Правительство Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
(НИУ ВШЭ)

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

ПРАКТИЧЕСКАЯ РАБОТА № 9

ТЕМА РАБОТЫ
«Web Scraping и анализ данных в RapidMiner»

Москва, 2025

Цель работы

Изучить методы сбора данных с веб-сайтов (web scraping) и последующего анализа текста с помощью платформы Altair AI Studio (RapidMiner). В ходе работы студенты:

- Познакомятся с инструментами RapidMiner для извлечения данных из веб-источников (например, оператором Read RSS Feed).
- Освоят этапы предобработки текстовых данных (очистка от HTML-разметки, приведение к нужному формату).
- Выполнят частотный анализ текста, выявят наиболее часто встречающиеся слова и темы (тренды) в данных.
- Научатся визуализировать результаты текстового анализа для интерпретации.

Целевая аудитория

Работа предназначена для студентов, начинающих изучение анализа данных и интересующихся практическими навыками извлечения информации из интернет-ресурсов. Она будет полезна всем, кто хочет научиться собирать данные с сайтов и проводить простой анализ текстовой информации без программирования, используя визуальную среду Altair AI Studio (RapidMiner version 2024.0).

Идея и концепция

В основе практической работы лежит изучение процесса получения данных из веб-среды и их анализ на примере реальных новостных данных. Студенты будут работать с актуальным русскоязычным веб-источником – например, новостной RSS-лентой сайта «Спорт-Экспресс», содержащей заголовки и краткое содержание последних новостей. Каждый студент самостоятельно:

- Извлечет данные с указанного веб-сайта с помощью RapidMiner (запрос RSS-ленты новостей).
- Преобразует неструктурированные веб-данные (текст новостей) в структурированный формат для анализа (очистка от лишних тегов, разделение текста на слова и т.д.).
- Проведет частотный анализ слов: определит, какие слова встречаются чаще всего, и сделает выводы о наиболее обсуждаемых темах (трендах) новостей на текущий момент.
- Построит визуализацию (например, диаграмму частот) для наглядного представления результатов и сформулирует выводы по итогам анализа.

Содержание практической работы

О наборе данных и задаче работы

Набор данных: RSS-лента новостного сайта (например, «Спорт-Экспресс» [https://www.sport-express.ru/services/materials/news/se/) – содержит список последних новостей с полями: дата публикации, заголовок, ссылка, содержание. В качестве содержания обычно приводится краткий анонс новости. RSS-лента обеспечивает доступ к данным в формате XML, удобном для машинного считывания.

Задача: Собрать данные новостной ленты и проанализировать текст новостей, чтобы определить наиболее часто упоминаемые слова и темы. Это позволит выявить, какие темы находятся в центре внимания в выбранный период. Необходимо продемонстрировать процесс веб-скрейпинга, очистки текста, расчет частот слов и визуализацию результатов в RapidMiner (Altair AI Studio).

Работа с данными

Получение данных (Web Scraping)

1) Чтение RSS-ленты:

- В панели Operators найдите оператор Read RSS Feed (расширение Web Mining). Перетащите его на рабочее поле процесса. В параметрах оператора укажите URL RSS-ленты сайта. Например, для sport-express.ru введите:
- url = https://www.sport-express.ru/services/materials/news/se/
- Оставьте остальные параметры по умолчанию. Этот оператор скачает данные RSS и выдаст таблицу с полями новости.

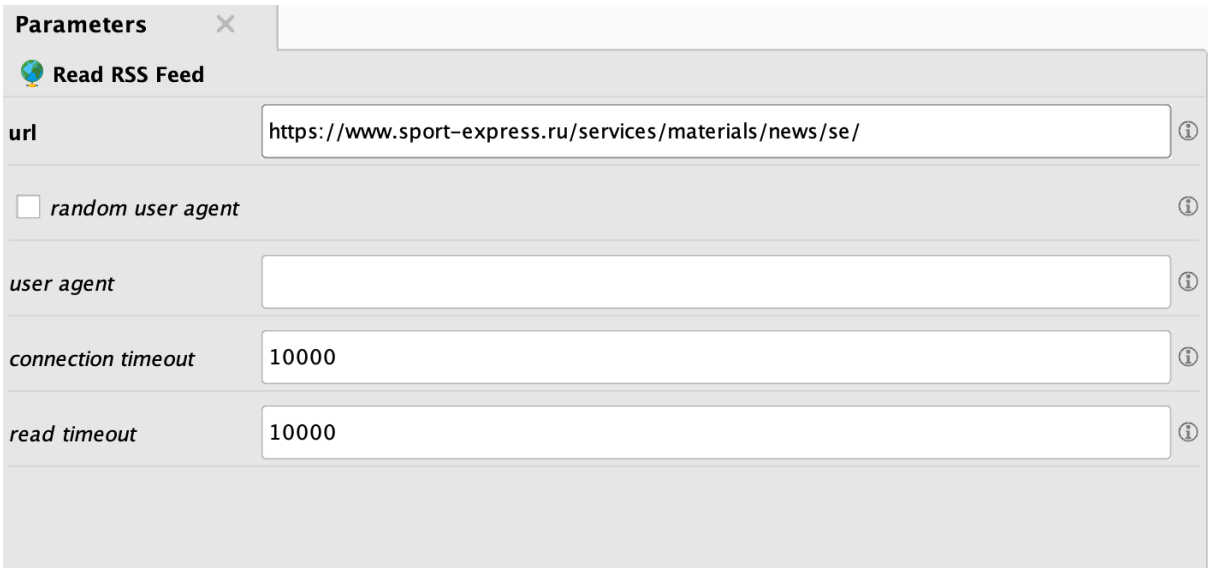


рис.1: Настройки параметров оператора Read RSS Feed

2) Просмотр загруженных данных:

- Подключите выходной порт оператора Read RSS Feed к результирующему порту процесса (или нажмите правой кнопкой на выходе и выберите Show Data).

- Убедитесь, что данные загрузились корректно: вы должны увидеть столбцы, такие как title (заголовок), published (дата), link, content, categories. Записи в таблице соответствуют новостным сообщениям из RSS.

