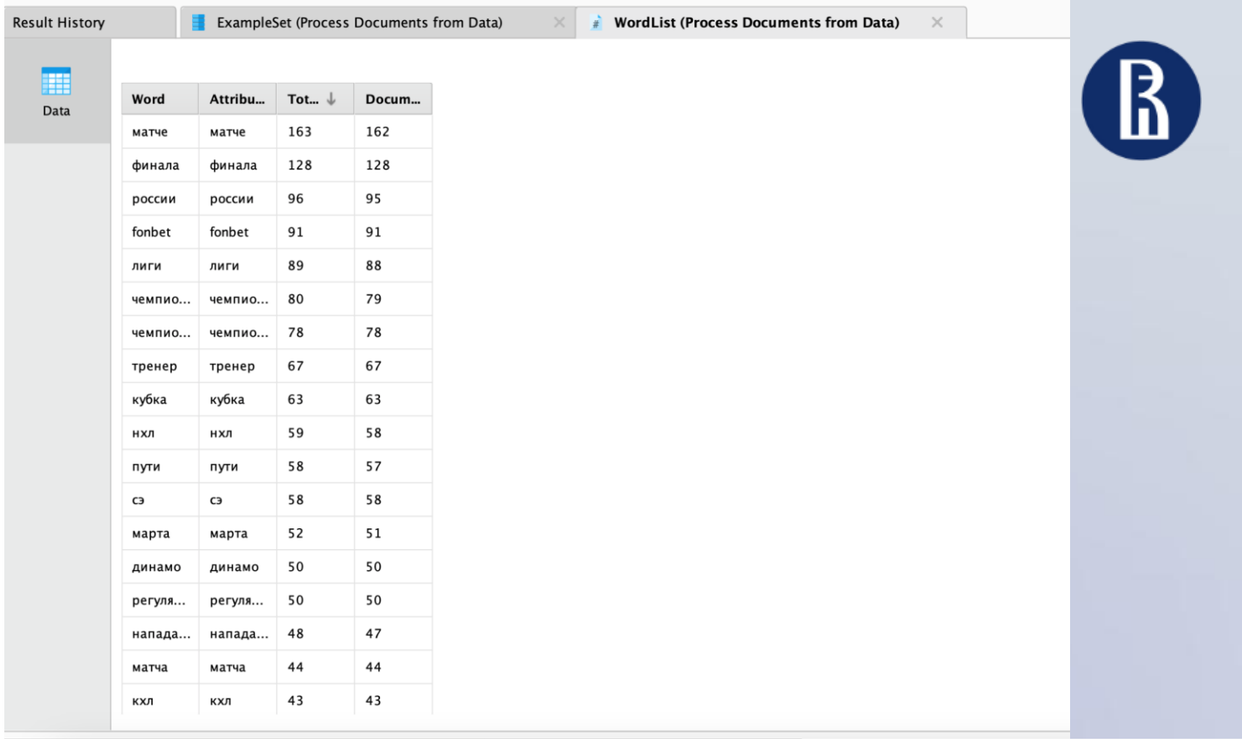


акция	атлетико	бельгии	большая	вторая	выгоняют	главное	голландца	деле	дэ	звезда	игру	изменил
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0.707	0	0	0	0	0
0	0	0	0	0	0.707	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0.707	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0
0	0.707	0	0	0	0	0	0	0	0	0	0	0.707
0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.447	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0.707	0	0	0	0	0	0	0	0	0	0
0.707	0	0	0	0	0	0	0	0	0	0	0	0

