# *Руководство пользователя IBM SPSS Modeler Text Analytics 18.0*

**Примечание**

Прежде чем использовать эту информацию и продукт, описанный в ней, прочтите сведения в разделе “Уведомления” на стр. 237.

**Информация о продукте**

Это издание применимо к версии 18, выпуск 0, модификация 0 IBM SPSS Modeler Text Analytics и ко всем последующим версиям и модификациям до тех пор, пока в новых изданиях не будет указано иное.

# Содержание

# Предисловие

IBM® SPSS Modeler Text Analytics предлагает мощные возможности аналитики текстовых данных, использующие расширенные лингвистические технологии и возможность обработки естественных языков (Natural Language Processing, NLP) для быстрой обработки самых разнообразных неструктурированных текстовых данных и извлечения и организации на их основе ключевых понятий. К тому же, IBM SPSS Modeler Text Analytics может сгруппировать эти понятия в категории.

Около 80% поддерживаемых в организации данных хранятся в виде текстовых документов, например, отчетов, веб-страниц, сообщений электронной почты и замечаний центра обработки вызовов. Текстовые данные - это ключевой фактор, позволяющий организации лучше понять поведение заказчиков ее продукции. Система с интегрированной в нее обработкой естественного языка (NLP) может аналитически извлекать понятия, включая сложные синтагмы. Более того, знание базового языка позволяет классифицировать синтаксические термы по смыслу и контексту, объединяя их в соответствующие группы, например, продуктов, организаций или людей. Благодаря этому, можно быстро определить значимость информации для ваших потребностей. Извлекаемые понятия и категории можно сочетать с существующими структурированными данными, такими как демографические, и применять к моделированию в имеющемся в IBM SPSS Modeler полном комплекте инструментов исследования данных для получения более качественных и специализированных решений.

Лингвистические системы восприимчивы к знаниям: чем больше информации в их словарях, тем выше качество результатов. IBM SPSS Modeler Text Analytics поставляется с набором лингвистических ресурсов, таких как словари терминов и синонимов, библиотеки и шаблоны. Данный продукт позволяет дополнительно разрабатывать и настраивать эти лингвистические ресурсы в соответствии с контекстом. Как правило, тонкая настройка лингвистических ресурсов представляет собой итеративный процесс и требуется для точного концептуального представления процессов получения и категоризации. В набор входят также пользовательские шаблоны, библиотеки и словари для конкретных областей знания, таких CRM и геномика.

## О бизнес аналитике IBM

Программное обеспечение IBM для бизнес аналитики предоставляет полную, последовательную и точную информацию, которая повышает эффективность ведения бизнеса. Полный набор программного обеспечения для business intelligence, прогностической аналитики, управления финансовой эффективностью и стратегией и аналитических приложений позволяет ясно видеть текущую ситуацию, а также делать прогнозы, позволяющие предпринимать практические действия. В сочетании с решениями для конкретных отраслей, проверенной практикой и услугами бизнес аналитика IBM позволяет организациям любых размеров достигать наивысшей производительности, уверенно автоматизировать процессы принятия решений и добиться лучших результатов.

Как составная часть этого набора, программное обеспечение IBM SPSS Predictive Analytics помогает организациям предсказывать будущие события и предпринимать практические действия непосредственно на основе этих предсказаний. Коммерческие, правительственные и академические организации всего мира, полагаются на технологию IBM SPSS, обеспечивающую конкурентное преимущество в привлечении, удержании и повышении отдачи от клиентов. Включая программное обеспечение IBM SPSS в свои ежедневные операции, организации могут прогнозировать будущие события, направлять и автоматизировать решения для соответствия бизнес-целям и достигать ощутимых конкурентных преимуществ. Чтобы получить дальнейшую информацию или связаться с представителем, зайдите на http://www.ibm.com/spss.

## Техническая поддержка

Техническая поддержка предоставляется клиентам, оплачивающим обновительные взносы. Пользователи могут обращаться в службу технической поддержки, если у них возникают какие-либо проблемы с использованием или установкой программного обеспечения IBM Corp.. За технической поддержкой обращайтесь на сайт IBM Corp.: http://www.ibm.com/support. При обращении за поддержкой будьте готовы назвать себя и организацию, в которой вы работаете.

# Глава 1. О программе IBM SPSS Modeler Text Analytics

IBM SPSS Modeler Text Analytics предлагает мощные возможности аналитики текстовых данных, использующие расширенные лингвистические технологии и возможность обработки естественных языков (Natural Language Processing, NLP) для быстрой обработки самых разнообразных неструктурированных текстовых данных и извлечения и организации на их основе ключевых понятий. К тому же, IBM SPSS Modeler Text Analytics может сгруппировать эти понятия в категории.

Около 80% поддерживаемых в организации данных хранятся в виде текстовых документов, например, отчетов, веб-страниц, сообщений электронной почты и замечаний центра обработки вызовов. Текстовые данные - это ключевой фактор, позволяющий организации лучше понять поведение заказчиков ее продукции. Система с интегрированной в нее обработкой естественного языка (NLP) может аналитически извлекать понятия, включая сложные синтагмы. Более того, знание базового языка позволяет классифицировать синтаксические термы по смыслу и контексту, объединяя их в соответствующие группы, например, продуктов, организаций или людей. Благодаря этому, можно быстро определить значимость информации для ваших потребностей. Извлекаемые понятия и категории можно сочетать с существующими структурированными данными, такими как демографические, и применять к моделированию в имеющемся в IBM SPSS Modeler полном комплекте инструментов исследования данных для получения более качественных и специализированных решений.

Лингвистические системы восприимчивы к знаниям: чем больше информации в их словарях, тем выше качество результатов. IBM SPSS Modeler Text Analytics поставляется с набором лингвистических ресурсов, таких как словари терминов и синонимов, библиотеки и шаблоны. Данный продукт позволяет дополнительно разрабатывать и настраивать эти лингвистические ресурсы в соответствии с контекстом. Как правило, тонкая настройка лингвистических ресурсов представляет собой итеративный процесс и требуется для точного концептуального представления процессов получения и категоризации. В набор входят также пользовательские шаблоны, библиотеки и словари для конкретных областей знания, таких CRM и геномика.

**Внедрение.** Вы можете внедрить потоки исследования текстовых данных при помощи IBM SPSS Modeler Solution Publisher для скоринга неструктурированных данных в реальном времени. Возможность внедрить эти потоки обеспечивает успешные реализации замкнутого исследования текста. Например, сейчас в вашей организации может производиться анализ блокнотных записей о входящих и исходящих звонках с применением предсказательных моделей для повышения точности маркетинговых сообщений в реальном времени.

*Примечание*: Чтобы запустить IBM SPSS Modeler Text Analytics с IBM SPSS Modeler Solution Publisher, добавьте каталог `<каталог_установки>/ext/bin/spss.TMWBServer` в переменную среды `$LD_LIBRARY_PATH`.

**Автоматизированный перевод поддерживаемых языков.** В сочетании с программой SDL SaaS (Software as a Service - Программное обеспечение как сервис), IBM SPSS Modeler Text Analytics позволяет перевести текст с указанных в списке поддерживаемых языков (включая арабский, китайский и персидский) на английский. Затем можно выполнить тестовый анализ переведенного текста и внедрить полученные результаты для сотрудников, которым не удалось понять текстовое содержимое на исходных языках. Поскольку результаты исследования текстовых данных автоматически связываются обратными ссылками с соответствующими текстовыми данными на иностранном языке, организация сможет после этого сфокусировать наиболее востребованные ресурсы (то есть сотрудников, владеющих исходным языком документов) только на самых важных результатах анализа. SDL предлагает перевод языков при помощи статистических алгоритмов перевода, полученных за 20 человеко-лет расширенных исследований переводов.

## Переход к IBM SPSS Modeler Text Analytics версии 18

Обновление от уровня предыдущих версий PASW Text Analytics или Text Mining for Clementine.

Перед установкой IBM SPSS Modeler Text Analytics версии 18 нужно сохранить и экспортировать из текущей версии все API, шаблоны и библиотеки, которые вы хотите использовать в новой версии. Мы рекомендуем сохранить эти файлы в каталоге, который не будет ни удален, ни перезаписан при установке последней версии.

После установки последней версии IBM SPSS Modeler Text Analytics можно загрузить сохраненный файл TAP, добавить любые сохраненные библиотеки или импортировать и загрузить любые сохраненные шаблоны, чтобы использовать их в последней версии.

**Важное замечание:** Если вы деинсталлируете текущую версию, не сохраняя и не экспортируя предварительно нужные файлы, все TAP, шаблоны и результаты работы с общедоступными библиотеками из предыдущей версии будут потеряны и недоступны для использования в IBM SPSS Modeler Text Analytics версии 18.

# Об исследовании текста

Сегодня все большее количество информации хранится в неструктурированных и полуструктурированных форматах, например, сообщения электронной почты заказчиков, замечания центра обработки вызовов, свободные ответы в соцопросах, ленты новостей, веб-формы и так далее. Это обилие информации ставит многие организации перед непростым вопросом "Как нам собирать, анализировать и с выгодой использовать эту информацию?"

*Исследование текста* - это процесс исследования наборов текстовых материалов с целью фиксации ключевых понятий и тем и выявления скрытых взаимосвязей и тенденций, при котором не требуется знать точные слова и термины, использованные авторами для формулировки этих понятий. Исследование текста иногда путают с поиском информации, хотя эти понятия существенно различны. Хотя точное получение и хранение информации - огромная проблема, извлечение контента высокого качества и управление им, терминология и взаимосвязи, содержащиеся в информации, - это решающие и критически важные процессы.

Исследование текста и исследование данных

Для каждого отдельного текста лингвистическое исследование текста возвращает индекс понятий вместе с информацией об этих понятиях. Эта очищенная и структурированная информация может быть объединена с другими источниками данных для работы с такими вопросами, как:

- Какие понятия встречаются вместе?
- С чем еще они связаны?
- Какие категории более высокого уровня можно создать на основе извлеченной информации?
- Что предсказывают понятия или категории?
- Как понятия или категории предсказывают поведение?

Объединение исследования текста с исследованием данных дает лучшее понимание, чем можно получить, исходя только из структурированных или неструктурированных данных. Этот процесс обычно включает следующие шаги:

1. **Определите, какой текст будет исследоваться.** Подготовьте текст для исследования. Если текст содержится в нескольких файлах, сохраните эти файлы в одном месте. Для баз данных определите поле, содержащее текст.
2. **Исследуйте текст и извлеките структурированные данные.** Примените алгоритмы исследования текста к исходному тексту.
3. **Постройте модели понятий и категорий.** Определите ключевые понятия и/или создайте категории. Число понятий, возвращенных из неструктурированных данных, обычно очень велико. Определите лучшие понятия и категории для оценки.
4. **Проанализируйте структурированные данные.** Используйте традиционные методы исследования данных, такие как кластеризация, классификация и прогностическое моделирование, для обнаружения взаимосвязей между понятиями. Объедините извлеченные понятия с другими структурированными данными для прогноза будущего поведения на основе понятий.

## Анализ и категоризация текста

Анализ текста, разновидность качественного анализа, - это извлечение полезной информации из текста таким образом, чтобы ключевые идеи и понятия, содержащиеся в этом тексте, можно было сгруппировать в приемлемое число категорий. Анализ текста может применяться к текстам любого типа и объема, хотя подходы к анализу могут несколько различаться.

Короткие записи или документы проще всего категоризировать, поскольку они менее сложны и обычно содержат меньше неоднозначных слов и ответов. Например, при свободных ответах на вопросы соцопроса, если людям предлагается назвать три любимых занятия во время отпуска, можно ожидать множества коротких ответов, таких как *ходить на пляж*, *посещать национальные парки* или *ничего не делать*. С другой стороны, более длинные свободные ответы могут оказаться весьма сложными и пространными, особенно если респонденты образованны, мотивированы и располагают достаточным временем для заполнения опросного листа. Если мы попросим людей рассказать об их политических взглядах в соцопросе или порассуждать о политике в блоге, можно ожидать объемных комментариев по самым разным вопросам и проблемам.

Способность очень быстро извлекать ключевые понятия и создавать содержательные категории для таких сравнительно длинных текстовых источников - это важнейшее преимущество работы с IBM SPSS Modeler Text Analytics. Это преимущество достигается посредством комбинирования лингвистических и статистических методов автоматической обработки для достижения максимально надежных результатов на каждой стадии процесса анализа текста.

## Обработка лингвистической информации и текстов на естественном языке

Главная проблема при управлении любыми неструктурированными текстовыми данными - отсутствие стандартных правил написания текстов, так чтобы компьютер мог понимать их. Язык, а тем самым и значения, варьируют для каждого документа и каждой части текста. Единственный способ точно получить и организовать такие неструктурированные данные - это проанализировать сам язык и таким образом раскрыть его значение. Существует несколько различных автоматизированных подходов к извлечению понятий из неструктурированной информации. Эти подходы можно подразделить на два вида - лингвистические и нелингвистические.

Некоторые организации пытались применить автоматизированные нелингвистические решения, основанные на статистике и нейронных сетях. Используя компьютерные технологии, эти решения могут просматривать данные и категоризировать ключевые понятия быстрее, чем это делает человек. К сожалению, точность таких решений довольно низка. Большинство систем, действующих на основе статистики, просто подсчитывает число вхождений слов и вычисляет их статистическую близость к связанным понятиям. Они выдают много бесполезных (нерелевантных) результатов или шума, пропуская при этом действительно ценную информацию (так называемое информационное молчание).

Чтобы компенсировать ограниченную точность, некоторые решения включают сложные нелингвистические правила, которые помогают отличить полезные результаты от бесполезных. Такой подход называется *исследованием текста на основе правил*.

С другой стороны, *исследование текста на лингвистической основе* применяет принципы обработки текста на естественном языке (natural language processing, NLP) — исследование человеческих языков компьютерными методами - к анализу слов, фраз, синтаксиса или структуры текста. Система с интегрированной в нее обработкой естественного языка (NLP) может аналитически извлекать понятия, включая сложные синтагмы. Более того, знание базового языка позволяет классифицировать понятия по смыслу и контексту, объединяя их в соответствующие группы, например, продуктов, организаций или людей.

Исследование текста на лингвистической основе находит в тексте значения во многом так же, как это делает человек - посредством распознавания различных словоформ с общим значением и анализа структуры предложений, создавая основу для понимания текста. Этот подход сочетает быстродействие и

экономическую эффективность систем на основе статистики со значительно более высокой точностью, требуя при этом намного меньшего участия персонала.

Как иллюстрацию отличия статистического подхода от лингвистического в процессе извлечения для текстов на всех языках, кроме японского, рассмотрим, каким образом тот и другой реагируют на запрос о `репродуцировании документов`. И статистическое, и лингвистическое решение должно раскрыть слово `репродуцирование`, чтобы включить такие синонимы, как `копирование` и `воспроизведение`. Без этого останется незамеченной важная информация. Однако статистическое решение пытается выполнить поиск синонимов - других слов с тем же смыслом - и захватывает термин `размножение`, генерируя множество посторонних результатов. Напротив, понимание языка пробивается через двусмысленность текста, так что лингвистическое исследование текста дает по определению более надежный подход.

Использование методов на основе лингвистики через анализатор настроений делает возможным извлечение более описательных выражений. Анализ и захват настроений преодолевает проблему неоднозначности текста и, в сущности, преобразует исследование текстовых данных на основе лингвистики в более надежный подход.

Понимая, как работает извлечение, вы сможете осознанно принимать ключевые решения при тонкой настройке лингвистических ресурсов (библиотек, типов, синонимов и других). Шаги в процессе извлечения включают в себя:

- Преобразование исходных данных в стандартный формат
- Идентификация терминов-кандидатов
- Идентификация классов эквивалентности и интеграция синонимов
- Назначение типа
- Индексация и (если затребовано) сопоставление с паттернами при помощи вторичного анализатора

Шаг 1. Преобразование исходных данных в стандартный формат

На этом первом шаге импортируемые данные преобразуются к единому формату, пригодному для дальнейшего анализа. Такое преобразование выполняется внутренним образом; исходные данные при этом не изменяются.

Шаг 2. Идентификация терминов-кандидатов

Важно понимать роль лингвистических ресурсов при идентификации терминов-кандидатов во время лингвистического извлечения. Лингвистические ресурсы используются при каждом выполнении извлечения. Они существуют в форме шаблонов, библиотек и скомпилированных ресурсов. Библиотеки включают в себя списки слов, взаимосвязей и других сведений, определяющих или уточняющих извлечение. Скомпилированные ресурсы нельзя просматривать и редактировать. Однако остальные ресурсы можно отредактировать в Редактор шаблонов или (находясь в сеансе интерактивной инструментальной среды) в Редактор ресурсов.

Скомпилированные ресурсы - это базовые внутренние компоненты механизма извлечения в IBM SPSS Modeler Text Analytics . В эти ресурсы входит общий словарь, содержащий список базовых форм с кодами частей речи (существительных, глаголов, прилагательных и так далее).

Помимо этих скомпилированных ресурсов с продуктом поставляются несколько библиотек, которые могут использоваться в скомпилированных ресурсах в дополнение к типам и определениям понятий, а также для предложения синонимов. Эти библиотеки, а также любые созданные вами пользовательские библиотеки, состоят из ряда словарей. Это словари типов, словари синонимов и словари исключения.

После импорта и преобразования данных механизм извлечения приступает к идентификации извлекаемых терминов-кандидатов. Термины-кандидаты - это слова или группы слов, которыми в тексте обозначаются те или иные понятия. Во время обработки текстовых данных отдельные слова (**отдельные термины**) и составные

слова (**составные термины**) выявляются при помощи модулей извлечения паттернов частей речи. Затем при помощи анализа текстовых связей на настроения выявляются ключевые слова-кандидаты настроений.

*Примечание*: Термины в упомянутом выше скомпилированном общем словаре представляют список всех слов, которые, вероятно, окажутся неинтересны или лингвистически неоднозначны как одиночные термины. Такие слова исключаются из извлечения при идентификации одиночных терминов. Однако они снова оцениваются при выявлении частей речи или поиске более длинных составных слов-кандидатов (терминов-словосочетаний).

Шаг 3. Идентификация классов эквивалентности и интеграция синонимов

После выявления отдельных терминов-кандидатов и составных терминов-кандидатов программный продукт при помощи словаря нормализации выявляет классы эквивалентности. Класс эквивалентности - это базовая форма словосочетания или один из двух вариантов форм одного и того же словосочетания. Чтобы определить, какое понятие следует использовать для класса эквивалентности , механизм извлечения применяет следующие правила в указанном порядке:

- Пользовательская форма в библиотеке.
- Самая частая форма, определяемая предварительно скомпилированными ресурсами.

Шаг 4. Задание типа

Далее извлеченным понятиям назначаются типы. Тип объединяет понятия по их смыслу. На этом шаге используются как скомпилированные ресурсы, так и библиотеки. К типа относятся понятия высокого уровня, слова с положительной и отрицательной оценкой, имена людей, названия мест и организаций и другое. Дополнительную информацию смотрите в разделе “Словари типов” на стр. 189.

Имейте в виду, что у ресурсов для японского языка особый набор типов.

Лингвистические системы восприимчивы к знаниям: чем больше информации в их словарях, тем выше качество результатов. Изменение содержимого словарей, например, определений синонимов, может упростить итоговую информацию. Этот процесс, который часто бывает итеративным, необходим для более точного извлечения понятий. Обработка текста на естественном языке - ключевой элемент IBM SPSS Modeler Text Analytics.

## Как работает извлечение

Извлекая ключевые понятия и понятия из ваших ответов, IBM SPSS Modeler Text Analytics опирается на лингвистический анализ текста. При помощи этого подхода достигается производительность и экономичность, как у систем на основе статистики. Однако этот подход обеспечивает гораздо более высокую точность при минимальном вмешательстве оператора. Лингвистический анализ текста основан на области исследований, называемой обработкой естественного языка или компьютерной лингвистикой.

**Важное замечание:** Для текста на японском процесс извлечения выполняет другой набор действий.

Понимая, как работает извлечение, вы сможете осознанно принимать ключевые решения при тонкой настройке лингвистических ресурсов (библиотек, типов, синонимов и других). Шаги в процессе извлечения включают в себя:

- Преобразование исходных данных в стандартный формат
- Идентификация терминов-кандидатов
- Идентификация классов эквивалентности и интеграция синонимов
- Назначение типа
- Индексация
- Сопоставление паттернов и извлечение событий

Шаг 1. Преобразование исходных данных в стандартный формат

На этом первом шаге импортируемые данные преобразуются к единому формату, пригодному для дальнейшего анализа. Такое преобразование выполняется внутренним образом; исходные данные при этом не изменяются.