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)

рис.2: Просмотр извлеченных с веб-страницы данных

Предобработка и структурирование данных

1) Настройка обработки текста:

- Для анализа текста новостей используйте оператор Process Documents from Data (находится в разделе Text Processing). Перенесите его на рабочее поле и подключите так, чтобы на его вход поступали данные из Read RSS Feed. В настройках Process Documents from Data выберите атрибут, содержащий текст новости – например, content. Укажите, чтобы на выходе сформировался список слов (WordList), частотная таблица (Term Frequency) или количество использования слова (Term Occurences).

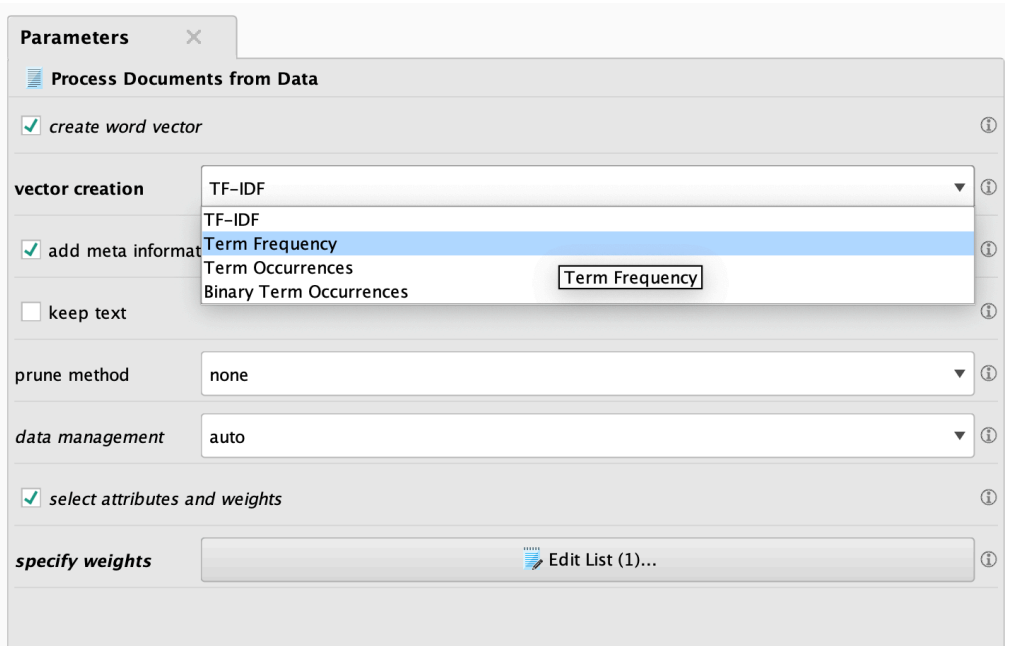


рис.3: Настройки параметров оператора Process Documents from Data

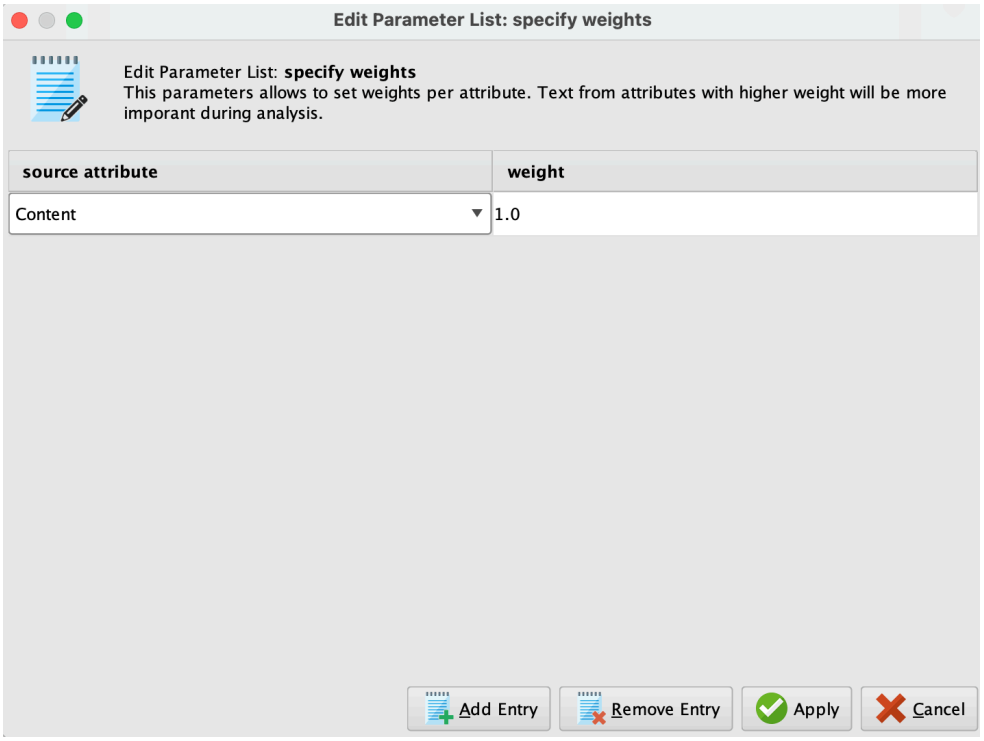


рис.4: Настройка specify weights внутри оператора Process Documents from Data

2) Конфигурация обработки текста (внутри субпроцесса). Двойным щелчком откройте оператор Process Documents from Data. Внутри него добавьте необходимые шаги обработки текста:

- Tokenize: разбивает текст на токены (слова). Добавьте оператор Tokenize и укажите разделение по пробелам и знакам препинания (настройка mode – non-letters, чтобы отделять слова, игнорируя знаки).

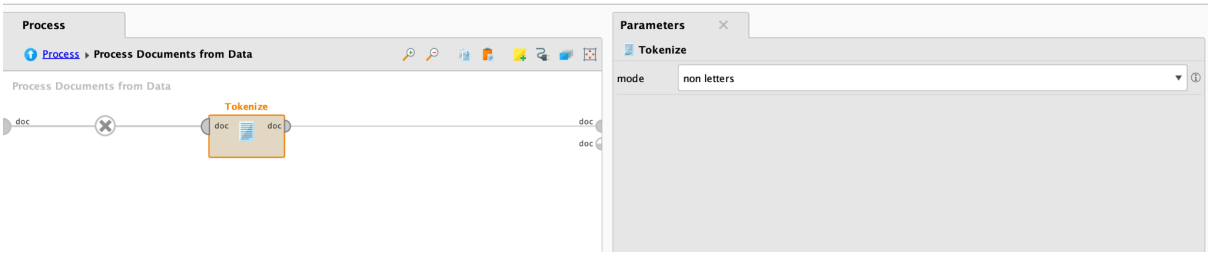


рис.5: Настройка субпроцесса Tokenize (токенизация) внутри оператора Process Documents from Data

- Transform Cases: приведите все слова к нижнему регистру (добавьте оператор Transform Cases, параметр to lower case = true). Это позволит считать "Футбол" и "футбол" одним словом.