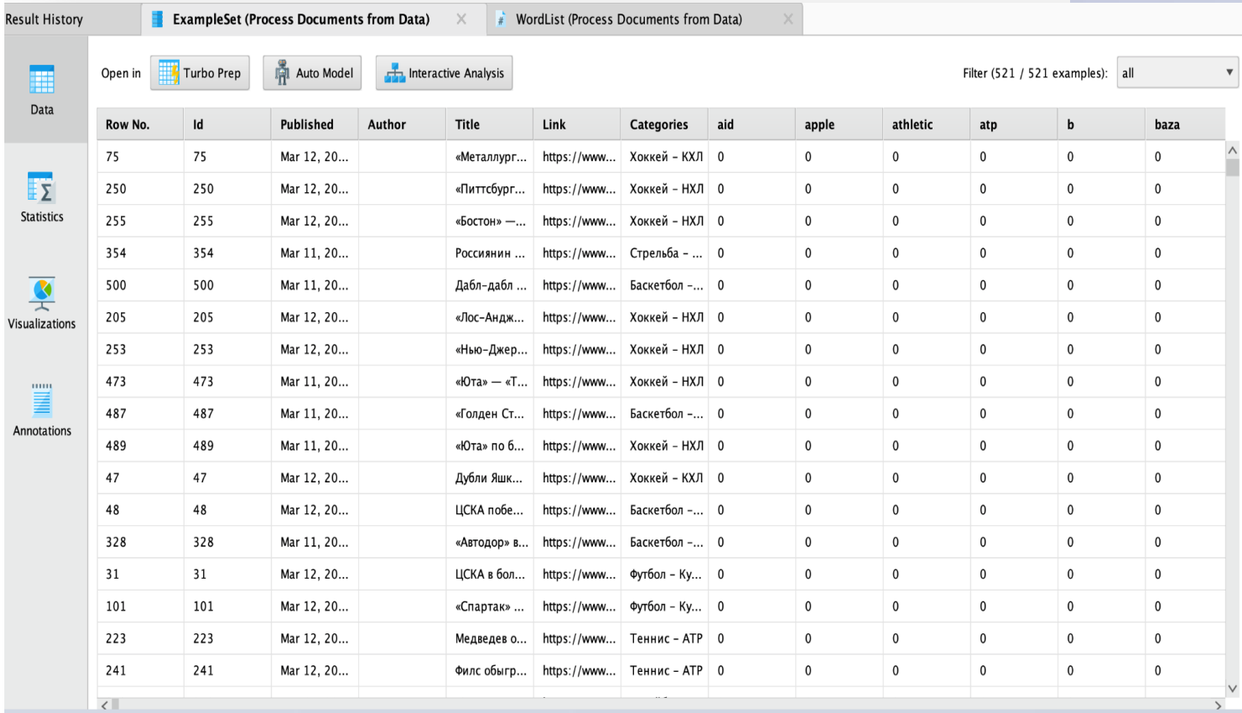
Использование оператора Parse Numbers преобразует текстовые данные в числовой формат, необходимый для проведения вычислений, фильтрации и последующего анализа.

Анализ данных после сбора

Извлечённые веб-данные подлежат дальнейшему анализу с помощью стандартных инструментов RapidMiner: статистических методов, кластеризации, классификации и прогнозного моделирования.

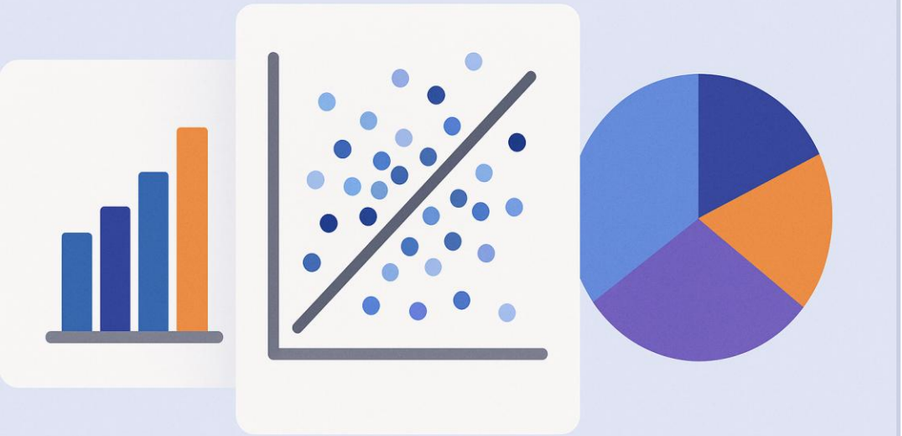


Word	Attribu...	Tot...	Docum...
матче	матче	163	162
финала	финала	128	128
россии	россии	96	95
fonbet	fonbet	91	91
лиги	лиги	89	88
чемпио...	чемпио...	80	79
чемпио...	чемпио...	78	78
тренер	тренер	67	67
кубка	кубка	63	63
нхл	нхл	59	58
пути	пути	58	57
сэ	сэ	58	58
марта	марта	52	51
динамо	динамо	50	50
регуля...	регуля...	50	50
нападa...	нападa...	48	47
матча	матча	44	44
кхл	кхл	43	43