Шаг 2. Идентификация терминов-кандидатов

Важно понимать роль лингвистических ресурсов при идентификации терминов-кандидатов во время лингвистического извлечения. Лингвистические ресурсы используются при каждом выполнении извлечения. Они существуют в форме шаблонов, библиотек и скомпилированных ресурсов. Библиотеки включают в себя списки слов, взаимосвязей и других сведений, определяющих или уточняющих извлечение. Скомпилированные ресурсы нельзя просматривать и редактировать. Но остальные ресурсы (шаблоны) можно редактировать в Редактор шаблонов или, если вы находитесь в интерактивном сеансе инструментальной среды, в Редактор ресурсов.

Скомпилированные ресурсы - это компоненты ядра, то есть внутренние компоненты механизма извлечения в IBM SPSS Modeler Text Analytics. Эти ресурсы включают в себя общий словарь, содержащий список базовых форм с кодом части речи (существительное, глагол, прилагательное, наречие, причастие, подчинительный союз, детерминатив, предлог). Кроме того, ресурсы включают в себя зарезервированные, встроенные типы, используемые для назначения большому числу извлекаемых терминов: `<Расположение>`, `<Организация>`, `<Личный>`. Дополнительную информацию смотрите в разделе “Встроенные типы” на стр. 190.

Кроме этих скомпилированных ресурсов с продуктом поставляется несколько библиотек, которые могут использоваться в дополнение к определениям типов и понятий в скомпилированных ресурсах, а также для поддержки других типов и синонимов. Эти библиотеки, а также любые созданные вами пользовательские библиотеки, состоят из ряда словарей. Они включают в себя словари типов, словари подстановок (синонимов и необязательных элементов) и словари исключения. Дополнительную информацию смотрите в разделе Глава 16, “Работа с библиотеками”, на стр. 179.

После импорта и преобразования данных механизм извлечения приступает к идентификации извлекаемых терминов-кандидатов. Термины-кандидаты - это слова или группы слов, которыми в тексте обозначаются те или иные понятия. При обработке текста отдельные слова (*одиночные термины*), отсутствующие в скомпилированных ресурсах, считаются терминами-кандидатами на извлечение. Составные слова-кандидаты (*термины-словосочетания*) выявляются при помощи средств извлечения паттернов частей речи. Например, термин `спортивный автомобиль`, следующий паттерну частей речи "прилагательное существительное", содержит два компонента. Термин `быстрый спортивный автомобиль`, следующий паттерну частей речи "прилагательное прилагательное существительное", содержит три компонента.

*Примечание*: Термины в упомянутом выше скомпилированном общем словаре представляют список всех слов, которые, вероятно, окажутся неинтересны или лингвистически неоднозначны как одиночные термины. Такие слова исключаются из извлечения при идентификации одиночных терминов. Однако они снова оцениваются при выявлении частей речи или поиске более длинных составных слов-кандидатов (терминов-словосочетаний).

Наконец, специальный алгоритм служит для обработки строк из заглавных букв, таких как должности сотрудников, чтобы была возможна извлечение таких специальных паттернов.

Шаг 3. Идентификация классов эквивалентности и интеграция синонимов

После идентификации кандидатов для отдельных терминов и терминов-словосочетаний программа использует набор алгоритмов для сравнения кандидатов и идентификации классов эквивалентности. Класс эквивалентности представлен базовой формой словосочетания или отдельного слова, для которых существует и другой вариант. Цель задания классов эквивалентности для словосочетаний - убедиться, что, например, `president of the company` и `company president` не будут считаться отдельными понятиями.

Чтобы решить, какое понятие будет представлять класс эквивалентности, например, выбрать главный термин из `president of the company` и `company president`, механизм извлечения применяет следующие правила, в указанном порядке:

- Пользовательская форма в библиотеке.
- Форма, наиболее частая во всем корпусе текста.
- Самая короткая форма во всем корпусе текста (обычно это базовая форма в данном языке).

Шаг 4. Задание типа

Далее извлеченным понятиям назначаются типы. Тип объединяет понятия по их смыслу. На этом шаге используются как скомпилированные ресурсы, так и библиотеки. К типа относятся понятия высокого уровня, слова с положительной и отрицательной оценкой, имена людей, названия мест и организаций и другое. Пользователь может определять дополнительные типы. Дополнительную информацию смотрите в разделе “Словари типов” на стр. 189.

Шаг 5. Индексация

Весь набор записей или документов индексируется - для каждого класса эквивалентности задается указатель из позиции в тексте на представляющий термин. Предполагается, что все вхождения флективных форм понятия-кандидата индексируются как базовая форма кандидата. Для каждой базовой формы вычисляется глобальная частота.

Шаг 6. Сопоставление паттернов и извлечение событий

IBM SPSS Modeler Text Analytics может обнаруживать не только типы и понятия, но и взаимосвязи между ними. Ряд алгоритмов и библиотек, доступных вместе с этим продуктом, поддерживают возможность извлекать паттерны взаимосвязей между типами и понятиями. Они особенно полезны при попытке обнаружить определенные мнения (например, реакцию на продукт) или взаимосвязи между людьми или объектами (например, связи между политическими группами или геномами).

## Как работает категоризация

При создании моделей категорий в IBM SPSS Modeler Text Analytics можно выбрать несколько различных методов создания категорий. Поскольку каждый набор данных уникален, число методов и порядок их применения может изменяться. Ваша интерпретация результатов может отличаться от чьей-либо еще, поэтому может потребоваться попробовать различные методы, чтобы понять, какой из них генерирует лучшие результаты для ваших текстовых данных. В IBM SPSS Modeler Text Analytics можно создать модели категорий в сеансе инструментальной среды, где вы можете исследовать ваши категории и далее настроить их.

В этом руководстве **построение категорий** относится к генерированию определений категорий и классификации при помощи одного или нескольких встроенных методов, а **категоризация** означает скоринг или ранжирование метками - процесс, в результате которого определениям категорий для каждой записи или документа назначаются уникальные идентификаторы (имя/ID/значение).

Во время построения категорий в качестве блоков их построения используются понятия и типы, полученные процессом извлечения. Когда вы строите категории, записи или документы автоматически назначаются категориям, если они содержат текст, совпадающий с одним из элементов определения категории.

IBM SPSS Modeler Text Analytics поддерживает несколько методов автоматического построения категорий, помогая быстро выполнять категоризацию ваших документов или записей.

Методы группирования

Каждый из доступных методов хорошо подходит для определенных типов данных и ситуаций, но часто для получения полного набора документов или записей бывает полезно сочетать методы в ходе одного анализа. Вы можете увидеть понятие в нескольких категориях или найти лишние категории.

**Вывод корня понятия**. Этот метод создает категории, исходя из некоторого понятия, путем поиска связанных с ним других понятий, когда соответствующие компоненты морфологически родственны или имеют общие корни. Этот метод очень полезен для выявления синонимичных понятий, выраженных сочетаниями слов, поскольку понятия в каждой созданной категории являются синонимами или близки по значению. Он работает с данными различной длины и создает небольшое число компактных категорий. Например, понятие возможности развиваться могло бы объединиться с понятиями возможности для развития и возможное развитие. Дополнительную информацию смотрите в разделе “Вывод корня понятия” на стр. 117. Для текста на японском эта опция недоступна.

**Семантическая сеть**. Этот метод начинает обработку с выявления возможных направлений обхода каждого понятия в его расширенном индексе взаимосвязей слов, а затем создает категории, группируя связанные понятия. Этот метод дает наилучшие результаты, когда понятия известны семантической сети и не слишком неоднозначны. Он менее полезен, когда текст содержит специальную терминологию или неизвестные сети жаргонизмы. Например, понятие яблоки Гренни Смит может быть объединено с понятиями яблоки Гала и яблоки Вайнсеп, поскольку все это представители одного класса объектов (сорта яблок). В другом примере понятие животное можно объединить с понятиями кошка и кенгуру, поскольку это гипонимы для понятия животное. Этот метод в данном выпуске применим только к английским текстам. Дополнительную информацию смотрите в разделе “Семантические сети” на стр. 119.

**Вложенные понятия**. Этот метод строит категории путем группировки составных терминов (словосочетаний) на основе наличия в них слов, являющихся поднаборами других или, наоборот, включающих в себя другие слова как поднаборы. Например, понятие сидение будет объединено с понятиями сидение с ремнем безопасности, ремень безопасности и пряжка ремня безопасности. Дополнительную информацию смотрите в разделе “Включение понятий” на стр. 118.

**Совместная встречаемость**. Этот метод создает категории, исходя из совместной встречаемости понятий в тексте. Идея состоит в том, что когда понятия или шаблоны понятий часто встречаются вместе в документах и записях, эта совместная встречаемость отражает некую базовую взаимосвязь, которая может оказаться полезной в ваших определениях категорий. Когда слова имеют существенную тенденцию к совместной встречаемости, создается правило совместной встречаемости, которое может выступать в качестве дескриптора для новой подкатегории. Например, если во многих записях встречаются слова цена и доступность (но по отдельности эти слова встречаются редко), соответствующие понятия можно сгруппировать, создав правило совместной встречаемости (цена & доступная) и включить их, например, в подкатегорию категории цена. Дополнительную информацию смотрите в разделе “Правила совместного появления” на стр. 120.

**Минимальное число документов.** Чтобы определить, насколько интересными могут быть совместные вхождения, задайте минимальное число документов или записей, в которых термины встречаются совместно, чтобы использовать эти случаи в качестве дескриптора категории.

## IBM SPSS Modeler Text Analytics узлов

Помимо стандартных узлов, поставляемых с IBM SPSS Modeler, можно также использовать узлы исследования текстовых данных для встраивания мощности исследования текста в ваши потоки. Чтобы сделать именно так, IBM SPSS Modeler Text Analytics предлагает несколько узлов исследования текстовых данных. Эти узлы хранятся на вкладке IBM SPSS Modeler Text Analytics палитры узлов.

Включены следующие узлы:

- **Исходный узел Список файлов** генерирует список имен документов в качестве входных данных для процесса исследования текстовых данных. Это полезно, если текст находится во внешних документах, а не в базе данных и не в других структурированных файлах. Выходные данные узла представляют собой одно

поле с одной записью для каждого документа или папки в списке, которое можно выбрать в качестве входных данных для последующего узла исследования текстовых данных. Дополнительную информацию смотрите в разделе “Узел списка файлов” на стр. 11.

- **Исходный узел веб-фидов** позволяет читать текст из веб-фидов (таких как блоги или ленты новостей в форматах RSS или HTML) и использовать эти данные в процессе исследования текстовых данных. Выходные данные этого узла представляют собой одно или несколько полей, которое можно выбрать в качестве входных данных для последующего узла исследования текстовых данных. Дополнительную информацию смотрите в разделе “Узел Веб-фид” на стр. 13.
- **Узел исследования текстовых данных** использует лингвистические методы для извлечения ключевых понятий из текста, позволяет создать категории с этими понятиями и другими данными и предоставляет возможность определения взаимосвязей и связей между понятиями на основе известных паттернов (так называемый анализ текстовых связей). Этот узел может использоваться для изучения содержимого текстовых данных либо для генерирования модели понятий (понятийной модели) или модели категорий. Понятия и категории можно сочетать с существующими структурированными данными, такими как демографические, и применять к моделированию. Дополнительную информацию смотрите в разделе “Режим моделирования Text Mining” на стр. 20.
- **Узел анализа текстовых связей** извлекает и выявляет взаимосвязи между понятиями на основе известных паттернов в текстовых данных. Извлечение паттернов может использоваться для выявления взаимосвязей между понятиями, а также между любыми мнениями или спецификаторами, присоединенными к этим понятиям. Узел анализа текстовых связей предлагает более прямой способ выявления паттернов и извлечения паттернов из текста и последующего добавления результатов паттернов в набор данных в потоке. Но анализ текстовых связей (Text Link Analysis, TLA) можно также выполнить при помощи сеанса интерактивной инструментальной среды на узле моделирования Исследование текстовых данных. Дополнительную информацию смотрите в разделе “Узел Text Link Analysis” на стр. 47.
- **Узел перевода** может использоваться для перевода текста с поддерживаемых языков (таких как арабский, китайский и персидский) на английский или другие языки в целях моделирования. Это разрешает исследование документов на двухбайтовых языках, поддержка которых в противном случае оказалась бы невозможной, и позволяет аналитикам извлекать из этих документов понятия, даже если они не владеют исследуемым языком. Такую же функциональную возможность можно вызвать с любых узлов моделирования текстовых данных, но использование отдельного узла перевода разрешает кэширование и повторное использование перевода на нескольких узлах. Дополнительную информацию смотрите в разделе “Узел перевода” на стр. 55.
- При исследовании текстовых данных из внешних документов с помощью **узла выходных данных исследования текста** можно сгенерировать страницу HTML, содержащую ссылки на документы, из которых были извлечены понятия. Дополнительную информацию смотрите в разделе “Узел программы просмотра файлов” на стр. 59.

## Программы

В общем случае все, кому требуется регулярно получать большие объемы документов в целях выявления ключевых элементов для дальнейших исследований, могут извлечь преимущества из IBM SPSS Modeler Text Analytics.

Вот некоторые конкретные примеры использования:

- **Научные и медицинские исследования.** Изучение вспомогательных материалов исследований, таких как отчеты о патентах, журнальные статьи и публикации протоколов. Выявление связей, ранее неизвестных (таких как связь добавки с конкретным продуктом), представляющих перспективы дальнейших исследований. Минимизация времени, требуемого на исследование новых лекарственных средств. Использование в качестве вспомогательного средства в исследованиях геномики.
- **Финансовый анализ.** Просмотр ежедневных аналитических отчетов, новых статей и пресс-релизов компаний с целью выявления ключевых позиций стратегии или изменений на рынке. Анализ тенденций такой информации показывает возникающие проблемы или возможности для фирмы или отрасли за определенный промежуток времени.

- **Выявление мошенничества.** Использование в сфере мошенничества в банковском деле и здравоохранении с целью выявления аномалий и обнаружения тревожных моментов в больших объемах текстовых данных.
- **Исследование рынка.** Использование в предприятиях изучения рынка с целью выявления ключевых тем в ответах опроса открытого характера.
- **Анализ блогов и веб-фидов.** Изучение и построение моделей при помощи ключевых идей, найденных в лентах новостей, блогах и так далее.
- **CRM.** (Customer Relationship Management - управление взаимосвязями с заказчиками) Построение моделей при помощи данных изо всех точек контактов с заказчиками, таких как электронная почта, транзакции и опросы.

# Глава 2. Чтение в исходном тексте

Данные для Text Mining могут находиться в любом из стандартных форматов, используемых IBM SPSS Modeler, включая базы данных или другие "прямоугольные" форматы, в которых данные представляются по строкам и столбцам, или форматы документов, такие как Microsoft Word, Adobe PDF или HTML, не обладающие такой структурой.

- Чтобы прочитать текст из документов, которые не соответствуют стандартной структуре данных, включая Microsoft Word, Microsoft Excel и Microsoft PowerPoint, а также Adobe PDF, XML, HTML и другие, можно использовать узел Список файлов, который генерирует список документов или папок в качестве входной информации для исследования текста. Дополнительную информацию смотрите в разделе “Узел списка файлов”.
- Чтобы прочитать текст из веб-фидов, например, блогов или лент новостей в форматах RSS или HTML, можно использовать узел Веб-фид, который форматирует данные веб-фидов для ввода в процесс Text Mining. Дополнительную информацию смотрите в разделе “Узел Веб-фид” на стр. 13.
- Чтобы прочитать текст из любого стандартного формата данных, используемого IBM SPSS Modeler, такого как база данных с одним или несколькими текстовыми полями для комментариев заказчиков, можно использовать любые собственные узлы источников IBM SPSS Modeler. Дополнительную информацию смотрите в документации по узлу IBM SPSS Modeler.

## Узел списка файлов

Для чтения неструктурированных документов, сохраняемых в форматах, таких как Microsoft Word, Microsoft Excel и Microsoft PowerPoint (а также в Adobe PDF, XML, HTML и других), при помощи узла Список файлов можно сгенерировать список документов или папок в качестве входных данных для процесса исследования текстовых данных. Необходимость этого обусловлена невозможностью представить неструктурированные текстовые документы полями и записями (строками и столбцами) тем же образом, что и другие данные, используемые IBM SPSS Modeler. Этот узел можно найти на палитре исследования текстовых данных.

Узел Список файлов работает как узел источника; однако возможен вариант прочитать при помощи этого узла (как и в случае чтения или вывода фактических данных исходных файлов) имена документов или каталогов в заданном корневом положении и сгенерировать из них список. При чтении с помощью этого узла имен документов или каталогов выходные данные представляют собой одно поле (с одной записью для каждого файла в списке), которое можно выбрать в качестве входных данных для последующего узла исследования текстовых данных (Text Mining) или узла анализа текстовых связей (Text Link Analysis, TLA).

Этот узел можно найти на вкладке IBM SPSS Modeler Text Analytics палитры узлов в нижней части окна IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “IBM SPSS Modeler Text Analytics узлов” на стр. 8.

**Важное замечание:** Никакие имена каталогов и файлов, содержащие символы, которые не включены в локальную кодировку компьютера, не поддерживаются. При попытке выполнить поток, содержащий узел Список файлов, любые имена файлов или каталогов, содержащие эти символы, приведут к неудачному завершению выполнения потока. Это может произойти с именами каталогов или файлов на иностранных языках, например, русским именем файлов во французской локали.

Поддержка локальных данных. Если вы соединяетесь с удаленным Сервер IBM SPSS Modeler Text Analytics и у вас есть поток с узлом Список файлов, данные должны располагаться на том же компьютере, что и Сервер IBM SPSS Modeler Text Analytics; либо убедитесь, что с компьютера сервера есть доступ к папке, где хранятся исходные данные узла Список файлов.

**Примечание:** В конфигурации IBM SPSS Collaboration and Deployment Services - Scoring узел Список файлов для скоринга использовать нельзя.

## Узел Список файлов: Вкладка Параметры

На этой вкладке можно определить каталоги, расширения файлов и получаемые с этого узла выходные данные.

*Примечание*: Операция извлечения исследования текстовых данных не может обрабатывать файлы Microsoft Office и Adobe PDF на платформах, иных чем Microsoft Windows. Однако текстовые файлы и файлы в формате XML и HTML можно обрабатывать всегда.

Никакие имена каталогов и файлов, содержащие символы, которые не включены в локальную кодировку компьютера, не поддерживаются. При попытке выполнить поток, содержащий узел Список файлов, любые имена файлов или каталогов, содержащие эти символы, приведут к неудачному завершению выполнения потока. Это может произойти с именами каталогов или файлов на иностранных языках, например, русским именем файлов во французской локали.

**Каталог.** Задает корневую папку, содержащую документы, список которых вы хотите получить.

- **Включить подкаталоги.** Указывает, что следует также просмотреть подкаталоги.

**Включаемые в список типы файлов:** Можно выбрать или отменить выбор типов файлов и расширений, которые вы хотите использовать. После отмены расширения файлы с этим расширением будут игнорироваться. Возможна фильтрация по следующим расширениям:

*Таблица 1. Фильтры типов файлов по расширениям файлов.*

| | | | |
|---|---|---|---|
| • *.rtf*, *.doc*, *.docx*, *.docm* | • *.xls, .xlsx, .xlsm* | • *.ppt, .pptx, .pptm* | • *.txt*, *.text* |
| • *.htm*, *.html*, *.shtml* | • *.xml* | • *.pdf* | • *.$* |

*Примечание*: Дополнительную информацию смотрите в разделе “Узел списка файлов” на стр. 11.

Для выбора файлов либо без расширения, либо с расширением в виде концевой точки (например, `File01` или `File01.`), если они есть, используется опция **Нет расширения**.

**Выходное поле представляет.** Выберите формат выходного поля. Варианты выбора:

- **Фактический текст.** Выберите эту опцию, если поле будет содержать точный текст. Тогда можно будет выбрать значение для **кодировки входных данных** в следующем списке:
  - Автоматически (европейский)
  - Автоматически (японский)
  - UTF-8
  - UTF-16
  - ISO-8859-1
  - US ASCII
  - CP850
  - Shift-JIS
- **Имена путей для документов.** Выберите эту опцию, если выходное поле будет содержать один или несколько имен путей для положений, где располагаются документы.

**Важно!** Начиная с версии 14, опция 'Список каталогов' более не доступна, и единственный тип выходных данных будет представлять собой список файлов.

## Узел Список файлов: Другие вкладки

Вкладка Типы - стандартная вкладка в узлах IBM SPSS Modeler, как и вкладка Аннотации.

## Использование узла Список файлов для исследования текстовых данных

Узел Список файлов используется, если текстовые данные располагаются во внешних неструктурированных документах в форматах, таких как Microsoft Word, Microsoft Excel и Microsoft PowerPoint, а также в Adobe PDF, XML, HTML и других. Кроме того, для вывода фактических текстовых данных при помощи этого узла можно сгенерировать список документов или папок в качестве входных данных для процесса исследования текстовых данных (например, для последующего узла исследования текстовых данных (Text Mining) или узла анализа текстовых связей (Text Link Analysis, TLA).