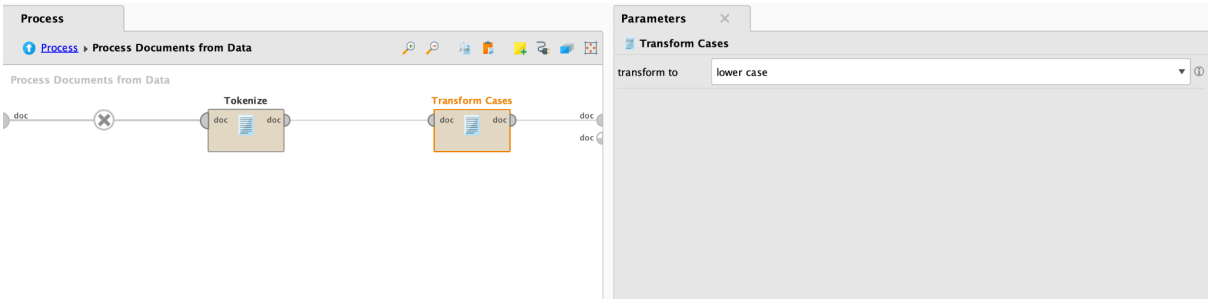


рис.6: Настройка субпроцесса Transform Cases (приведение к регистру) внутри оператора Process Documents from Data

- Оператор Generate n-Grams (Tokens) для создания пар слов: параметр max length = 2 (для создания биграмм).

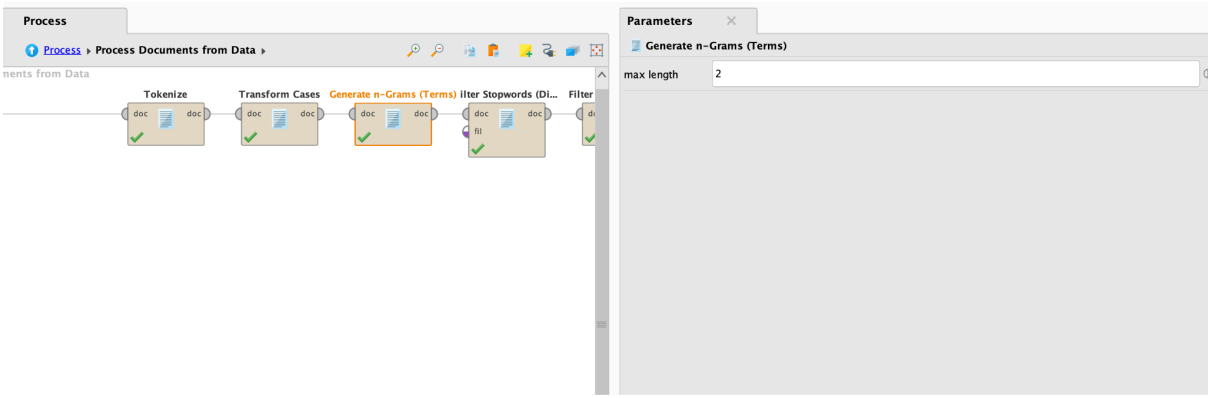


рис. 7: Настройка субпроцесса Generate n-Grams (создание комбинаций n-слов) внутри оператора Process Documents from Data

- Удаление стоп-слов: также можно удалить стоп-слова – часто встречающиеся служебные слова, не несущие смысл (союзы, предлоги и т.п.). Добавьте Filter Stopwords. В выпадающем списке языков встроенных слов отсутствует русский язык, поэтому в рамках данной работы мы будем использовать “собственный” список

русских стоп-слов. (Загрузим его по ссылке: https://github.com/stopwords-iso/stopwords-ru/blob/master/stopwords-ru.txt и подключим через оператор Filter Stopwords (Dictionary).

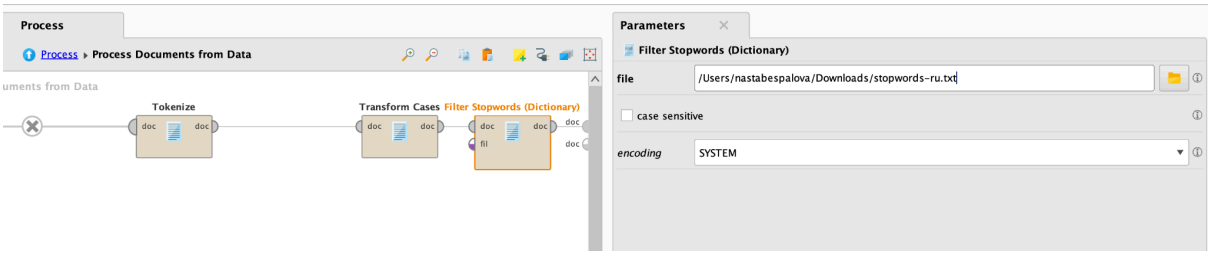


рис. 8: Настройка субпроцесса Filter Stopwords (удаление стоп-слов) внутри оператора Process Documents from Data

- Filter Tokens by Length: можно отбросить слишком короткие токены, например, длиной в 1 символ (это могут быть знаки или отдельные буквы). Добавьте Filter Tokens by Length и установите min chars = 2.