Row No.	Id	Published	Author	Title	Link	Categories	aid	apple	athletic	atp	b	baza
75	75	Mar 12, 20...		«Металлург...	https://www...	Хоккей - КХЛ	0	0	0	0	0	0
250	250	Mar 12, 20...		«Питтсбург...	https://www...	Хоккей - НХЛ	0	0	0	0	0	0
255	255	Mar 12, 20...		«Бостон» —...	https://www...	Хоккей - НХЛ	0	0	0	0	0	0
354	354	Mar 11, 20...		Россиянин ...	https://www...	Стрельба - ...	0	0	0	0	0	0
500	500	Mar 11, 20...		Дабл-дабл ...	https://www...	Баскетбол -...	0	0	0	0	0	0
205	205	Mar 12, 20...		«Лос-Андж...	https://www...	Хоккей - НХЛ	0	0	0	0	0	0
253	253	Mar 12, 20...		«Нью-Джер...	https://www...	Хоккей - НХЛ	0	0	0	0	0	0
473	473	Mar 11, 20...		«Юта» — «Т...	https://www...	Хоккей - НХЛ	0	0	0	0	0	0
487	487	Mar 11, 20...		«Голден Ст...	https://www...	Баскетбол -...	0	0	0	0	0	0
489	489	Mar 11, 20...		«Юта» по б...	https://www...	Хоккей - НХЛ	0	0	0	0	0	0
47	47	Mar 12, 20...		Дубли Яшк...	https://www...	Хоккей - КХЛ	0	0	0	0	0	0
48	48	Mar 12, 20...		ЦСКА побе...	https://www...	Баскетбол -...	0	0	0	0	0	0
328	328	Mar 11, 20...		«Автодор» в...	https://www...	Баскетбол -...	0	0	0	0	0	0
31	31	Mar 12, 20...		ЦСКА в бол...	https://www...	Футбол - Ку...	0	0	0	0	0	0
101	101	Mar 12, 20...		«Спартак» ...	https://www...	Футбол - Ку...	0	0	0	0	0	0
223	223	Mar 12, 20...		Медведев о...	https://www...	Теннис - ATP	0	0	0	0	0	0
241	241	Mar 12, 20...		Филс обыгр...	https://www...	Теннис - ATP	0	0	0	0	0	0