Когда с помощью узла Список файлов генерируется список документов, а не фактические данные, если в дальнейшем будет использоваться либо узел Исследование текста, либо узел Анализ текстовых связей, укажите, что выбранное поле будет содержать не фактические текстовые данные для исследования, а пути к документам, где находится текст, указав для этого, что в **текстовом поле** будут представлены **Имена путей для документов**.

В качестве примера допустим, что мы соединили узел Список файлов с узлом Исследование текстовых данных для подачи текста, находящегося во внешних документах.

1. **Узел Список файлов (вкладка параметров).** Первым действием мы добавили этот узел в поток, чтобы задать, где хранятся текстовые документы. Мы выбрали каталог, содержащий все документы, для которых нужно выполнить исследование текста.
2. **Узел Text Mining (вкладка полей).** Затем мы добавили и подключили к узлу Список файлов узел Text Mining. В этом узле мы определили формат ввода, шаблон ресурса и формат вывода. Мы выбрали имя поля, созданного узлом список файлов, и выбрали опцию, в которой текстовое поле представляет **имена и пути к документам**, а также другие параметры. Дополнительную информацию смотрите в разделе “Использование узла Text Mining в потоке” на стр. 30.

Дополнительную информацию об использовании узла Исследование текста смотрите в разделе “Режим моделирования Text Mining” на стр. 20.

---

## Узел Веб-фид

Узел веб-фидов может использоваться для подготовки текстовых данных из веб-фидов для процесса исследования текстовых данных. Это узел принимает веб-фиды в двух форматах:

- Формат RSS. RSS - это простой стандартизованный формат на основе XML для веб-содержимого. URL для этого формата указывает на страницу, содержащую связанные ссылками статьи, например, централизованные источники новостей и блоги. Поскольку RSS - стандартизованный формат, каждая связанная ссылкой статья автоматически определяется и обрабатывается как отдельная запись в итоговом наборе данных. Для возможности выявления важных текстовых данных и записей из фида никаких дополнительных входных данных не требуется, если только вы не захотите применить к тексту метод фильтрации.
- Формат HTML. На вкладке входных данных можно задать один или несколько URL для страниц HTML. Далее, на вкладке Записи задайте открывающий тег записи, а также определите теги, разделяющие содержимое назначения, и назначьте эти теги выходным полям по вашему выбору (полю описания, заголовка, даты изменения и так далее). Дополнительную информацию смотрите в разделе “Узел веб-фидов: Вкладка Записи” на стр. 15.

**Важно!** Если вы пытаетесь получить информацию по сети через прокси-сервер, нужно включить использование прокси-сервера в файле `net.properties` и для клиента, и для сервера IBM SPSS Modeler Text Analytics. Следуйте подробным указаниям в этом файле. Они применимы при доступе к сети через узел Веб-фид или при получении лицензии программы как службы (Software as a Service, SaaS) для SDL, поскольку такие соединения проходят через Java™. По умолчанию этот файл расположен в каталоге *C:\Program Files\IBM\SPSS\Modeler\18\jre\lib\net.properties*.

Выходные данные этого узла - набор полей, используемых для описания записей. Поле **Описание** используется чаще всего, поскольку содержит подавляющий объем текстового содержимого. Однако могут оказаться интересны и другие поля, такие как поле краткого описания записи (**Краткое описание**) или поле заголовка записи (**Заголовок**). Любые из этих выходных полей можно выбрать в качестве входных данных для последующего поля исследования текстовых данных.

**Примечание:** Узел веб-фидов для скоринга в конфигурации IBM SPSS Collaboration and Deployment Services - Scoring использовать нельзя.

Этот узел можно найти на вкладке IBM SPSS Modeler Text Analytics палитры узлов в нижней части окна IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “IBM SPSS Modeler Text Analytics узлов” на стр. 8.

## Узел Веб-фид: Вкладка Входные данные

С помощью вкладки Входные данные задается один или несколько веб-адресов (или URL) для захвата текстовых данных. В контексте исследования текста можно задать адреса URL для фидов, содержащих текстовые данные.

**Важно!** При работе с данными в формате, ином чем RSS, возможно, вы предпочтете использовать веб-инструмент утилизации, такой как WebQL®, для автоматизации сбора содержимого и последующего обращения к выходным данным из этого инструмента с применением другого исходного узла.

Можно задать следующие параметры:

**Введите или вставьте адреса URL.** В этом поле можно ввести или вставить один или несколько адресов URL. При вводе нескольких URL вводите их только по одному в каждой строке и разделяйте строки при помощи клавиши **Enter/Return**. Вводите полный путь URL к файлу. Эти URL могут быть для фидов одного из двух форматов:

- *Формат RSS.* RSS - это простой стандартизованный формат на основе XML для веб-содержимого. URL для этого формата указывает на страницу, содержащую связанные ссылками статьи, например, централизованные источники новостей и блоги. Поскольку RSS - стандартизованный формат, каждая связанная ссылкой статья автоматически определяется и обрабатывается как отдельная запись в итоговом наборе данных. Для возможности выявления важных текстовых данных и записей из фида никаких дополнительных входных данных не требуется, если только вы не захотите применить к тексту метод фильтрации.
- *Формат HTML.* На вкладке входных данных можно задать один или несколько URL для страниц HTML. Далее, на вкладке Записи задайте открывающий тег записи, а также определите теги, разделяющие содержимое назначения, и назначьте эти теги выходным полям по вашему выбору (полю описания, заголовка, даты изменения и так далее). При работе с данными в формате, ином чем RSS, возможно, вы предпочтете использовать веб-инструмент утилизации, такой как WebQL®, для автоматизации сбора содержимого и последующего обращения к выходным данным из этого инструмента с применением другого исходного узла. Дополнительную информацию смотрите в разделе “Узел веб-фидов: Вкладка Записи” на стр. 15.

**Число наиболее новых записей, читаемых из одного URL.** Это поле задает максимальное число читаемых записей для каждого указанного в поле адреса URL, начиная с первой записи, найденной в фиде. Объем текстовых данных влияет на скорость обработки нисходящего потока извлечения на узле Исследование текстовых данных или узле Анализ текстовых связей.

**Сохранять и по возможности повторно использовать предыдущие веб-фиды.** Если включить эту опцию, веб-фиды будут просматриваться, а обработанные результаты кэшироваться. Далее, при последующих выполнениях в потоке, если содержимое данного фида не изменилось или если фид недоступен (например, из-за отключения Интернета), будет использоваться кэшированная версия, сокращающая время обработки. Любое содержимое, обнаруженное в этих фидах, также будет кэшироваться при следующем выполнении обработки узла.

- **Метка**. При выборе опции **Сохранять и по возможности повторно использовать предыдущие веб-фиды** нужно задать имя метки для результатов. Эта метка будет использоваться для описания кэшируемых фидов на сервере. Если метка не будет задана или будет нераспознана, повторное использование будет невозможно. Кэшированием веб-фидов можно управлять в таблице сеанса IBM SPSS Text Analytics Administration Console . Посмотрите дополнительную информацию в Руководстве пользователя IBM SPSS Text Analytics Administration Console .

## Узел веб-фидов: Вкладка Записи

На вкладке Записи задается текстовое содержимое фидов в формате, ином чем RSS, для чего определяется, где начинается каждая новая запись, а также другая нужная информация о каждой записи. Если известно, что фид в формате, ином чем RSS, (HTML) содержит текст, находящийся в нескольких записях, здесь нужно определить открывающий тег записи, иначе этот текст будет обработан как одна запись. Хотя фиды RSS стандартизованы и не требуют никакой спецификации на этой вкладке, но содержимое все равно можно будет предварительно просмотреть на вкладке Предварительный просмотр.

**Важно!** При работе с данными в формате, ином чем RSS, возможно, вы предпочтете использовать веб-инструмент утилизации, такой как WebQL®, для автоматизации сбора содержимого и последующего обращения к выходным данным из этого инструмента с применением другого исходного узла.

**URL.** Этот выпадающий список содержит адреса URL, вводимые на вкладке Входные данные. Представлены фиды и в формате HTML, и в формате RSS. Если адрес URL окажется слишком длинным для этого выпадающего списка, он автоматически будет сокращен в середине при помощи многоточия, заменяющего урезаемый текст, например: *http://www.ibm.com/example/start-of-address...rest-of-address/path.htm*.

- В опции **Фиды в формате HTML**, если фид содержит несколько записей (или статей), можно определить теги HTML, содержащие данные, которые соответствуют полю, указанному в приведенной таблице. Например, можно определить открывающий тег, указывающий, что начата новая запись, тег даты изменения или имя автора.
- В опции **Фиды в формате RSS** не предлагается вводить никакие теги, поскольку RSS - стандартизованный формат. Однако выборочные результаты при желании можно просмотреть на вкладке Предварительный просмотр. Всем распознанным фидам RSS будет предшествовать изображение логотипа RSS.

**Вкладка Источник.** На этой вкладке можно просмотреть исходный код для любых фидов HTML. Это нередактируемый код. При помощи поля Найти на этой странице можно найти конкретные теги или информацию, которые можно затем скопировать и вставить в таблицу ниже. Поле Найти регистронезависимо и будет соответствовать частично вводимым строкам.

**Вкладка Предварительный просмотр.** На этой вкладке можно предварительно просмотреть, как запись будет прочитана узлом веб-фидов. Особенно это полезно для фидов HTML, поскольку можно изменить способ чтения записи, определив теги HTML в таблице ниже вкладки Предварительный просмотр.

**Открывающий тег записей в формате, ином чем RSS.** Эта опция применяется только к фидам в формате, ином чем RSS. Если фид HTML содержит сложные текстовые данные, которые вы хотите разбить на несколько записей, задайте здесь тег HTML, сигнализирующий о начале записи (такой как статья или запись в блоге). Если не определить этот тег для фида в формате, ином чем RSS, тогда вся страница будет обработана как всего одна запись, все содержимое будет выведено в поле **Описание** и дата выполнения для узла будет использоваться и как **Дата изменения**, и как **Дата публикации**.

**Вкладка Поле.** Эта опция применяется только к фидам в формате, ином чем RSS. В этой таблице текстовое содержимое можно разбить на конкретные выходные поля, введя открывающий тег для любых предопределенных выходных полей. Вводится только открывающий тег. Все соответствия устанавливаются посредством синтаксического анализа HTML и сопоставления содержимого таблицы с обнаруженными в HTML именами и атрибутами тегов. При помощи кнопок в нижней части вкладки можно скопировать определенные теги и использовать их повторно для других фидов.

*Таблица 2. Возможные выходные поля для фидов в формате, ином чем RSS (в форматах HTML)*

| Имя поля выходных данных | Ожидаемое содержимое тега |
|---|---|
| **Заголовок** | Разделительный тег заголовка записи. (необязательно) |
| **Краткое описание** | Разделительный тег краткого описания или метки. (необязательно) |
| **Описание** | Разделительный тег основного текста. Если оставить пустым, все остальное содержимое этого поля будет находиться либо в теге `<body>` (в случае одной записи), либо в содержимом в текущей записи (если был задан разделитель записей). |
| **Автор** | Разделительный тег автора текста. (необязательно) |
| **Участники** | Тег, разделяющий имена участников. (необязательно) |
| **Дата публикации** | Разделительный тег даты публикации текстовых данных. Если оставить пустым, это поле будет содержать дату чтения данных узлом. |
| **Дата изменения** | Разделительный тег даты изменения текстовых данных. Если оставить пустым, это поле будет содержать дату чтения данных узлом. |

После ввода в таблицу тега фид будет сканироваться с применением этого тега в качестве минимального для установления соответствия, а не для установления точного совпадения. То есть если для поля Заголовок ввести `<div>`, он будет соответствовать любому тегу `<div>` в фиде, включая теги с заданными атрибутами (такими как `<div class="post three">`), так что тег `<div>` будет эквивалентен корневому тегу (`<div>`) и любому дериативу, включающему в себя атрибут, и это содержимое будет использоваться для выходного поля Заголовок. Если ввести корневой тег, будут также включены и все дополнительные атрибуты.

*Таблица 3. Примеры тегов HTML, используемых для идентификации текста для выходных полей.*

| Если вы вводите: | Будет соответствовать: | А также соответствовать: | Но не будет соответствовать: |
|---|---|---|---|
| `<div>` | `<div>` | `<div class="post">` | любому другому тегу |
| `<p class="auth">` | `<p class="auth">` | `<p color="black" class="auth" id="85643">` | `<p color="black">` |

## Узел веб-фидов: Вкладка Фильтр содержимого

Вкладка Фильтр содержимого используется для применения метода фильтрации к содержимому фидов RSS. К фидам HTML эта вкладка не применяется. Вам может потребоваться применить фильтр, если фид содержит множество тестовых данных в виде верхних и нижних колонтитулов, меню, рекламы и так далее. С помощью этой вкладки можно отсечь от содержимого ненужные теги HTML, JavaScript и короткие слова или строки.

**Фильтрация содержимого**. Если вы не хотите применять метод очистки, выберите **Нет**. В противном случае выберите **Очистка содержимого RSS**.

**Опции содержимого RSS**. Если выбрать **Очистка содержимого RSS**, можно будет выбрать вариант отбрасывания строк на основе определенных критериев. Строка ограничивается тегом HTML, таким как `<p>` или `<li>`, но с исключением встроенных в нее тегов, таких как `<span>`, `<b>` или `<font>`. Имейте в виду, что теги `<br>` обрабатываются как символы перевода строки.

- **Отбрасывать короткие строки**. Эта опция игнорирует строки, не содержащие определенное здесь **минимальное число слов**.
- **Отбрасывать строки с короткими словами**. Эта опция игнорирует строки, длина которых превышает определенную здесь **минимальную среднюю длину слов**.
- **Отбрасывать строки с множеством односимвольных слов**. Эта опция игнорирует строки, содержащие определенную **долю односимвольных слов**.
- **Отбрасывать строки, содержащие конкретные теги**. Эта опция игнорирует текст в строках, содержащих любые из тегов, указанных в этом поле.

- **Отбрасывать строки, содержащие конкретный текст**. Эта опция игнорирует строки, содержащие любые текстовые данные, указанные в этом поле.

## Использование узла веб-фидов для исследования текстовых данных

Узел веб-фидов может использоваться для подготовки текстовых данных из веб-фидов Интернета для процесса исследования текстовых данных. Это узел принимает веб-фиды либо в формате HTML, либо в формате RSS. Эти фиды выступают в качестве входных данных, вводимых в процесс исследования текста (последующего узла исследования текстовых данных или узла анализа текстовых связей).

Если используется узел веб-фидов, нужно обязательно указать, что текстовое поле представляет **фактический текст** на узле исследования текстовых данных или на узле анализа текстовых связей, тем самым указав, что эти фиды ссылаются непосредственно на каждую статью или запись в блоге.

**Важно!** Если вы пытаетесь получить информацию по сети через прокси-сервер, нужно включить использование прокси-сервера в файле `net.properties` и для клиента, и для сервера IBM SPSS Modeler Text Analytics. Следуйте подробным указаниям в этом файле. Они применимы при доступе к сети через узел Веб-фид или при получении лицензии программы как службы (Software as a Service, SaaS) для SDL, поскольку такие соединения проходят через Java. По умолчанию этот файл расположен в каталоге *C:\Program Files\IBM\SPSS\Modeler\18\jre\lib\net.properties*.

Пример: Узел Веб-фид (фид RSS) с узлом моделирования Исследование текстовых данных

В качестве примера допустим, что мы соединяем узел Веб-фид с узлом Исследование текстовых данных для подачи текстовых данных из фида RSS в процесс исследования текстовых данных.

1. **Узел веб-фид (вкладка Входные данные).** Сначала мы добавляем этот узел в поток, чтобы указать, где находится содержимое веб-фида, и проверить структуру содержимого. На первой вкладке мы задаем URL для фида RSS. Поскольку наш пример предназначен для фида RSS, форматирование уже определено, и нам не нужно вносить никаких изменений на вкладке Записи. Для фидов RSS доступен необязательный алгоритм фильтрации содержимого, однако в этом случае мы его не применяем.
2. **Узел Text Mining (вкладка полей).** Далее мы добавляем и подключаем к узлу веб-фидов узел исследования текстовых данных. На этой вкладке мы определяем текстовое поле, выводимое узлом веб-фидов. В этом случае мы хотим использовать поле **Описание**. Также мы выбираем опцию Текстовое поле представляет **фактический текст**, а также задаем другие параметры.
3. **Узел Text Mining (вкладка Модель).** Далее (на вкладке Модель) мы выбираем режим и ресурсы построения. В этом случае мы выбираем построение модели понятий непосредственно с этого узла при помощи шаблона ресурсов по умолчанию.

Дополнительную информацию об использовании узла Исследование текста смотрите в разделе “Режим моделирования Text Mining” на стр. 20.

# Глава 3. Исследование концепций и категорий

Узел моделирования исследования текста используется для создания одного из двух слепков модели исследования текста:

- *Слепки модели понятий* вскрывают и извлекают основные концепты из структурированных или неструктурированных текстовых данных.
- *Слепки модели категорий* оценивают документы и записи и распределяют их по категориям, созданным из извлеченных понятий (и паттернов).

Все извлеченные концепты и паттерны, а также категории из ваших слепков моделей, можно объединять с существующими структурированными данными, такими как демографические, и применять к моделированию при помощи полного комплекта инструментов исследования данных IBM SPSS Modeler для получения более качественных и специализированных решений. Например, если заказчики часто указывают на проблемы с регистрацией как на главное затруднение в завершении задач управления оперативной учетной записью, можно вставить в ваши модели “проблемы с регистрацией”.

Кроме этого, узел моделирования исследования текста полностью интегрирован в IBM SPSS Modeler, поэтому можно внедрить потоки исследования текста через IBM SPSS Modeler Solution Publisher для оценки в реальном времени неструктурированных данных в прикладных программах, таких как PredictiveCallCenter. Возможность внедрить эти потоки обеспечивает успешные реализации замкнутого исследования текста. Например, сейчас в вашей организации может производиться анализ блокнотных записей о входящих и исходящих звонках с применением предсказательных моделей для повышения точности маркетинговых сообщений в реальном времени. Показано, что использование результатов исследования текста в потоках повышает точность предсказательных моделей данных.

*Примечание*: Чтобы запустить IBM SPSS Modeler Text Analytics с IBM SPSS Modeler Solution Publisher, добавьте каталог `<каталог_установки>/ext/bin/spss.TMWBServer` в переменную среды `$LD_LIBRARY_PATH`.

В IBM SPSS Modeler Text Analytics мы часто говорим об извлеченных понятиях и категориях. Важно понимать значение этих терминов, понятие и категория, так как они помогут вам принимать более осмысленные решения при исследовательской работе и при построении моделей.

Понятия и слепки модели понятий

В процессе извлечения текстовые данные сканируются и анализируются, чтобы идентифицировать отдельные интересные или значимые слова, такие как `выборы` или `мир`, и словосочетания, такие как `президентские выборы`, `выборы президента` или `мирный договор`. Эти слова и словосочетания вместе называются *терминами*. С использованием лингвистический ресурсов соответствующие термины извлекаются, а аналогичные термины собираются вместе с главным термином, называемым **концепт**.

Таким образом, понятие может представлять несколько терминов в зависимости от текста и используемого набора лингвистических ресурсов. Например, рассмотрим исследование удовлетворенности сотрудников, в котором извлечено понятие `зарплата`. Просматривая записи, связанные с понятием `зарплата`, вы замечаете, что термина `зарплата` в тексте часто нет, зато записи содержат сходные термины, такие как `оклад`, `ставка` и `заработная плата`. Эти термины группируются вокруг термина `зарплата`, поскольку механизм извлечения определил их как сходные или синонимичные, опираясь на правила обработки или на лингвистические ресурсы. В этом случае любые документы или записи, содержащие любые из этих терминов, обрабатываются так, как будто они содержат слово `зарплата`.

Если нужно увидеть, какие термины группируются в концепт, этот концепт можно исследовать в интерактивной инструментальной среде или посмотреть, какие синонимы показаны в модели концепта. Дополнительную информацию смотрите в разделе “Составные термины в моделях понятий” на стр. 34.

**Слепок модели понятий** содержит набор понятий, которые можно использовать для идентификации записей или документов, также содержащих концепт (включая все все его синонимы или сгруппированные термины). Модель концепта можно использовать двумя способами. Первый способ - исследовать и проанализировать концепты, обнаруженные в исходном тексте, для быстрой идентификации интересных документов. Вторая возможность - применить эту модель к записям или документам для быстрой идентификации тех же ключевых понятий в данных блокнотов из колл-центров.