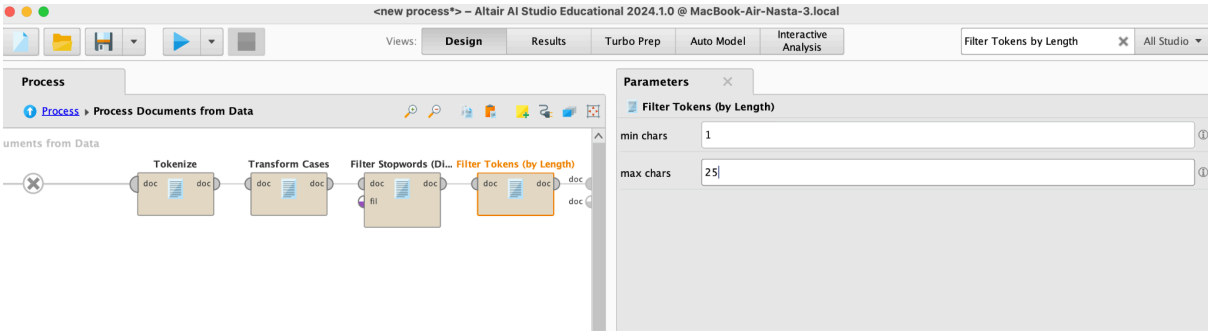
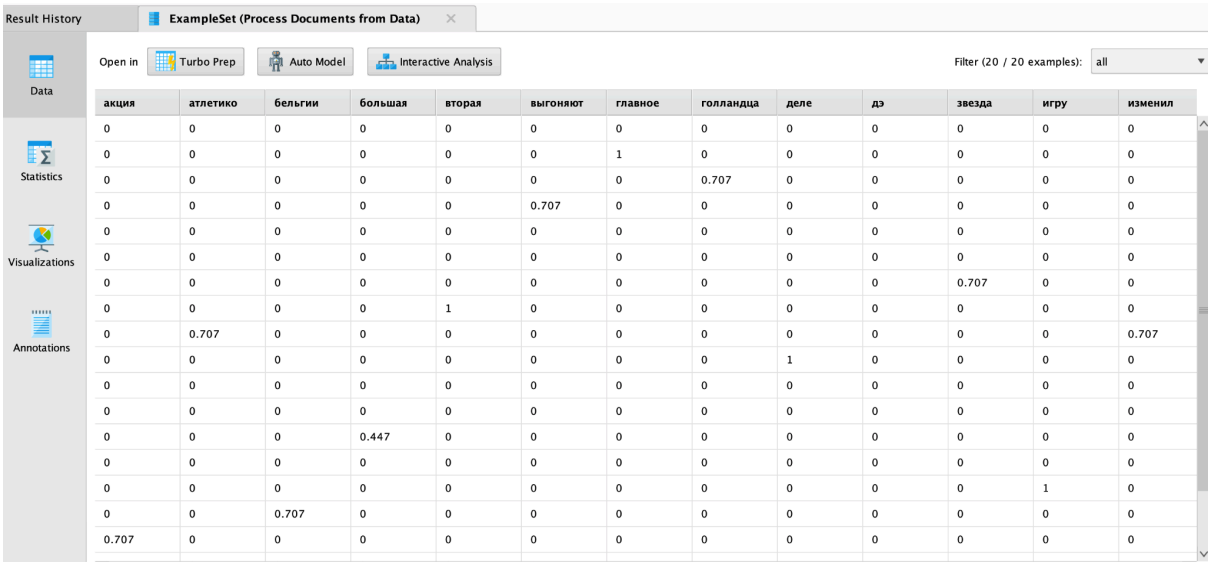


рис. 9: Настройка субпроцесса Filter Tokens by Length (удаление коротких токенов) внутри оператора Process Documents from Data

- После настройки последовательности операторов внутри Process Documents, вернитесь в основной процесс (кнопка Return to parent).

Вывод результатов обработки текста

У оператора Process Documents from Data имеется несколько выходов. Подключите соответствующий выход к result-порту процесса. Это позволит получить на выходе таблицу со списком слов и их частотами. Теперь выполните процесс (кнопка Run).



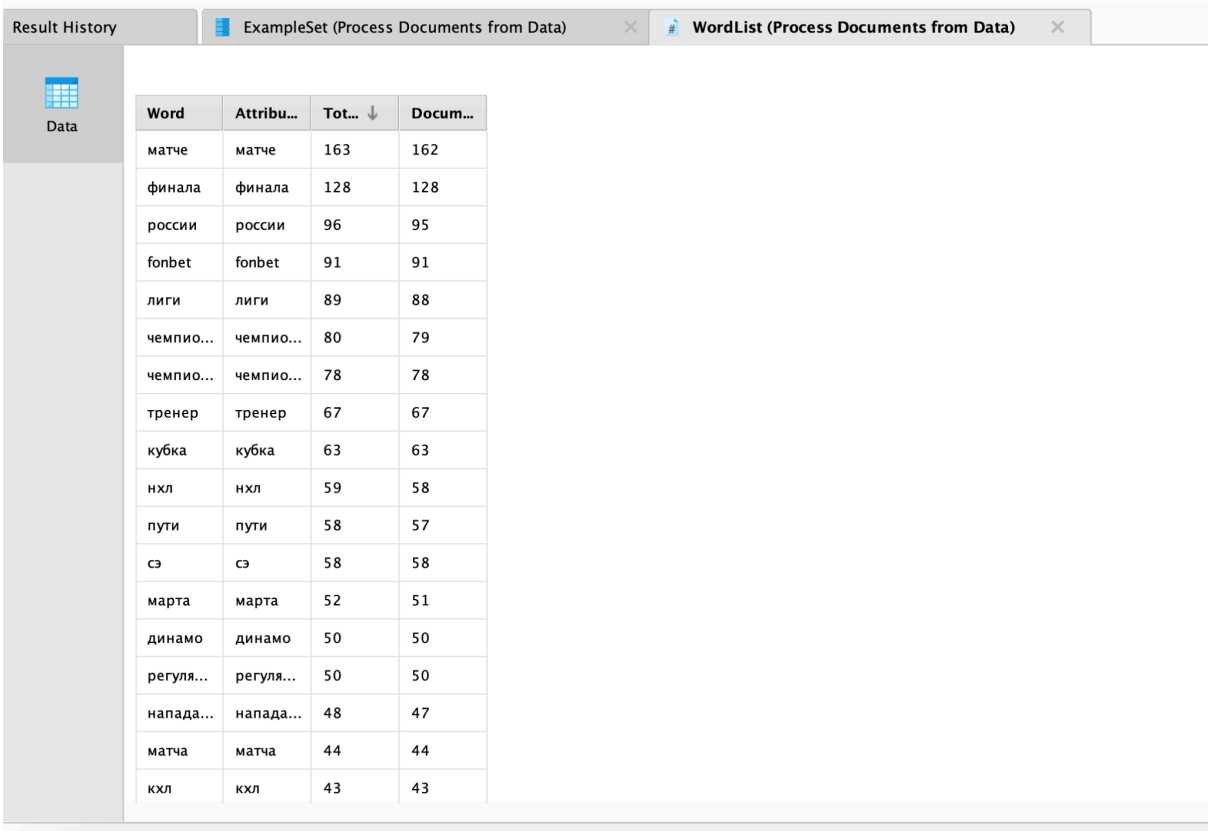
акция	атлетико	бельгии	большая	вторая	выгоняют	главное	голландца	деле	дэ	звезда	игру	изменил
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0.707	0	0	0	0	0
0	0	0	0	0	0.707	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0.707	0	0
0	0	0	0	1	0	0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