Построение визуализаций

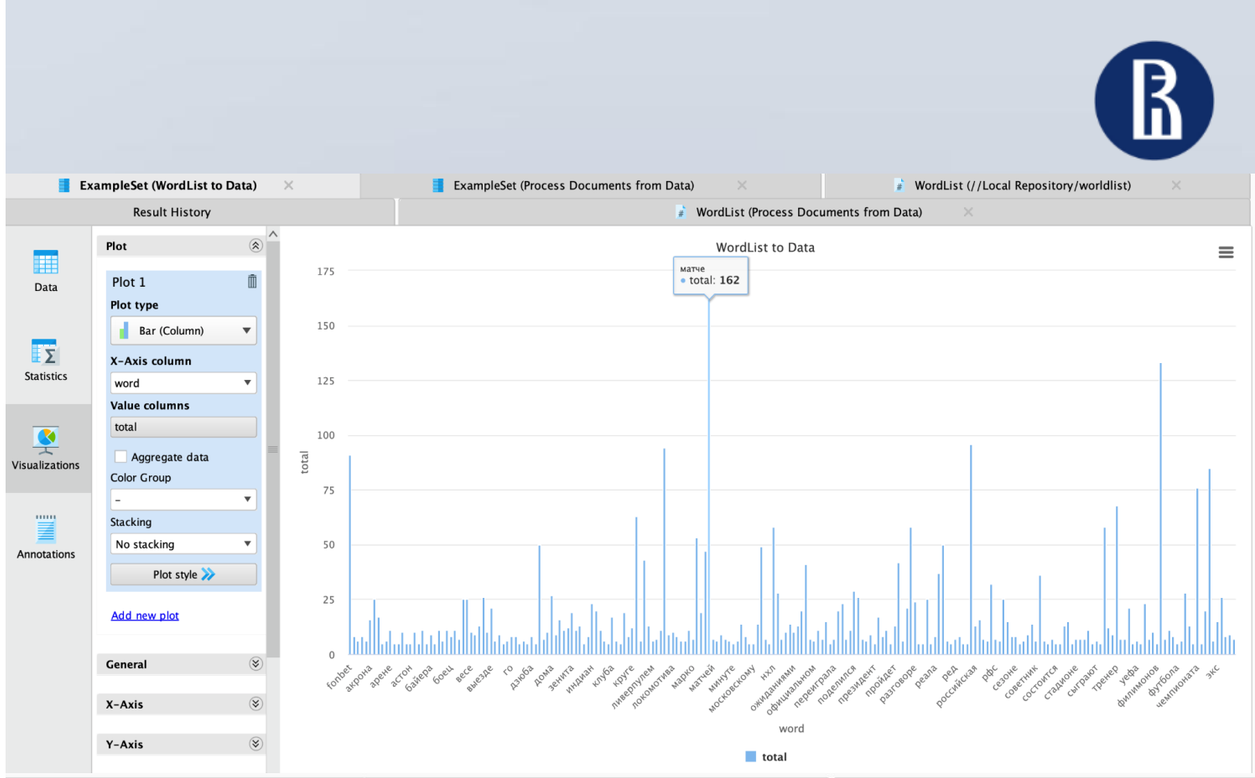


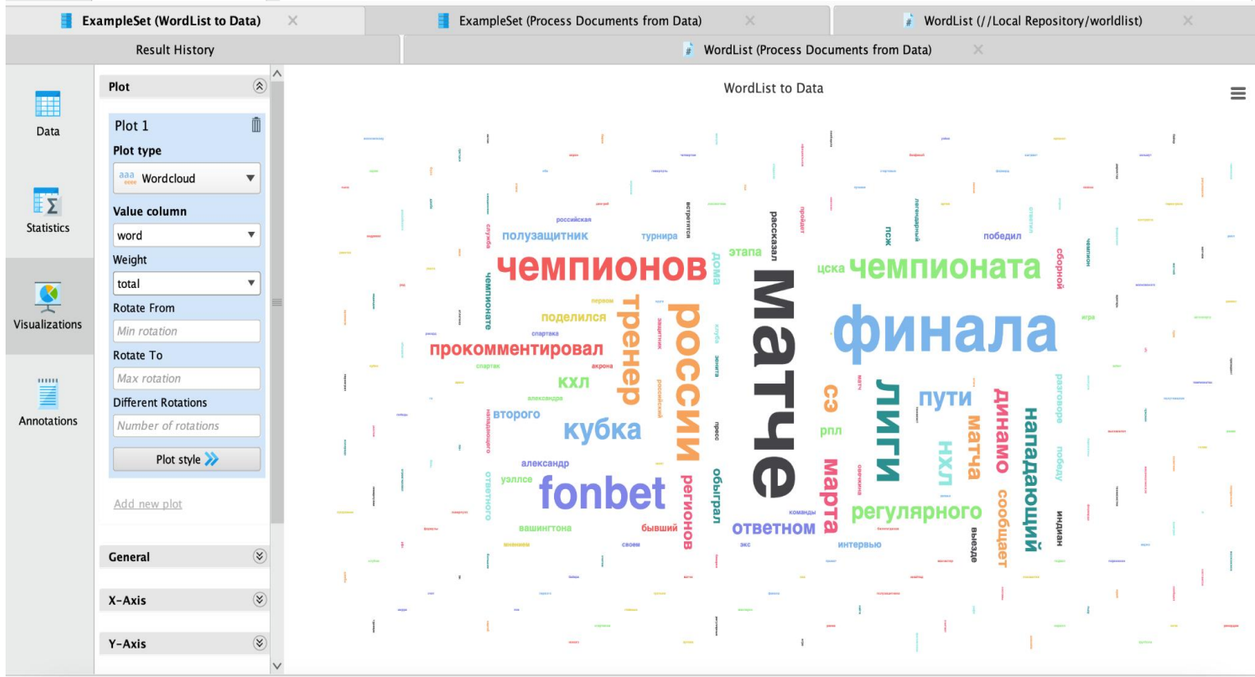
RapidMiner предоставляет инструменты для визуализации собранных данных. Гистограммы, диаграммы рассеивания и столбчатые графики позволяют быстро оценить распределение и структуру информации.