Дополнительную информацию смотрите в разделе “Слепок Text Mining: модель понятий” на стр. 31.

Категории и слепки модели категорий

Можно создавать **категории**, по сути представляющие собой концепты более высокого уровня или темы для захвата ключевых идей, знаний и выраженных в тексте позиций. Категории состоят из набора дескрипторов, таких как *концепты*, *типы* и *правила*. Все вместе дескрипторы применяются для идентификации, принадлежит ли запись или документ к данной категории. Документ или запись можно просканировать для определения, существуют ли текстовые совпадения с дескриптором. Если совпадение найдено, документ/запись назначается данной категории. Этот процесс называется **категоризацией**.

Категории можно построить автоматически, используя надежный набор автоматизированных средств этого продукта, и вручную, используя дополнительное понимание этих данных, которое может у вас быть, а также используя оба этих подхода. На вкладке Модель этого узла можно загрузить также набор предварительно построенных категорий из пакета текстового анализа. Создание категорий вручную или уточнение категорий можно произвести только в интерактивной инструментальной среде. Дополнительную информацию смотрите в разделе “Узел Text Mining: вкладка Модель” на стр. 24.

**Слепок модели категорий** содержит набор категорий вместе с их дескрипторами. Эту модель можно использовать для категоризации набора документов или записей на основе текста каждого документа/записи. Всякий документ или запись читается, а затем назначается любой категории, для которой обнаружено совпадение дескриптора. Это значит, что документ или запись можно назначить нескольким категориям. Слепки модели категорий можно использовать для нахождения существенного нового содержания в ответах продолжающегося опроса или, например, в наборе записей блога.

Дополнительную информацию смотрите в разделе “Слепок Text Mining: модель категорий” на стр. 40.

## Режим моделирования Text Mining

Узел Text Mining использует лингвистические и частотные методы для извлечения ключевых понятий из текста и создания категорий при помощи этих понятий и других данных. При помощи этого узла можно изучить содержимое текстовых данных и создать слепок модели понятий или слепок модели категорий. При выполнении этого узла моделирования внутренний лингвистический механизм извлечения извлекает и группирует понятия, паттерны и/или категории, используя методы обработки естественного языка.

Можно выполнить узел Text Mining и автоматически создать слепок модели понятий или категорий с использованием опции **Генерировать непосредственно**. Другой вариант - более ручной, исследовательский подход в режиме **Построить интерактивно**, в котором можно не только извлечь понятия, создать категории и уточнить лингвистические ресурсы, но также выполнить анализ текстовых связей и изучить кластеры. Дополнительную информацию смотрите в разделе “Узел Text Mining: вкладка Модель” на стр. 24.

Этот узел можно найти на вкладке IBM SPSS Modeler Text Analytics палитры узлов в нижней части окна IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “IBM SPSS Modeler Text Analytics узлов” на стр. 8.

**Требования.** Узлы моделирования Text Mining принимают текстовые данные от узла Веб-фид, узла Список файлов или любого узла стандартного источника. Этот узел устанавливается вместе с IBM SPSS Modeler Text Analytics и доступен на палитре IBM SPSS Modeler Text Analytics.

**Примечание:** Этот узел замещает узел Извлечение текста для всех пользователей и старый узел Text Mining для пользователей японского языка, который предлагался в прошлых версиях Text Mining для Clementine. Если у вас есть старые потоки, использующие эти узлы или слепки моделей, нужно перестроить эти потоки с использованием нового узла Text Mining.

## Узел Text Mining: вкладка Поля

На вкладке Поля можно задать параметры полей для данных, из которых будут извлекаться понятия. При работе с большими наборами данных есть смысл выше этого узла поместить узел выборки, чтобы ускорить обработку. Дополнительную информацию смотрите в разделе “Добавление расположенного выше узла выборки для экономии времени” на стр. 30.

Можно задать следующие параметры:

**Текстовое поле.** Выберите поле, содержащее текст для исследования, имя пути к документу или имя пути к каталогу документов. Это поле зависит от источника данных.

**Что представляет текстовое поле.** Укажите, что именно содержит текстовое поле, заданное в предыдущем параметре. Варианты выбора:

- **Фактический текст.** Выберите эту опцию, если данное поле содержит именно тот текст, из которого должны извлекаться понятия.
- **Имена путей для документов.** Выберите эту опцию, если данное поле содержит одно или несколько имен путей к положениям, в которых находятся текстовые документы.

**Тип документа.** Эта опция доступна, только если вы задали, что текстовое поле представляет **Пути к документам**. Тип документа определяет структуру текста. Выберите один из следующих типов:

- **Полный текст.** Используется для большинства документов или текстовых источников. Для извлечения просматривается весь массив текста. В отличие от других опций, для этой опции нет никаких дополнительных параметров.
- **Структурированный текст.** Используется для библиографических форм, патентов и любых файлов, содержащих регулярные структуры, которые могут быть выявлены и проанализированы. Этот тип документов используется для полного или частичного пропуска процесса извлечения. Это позволяет определять разделители терминов, назначать типы и задавать минимальное значение частоты. Если выбрана эта опция, нажмите кнопку **Параметры** и введите текстовые разделители в области **Форматирование структурированного текста** диалогового окна Параметры документа. Дополнительную информацию смотрите в разделе “Вкладка Параметры документов для полей” на стр. 22.
- **Текст XML.** Используется для задания тегов XML, содержащих извлекаемый текст. Все остальные теги игнорируются. Если выбрана эта опция, нажмите кнопку **Параметры** и укажите явным образом элементы XML, содержащие читаемый текст, в области **Форматирование текста XML** диалогового окна Параметры документа. Дополнительную информацию смотрите в разделе “Вкладка Параметры документов для полей” на стр. 22.

**Общность текста.** Эта опция доступна только в том случае, если вы указали, что текстовое поле представляет **Имена путей к документам**, и выбрали **Полный текст** как тип документа. Выберите режим извлечения из числа следующих возможностей:

- **Режим документов.** Используйте этот режим для коротких и семантически однородных документов, например, для сообщений информационных агентств.
- **Режим абзацев.** Используйте для веб-страниц и документов без тегов. Процесс извлечения семантически разделяет документы, используя преимущества таких характеристик, как внутренние теги и синтаксис. При выборе этого режима скоринг выполняется по абзацам. Поэтому, например, правило `яблоко & апельсин` выполняется только в том случае, если `яблоко` и `апельсин` найдены в одном абзаце.

  **Примечание:** Из-за того способа, которым текст извлекается из документов PDF, **Режим абзацев** для этих документов не работает. Это связано с тем, что при извлечении подавляется маркер возврата каретки.

**Параметры режима абзацев.** Эта опция доступна только в том случае, если вы указали, что текстовое поле представляет **Имена путей к документам**, и выбрали для опции текстовой однородности значение **Режим абзацев**. Задайте пороговые значения числа символов, которые будут использоваться при любом извлечении. Фактический размер будет округляться до ближайшей точки. Для обеспечения репрезентативности связывания слов, полученного из текста собрания документов, не задавайте слишком маленький размер извлечения.

- **Минимум.** Задайте минимальное число символов, используемых для извлечения.
- **Максимум.** Задайте максимальное число символов, используемых для извлечения.

**Кодировка ввода.** Эта опция доступна, только если вы указали, что текстовое поле представляет **Пути к документам**. Она задает кодировку текста по умолчанию. Для всех языков, кроме японского, выполняется преобразование из заданной или распознанной кодировки в кодировку `ISO-8859-1`. Таким образом, даже если вы зададите другую кодировку, механизм извлечения перед обработкой преобразует ее в `ISO-8859-1`. Все символы, которые не соответствуют формату кодировки `ISO-8859-1`, будут преобразованы в пробелы. Для японского текста можно выбрать одну из нескольких опций кодировки: `SHIFT_JIS`, `EUC_JP`, `UTF-8` или `ISO-2022-JP`.

**Режим разделения.** При помощи режима разделения можно выбирать использование разделов с учетом параметров узла типа или выбрать другое разделение. Разделение помещает одну часть данных в обучающую выборку, а другую - в тестовую выборку.

## Вкладка Параметры документов для полей

Форматирование структурированного текста

Если требуется полностью или частично пропустить процесс извлечения в связи с тем, что вы работаете со структурированными данными, или нужно применить правила обработки текста, воспользуйтесь опцией типа документа **Структурированный текст** и объявите поля или теги, содержащие текст, в разделе **Форматирование структурированного текста** диалогового окна Параметры документа. Извлеченные термины будут браться только из текста, содержащегося внутри объявленных полей или тегов (включая дочерние теги). Все необъявленные поля и теги будут проигнорированы.

В некоторых контекстах лингвистическая обработка не требуется, и лингвистический механизм извлечения можно заменить явными объявлениями. В файле библиографии, где поля ключевых слов разделяются такими разделителями, как точка с запятой (;) или запятая (,), достаточно извлечь строку, содержащуюся между двумя разделителями. Поэтому можно приостановить процесс полного извлечения и вместо этого задать специальные правила обработки для объявления разделителей терминов, назначить типы для извлеченного текста или же задать минимальное значение частоты для извлечения.

Используйте следующие правила при объявлении элементов структурированного текста:

- Можно объявить только одно поле, тег или элемент на строку. Они не обязательно должны присутствовать в данных.
- Объявления регистрозависимы.
- Если объявляется тег с атрибутами, например, `<title id="1234">`, и требуется включить все вариации или, в данном случае, все ID, добавьте этот тег без атрибута или закрывающей угловой скобки (`>`), то есть в виде `<title`
- Добавьте двоеточие после имени поля или тега, чтобы указать, что это структурированный текст. Двоеточие можно добавить непосредственно после поля или тега, но обязательно перед разделителями, типами и значениями частот, например, `автор:` или `<место>:`.
- Чтобы указать, что в поле или теге содержится несколько терминов, и что для обозначения самостоятельных терминов используются разделители, объявите разделитель после двоеточия, например, `автор:,` или `<раздел>:;`.
- Чтобы назначить тип содержимому, найденному в теге, объявите имя типа после двоеточия и разделителя, например, `автор:,Сотрудник` или `<место>:;Положение`. При объявлении типов используйте имена в том виде, как они выводятся в редакторе ресурсов.

- Чтобы задать минимальную частоту для поля или тега, объявите ее значение в конце строки, например, `автор:,Сотрудник1` или `<место>:;Положение5`. Здесь `n` - значение заданной вами частоты, термины, найденные в поле или теге, должны для их извлечения встречаться, как минимум, с частотой `n` во всем наборе документов или записей. Для этого также требуется задать используемый разделитель.
- Если имеется тег, содержащий двоеточие, надо поставить перед двоеточием символ обратной косой черты, чтобы объявление не было проигнорировано. Например, если у вас есть поле с именем `<тема:источник>`, введите его как `<тема\:источник>`.

Для иллюстрации синтаксиса предположим, что у вас есть следующие повторяющиеся библиографические поля:

```
автор:Морел, Кавасима
аннотация:В этой статье описывается, как надо объявлять поля.
публикация:Документация по исследованию текстов
дата публикации:Март, 2010
```

В этом примере, если требуется в процессе извлечения сосредоточить внимание на авторах и аннотации и проигнорировать остальное содержимое, надо объявить только следующие поля:

```
автор:,Сотрудник1
аннотация:
```

В данном примере объявление поля `автор:,Сотрудник1` указывает, что лингвистическая обработка для содержимого этого поля приостановлена. Вместо этого констатируется, что поле "автор" содержит несколько имен, разделенных запятыми, что эти имена требуется назначить типу Сотрудник и что если такое имя встретится хотя бы однажды во всем наборе документов или записей, его надо извлечь. Поскольку поле `аннотация:` выводится без каких-либо других объявлений, это поле при извлечении будет просмотрено и к нему будут применены стандартная лингвистическая обработка и назначение типа.

Форматирование текста XML

Если требуется ограничить процесс извлечения только текстом внутри определенных тегов XML, выберите опцию типа документа **текст XML** и объявите теги, содержащие этот текст, в разделе **Форматирование текста XML** диалогового окна Параметры документа. Извлеченные термины будут получены только из текста, содержащегося внутри этих тегов или их дочерних тегов.

**Важно!** Если нужно пропустить процесс извлечения и задать правила для разделителей терминов, назначить типы для извлеченного текста или задать значение частоты для извлекаемых терминов, воспользуйтесь описанной ниже опцией **Структурированный текст**.

Используйте следующие правила при объявлении тегов для форматирования текста XML:

- Можно объявить только один тег XML на строку.
- Элементы тегов регистрозависимы.
- Если у тега есть атрибуты, например, `<title id="1234">`, и требуется включить все вариации или, в данном случае, все ID, добавьте этот тег без атрибута или закрывающей угловой скобки (`>`), то есть в виде `<title`

Для иллюстрации синтаксиса предположим, что у вас есть следующий документ XML:

```
<section>Правила дорожного движения
     <title id="01234">Сигналы светофора</title>
     <p>Дорожные знаки полезны.</p>
</section>
<p>Изучение правил важно.</p>
```

В этом примере надо объявить следующие теги:

```
<section>
<title
```

В этом примере, поскольку вы объявили тег <section>, текст в этом теге и его вложенных тегах, Сигналы светофора и Дорожные знаки полезны, будет просматриваться в процессе извлечения. Однако текст Изучение правил важно будет проигнорирован, поскольку тег <p> не был объявлен явным образом и он не вложен в объявленный тег.

## Узел Text Mining: вкладка Модель

На вкладке Модель можно задать метод построения и общие параметры модели, выводимой узлом.

Можно задать следующие параметры:

**Имя модели** Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

**Использовать разделенные данные.** Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

**Режим построения.** Задает способ создания слепков модели при выполнении потока с этим узлом Text Mining. Другой вариант - более ручной, исследовательский подход в режиме **Построить интерактивно**, в котором можно не только извлечь понятия, создать категории и уточнить лингвистические ресурсы, но также выполнить анализ текстовых связей и изучить кластеры.

- **Построить интерактивно.** Во время выполнения потока эта опция запускает интерактивный интерфейс, в котором можно извлечь понятия и паттерны, изучить и тонко настроить результаты извлечения, построить и уточнить категории, тонко настроить лингвистические ресурсы (паттерны, синонимы, типы, библиотеки и т. д.) и построить слепки модели категорий. Дополнительную информацию смотрите в разделе “Построить интерактивно” на стр. 25.
- **Генерировать непосредственно.** Эта опция показывает, что во время выполнения потока нужно автоматически создать модель и добавить ее на палитру моделей. В отличие от интерактивной инструментальной среды, от вас не требуется дополнительных манипуляций во время выполнения, кроме настроек, задаваемых в узле. При выборе этой опции выводятся те опции, которые относятся к конкретной модели, при помощи которых можно определить тип модели, которую нужно создать. Дополнительную информацию смотрите в разделе “Генерировать непосредственно” на стр. 26.

**Копировать ресурсы из.** При исследовании текста извлечение учитывает не только параметры на вкладке Эксперт, но и лингвистические ресурсы. Эти ресурсы служат источником базовых сведений при извлечении из текста понятий, типов и, в некоторых случаях, паттернов. Ресурсы можно скопировать в этот узел из шаблона ресурса или из пакета анализа текста (text analysis package, TAP). Выберите одно из этого и щелкните по **Загрузить**, чтобы определить пакет или шаблон, из которого будет скопирован ресурс. В момент загрузки копия ресурсов сохраняется в узле. Поэтому, если понадобится использовать обновленный шаблон или TAP, нужно будет перезагрузить его, здесь или в сеансе интерактивной инструментальной среды. Для вашего удобства в узле выводится дата и время копирования и загрузки ресурсов. Дополнительную информацию смотрите в разделе “Копирование ресурсов из шаблонов и файлов TAP” на стр. 26.

**Язык текста.** Идентифицирует язык исследуемого текста. Скопированные в данный узел ресурсы управляют представленными опциями языков. Можно или выбрать язык, для которого были настроены эти ресурсы, или выбрать опцию **ВСЕ**. Настоятельно рекомендуется точно задавать язык для текстовых данных, однако при неуверенности можно выбрать опцию **ВСЕ**. Для текстов на японском языке опция **ВСЕ** недоступна. Опция **ВСЕ** увеличивает время извлечения, так как при ее применении сначала используется автоматическое распознавание языка с просмотром всех документов и записей для идентификации языка текстов. При включении этой опции все записи или документы на поддерживаемых и лицензированных языках читаются механизмом извлечения с использованием внутренних словарей на конкретных языках. Дополнительную информацию смотрите в разделе “Идентификатор языка” на стр. 212. Обратитесь к торговому представителю, если хотите приобрести лицензию на поддерживаемый язык, к которому в настоящее время не имеете доступа.

## Построить интерактивно

На вкладке Модель узла моделирования исследования текста можно выбрать режим построения для слепков модели. Если выбрать **Построить интерактивно**, при выполнении потока откроется интерактивный интерфейс. В этой интерактивной инструментальной среде можно:

- Извлекать и изучать результаты, включая понятия и назначение типов в поисках значимых идей в текстовых данных.
- Использовать ряд методов для построения и расширения категорий по понятиям, типам, паттернам TLA и правилам, чтобы оценить документы и записи по отношению к этим категориям.
- Уточнять лингвистические ресурсы (шаблоны ресурсов, библиотеки, словари, синонимы и другие), чтобы улучшить результаты в интерактивном процессе, в котором понятия извлекаются, изучаются и уточняются.
- Выполнять анализ текстовых связей (Text Link Analysis, TLA) и использовать обнаруженные паттерны TLA для построения лучших слепков модели категорий. Узел Text Link Analysis не поддерживает аналогичные опции изучения или возможности моделирования.
- Генерировать кластеры, чтобы обнаружить новые взаимосвязи и изучить взаимосвязи между понятиями, типами, паттернами и категориями на панели визуализации.
- Генерировать уточненные слепки модели категорий для палитры моделей в IBM SPSS Modeler и использовать их в других потоках.

**Примечание:** Интерактивную модель нельзя построить при создании задания IBM SPSS Collaboration and Deployment Services.

**Использовать работу сеанса (категории, TLA, ресурсы и т.д.) из последнего изменения узла.** При работе в сеансе интерактивной инструментальной среды можно изменить данные сеанса для узла (параметры извлечения, ресурсы, определения категорий и т. д.). При помощи опции **Использовать работу сеанса** можно перезапустить интерактивную инструментальную среду с использованием сохраненных данных сеанса. Эта опция недоступна при первом использовании этого узла, поскольку сохраненных данных сеанса еще нет. Как изменить данные сеанса для узла с использованием этой опции, описано в разделе “Обновление узлов моделирования и сохранение” на стр. 84.

Если сеанс запущен *с этой опцией*, то при следующем запуске сеанса будут доступны параметры извлечения, категории, ресурсы и любая другая работа, сохраненная при последнем изменении узла в сеансе интерактивной интерактивный инструментальной среды. Поскольку при этой опции используются сохраненные данные сеанса, некоторое содержимое, такое как ресурсы, скопированные из приведенного ниже шаблона, и другие вкладки недоступны и игнорируются. Но если запустить сеанс *без этой опции*, то используется только содержимое узла, определенное сейчас, и тем самым прошлая работа, выполненная в инструментальной среде, недоступна.

*Примечание*: Если изменить узел источника для потока после кэширования результатов извлечения при включенной опции **Использовать работу сеанса...**, потребуется новая извлечение после запуска сеанса интерактивной инструментальной среды, если будут нужны обновленные результаты извлечения.

**Пропустить извлечение и повторно использовать кэшированные данные и результаты.** В сеансе интерактивной инструментальной среды можно повторно использовать любые результаты извлечения и данные. Эта опция особенно полезна, когда нужно сэкономить время и повторно использовать результаты извлечения, а не дожидаться полного нового извлечения после запуска сеанса. Чтобы воспользоваться этой опцией, нужно сначала обновить этот узел из сеанса интерактивной инструментальной среды при выбранной опции **Сохранить работу сеанса и кэшировать текстовые данные вместе с результатами извлечения для повторного использования**. Как изменить данные сеанса для узла с использованием этой опции, описано в разделе “Обновление узлов моделирования и сохранение” на стр. 84.

**Начать сеанс с.** Выберите опцию, указывающую представление и действие, с которого нужно начинать после запуска сеанса интерактивной инструментальной среды. Независимо от того, в каком представлении начат сеанс, вы сможете переключиться на любое представление после запуска сеанса.