рис. 10: Просмотр данных после выполнения процесса с включением оператора Process Documents from Data

Анализ данных (частотный анализ, тренды)

Изучение частот слов

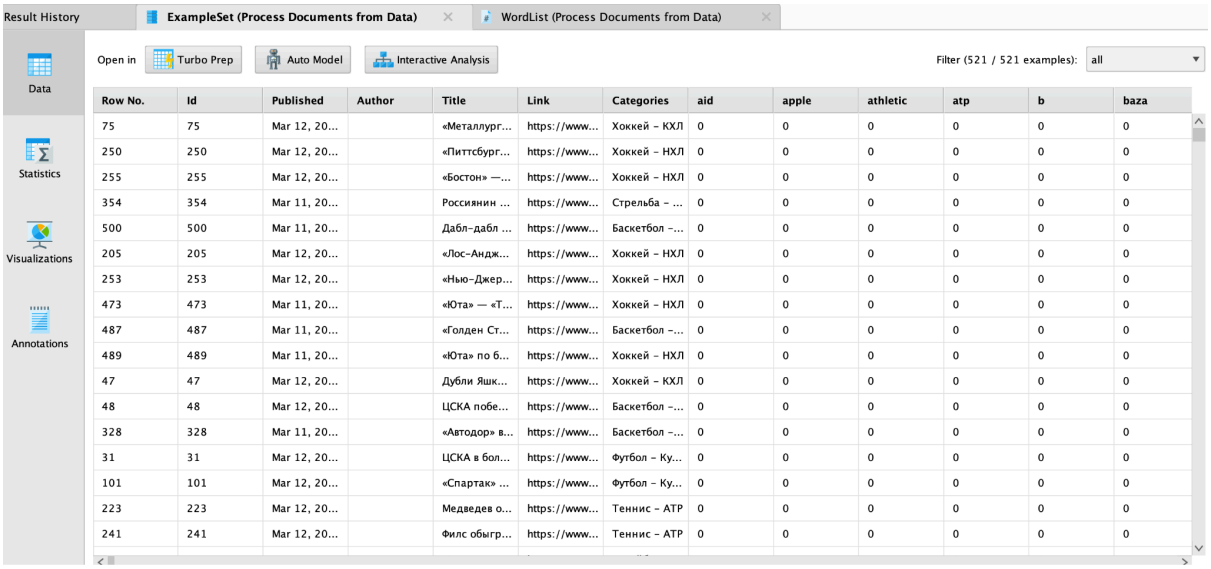
Просмотрите полученный список слов в результатах (Word List). Для каждого уникального слова вы видите количество его упоминаний (term frequency). Отсортируйте таблицу по убыванию частоты (щелкните на заголовок столбца с частотой). Выявите топ-10 самых частотных слов. Например, учитывая спортивную направленность сайта, слова, связанный с термином “матч” оказалось самым частым.



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

рис. 11: Просмотр созданного листа слов после выполнения процесса с включением оператора Process Documents from Data

Далее можно перейти во вкладку ExampleSet (Process Document from Data) – здесь вы увидите исходную таблицу, составленную в результате получения информации с сайта, дополненную колонками с распознанными словами. Например, найдем слово “победил” (при выборе Term Frequency для параметра Vector Creation в Process Document from Data).



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

рис. 12: Таблица, составленная в результате получения информации с сайта, дополненная колонками с распознанными словами

период	победил ↓	периоде
0	0.354	0
0	0.354	0
0	0.354	0
0	0.354	0
0	0.354	0
0	0.333	0
0	0.333	0
0	0.333	0
0	0.333	0
0	0.333	0
0	0.316	0
0	0.316	0
0	0.316	0
0	0.289	0
0	0.277	0
0	0.277	0
0	0.277	0

рис. 13: Распределение частоты упоминания слова "победил" в различных статьях, извлеченных с веб-ресурса, сортировка – по убыванию

Визуализация и анализ результатов

1) Столбчатая диаграмма (Bar Chart) – частота слов:
    - Цель: показать, какие слова встречаются чаще всего в заголовках/анонсе новостей.
    - Как построить: Вкладка Charts → Bar Chart, по оси X – слова, по оси Y – их частота.
    - Выводы: позволяет увидеть самые популярные слова в новостях (например, слово «матч» оказалось самым часто используемым).