Bar Chart и Word Cloud

На вкладке Charts можно настроить Bar Chart (слово – ось X, частота – ось Y). Самые популярные слова будут на вершине столбцов.

Word Cloud — размер шрифта слова зависит от его частоты.

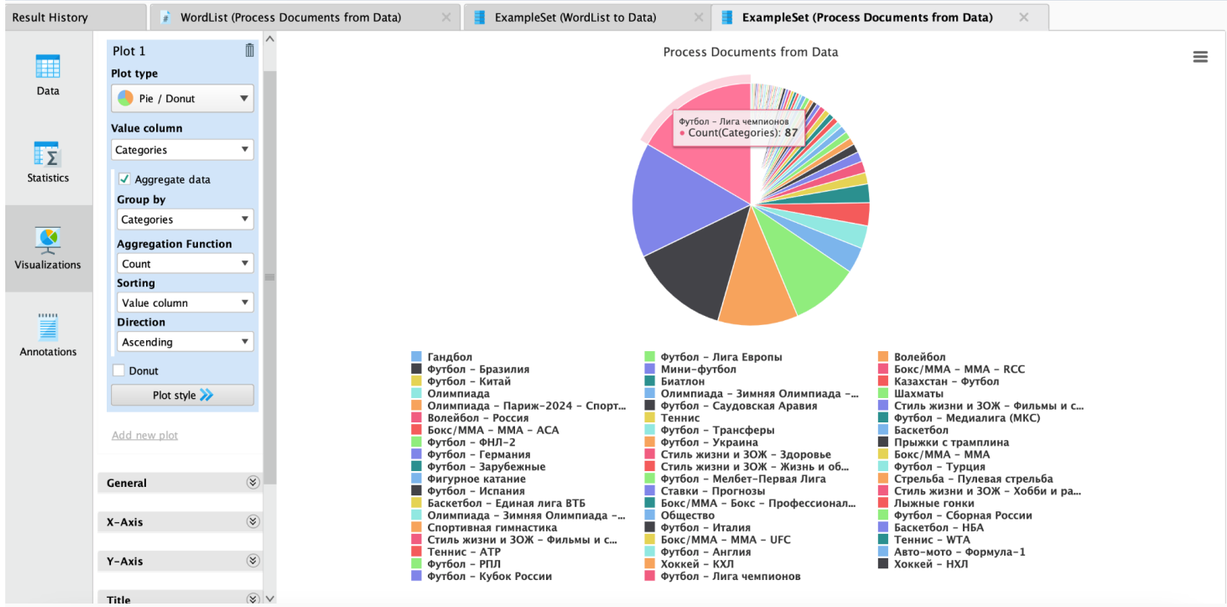




Pie Chart и Line Chart

Если при парсинге удалось получить категорию каждой новости, то можно проанализировать, какие направления спорта преобладают.