- **Использование результатов извлечения для построения категорий.** Эта опция запускает интерактивную инструментальную среду в представлении Категории и понятия и, если применимо, выполняет извлечение. В этом представлении можно создать категории и сгенерировать модель категорий. Кроме того, можно переключиться в другое представление. Дополнительную информацию смотрите в разделе Глава 8, “Режим интерактивного сеанса инструментальной среды”, на стр. 73.
- **Изучение результатов Text Link Analysis (TLA).** Эта опция запускает представление анализа текстовых связей (Text Link Analysis, TLA) и начинает его с извлечения и идентификации взаимосвязей между понятиями в тексте, таких как мнения и другие связи. Чтобы использовать эту опцию и получить результаты, нужно выбрать паттерн или пакет Text Analysis Package, содержащий правила паттернов TLA. Если вы работаете с большими наборами данных, извлечение TLA может занять некоторое время. В этом случае есть смысл использовать расположенный выше узел выборки. Дополнительную информацию смотрите в разделе Глава 12, “Изучаем анализ текстовых связей (Text Link Analysis, TLA)”, на стр. 153.
- **Анализ кластеров совместного появления.** Эта опция запускается в представлении кластеров и обновляет все устаревшие результаты извлечения. В этом представлении можно выполнить анализ кластеров совместного появления, при котором создается набор кластеров. Кластеризация совместного появления начинается с оценки силы связи между двумя понятиями, исходя из числа их совместных появлений в данной записи или документе, и завершается группировкой сильно связанных понятий в кластеры. Дополнительную информацию смотрите в разделе Глава 8, “Режим интерактивного сеанса инструментальной среды”, на стр. 73.

## Генерировать непосредственно

На вкладке Модель узла моделирования исследования текста можно выбрать режим построения для слепков модели. Если выбрать **Генерировать непосредственно**, можно задать опции в узле и просто выполнить поток. На выходе будет слепок модели понятия, который непосредственно поместить на палитру моделей. В отличии от интерактивной инструментальной среды, от вас не требуется дополнительных манипуляций во время выполнения, кроме настроек частоты, задаваемых для этой опции в узле.

**Максимальное число понятий, которые можно включить в модель.** Эта опция, применяемая только при автоматическом (не интерактивном) построении модели, показывает, что нужно создать модель понятий. Кроме того, она объявляет о том, что модель должна содержать не более указанного числа понятий.

- **Включать понятия на основе высокой частоты. Максимальное число понятий.** Включено будет заданное число понятий, начиная с понятия с наибольшей частотой. Здесь частота - это число раз, которое понятие (и все его термины) встречается во всем наборе документов/записей. Это число может оказаться выше числа записей, поскольку понятие может встретиться несколько раз в одной записи.
- **Исключать понятия со слишком большим числом записей. Процент записей.** Исключает понятия с процентом числа записей выше указанного числа. Эта опция полезна для исключения понятий, которые часто встречаются в тексте или в записях, но не имеют значения для анализа.

**Оптимизировать для скорости скоринга.** Эта опция, которая по умолчанию выбрана, нацелена на создание компактное модели с быстрыми оценками. Если отменить выбор этой опции, будет создана модель большего размера с более медленными оценками. Зато в модели большего размера начальные оценки сгенерированной модели понятия совпадают с оценками, полученными при скоринге того же текста с использованием слепка модели.

## Копирование ресурсов из шаблонов и файлов TAP

При исследовании текста извлечение учитывает не только параметры на вкладке Эксперт, но и лингвистические ресурсы. Эти ресурсы служат источником базовых сведений при извлечении из текста понятий, типов и, в некоторых случаях, паттернов. Ресурсы можно скопировать в этот узел из *шаблона ресурса*, а если вы находитесь в узле Text Mining, можно также выбрать *пакет анализа текста* (text analysis package, TAP).

По умолчанию при добавлении узла на холст в узел копируются ресурсы из базового шаблона для языка, лицензированного для вашего продукта. Если у вас лицензия на несколько языков, для автоматической загрузки используется шаблон первого выбранного языка.

В момент загрузки копия выбранных ресурсов сохраняется в узле. Содержимое шаблона или TAP просто копируется, без связывания шаблона или TAP с узлом. Это значит, что в дальнейшем в случае изменения этого шаблона или TAP изменения не станут автоматически доступны в узле. Другими словами, ресурсы, загруженные в узел, продолжают использоваться, пока вы не перезагрузите копию шаблона или TAP или не измените узел Text Mining, выбрав опцию **Использовать работу сеанса**. Дополнительную информацию о том, как **Использовать работу сеанса**, смотрите далее в этом разделе.

Выбирая шаблон или TAP, используйте тот, язык которого совпадает с языком текстовых данных. Вам доступны шаблоны и файлы TAP только на тех языках, на которые у вас есть лицензия. Если нужно выполнить анализ текстовых связей, выберите паттерн, содержащий паттерны TLA. Если паттерн содержит паттерны TLA, выводится значок в столбце TLA в диалоговом окне Загрузить паттерн ресурса.

**Примечание:** Нельзя загрузить файлы TAP в узел Text Link Analysis.

Шаблоны ресурсов

Шаблон ресурса - это заранее определенный набор библиотек и дополнительных лингвистические и нелингвистических ресурсов, тонко настроенных на конкретный домен использования. В узле моделирования Text Mining копия ресурсов из базового шаблона уже загружена при добавлении узла в поток, но вы можете изменить шаблоны или загрузить пакет Text Analysis Package, выбрав **Шаблон ресурса** или **Text Analysis Package** и щелкнув по **Загрузить**. Шаблон можно выбрать диалоговом окне Загрузить шаблон ресурсов.

*Примечание*: Если вы не видите нужный шаблон в списке, но на компьютере есть экспортированная копия, можно импортировать шаблон теперь. Кроме того, в этом диалоговом окне шаблон можно экспортировать для совместного использования с другими пользователями. Дополнительную информацию смотрите в разделе “Импорт и экспорт шаблонов” на стр. 177.

Пакеты Text Analysis Package (TAP)

Пакет анализа текста (text analysis package, TAP) - это заранее определенный набор библиотек и дополнительных лингвистические и нелингвистических ресурсов, увязанных с одним или несколькими наборами заранее определенных категорий. IBM SPSS Modeler Text Analytics предлагает ряд заранее построенных TAP для текстов на английском и японском, тонко настроенных для конкретных доменов. Эти TAP редактировать нельзя, но можно использовать как старт для построения своей модели категорий. Кроме того, вы можете создавать свои TAP в интерактивном сеансе. Дополнительную информацию смотрите в разделе “Загрузка пакетов анализа текста” на стр. 142.

**Примечание:** Нельзя загрузить файлы TAP в узел Text Link Analysis.

Использование опции "Использовать работу сеанса" (вкладка Модель)

Хотя ресурсы копируются в узел на вкладке Модель, иногда в дальнейшем нужно вносить изменения в ресурсы в интерактивном сеансе и обновлять узел моделирования Text Mining с учетом этих новейших изменений. В этом случае следует выбрать опцию **Использовать работу сеанса** на вкладке Модель узла моделирования Text Mining.

Если выбрать **Использовать работу сеанса**, становится недоступна кнопка **Загрузить** в узле, что показывает, что вместо ресурсов, загруженных сюда ранее, будут использоваться эти ресурсы, полученные из интерактивной инструментальной среды.

Чтобы внести изменения в ресурсы после выбора опции **Использовать работу сеанса**, можно отредактировать или непосредственно переключить ресурсы в сеансе интерактивной инструментальной среды в представлении Редактор ресурсов. Дополнительную информацию смотрите в разделе “Изменение ресурсов узла после загрузки” на стр. 175.

## Слепок Text Mining: вкладка Эксперт

Вкладка Эксперт содержит некоторые дополнительные параметры, которые влияют на извлечение и обработку текста. Параметры в этом диалоговом окне управляют базовым поведением, а также некоторыми дополнительными стратегиями извлечения. Но они представляют лишь часть доступных вам опций. Есть еще целый ряд лингвистических ресурсов и опций, влияющих на результаты извлечения и управляемых шаблоном ресурса, который выбирается на вкладке Модель. Дополнительную информацию смотрите в разделе “Узел Text Mining: вкладка Модель” на стр. 24.

**Примечание:** Вся эта вкладка недоступна, если выбран режим **Построить интерактивно** с использованием сохраненной информации интерактивной инструментальной среды на вкладке Модель; в этом случае параметры извлечения берутся из последнего сохраненного сеанса инструментальной среды.

**Для текста на голландском, английском, французском, немецком, итальянском, португальском и испанском**

Приведенные ниже параметры можно задать при каждой извлечения для языков, кроме японского, таких как английский, испанский, французский, немецкий и так далее:

**Примечание:** Информацию об экспертных параметрах для текстов на японском смотрите ниже в этом разделе.

**Ограничить извлечение понятиями с глобальной частотой не менее [n].** Задает минимальное число раз, которое слово или фраза должны встретиться в тексте для их извлечения. В данном случае значение 5 ограничивает извлечение словами и фразами, которые встретились во всем массиве записей или документов не менее пяти раз.

В некоторых случаях изменение этого предела может сильно повлиять на результаты извлечения и, соответственно, на создаваемые вами категории. Допустим, что вы обрабатываете данные о ресторанах и ограничились для этой опции значением 1. В этом случае в результатах извлечения вы можете обнаружить: *пицца (1), тонкая пицца (2), пицца со шпинатом (2)* и *любимая пицца (2)*. Однако если вы задали предел 5 или более для глобальной частоты для извлечения и выполните извлечение повторно, то больше не увидите три последние понятия. Вместо этого вы получите *пицца (7)*, поскольку *пицца* - это наиболее простая форма и это слово уже фигурировало в качестве возможного кандидата. Не исключено, что в тексте могут быть и другие фразы со словом "пицца", и тогда вы получите фактическую частоту больше семи. Кроме того, если выражение *пицца со шпинатом* уже использовалось в качестве дескриптора категории, вам, возможно, надо будет добавить дескриптор *пицца*, чтобы не захватывать все записи. Поэтому если категории уже созданы, изменять значение этого предела следует с осторожностью.

Обратите внимание, что это эта функция относится только к извлечению. Если шаблон содержит термины (обычно это так) и в тексте обнаружен термин из этого шаблона, этот термин будет проиндексирован независимо от его частоты.

Например, пусть вы работаете с шаблоном Базовые ресурсы, который содержит в типе <Положение> в главной библиотеке термин "Лос-Анджелес". Даже если в вашем документе Лос-Анджелес упоминается только один раз, он будет включен в список понятий. Чтобы избежать этого, надо задать фильтр для вывода только понятий, которые встречаются в тексте, как минимум, столько раз, сколько задано значением в поле **Ограничить извлечение понятиями с глобальной частотой не менее [n]**.

**Допускать ошибки пунктуации.** Эта опция временно нормализует текст, содержащий ошибки пунктуации (например, неверно используемые знаки препинания) во время извлечения, чтобы повысить извлекаемость понятий. Эта опция особенно полезна для коротких текстов низкого качества (например, ответы при опросе с произвольным ответом, электронная переписка, данные CRM), а также для текста, содержащего много сокращений.

**Допускать орфографические ошибки при минимальном числе символов корня [n].** Эта опция применяет метод нечеткой группировки, который помогает группировать в одну концепцию слова, которые часто пишутся с

ошибками, а также вариативные написания слова. Алгоритм нечеткой группировки перед сравнением временно удаляет из извлеченных слов все гласные (кроме первой) и двойные или тройные согласные, так что тунель и тоннель попадут в одну группу. Методы нечеткой группировки, однако, не применяются, если различным терминам назначены различные типы, кроме типа <Неизвестный>.

Кроме того, можно задать минимально необходимое число символов *корня* при использовании нечеткой группировки. Число символов корня в термине рассчитывается как общее число символов минус число символов окончания; кроме того, в случае термина-словосочетания вычитаются детерминативы и предлоги. Например, в термине упражнения будет насчитано 9 символов корня "упражнени", поскольку буква *я* на конце слова относится к окончанию множественного числа. Аналогичным образом в пакет яблок насчитывается 10 символов корня ("пакет яблок"), а в магнитола для автомобиля насчитывается 17 символов корня ("магнитол автомобил"). Этот метод подсчета используется только при проверке применимости нечеткой группировки и не используется в алгоритмах сравнения слов.

*Примечание*: Если окажется, что некоторые слова группируются неправильно, такие пары слов можно исключить из метода при помощи явного объявления в разделе **Нечеткая группировка: исключения** на вкладке Расширенные ресурсы. Дополнительную информацию смотрите в разделе "Нечеткая группировка" на стр. 206.

**Извлечь одиночные термины.** Эта опция извлекает отдельные слова (одиночные термины), если слово не входит в словосочетание и если это существительное или нераспознанная часть речи.

**Извлечь нелингвистические объекты.** Эта опция извлекает нелингвистические объекты, такие как номера телефонов, номера социального страхования, время, даты, валюты, цифры, проценты, адреса электронной почты и HTTP-адреса. Вы можете включить или исключить те или иные типы нелингвистических объектов в разделе **Нелингвистические объекты: конфигурация** на вкладке Расширенные ресурсы. Выключив ненужные объекты, вы сэкономите время обработки механизмом извлечения. Дополнительную информацию смотрите в разделе "Конфигурация" на стр. 210.

**Алгоритм верхнего регистра.** Эта опция извлекает простые и составные термины, не входящие во встроенные словари, если первая буква термина - в верхнем регистре. Это хороший способ извлечь большинство имен собственных.

**Группировать частичные и полные личные имена, где возможно.** Эта опция группирует имена, которые по-разному появляются в тексте. Эта возможность полезна, поскольку имена часто употребляются в начале текста в полной форме, а затем - в краткой. Эта опция пытается сопоставить каждый одиночный термин с типом <Неизвестный> последнему слову в любом составном термине, типизированном как <Личный>. Например, если найден терм *иванов*, получивший вначале тип <Неизвестный>, механизм извлечения поищет составные термины в типе <Личный>, содержащие *иванов* как последнее слово, например, *александр иванов*. Эта опция применяется только к фамилии, поскольку первое имя почти никогда не извлекается как одиночный термин.

**Максимум неслужебных слов при перестановке.** Эта опция задает максимально допустимое число неслужебных слов при применении метода перестановки. Этот метод перестановок группирует как близкие словосочетания, содержащие в своем составе одни и те же неслужебные слова, если игнорировать форму слова. Например, если задать ограничение в два неслужебных слова, будут обработаны такие извлеченные словосочетания, как компания клиенту и клиенту от нашей компании. В этом примере такие словосочетания будут сгруппированы в итоговом списке понятий, поскольку считаются одинаковыми, если проигнорировать слова от нашей.

**Примечание:** Чтобы была доступна извлечение результатов Text Link Analysis, нужно начать сеанс с опцией **Изучение результатов анализа текстовых связей**; кроме того, нужно выбрать ресурсы, содержащие определения TLA. Результаты TLA можно извлечь и позже, в сеансе интерактивной инструментальной среды, в диалоговом окне Параметры извлечения. Дополнительную информацию смотрите в разделе "Извлечение данных" на стр. 88.

Для текста на японском

Для текста на японском опции этого диалогового окна другие, что отражает отличия при обработке японского языка. Кроме того, для работы с японскими текстами нужно выбрать шаблон или Text Analysis Package, настроенные на японский язык, на вкладке Модель этого узла. Дополнительную информацию смотрите в разделе “Копирование ресурсов из шаблонов и файлов TAP” на стр. 26.

**Вторичный анализ.** При извлечения базовые ключевые слова извлекаются при помощи набора типов по умолчанию. Но, если выбрать тот или иной вторичный анализатор, можно получить много дополнительных и более богатых понятий, поскольку теперь экстрактор будет учитывать частицы и вспомогательные глаголы как часть концепции. Кроме того, при анализе эмоциональной окраски можно включить большое число дополнительных типов. После выбора вторичного анализатора можно сгенерировать результаты Text Link Analysis.

**Примечание:** При вызове вторичного анализатора извлечение занимает больше времени.

- **Анализ зависимостей**. При выборе этой опции извлечение понятий производится с дополнительным учетом частиц по сравнению с извлечением базовых типов и ключевых слов. Кроме того, при анализе зависимостей можно получить более богатые результаты паттернов TLA.
- **Анализ эмоциональной окраски**. При выборе этого анализатора извлекаются дополнительные понятия и, если применимо, результаты паттернов TLA. Помимо базовых типов можно воспользоваться более чем 80 типами эмоциональной окраски . При помощи таких типов можно раскрывать в тексте понятия и паттерны, выражающие эмоции, настроения и мнения. Фокус анализа эмоциональной окраски управляется тремя опциями: **Все эмоциональные окраски**, **Только репрезентативные эмоциональные окраски** и **Только заключения**.
- **Без вторичного анализатора**. Эта опция отключает все вторичные анализаторы. Эта опция скрыта, если выбрана опция **Изучение результатов Text Link Analysis (TLA)** на вкладке Модель, поскольку вторичный анализатор необходим для получения результатов. Если выбрать эту опцию, но позже выбрать опцию **Изучение результатов Text Link Analysis (TLA)**, при выполнении потока возникнет ошибка.

## Добавление расположенного выше узла выборки для экономии времени

Если у вас большие объемы данных, обработка может занимать минуты и даже часы, особенно в сеансе интерактивной инструментальной среды. Чем больше размер данных, тем больше времени уходит на извлечение и категоризацию. Для эффективной работы можно выше вашего узла Text Mining добавить узел выборки IBM SPSS Modeler. При помощи узла выборки можно выполнить случайную выборку и использовать для первых прогонов лишь часть документов или записей.

Небольшая выборка часто подходит для решений по редактированию ресурсов и даже созданию многих, а то и всех категорий. Добившись удовлетворительных результатов для небольшого набора данных, вы можете применить те же методы для создания категорий всего набора данных. Затем можно найти документы или записи, не отвечающие созданным категориям, и произвести необходимые корректировки.

**Примечание:** Узел выборки - стандартный узел IBM SPSS Modeler.

## Использование узла Text Mining в потоке

Узел моделирования Text Mining служит в потоке для доступа к данным и извлечения понятий. Для доступа к данным можно использовать любой узел источника, например, узел База данных, узел Файл переменных, узел Веб-фид или узел Фиксированный файл. Для текста во внешних документах можно использовать узел Список файлов.

Пример 1: Узел Список файлов и узел Text Mining используются для непосредственного построения слепка модели понятий

В следующем примере показано, как Использовать узел Список файлов и узел моделирования Text Mining, чтобы сгенерировать слепок модели понятий. Дополнительную информацию об использовании узла Список файлов смотрите в разделе “Узел списка файлов” на стр. 11.

1. **Узел Список файлов (вкладка параметров).** Первым действием мы добавили этот узел в поток, чтобы задать, где хранятся текстовые документы. Мы выбрали каталог, содержащий все документы, для которых нужно выполнить исследование текста.
2. **Узел Text Mining (вкладка полей).** Затем мы добавили и подключили к узлу Список файлов узел Text Mining. В этом узле мы определили формат ввода, шаблон ресурса и формат вывода. Мы выбрали имя поля, созданного узлом список файлов, и выбрали опцию, в которой текстовое поле представляет **имена и пути к документам**, а также другие параметры. Дополнительную информацию смотрите в разделе “Использование узла Text Mining в потоке” на стр. 30.
3. **Узел Text Mining (вкладка Модель)**. Затем на вкладке Модель мы выбрали режим построения, чтобы сгенерировать слепок модели понятий непосредственно из этого узла. Вы можете выбрать другой шаблон ресурсов или оставить базовые ресурсы.

Пример 2: Узлы Файл Excel и Text Mining используются для интерактивного построения модели категорий

Этот пример показывает, как при помощи узла Text Mining можно запустить сеанс интерактивной инструментальной среды. Дополнительную информацию об интерактивной инструментальной среде смотрите в разделе Глава 8, “Режим интерактивного сеанса инструментальной среды”, на стр. 73.

1. **Узел источника Excel (вкладка данных).** Во-первых, мы добавили этот узел в поток, чтобы задать, где хранится текст.
2. **Узел Text Mining (вкладка полей).** Затем мы добавили и подключили узел Text Mining. На этой первой вкладке мы определили формат ввода. Мы выбрали имя поля из узла источника и выбрали опцию Текстовое поле представляет **Фактический текст**, поскольку данные берутся непосредственно из узла источника Excel.
3. **Узел Text Mining (вкладка Модель).** Затем на вкладке Модель мы выбрали интерактивное построение слепка модели категорий и использование результатов извлечения для автоматического построения категорий. В этом примере мы загрузили копию ресурсов и набор категорий из пакета Text Analysis Package.
4. **Сеанс интерактивной инструментальной среды**. Затем мы выполнили поток, и открылся интерфейс интерактивной инструментальной среды. После выполнения извлечения мы приступили к изучению данных и совершенствованию категорий.

## Слепок Text Mining: модель понятий

Слепок модели понятий создается при каждом успешном выполнении узла модели Text Mining, в котором вы выбрали опцию **Сгенерировать модель непосредственно** на вкладке Модель. Слепок модели понятий Text Mining служит для обнаружения в реальном времени ключевых понятий в других текстовых данных, например, в данных блокнотов из колл-центров.