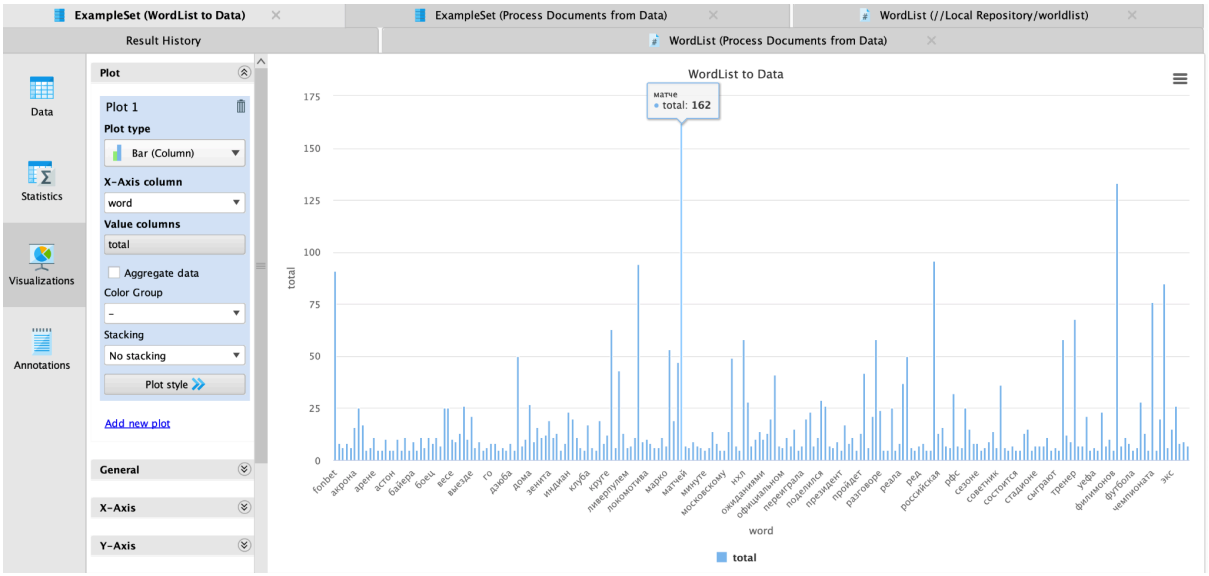


рис. 14: Столбчатая диаграмма, показывающая частоту упоминания различных слов

2) Облако слов (Word Cloud):
   - Цель: визуально представить наиболее часто встречающиеся слова в текстах новостей.-
   - Как построить: Word Cloud → Входные данные – список слов и их частоты.
   - Выводы: Более наглядное представление популярных тем (чем крупнее слово, тем оно более частотное).Например, в нашем случае крупными буквами выделены «Матче», «Финала», «Чемпионата», «Кубка» – значит, эти темы особенно актуальны.
   -

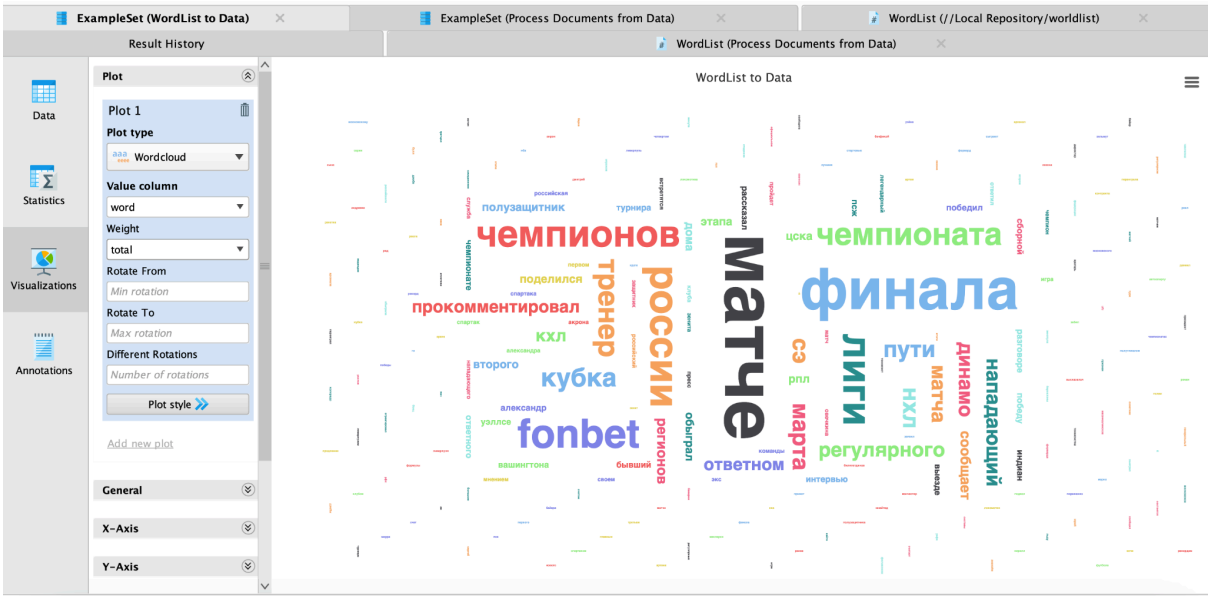


рис. 15: Облако слов, показывающее наиболее частые слова в собранных статьях

3) Круговая диаграмма (Pie Chart) – распределение категорий новостей:
   - Цель: показать процентное соотношение различных категорий спортивных новостей.
   - Как построить: Вкладка Charts → Pie Chart, по оси Категории – типы новостей (categories) → Aggregate Data, Aggregation Function – Count (количество новостей в каждой категории).
   - Выводы: Помогает понять, какие виды спорта наиболее освещаются в новостях. Например, в данном примере доминируют статьи, посвященные Лиге Чемпионов и Кубку России по футболу.