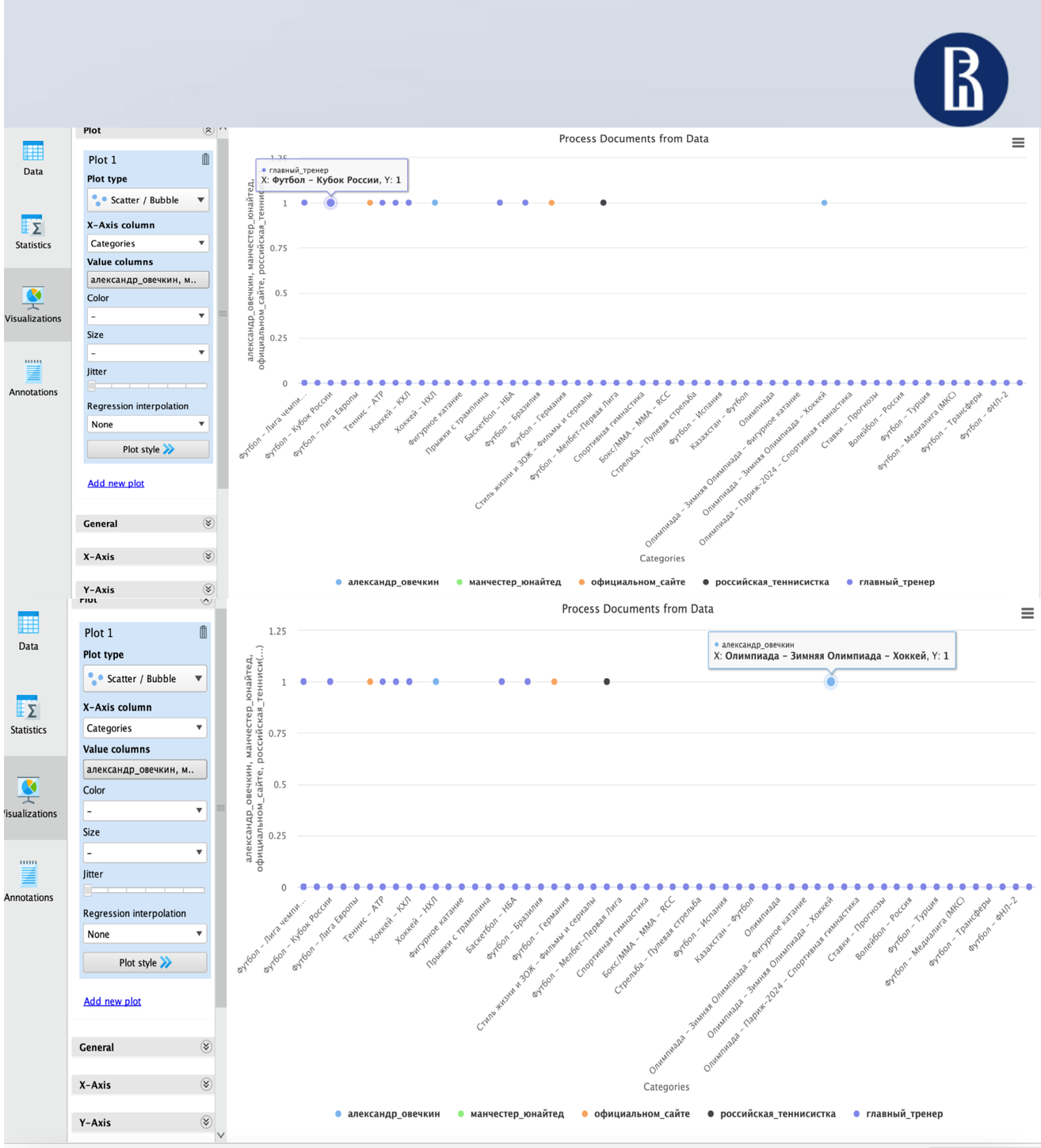
Pie Chart помогает понять процентное соотношение. Line Chart подойдет для просмотра динамики во времени.



Диаграммы рассеяния

Scatter/Bubble можно использовать, когда требуется сравнить несколько категорий и упоминаемость набора биграмм.

Подобная диаграмма иллюстрирует в каких рубриках чаще всего всплывают конкретные фразы.



Интерпретация визуализаций

Графики помогают выявить тенденции, аномалии или группировки в собранных данных, облегчая принятие решений и ускоряя процесс анализа больших массивов информации.

Особенности анализа текстовых данных

Текстовые данные требуют специальной обработки: токенизация, удаление стоп-слов, подсчёт частот слов, лемматизация. Это улучшает точность анализа и интерпретации.