Сам слепок модели понятий состоит из списка понятий, для которых заданы типы. Для скоринга других данных можно выбрать любые или все понятия в этой модели. Во время выполнения потока, содержащего слепок модели Text Mining, в данные добавляются новые поля согласно режиму построения, выбранному на вкладке Модель узла моделирования Text Mining перед построением модели. Дополнительную информацию смотрите в разделе “Модель понятий: вкладка Модель” на стр. 32.

Если слепок модели создан с использованием переведенных документов, оценка будет выполнена в языке перевода. Аналогично этому, если слепок модели сгенерирован для английского языка, можно задать в этом слепке язык перевода, поскольку эти документы будут затем переведены на английский язык.

Сгенерированные слепки моделей исследования текста помещаются на палитре слепков моделей (расположенной на вкладке Модели в верхней правой части) окна IBM SPSS Modeler).

Просмотр результатов

Чтобы получить информацию о слепке модели, щелкните правой кнопкой мыши по узлу в палитре слепков моделей и выберите **Обзор** из контекстного меню (или **Правка** для узлов в потоке).

Добавление моделей в потоки

Чтобы добавить слепок модели в поток, щелкните по значку на палитре слепков моделей и затем щелкните по тому месту холста потока, в которое хотите поместить узел. Другой вариант - щелкните правой кнопкой по значку и выберите в контекстном меню **Добавить в поток**. Затем подключите свой поток к этому узлу - теперь вы готовы к передаче данных для генерирования прогнозов.

**Внимание**: Если нужно использовать слепок скоринга, чтобы сгенерировать заново узел моделирования, содержащий и модель категорий, и используемый шаблон, вместо узла моделирования рекомендуется создать TAP и использовать его в интерактивном сеансе перед генерированием слепка скоринга.

## Модель понятий: вкладка Модель

В моделях понятий вкладка Модель содержит набор извлеченных понятий. Понятия представлены в табличном формате, каждое в своей строке таблицы. Цель этой вкладки - выбрать нужные понятия для скоринга.

*Примечание*: Информация на этой вкладке будет другой, если сгенерировать слепок модели категорий. Дополнительную информацию смотрите в разделе “Слепок модели категорий: вкладка Модель” на стр. 40.

По умолчанию выбраны понятия из начала списка, что показывают переключатели в крайнем слева столбце. Включенный переключатель означает, что понятие будет использоваться для скоринга. Выключенный переключатель означает, что понятие будет исключено из скоринга. Можно включить сразу несколько строк, выделив их и щелкнув по одному из переключателей в выделении.

Дополнительную информацию о понятии можно найти в следующих столбцах:

**Понятие.** Это извлеченное ведущее слово или словосочетание. В некоторых случаях понятие представляет имя понятия и некоторые связанные с ним термины. Чтобы вывести участвующие в понятии термины, откройте панель терминов на этой вкладке и выберите понятие; соответствующие термины появятся в нижней части диалогового окна. Дополнительную информацию смотрите в разделе “Составные термины в моделях понятий” на стр. 34.

**Глобальная.** Это глобальная частота; столько раз встречается понятие (и все его термины) во всем наборе документов/записей.

- **Столбчатая диаграмма.** Глобальная частота этого понятия в текстовых данных представлена как столбчатая диаграмма. Для наглядного разделения типов тип понятия показан цветом столбика.
- **%.** Глобальная частота этого понятия в текстовых данных представлена как процент.
- **N.** Фактическое число вхождений этого понятия в текстовых данных.

**Документы.** Здесь Документы - это число документов или записей, в которых встречается понятие и все его термины.

- **Столбчатая диаграмма.** Число документов для этого понятия в виде столбчатой диаграммы. Для наглядного разделения типов тип понятия показан цветом столбика.
- **%.** Число документов для этого понятия как процент.
- **N.** Фактическое число документов или записей, содержащих это понятие.

**Тип.** Тип, назначенный для понятия. Для каждого понятия цвет столбцов Глобальная и Документы показывает тип, заданный для этого понятия. **Тип** объединяет понятия по их смыслу. Дополнительную информацию смотрите в разделе “Словари типов” на стр. 189.

Щелкнув правой кнопкой по ячейке в таблице, можно вывести контекстное меню, в котором можно:

- **Выделить все.** Выделяются все строки в таблице.
- **Копировать.** Выделенные понятия копируются в буфер обмена.
- **Копировать с полями** Выделенные понятия копируются в буфер обмена вместе с заголовками столбцов.
- **Включить выделенное.** Включает переключатели во всех выделенных строках таблицы; этим понятия включаются в скоринг.
- **Выключить выделенное.** Выключает переключатели во всех выделенных строках таблицы.
- **Включить все.** Включает все переключатели в таблице. В результате все понятия будут использоваться в окончательном выводе.
- **Выключить все.** Выключает все переключатели в таблице. Выключение понятия означает, что он не используется в окончательном выводе.
- **Включить понятия.** Открывает диалоговое окно Включить понятия. Дополнительную информацию смотрите в разделе “Опции для включения понятий в скоринг”.

## Опции для включения понятий в скоринг

Чтобы быстро включить или выключить понятия для использования в скоринге, нажмите на панели инструментов кнопку **Включить понятия**.

*Рисунок 1. Панель инструментов - кнопка Включить понятия*

Щелчок по этой кнопке на панели инструментов откроет диалоговое окно Включить понятия, в котором можно выбрать понятия на основе правил. В скоринг будут включены все понятия, у которых включен переключатель на вкладке Модель. Примените правило в этом диалоговом подокне, чтобы изменить понятия, используемые в скоринге.

Для выбора доступны следующие опции:

**Включать понятия на основе высокой частоты. Максимальное число понятий.** Включено будет это число понятий, начиная с понятия с наибольшей глобальной частотой. Здесь частота - это число раз, которое понятие (и все его термины) встречается во всем наборе документов/записей. Это число может оказаться больше числа записей, поскольку понятие может встретиться несколько раз в одной записи.

**Включать понятия на основе числа документов. Минимальное число.** Это минимально необходимое число документов для включения понятия. Здесь число документов - это число документов/записей, в которых встречается понятие и все его термины.

**Включать понятия, для которых задан тип.** Выберите тип в выпадающем списке, чтобы включить все понятия, для которых задан этот тип. Тип назначается понятиям автоматически во время извлечения. **Тип** объединяет понятия по их смыслу. К типам относятся понятия высокого уровня, слова и спецификаторы с положительной и отрицательной оценкой, контекстные спецификаторы, имена людей, названия мест и организаций и другое. Дополнительную информацию смотрите в разделе “Словари типов” на стр. 189.

**Исключать понятия со слишком большим числом записей. Процент записей.** Исключает понятия с процентом числа записей выше указанного числа. Эта опция полезна для исключения понятий, которые часто встречаются в тексте или в записях, но не имеют значения для анализа.

**Исключать понятия, для которых задан тип.** Исключает понятия, соответствующие типу, который вы выбрали в выпадающем списке.

### Составные термины в моделях понятий

Можно просмотреть составные термины, определенные для понятий, которые выбрали в таблице. Щелкая на панели инструментов по переключателю составных терминов, можно выводить таблицу составных терминов на разделенной панели в нижней части диалогового окна.

К таким терминам относятся синонимы, определенные в лингвистических ресурсах (независимо от того, найдены ли они в тексте), а также извлеченные формы единственного и множественного числа, найденные в тексте, по которому генерируется слепок модели, переставленные термины, термины из нечеткой группировки и так далее.

*Рисунок 2. Панель инструментов - кнопка Вывести составные термины*

*Примечание*: Редактировать список составных терминов нельзя. Этот список генерируется через подстановки, определения синонимов (в словаре подстановок), нечеткую группировку и другие средства, и все они определены в лингвистических ресурсах. Чтобы внести изменения в способ группировки терминов в концепцию или способ их обработки, нужно внести изменения в сами ресурсы (которые можно отредактировать в Редактор ресурсов в интерактивной инструментальной среде или в Редактор шаблонов, а затем перезагрузить в узле), а затем еще раз выполнить поток, чтобы получить новый слепок модели, с новыми результатами.

Щелкнув правой кнопкой по ячейке, содержащей составной термин или понятие, можно вывести контекстное меню со следующими пунктами:

- **Копировать.** Выбранная ячейка копируется в буфер обмена.
- **Копировать с полями.** Выбранная ячейка копируется в буфер обмена вместе с заголовками столбцов.
- **Выделить все.** Выделяются все ячейки в таблице.

## Модель понятий: вкладка параметров

Вкладка параметров служит для задания значений полей для новых входных данных, если потребуется. Кроме этого, здесь определяется модель данных для вывода (режим скоринга).

**Примечание:** Эта вкладка выводится только тогда, когда слепок модели помещается на холст. Если открыть это диалоговое окно непосредственно на палитре моделей, она не выводится.

### Режим скоринга: понятия как записи

При использовании этого режима скоринга новая запись создается для каждой пары `понятие/документ`. Обычно на выходе записей больше, чем на входе.

Кроме того, к данным, помимо входных полей, добавляются новые следующие поля:

*Таблица 4. Выходные поля в режиме "Понятия как записи".*

| Поле | Описание |
|---|---|
| `Понятие` | Содержит имя извлеченного понятия, найденное в поле текстовых данных. |
| `Тип` | Хранит тип понятия как полное имя типа, такое как *Расположение* или *Персональный*. **Тип** объединяет понятия по их смыслу. Дополнительную информацию смотрите в разделе “Словари типов” на стр. 189. |
| `Количество` | Выводит число вхождений этого понятия (и его терминов) в теле текста (запись/документ). |

При выборе этой опции становятся недоступны остальные опции, кроме **Допускать ошибки пунктуации**.

### Режим скоринга: понятия категории как поля

В моделях понятий для каждой входной записи создается новая запись для каждого понятия, найденного в данном документе. Поэтому выходных записей ровно столько, сколько входных. Но каждая запись, или строка таблицы содержит по одному новому полю, или столбцу для каждого выбранного (включенного переключателем) понятия на вкладке Модель. Значение для каждого поля понятия зависит от выбора одной из опций - **Флаги** или **Количества** - для значения поля на этой вкладке.

**Примечание:** Если вы работаете с очень большими наборами данных, например, при помощи базы данных DB2, в режиме **Понятия как поля** могут возникать ошибки обработки большого объема данных. В таком случае рекомендуется перейти на режим **Понятия как записи**.

**Значения полей.** Выберите использование количества или флага в новом поле для каждого понятия.

- **Флаги.** Эта опция служит для получения на выходе флагов с двумя различными значениями, например, *Да*/*Нет*, *Истина*/*Ложь*, *И*/*Л*, *1* и *2*. Типы хранения задаются автоматически с учетом выбранных значений. Например, если как флаговые значения ввести числа, они будут обрабатываться как целые значения. Типом хранения для флагов не может быть строка, целое число, действительное число или дата/время. Введите флаговое значение в качестве **True** и **False**.
- **Количества.** Служит для получения количества вхождений понятия в данной записи.

**Расширение имени поля.** Задайте расширение для имени поля. Имена полей генерируются с использованием имени понятия и этого расширения.

- **Добавить как.** Укажите, с какой стороны добавить расширение к имени поля. Выберите **Префикс**, чтобы добавить расширение в начало строки. Выберите **Суффикс**, чтобы добавить расширение в конец строки.

**Допускать ошибки пунктуации.** Эта опция временно нормализует текст, содержащий ошибки пунктуации (например, неверно используемые знаки препинания) во время извлечения, чтобы повысить извлекаемость понятий. Эта опция особенно полезна для коротких текстов низкого качества (например, ответы при опросе с произвольным ответом, электронная переписка, данные CRM), а также для текста, содержащего много сокращений.

**Примечание:** Опция **Допускать ошибки пунктуации** не применяется при работе с текстом на японском языке.

## Модель понятий: вкладка полей

Вкладка полей служит для задания значений полей для новых входных данных, если потребуется.

*Примечание*: Эта вкладка выводится только тогда, когда слепок модели помещается в поток. Если открыть вывод непосредственно на палитре моделей, она не выводится.

**Текстовое поле.** Выберите поле, содержащее текст для исследования, имя пути к документу или имя пути к каталогу документов. Это поле зависит от источника данных.

**Что представляет текстовое поле.** Укажите, что именно содержит текстовое поле, заданное в предыдущем параметре. Варианты выбора:

- **Фактический текст.** Выберите эту опцию, если данное поле содержит именно тот текст, из которого должны извлекаться понятия.
- **Имена путей для документов.** Выберите эту опцию, если данное поле содержит одно или несколько имен путей к положениям, в которых находятся текстовые документы.

**Тип документа.** Эта опция доступна, только если вы задали, что текстовое поле представляет **Пути к документам**. Тип документа определяет структуру текста. Выберите один из следующих типов:

- **Полный текст.** Используется для большинства документов или текстовых источников. Для извлечения просматривается весь массив текста. В отличие от других опций, для этой опции нет никаких дополнительных параметров.

- **Структурированный текст.** Используется для библиографических форм, патентов и любых файлов, содержащих регулярные структуры, которые могут быть выявлены и проанализированы. Этот тип документов используется для полного или частичного пропуска процесса извлечения. Это позволяет определять разделители терминов, назначать типы и задавать минимальное значение частоты. Если выбрана эта опция, нажмите кнопку **Параметры** и введите текстовые разделители в области **Форматирование структурированного текста** диалогового окна Параметры документа. Дополнительную информацию смотрите в разделе “Вкладка Параметры документов для полей” на стр. 22.
- **Текст XML.** Используется для задания тегов XML, содержащих извлекаемый текст. Все остальные теги игнорируются. Если выбрана эта опция, нажмите кнопку **Параметры** и укажите явным образом элементы XML, содержащие читаемый текст, в области **Форматирование текста XML** диалогового окна Параметры документа. Дополнительную информацию смотрите в разделе “Вкладка Параметры документов для полей” на стр. 22.

**Кодировка ввода.** Эта опция доступна, только если вы указали, что текстовое поле представляет **Пути к документам**. Она задает кодировку текста по умолчанию. Для всех языков, кроме японского, выполняется преобразование из заданной или распознанной кодировки в кодировку `ISO-8859-1`. Таким образом, даже если вы зададите другую кодировку, механизм извлечения перед обработкой преобразует ее в `ISO-8859-1`. Все символы, которые не соответствуют формату кодировки `ISO-8859-1`, будут преобразованы в пробелы. Для японского текста можно выбрать одну из нескольких опций кодировки: `SHIFT_JIS`, `EUC_JP`, `UTF-8` или `ISO-2022-JP`.

**Язык текста.** Идентифицирует язык исследуемого текста; это основной язык, обнаруженный при извлечении. Обратитесь к торговому представителю, если хотите приобрести лицензию на поддерживаемый язык, к которому в настоящее время не имеете доступа.

## Модель понятий: вкладка Сводка

На вкладке Сводка представлена информация о самой модели (папка *Анализ*), об используемых в ней полях (папка *Поля*), значениях параметров, используемых при построении модели (папка *Параметры построения*) и об обучении модели (папка *Сводка по обучению*).

При первом просмотре узла моделирования папки на вкладке Сводка свернуты. Чтобы увидеть нужные результаты, выведите их при помощи переключателя раскрытия и сворачивания слева от папки или выведите все результаты, нажав кнопку **Раскрыть все**. Чтобы скрыть результаты после просмотра, сверните конкретную папку при помощи переключателя раскрытия и сворачивания или нажмите кнопку **Свернуть все**, чтобы свернуть все результаты.

## Использование слепков модели понятий в потоке

При использовании узла моделирования Text Mining можно сгенерировать либо слепок модели понятий, либо слепок модели категорий (из сеанса интерактивной инструментальной среды). В следующем примере показано, как использовать слепок модели понятий в простом потоке.

### Пример: Узел Файл статистики со слепком модели понятий

В следующем примере показано, как использовать слепок модели понятий Text Mining.

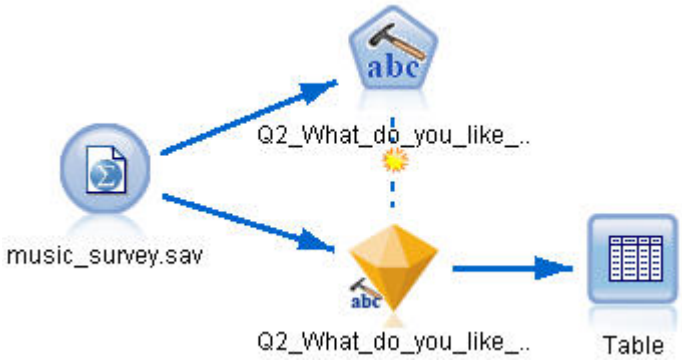



*Рисунок 3. Поток примера: Узел Файл статистики со слепком модели понятий Text Mining*

1. **Узел Файл статистики (вкладка данных).** Первым действием мы добавили этот узел в поток, чтобы задать, где хранятся текстовые документы.

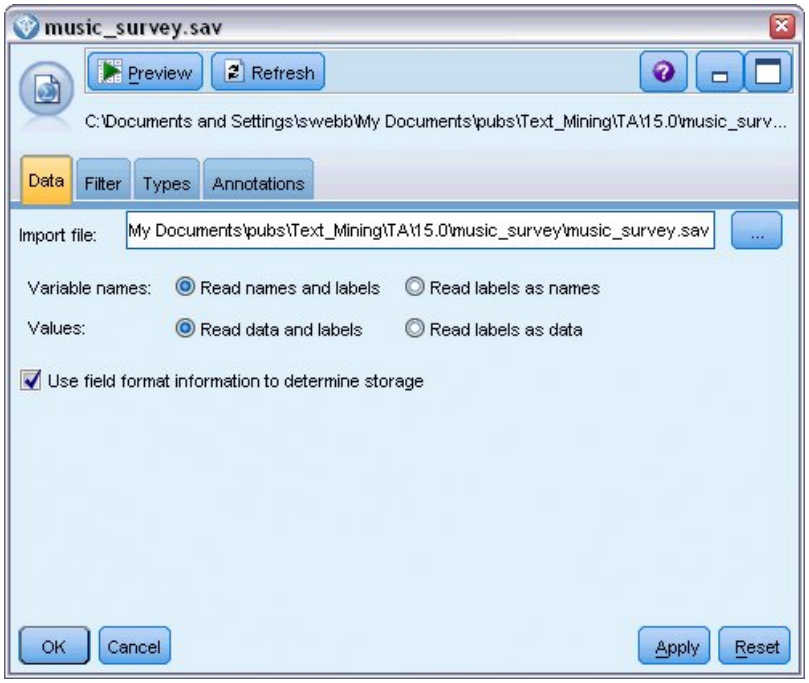



*Рисунок 4. Диалоговое окно узла Файл статистики: вкладка данных*

2. **Слепок модели понятий Text Mining (вкладка данных).** Затем мы добавили и подключили к узлу Файл статистики слепок модели понятий. Мы выбрали нужные понятия для скоринга наших данных.


*Рисунок 5. Диалоговое окно слепка модели категорий Text Mining: вкладка Модель*

3. **Слепок модели понятий Text Mining (вкладка параметров).** Затем мы определили формат вывода и выбрали вывод *понятий как полей*. По одному новому полю будет выведено для каждого понятия, выбранного на вкладке Модель. Каждое имя поля будет составлено из имени понятия и префикса "*Concept_*"

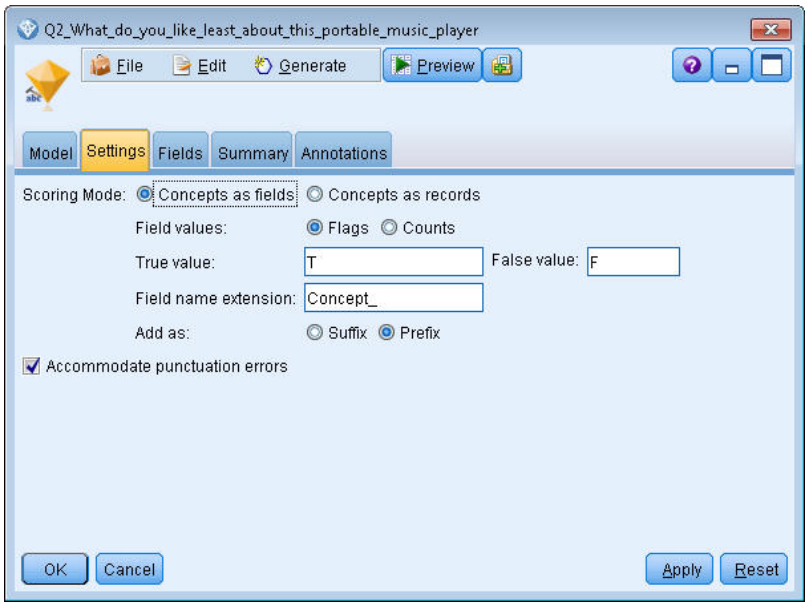



*Рисунок 6. Диалоговое окно слепка модели понятий Text Mining: вкладка параметров*

4. **Слепок модели понятий Text Mining (вкладка полей).** Затем мы выбрали текстовое поле с именем **Q2_What_do_you_like_least_about_this_portable_music_player** (вопрос 2, что вам меньше всего понравилось в этом портативном плеере), полученном из узла Файл статистики. Кроме того, мы выбрали опцию **Текстовое поле представляет: фактический текст**.