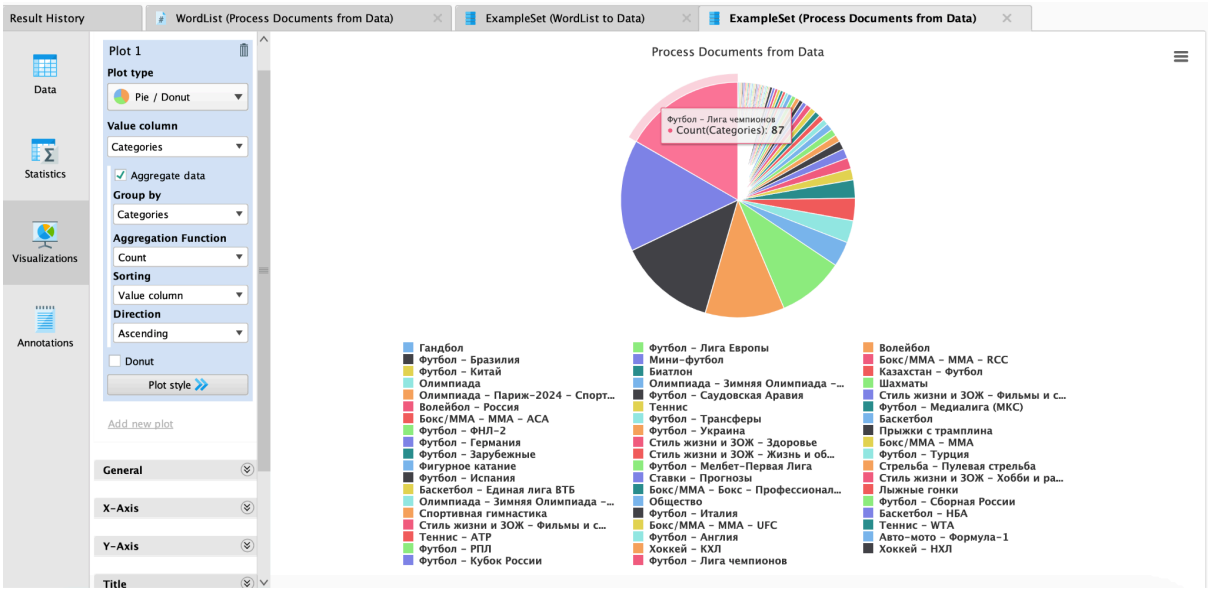


рис. 16: Круговая диаграмма, показывающая процентное соотношение

различных категорий спортивных новостей

4) Аналогичную информацию можно получить с помощью линейного графика (Line Chart):

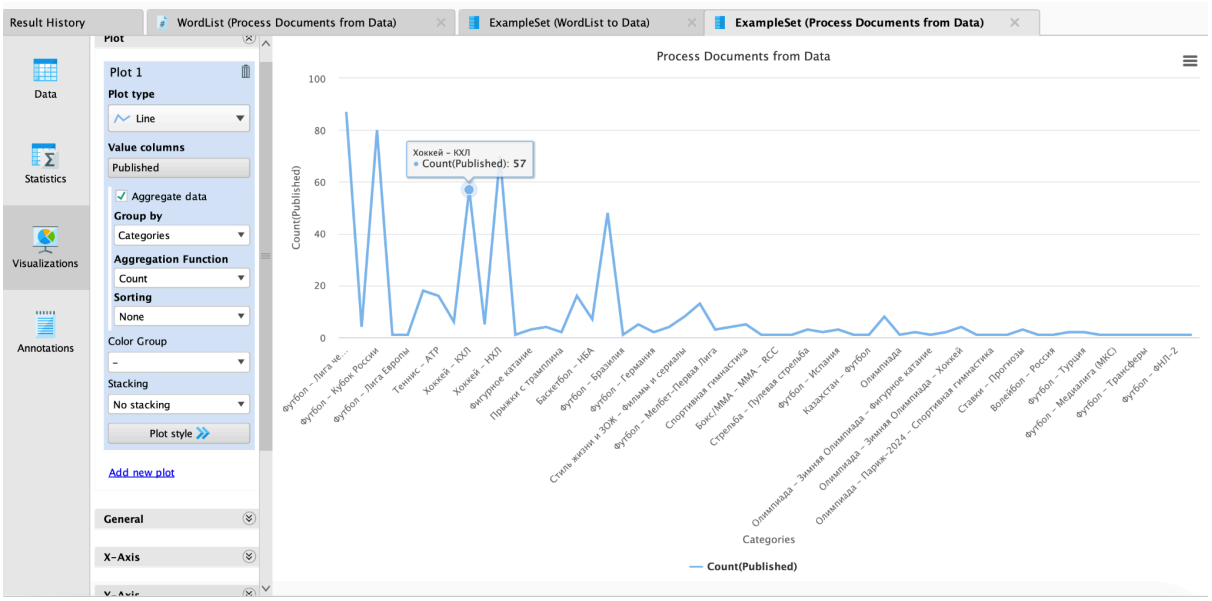


рис. 17: Линейный график, показывающий количество статей в различных категориях спортивных новостей

5) Диаграмма рассеяния (Scatter/Bubble): ранее мы создавали пары слов (биграммы). Теперь предлагается посмотреть, в каких категориях эти пары упоминаются.
   - Как построить: Вкладка Charts → Scatter/Bubble, по оси X – типы новостей (categories), Value Columns – выбираем интересующие биграммы.
   - Выводы: Помогает понять, в каких новостных разделах упомянуты пары слов. Например, выберем связки «александр_овечкин», «главный_тренер», «манчестер_юнайтед», «российская_теннисистка», «официальном_сайте». Так, «главный_тренер» упоминается в разделах: Футбол – Лига Чемпионов, Футбол – Кубок России, Футбол – Италия, Теннис – АТР, Хоккей – КХЛ, Футбол – Англия. Александ Овечкин упомянут: Хоккей – НХЛ, Олимпиада – Зимняя Олимпиада – Хоккей.

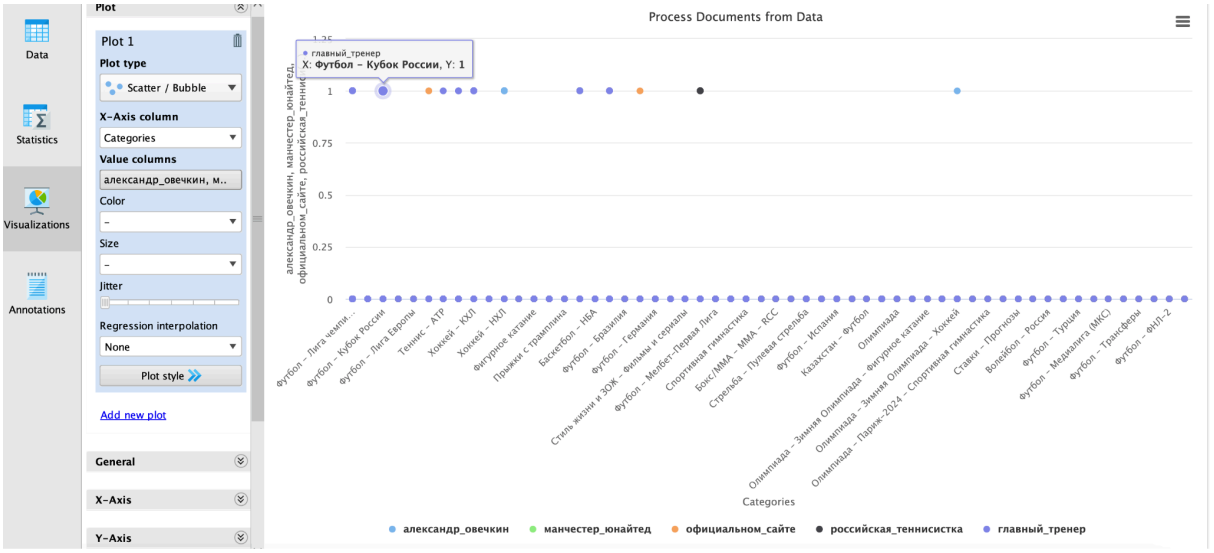


рис. 18: Диаграмма рассеяния, показывающая упоминания пар слов в статьях различных категорий (главный тренер: Футбол – Кубок России)