Интеграция RapidMiner с другими сервисами

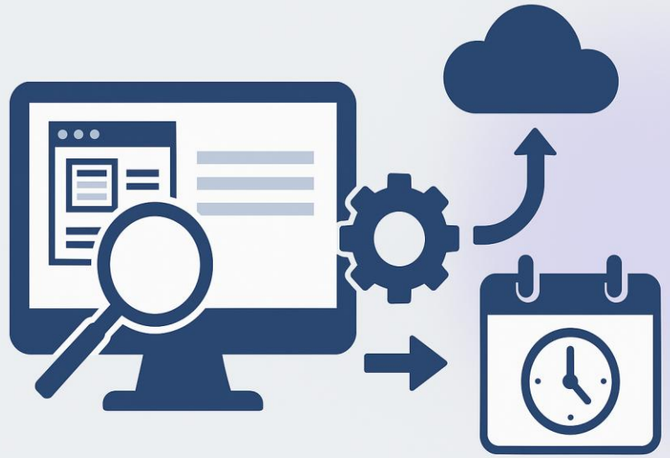
RapidMiner поддерживает интеграцию с внешними API (Google, Twitter, OpenAI), что позволяет расширять возможности анализа и обработки данных с веб-ресурсов в режиме реального времени.

Хранение извлечённых данных

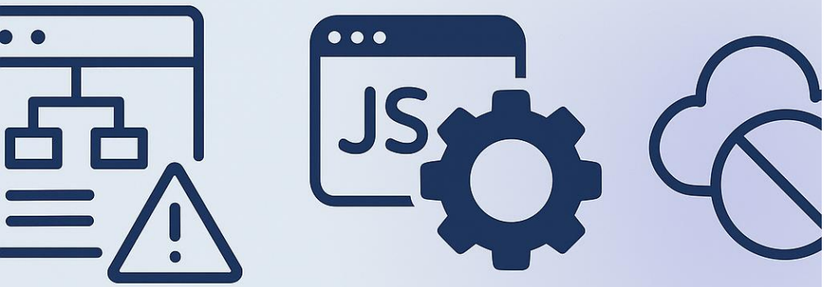
Данные после сбора сохраняются в табличных форматах CSV, Excel или базы данных. Это упрощает дальнейшее использование и позволяет интегрировать результаты анализа с другими программами.

Автоматизация процессов Web Scraping

RapidMiner Server позволяет автоматизировать регулярные процессы веб-скрейпинга и анализа данных, выполняя задачи по заданному расписанию, что экономит ресурсы и время.



Проблемы и ограничения Web Scraping

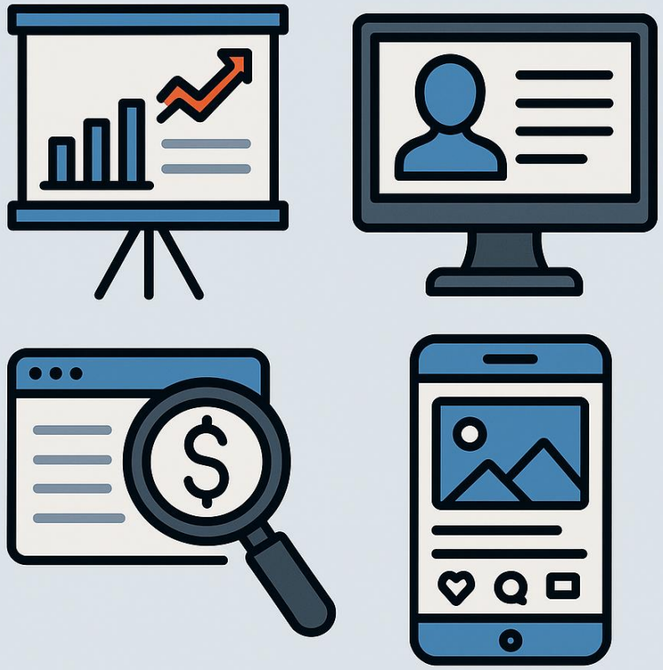


- Частые изменения структуры сайтов.
- Использование JavaScript, затрудняющее извлечение.
- Блокировка IP-адресов при частых запросах к серверу.

Решение сложностей Web Scraping

Использование прокси-серверов, сервисов для обработки JavaScript (например, Selenium) и регулярный мониторинг изменений на сайте помогают эффективно справляться с трудностями веб-скрейпинга.

Примеры успешного применения Web Scraping



Web Scraping используется в маркетинговых исследованиях, мониторинге отзывов клиентов, анализе цен конкурентов и прогнозировании трендов на основе данных из социальных сетей.

Заключение

RapidMiner — мощный инструмент для Web Scraping и анализа данных, не требующий написания кода. Грамотное применение методов извлечения и анализа веб-информации значительно повышает эффективность аналитических проектов.