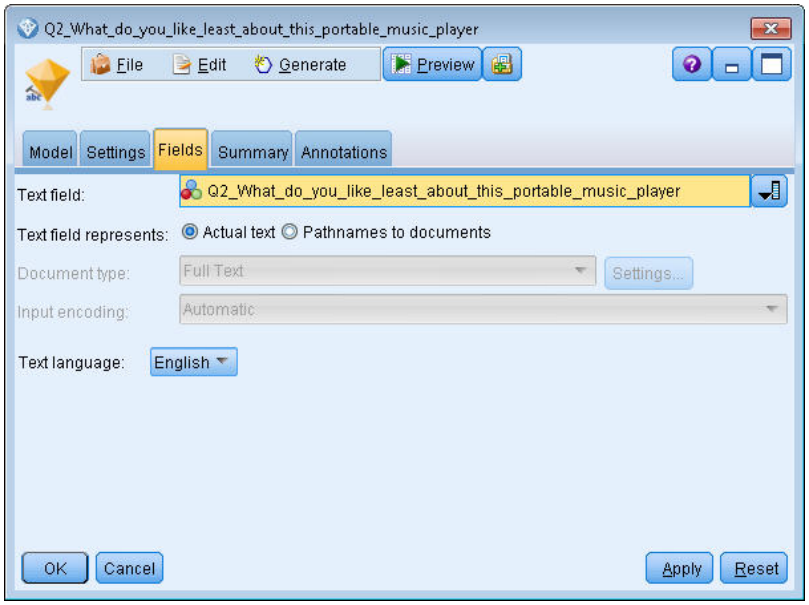
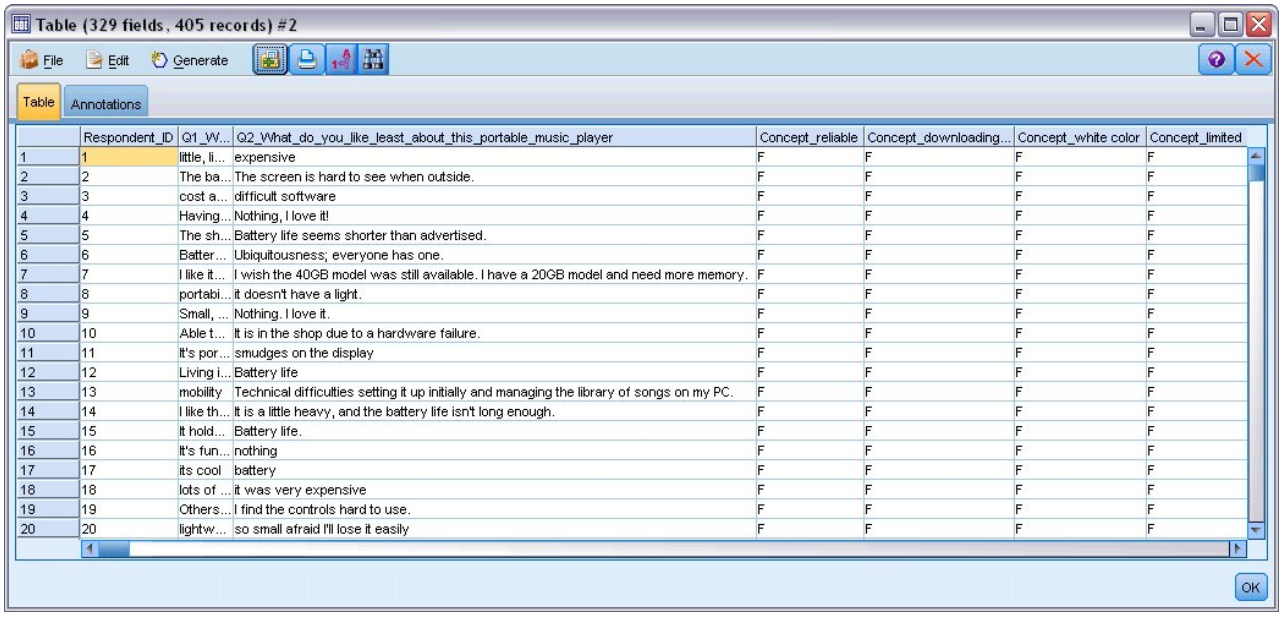



*Рисунок 7. Диалоговое окно слепка модели понятий Text Mining: вкладка полей*

5. **Узел таблицы.** Затем мы подключили узел таблицы, чтобы увидеть результаты, и выполнили поток. На экране откроется окно табличного вывода.




| | Respondent_ID | Q1_W... | Q2_What_do_you_like_least_about_this_portable_music_player | Concept_reliable | Concept_downloading... | Concept_white color | Concept_limited |
|---|---|---|---|---|---|---|---|
| 1 | 1 | little, li... | expensive | F | F | F | F |
| 2 | 2 | The ba... | The screen is hard to see when outside. | F | F | F | F |
| 3 | 3 | cost a... | difficult software | F | F | F | F |
| 4 | 4 | Having... | Nothing, I love it! | F | F | F | F |
| 5 | 5 | The sh... | Battery life seems shorter than advertised. | F | F | F | F |
| 6 | 6 | Batter... | Ubiquitousness; everyone has one. | F | F | F | F |
| 7 | 7 | I like it... | I wish the 40GB model was still available. I have a 20GB model and need more memory. | F | F | F | F |
| 8 | 8 | portabi... | it doesn't have a light. | F | F | F | F |
| 9 | 9 | Small, ... | Nothing. I love it. | F | F | F | F |
| 10 | 10 | Able t... | It is in the shop due to a hardware failure. | F | F | F | F |
| 11 | 11 | It's por... | smudges on the display | F | F | F | F |
| 12 | 12 | Living i... | Battery life | F | F | F | F |
| 13 | 13 | mobility | Technical difficulties setting it up initially and managing the library of songs on my PC. | F | F | F | F |
| 14 | 14 | I like th... | It is a little heavy, and the battery life isn't long enough. | F | F | F | F |
| 15 | 15 | It hold... | Battery life. | F | F | F | F |
| 16 | 16 | It's fun... | nothing | F | F | F | F |
| 17 | 17 | its cool | battery | F | F | F | F |
| 18 | 18 | lots of ... | it was very expensive | F | F | F | F |
| 19 | 19 | Others... | I find the controls hard to use. | F | F | F | F |
| 20 | 20 | lightw... | so small afraid I'll lose it easily | F | F | F | F |

*Рисунок 8. Табличный вывод, прокрученный до флагов понятий*

## Слепок Text Mining: модель категорий

Слепок модели категорий Text Mining создается каждый раз, когда вы генерируете модель из интерактивной инструментальной среды. Этот слепок модели содержит набор категорий, определения которых состоят из понятий, типов, паттернов TLA и/или правил категории. При помощи слепка категоризуются ответы на опросы, записи в блогах, веб-фиды и любые другие текстовые данные.

Если вы запустили сеанс интерактивной инструментальной среды в узле моделирования, то можете исследовать результаты извлечения, уточнять ресурсы, выполнять тонкую настройку категорий перед тем, как генерировать модели категорий. Во время выполнения потока, содержащего слепок модели Text Mining, в данные добавляются новые поля согласно режиму построения, выбранному на вкладке Модель узла моделирования Text Mining перед построением модели. Дополнительную информацию смотрите в разделе “Слепок модели категорий: вкладка Модель”.

Если слепок модели создан с использованием переведенных документов, оценка будет выполнена в языке перевода. Аналогично этому, если слепок модели сгенерирован для английского языка, можно задать в этом слепке язык перевода, поскольку эти документы будут затем переведены на английский язык.

Сгенерированные слепки моделей исследования текста помещаются на палитре слепков моделей (расположенной на вкладке Модели в верхней правой части) окна IBM SPSS Modeler).

Просмотр результатов

Чтобы получить информацию о слепке модели, щелкните правой кнопкой мыши по узлу в палитре слепков моделей и выберите **Обзор** из контекстного меню (или **Правка** для узлов в потоке).

Добавление моделей в потоки

Чтобы добавить слепок модели в поток, щелкните по значку на палитре слепков моделей и затем щелкните по тому месту холста потока, в которое хотите поместить узел. Другой вариант - щелкните правой кнопкой по значку и выберите в контекстном меню **Добавить в поток**. Затем подключите свой поток к этому узлу - теперь вы готовы к передаче данных для генерирования прогнозов.

**Внимание**: Если нужно использовать слепок скоринга, чтобы сгенерировать заново узел моделирования, содержащий и модель категорий, и используемый шаблон, вместо узла моделирования рекомендуется создать TAP и использовать его в интерактивном сеансе перед генерированием слепка скоринга.

## Слепок модели категорий: вкладка Модель

Для моделей категорий на вкладке Модель слева выводится список категорий в модели категорий, а справа - дескрипторы для выбранной категории. Каждая категория состоит из некоторого числа дескрипторов. Для каждой категории, которую вы выберите, соответствующие дескрипторы выводятся в таблице. Эти дескрипторы могут включать в себя понятия, правила категории, типы и паттерны TLA. Кроме того, показан тип каждого дескриптора, а также некоторые примеры того, что представляет дескриптор.

На этой вкладке ваша цель - выбрать категории, полезные для скоринга. Для модели категорий документы и записи оцениваются по категориям. Если документ или запись содержит в своем тексте или в любом термине один или несколько дескрипторов, такому документу или записи назначается категория, которой принадлежит этот дескриптор. К таким терминам относятся синонимы, определенные в лингвистических ресурсах (независимо от того, найдены ли они в тексте), а также извлеченные формы единственного и множественного числа, найденные в тексте, по которому генерируется слепок модели, термины с перестановками, термины из нечеткой группировки и так далее.

*Примечание*: Результаты на этой вкладке будут другими, если сгенерировать слепок модели понятий. Дополнительную информацию смотрите в разделе “Модель понятий: вкладка Модель” на стр. 32.

Дерево категорий

Чтобы вывести дополнительную информацию о каждой категории, выберите категорию и просмотрите сведения, которые появятся для дескрипторов в этой категории. Для каждого дескриптора можно просмотреть следующую информацию:

- Имя **дескриптора**. Это поле содержит значок, представляющий разновидность дескриптора и его имя.

*Таблица 5. Значки дескрипторов*

| | | | |
|---|---|---|---|
|  | Концепции |  | Паттерны TLA |
|  | Типы |  | Правила категории |

- **Тип.** Это поле содержит имя типа для дескриптора. Типы - это собрания близких понятий (семантических групп), например, названий организаций, продукты или положительные мнения. Правил не задаются для типов.
- **Подробности.** Это поле содержит список включенного в дескриптор. В зависимости от числа соответствий, для некоторых дескрипторов виден не весь список из-за ограничений на размер диалогового окна.

Выбор и копирование категорий

По умолчанию выбраны категории из начала списка, что показывают их переключатели на левой панели. Включенный переключатель означает, что категория будет использоваться для скоринга. Выключенный переключатель означает, что категория будет исключена из скоринга. Можно включить сразу несколько, выделив несколько строк и щелкнув по одному из переключателей в выделении. Кроме того, если некоторая категория или подкатегория выбрана, но одна из ее подкатегорий не выбрана, переключатель выводится с синим фоном, показывая, что выбрана только часть дочерних категорий выбранной категории.

Щелкнув правой кнопкой по категории в дереве, можно вывести контекстное меню, в котором можно:

- **Включить выделенное.** Включает переключатели во всех выделенных строках таблицы.
- **Выключить выделенное.** Выключает переключатели во всех выделенных строках таблицы.
- **Включить все.** Включает все переключатели в таблице. В результате все категории будут использоваться в окончательном выводе. Так же действует соответствующий значок переключателя на панели инструментов.
- **Выключить все.** Выключает все переключатели в таблице. Выключение категории означает, что она не используется в окончательном выводе. Так же действует соответствующий значок выключенного переключателя на панели инструментов.

Щелкнув правой кнопкой по ячейке в таблице дескрипторов, можно вывести контекстное меню со следующими действиями:

- **Копировать.** Выделенные понятия копируются в буфер обмена.
- **Копировать с полями.** Выделенный дескриптор копируется в буфер обмена вместе с заголовками столбцов.
- **Выделить все.** Выделяются все строки в таблице.

## Слепок модели категорий: вкладка Параметры

Вкладка параметров служит для задания значений полей для новых входных данных, если потребуется. Кроме этого, здесь определяется модель данных для вывода (режим скоринга).

**Примечание:** Эта вкладка выводится в диалоговом окне узла, только когда слепок модели помещается на холст или в поток. Если открыть слепок непосредственно на палитре моделей, она не выводится.

## Режим скоринга: категории как поля

При выборе этой опции выходных записей ровно столько, сколько входных. Но каждая запись содержит по одному новому полю для каждой выбранной (включенным переключателем) категории на вкладке Модель. Для каждого поля введите значения флага в качестве **True** и **False**, например, *Да*/*Нет*, *Истина*/*Ложь*, *И*/*Л*, *1* и *2*. Типы хранения задаются автоматически с учетом выбранных значений. Например, если как флаговые значения ввести числа, они будут обрабатываться как целые значения. Типом хранения для флагов не может быть строка, целое число, действительное число или дата/время.

**Примечание:** Если вы работаете с очень большими наборами данных, например, при помощи базы данных DB2, в режиме **Категории как поля** могут возникать ошибки обработки большого объема данных. В таком случае рекомендуется перейти на режим **Категории как записи**.

**Расширение имени поля.** Можно задать использование расширяющего префикса/суффикса для имени поля или использование кодов категорий. Имена полей генерируются с использованием имени категории и этого расширения.

- **Добавить как.** Укажите, с какой стороны добавить расширение к имени поля. Выберите **Префикс**, чтобы добавить расширение в начало строки. Выберите **Суффикс**, чтобы добавить расширение в конец строки.

**Если выбор подкатегории отменен.** При помощи этой опции можно задать, как нужно обработать дескрипторы, принадлежащие тем подкатегориям, которые не были включены в скоринг. Есть две опции.

- Опция **Полностью исключить из оценки ее дескрипторы** заставляет игнорировать и не использовать при оценке дескрипторы подкатегорий со снятыми отметками переключателей (отмена выбора).
- Опция **Агрегировать дескрипторы с дескрипторами в родительской категории** заставляет использовать дескрипторы подкатегорий со снятыми отметками переключателей (отмена выбора) в качестве дескрипторов родительской категории (категория, включающая данную подкатегорию). В случае отмены выбора нескольких уровней подкатегорий, их дескрипторы будут свернуты в ближайшую доступную родительскую категорию.

**Допускать ошибки пунктуации.** Эта опция временно нормализует текст, содержащий ошибки пунктуации (например, неверно используемые знаки препинания) во время извлечения, чтобы повысить извлекаемость понятий. Эта опция особенно полезна для коротких текстов низкого качества (например, ответы при опросе с произвольным ответом, электронная переписка, данные CRM), а также для текста, содержащего много сокращений.

**Примечание:** Опция **Сглаживать ошибки пунктуации** не применяется при работе с текстом на японском языке.

## Режим скоринга: категории как записи

При использовании этой опции новая запись создается для каждой пары `категория`, `документ`. Обычно на выходе записей больше, чем на входе. Кроме того, в зависимости от разновидности модели к данным, помимо входных полей, добавляются новые поля.

*Таблица 6. Выходные поля в режиме "Категории как записи".*

| Новое выходное поле | Описание |
|---|---|
| `Категория` | Содержит имя категории, назначенной для текстового документа. Если категория - подкатегория другой категории, то полный путь к имени категории зависит от значения, выбираемого в этом диалоговом окне. |

**Значения для иерархических категорий.** Эта опция управляет тем, как в результатах будут выводиться имена подкатегорий.

- **Полный путь категории.** При выборе этой опции имена категорий будут выводиться с полным путем родительских категорий, если они есть, а для разделения имен категорий и подкатегорий будет использоваться дробная черта.
- **Короткий путь категории.** При выборе этой опции будут выводиться только имена категорий, но при наличии родительских категорий будет вставляться многоточие.
- **Категории нижнего уровня.** При выборе этой опции будет выводиться только имя категория без полного пути и без обозначения наличия родительских категорий.

**Если выбор подкатегории отменен.** При помощи этой опции можно задать, как нужно обработать дескрипторы, принадлежащие тем подкатегориям, которые не были включены в скоринг. Есть две опции.

- Опция **Полностью исключить из оценки ее дескрипторы** заставляет игнорировать и не использовать при оценке дескрипторы подкатегорий со снятыми отметками переключателей (отмена выбора).
- Опция **Агрегировать дескрипторы с дескрипторами в родительской категории** заставляет использовать дескрипторы подкатегорий со снятыми отметками переключателей (отмена выбора) в качестве дескрипторов родительской категории (категория, включающая данную подкатегорию). В случае отмены выбора нескольких уровней подкатегорий, их дескрипторы будут свернуты в ближайшую доступную родительскую категорию.

**Допускать ошибки пунктуации.** Эта опция временно нормализует текст, содержащий ошибки пунктуации (например, неверно используемые знаки препинания) во время извлечения, чтобы повысить извлекаемость понятий. Эта опция особенно полезна для коротких текстов низкого качества (например, ответы при опросе с произвольным ответом, электронная переписка, данные CRM), а также для текста, содержащего много сокращений.

**Примечание:** Опция **Допускать ошибки пунктуации** не применяется при работе с текстом на японском языке.

## Слепок модели категорий: вкладка Другое

Вкладка полей и вкладка параметров для слепка модели категорий такие же, как для слепка модели понятий.

- Вкладка полей. Дополнительную информацию смотрите в разделе “Модель понятий: вкладка полей” на стр. 35.
- Вкладка сводки. Дополнительную информацию смотрите в разделе “Модель понятий: вкладка Сводка” на стр. 36.

## Использование слепков модели категорий в потоке

Слепок модели категорий Text Mining генерируется из сеанса интерактивной инструментальной среды. Этот слепок модели можно использовать в потоке.

### Пример: Узел Файл статистики со слепком модели категорий

В следующем примере показано, как использовать слепок модели Text Mining.

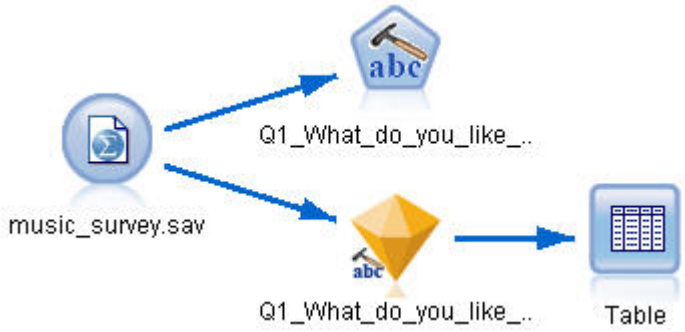



*Рисунок 9. Поток примера: Узел Файл статистики со слепком модели категорий Text Mining*

1. **Узел Файл статистики (вкладка данных).** Первым действием мы добавили этот узел в поток, чтобы задать, где хранятся текстовые документы.

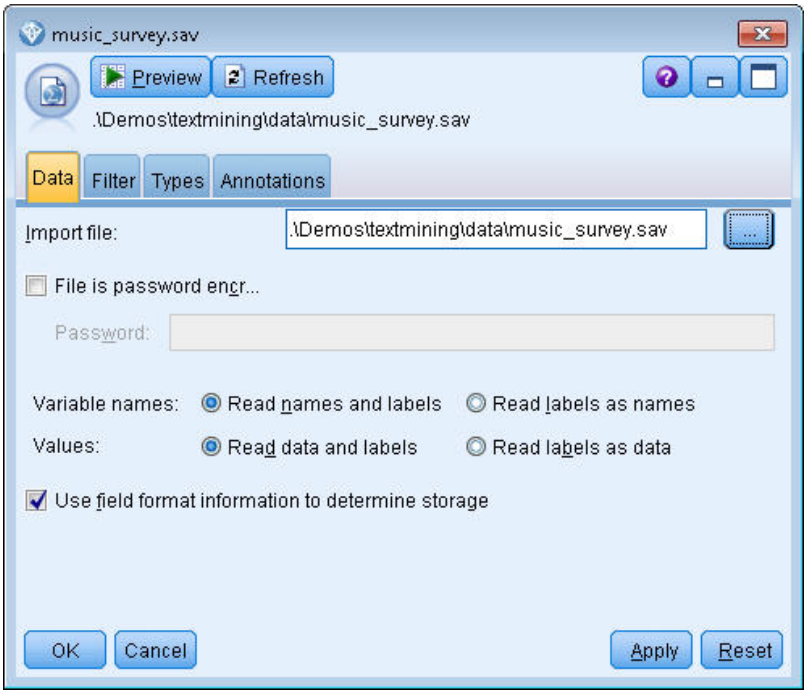



*Рисунок 10. Диалоговое окно узла Файл статистики: вкладка данных*

2. **Слепок модели категорий Text Mining (вкладка данных).** Затем мы добавили и подключили к узлу Файл статистики слепок модели категорий. Мы выбрали нужные категории для скоринга наших данных.