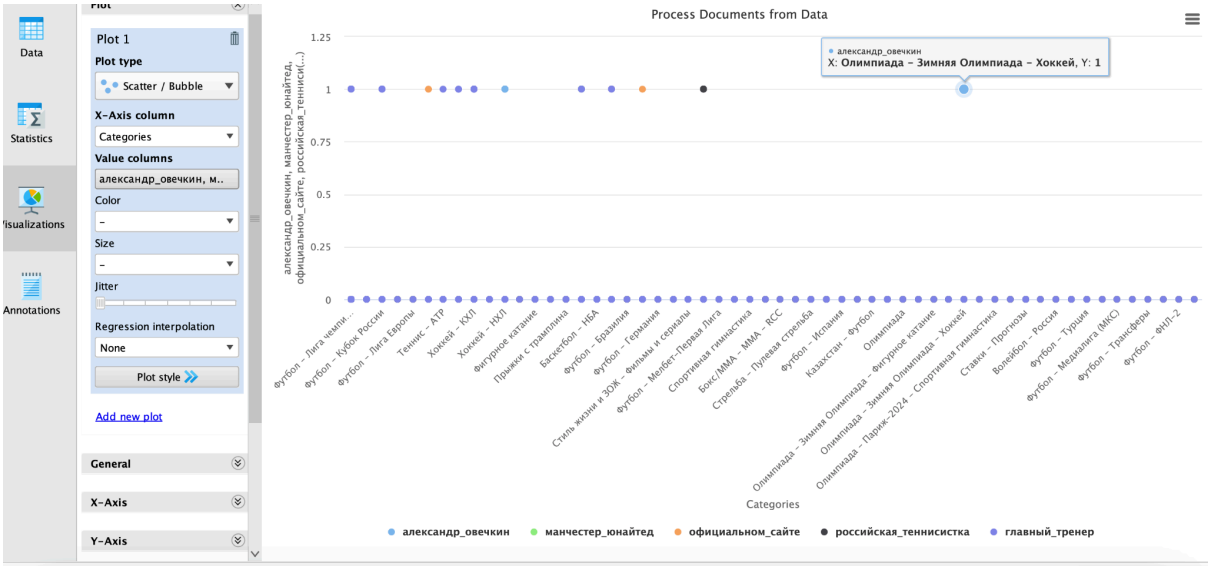


рис. 19: Диаграмма рассеяния, показывающая упоминания пар слов в статьях различных категорий (Александр Овечкин: Олимпиада – Зимняя Олимпиада – Хоккей)

Пример выполненной работы

- Шаг 1: С помощью оператора Read RSS Feed загружены данные RSS-ленты новостей. На выходе получена таблица с последними новостями, содержащая столбцы даты, заголовка, ссылки и текста

анонса. Данные проверены на корректность отображения в RapidMiner.
- Шаг 2: Выполнена предобработка текста новостей. С помощью оператора Process Documents from Data текст всех анонсов разбит на отдельные слова (Tokenize), приведён к нижнему регистру и частично очищен от стоп-слов (например, удалены «в», «и», «на» и др.). Получен список уникальных слов с подсчетом общего количества вхождений каждого слова в новости.
- Шаг 3: Проведен частотный анализ слов. Выявлен список наиболее часто встречающихся слов.
- Шаг 4: Построены визуализации различных типов, позволяющие интерпретировать результаты. На них наглядно показано, что некоторые существенно опережают остальные по частоте.
- Шаг 5: Сделаны выводы по результатам анализа. Определены основные темы дня в новостной ленте. Студенты сформулировали отчет, описав процесс получения и обработки данных, а также перечислив выявленные тренды в новостях.

Приобретенные навыки

В результате выполнения данной лабораторной работы студенты:

- Научились использовать инструменты Altair AI Studio (RapidMiner) для веб-скрейпинга: освоили загрузку данных с веб-сайтов через RSS.
- Получили опыт обработки неструктурированного текстового контента: очистка текста, разбиение на токены, удаление лишних слов.
- Освоили базовые методы текстовой аналитики (частотный анализ слов) для выявления ключевой информации и трендов в текстовых данных.
- На практике применили навыки визуализации результатов анализа, научились интерпретировать графики распределения частот.
- Расширили понимание процесса анализа данных: от добычи сырых данных в интернете до получения осмысленных выводов.

Обобщенная задача для выполнения индивидуального варианта

Разработать проект, направленный на сбор и анализ текстовых данных из веб-источников. В рамках задания необходимо:

Сбор данных:

- Настроить процесс веб-скрейпинга для автоматизированного извлечения данных с выбранного интернет-ресурса (RSS-лента новостного сайта, блогов или форумов).
- Получить данные в структурированном формате, включающем такие поля, как заголовок, дата публикации, ссылка и содержание.

Предобработка и структурирование текста:

- Очистить полученные данные от HTML-разметки и лишних символов.
- Выполнить базовую обработку текста: токенизацию, приведение к нижнему регистру, удаление стоп-слов и, при необходимости, лемматизацию или стемминг.
- Преобразовать текст в числовые признаки с помощью методов векторизации.

Анализ данных:

- Провести частотный анализ терминов для выявления наиболее часто встречающихся слов и биграмм, позволяющих определить актуальные темы и тренды в новостной ленте.
- Оценить распределение категорий новостей, если данные содержат тематические метки или теги.

Визуализация и интерпретация результатов:

- Построить различные виды графиков для наглядного представления результатов анализа: столбчатые диаграммы частот, облака слов, круговые диаграммы для распределения категорий, линейные графики для отображения трендов во времени.
- Сделать выводы о доминирующих темах и трендах, а также проанализировать полученные результаты в контексте специфики выбранного ресурса.

Распределение вариантов


RAPIDMINER