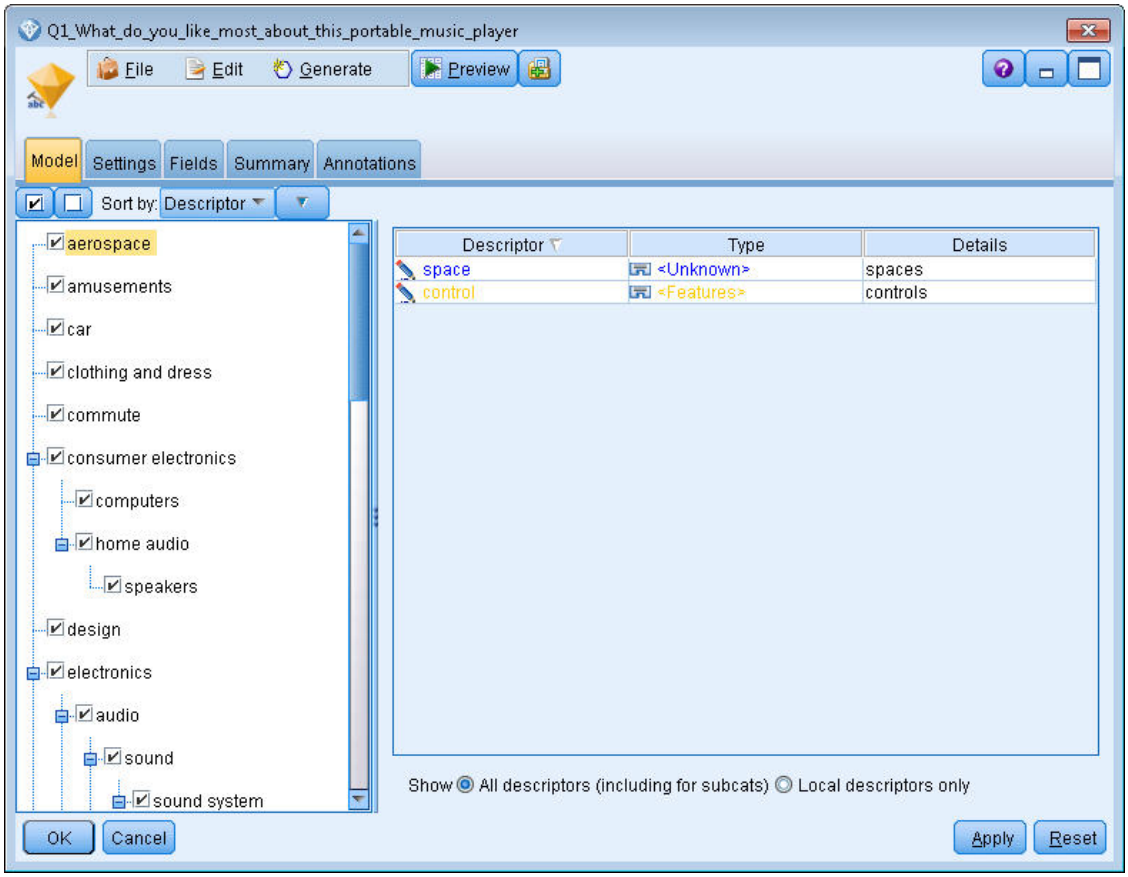



*Рисунок 11. Диалоговое окно слепка модели категорий Text Mining: вкладка Модель*

3. **Слепок модели Text Mining (вкладка параметров).** Затем мы определили формат вывода **категорий как полей**.

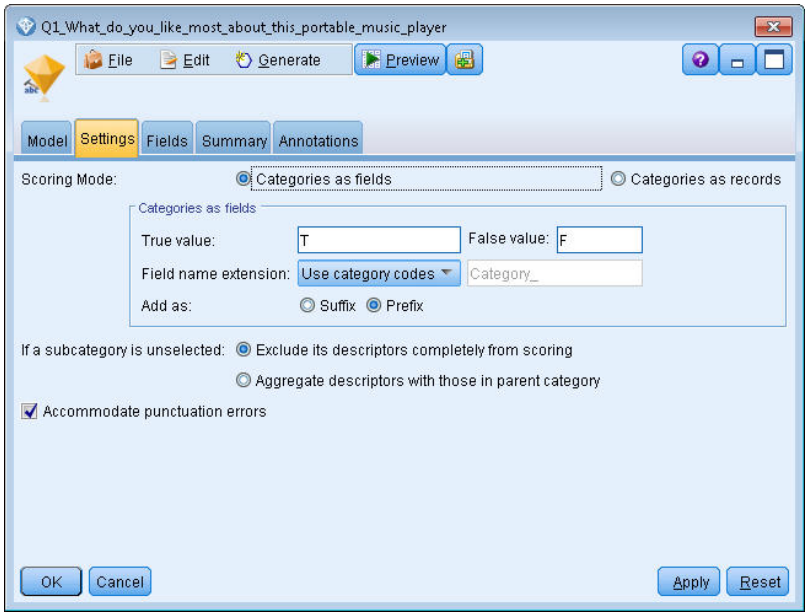



*Рисунок 12. Диалоговое окно слепка модели категорий: вкладка параметров*

4. **Слепок модели категорий Text Mining (вкладка полей).** Затем мы выбрали переменную текстового поля, представляющую собой имя поля из узла Файл статистики, выбрали ту опцию, что текстовое поле представляет **Фактический текст**, а также другие значения параметров.

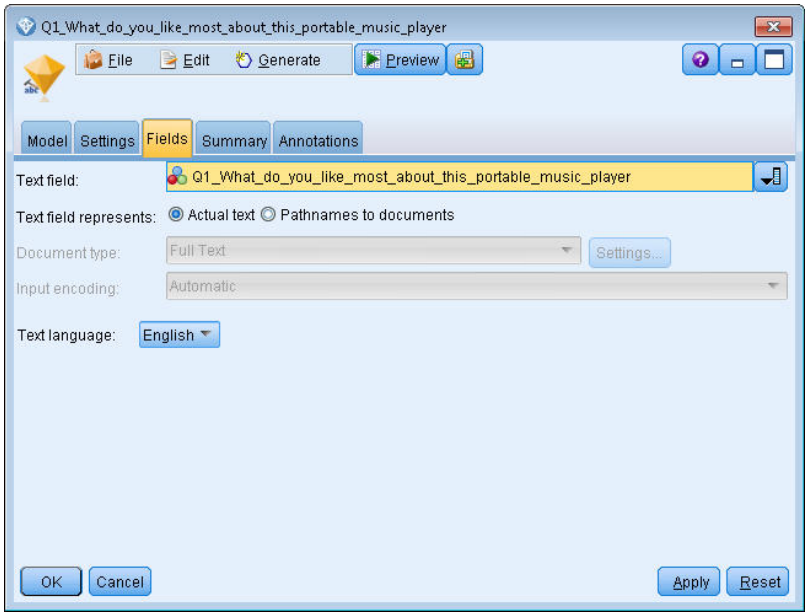
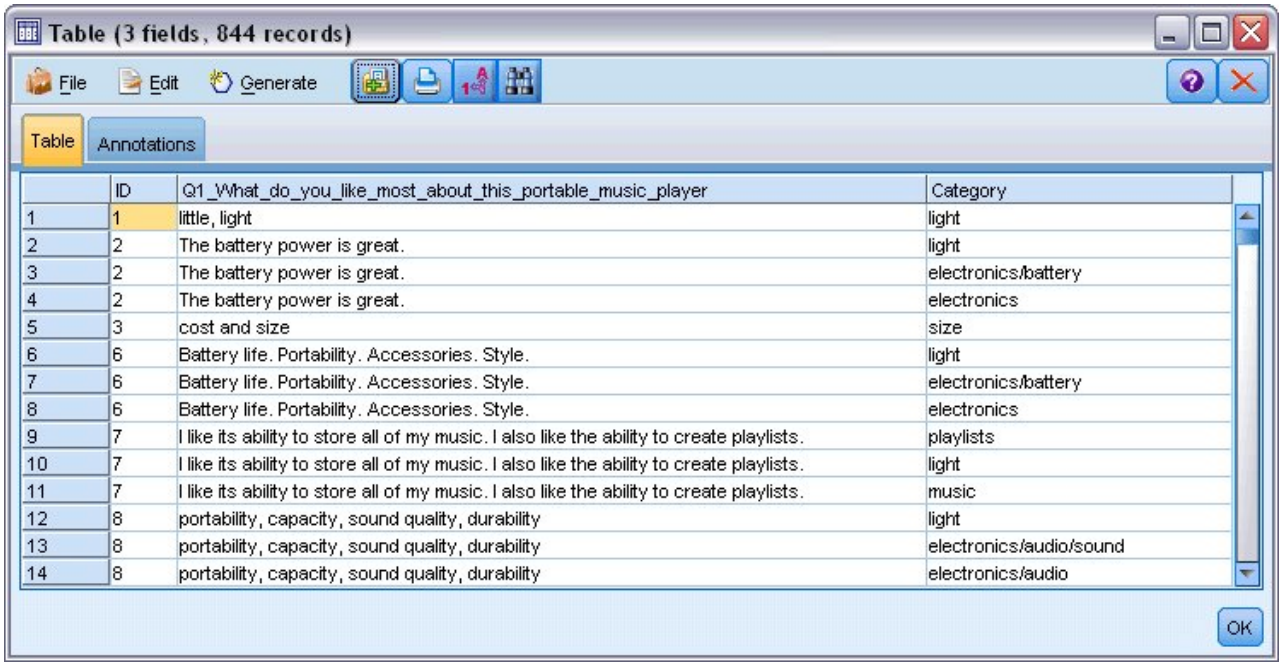



*Рисунок 13. Диалоговое окно слепка модели категорий Text Mining: вкладка полей*

5. **Узел таблицы.** Затем мы подключили узел таблицы, чтобы увидеть результаты, и выполнили поток.




| | ID | Q1_What_do_you_like_most_about_this_portable_music_player | Category |
|---|---|---|---|
| 1 | 1 | little, light | light |
| 2 | 2 | The battery power is great. | light |
| 3 | 2 | The battery power is great. | electronics/battery |
| 4 | 2 | The battery power is great. | electronics |
| 5 | 3 | cost and size | size |
| 6 | 6 | Battery life. Portability. Accessories. Style. | light |
| 7 | 6 | Battery life. Portability. Accessories. Style. | electronics/battery |
| 8 | 6 | Battery life. Portability. Accessories. Style. | electronics |
| 9 | 7 | I like its ability to store all of my music. I also like the ability to create playlists. | playlists |
| 10 | 7 | I like its ability to store all of my music. I also like the ability to create playlists. | light |
| 11 | 7 | I like its ability to store all of my music. I also like the ability to create playlists. | music |
| 12 | 8 | portability, capacity, sound quality, durability | light |
| 13 | 8 | portability, capacity, sound quality, durability | electronics/audio/sound |
| 14 | 8 | portability, capacity, sound quality, durability | electronics/audio |

*Рисунок 14. Табличный вывод*

# Глава 4. Исследование текстовых связей

## Узел Text Link Analysis

Узел анализа текстовых связей (Text Link Analysis, TLA) добавляет в извлечение понятий исследования текстовых данных методы сопоставления паттернов для выявления взаимосвязей между понятиями в текстовых данных на основе известных паттернов. Это могут быть взаимосвязи, описывающие, что заказчик думает про продукт, какие компании занимаются совместным бизнесом, или даже взаимосвязи между генами или фармацевтическими веществами.

Например, извлечение названия продукта вашего конкурента, вероятно, не сильно вас заинтересует. С помощью указанного узла можно будет также узнать, что люди думают об этом продукте (если такие мнения окажутся в данных). Взаимосвязи и связи выявляются и извлекаются посредством сопоставления известных паттернов с используемыми текстовыми данными.

Вы можете использовать правила паттернов TLA в определенных шаблонах ресурсов, поставляемых с IBM SPSS Modeler Text Analytics, или создать/отредактировать свои собственные правила. Правила паттернов составляются из макросов, списков слов и промежутков между словами, образуя логический запрос или правило, которое сравнивается с входными текстовыми данными. При каждом совпадении паттерна TLA с текстовыми данными эти текстовые данные можно будет извлечь в качестве результата TLA и переструктурировать как выходные данные. Дополнительную информацию смотрите в разделе Глава 19, “О правилах текстовых связей”, на стр. 215.

Узел анализа текстовых связей предлагает более прямой способ выявления и извлечения результатов паттернов TLA из текста и последующего добавления этих результатов в набор данных в потоке. Но узел анализа текстовых связей - не единственный возможный способ выполнения анализа текстовых связей. Можно также использовать сеанс интерактивной инструментальной среды на узле моделирования Исследование текстовых данных.

В интерактивной инструментальной среде можно исследовать результаты паттернов TLA и использовать их в качестве дескрипторов категорий и/или для получения дополнительной информации о результатах с применением детализации и графиков. Дополнительную информацию смотрите в разделе Глава 12, “Изучаем анализ текстовых связей (Text Link Analysis, TLA)”, на стр. 153. По существу, использование узла исследования текстовых данных для извлечения результатов анализа текстовых связей - замечательный способ исследования и точной настройки шаблонов для ваших данных с целью последующего их использования непосредственно на узле TLA.

Представление выходных данных может составлять до 6 слотов (или частей). Паттерны для японского языка выводятся только в виде одного или двух слотов. Дополнительную информацию смотрите в разделе “Выходные данные узла TLA” на стр. 51.

Этот узел можно найти на вкладке IBM SPSS Modeler Text Analytics палитры узлов в нижней части окна IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “IBM SPSS Modeler Text Analytics узлов” на стр. 8.

**Требования.** Узел анализа текстовых связей принимает текстовые данные, считываемые в поле, где используется любой из стандартных узлов (узел базы данных, узел плоского файла и так далее) или считываемых в поле, где указываются пути к внешним документам, генерируемым узлом списка файлов или узлом веб-каналов.

**Достоинства.** Узел анализа текстовых связей выходит за рамки базового извлечения понятий, предоставляя информацию о взаимосвязях *между* понятиями, а также связанными с ними мнениями или спецификаторами, которые могут быть выявлены в данных.

## Узел Анализ текстовых связей: Вкладка Поля

На вкладке Поля можно задать параметры полей для данных, из которых будут извлекаться понятия. Можно задать следующие параметры:

**Поле ID.** Выберите поле, содержащее идентификатор для текстовых записей. Идентификаторы должны быть целыми числами. Поле ID выполняет роль индекса для отдельных текстовых записей. Поле ID надо использовать, если текстовое поле представляет текст, подлежащий исследованию. Не используйте поле ID, если текстовое поле представляет **Пути имен к документам**.

**Текстовое поле.** Выберите поле, содержащее текст для исследования, имя пути к документу или имя пути к каталогу документов. Это поле зависит от источника данных.

**Что представляет текстовое поле.** Укажите, что именно содержит текстовое поле, заданное в предыдущем параметре. Варианты выбора:

- **Фактический текст.** Выберите эту опцию, если данное поле содержит именно тот текст, из которого должны извлекаться понятия.
- **Имена путей для документов.** Выберите эту опцию, если данное поле содержит одно или несколько имен путей к положениям, в которых находятся текстовые документы.

**Тип документа.** Эта опция доступна, только если вы задали, что текстовое поле представляет **Пути к документам**. Тип документа определяет структуру текста. Выберите один из следующих типов:

- **Полный текст.** Используется для большинства документов или текстовых источников. Для извлечения просматривается весь массив текста. В отличие от других опций, для этой опции нет никаких дополнительных параметров.
- **Структурированный текст.** Используется для библиографических форм, патентов и любых файлов, содержащих регулярные структуры, которые могут быть выявлены и проанализированы. Этот тип документов используется для полного или частичного пропуска процесса извлечения. Это позволяет определять разделители терминов, назначать типы и задавать минимальное значение частоты. Если выбрана эта опция, нажмите кнопку **Параметры** и введите текстовые разделители в области **Форматирование структурированного текста** диалогового окна Параметры документа. Дополнительную информацию смотрите в разделе “Вкладка Параметры документов для полей” на стр. 22.
- **Текст XML.** Используется для задания тегов XML, содержащих извлекаемый текст. Все остальные теги игнорируются. Если выбрана эта опция, нажмите кнопку **Параметры** и укажите явным образом элементы XML, содержащие читаемый текст, в области **Форматирование текста XML** диалогового окна Параметры документа. Дополнительную информацию смотрите в разделе “Вкладка Параметры документов для полей” на стр. 22.

**Общность текста.** Эта опция доступна только в том случае, если вы указали, что текстовое поле представляет **Имена путей к документам**, и выбрали **Полный текст** как тип документа. Выберите режим извлечения из числа следующих возможностей:

- **Режим документов.** Используйте этот режим для коротких и семантически однородных документов, например, для сообщений информационных агентств.
- **Режим абзацев.** Используйте для веб-страниц и документов без тегов. Процесс извлечения семантически разделяет документы, используя преимущества таких характеристик, как внутренние теги и синтаксис. При выборе этого режима скоринг выполняется по абзацам. Поэтому, например, правило яблоко & апельсин выполняется только в том случае, если яблоко и апельсин найдены в одном абзаце.

  **Примечание:** Из-за того способа, которым текст извлекается из документов PDF, **Режим абзацев** для этих документов не работает. Это связано с тем, что при извлечении подавляется маркер возврата каретки.

**Параметры режима абзацев.** Эта опция доступна только в том случае, если вы указали, что текстовое поле представляет **Имена путей к документам**, и выбрали для опции текстовой однородности значение **Режим абзацев**. Задайте пороговые значения числа символов, которые будут использоваться при любом извлечении.

Фактический размер будет округляться до ближайшей точки. Для обеспечения репрезентативности связывания слов, полученного из текста собрания документов, не задавайте слишком маленький размер извлечения.

- **Минимум.** Задайте минимальное число символов, используемых для извлечения.
- **Максимум.** Задайте максимальное число символов, используемых для извлечения.

**Кодировка ввода.** Эта опция доступна, только если вы указали, что текстовое поле представляет **Пути к документам**. Она задает кодировку текста по умолчанию. Для всех языков, кроме японского, выполняется преобразование из заданной или распознанной кодировки в кодировку `ISO-8859-1`. Таким образом, даже если вы зададите другую кодировку, механизм извлечения перед обработкой преобразует ее в `ISO-8859-1`. Все символы, которые не соответствуют формату кодировки `ISO-8859-1`, будут преобразованы в пробелы. Для японского текста можно выбрать одну из нескольких опций кодировки: `SHIFT_JIS`, `EUC_JP`, `UTF-8` или `ISO-2022-JP`.

**Копировать ресурсы из.** При исследовании текста извлечение учитывает не только параметры на вкладке Эксперт, но и лингвистические ресурсы. Эти ресурсы служат основной того, как управлять текстом и обрабатывать его во время извлечения с целью получения понятий, типов и паттернов анализа текстовых связей (Text Link Analysis, TLA). Ресурсы можно скопировать на этот узел из шаблона ресурсов.

Шаблон ресурса - это заранее определенный набор библиотек и дополнительных лингвистические и нелингвистических ресурсов, тонко настроенных на конкретный домен использования. Эти ресурсы служат основной способов управления данными и их обработки во время извлечения. Нажмите кнопку **Загрузить** и выберите шаблон, из которого следует скопировать ресурсы.

Шаблоны загружаются при их выборе, а не при выполнении потока. В момент загрузки копия ресурсов сохраняется на узле. Поэтому, если вы когда-либо решите использовать обновленный шаблон, в этот момент его нужно будет перезагрузить. Дополнительную информацию смотрите в разделе “Копирование ресурсов из шаблонов и файлов TAP” на стр. 26.

**Язык текста.** Идентифицирует язык исследуемого текста. Скопированные в данный узел ресурсы управляют представленными опциями языков. Можно или выбрать язык, для которого были настроены эти ресурсы, или выбрать опцию **ВСЕ**. Настоятельно рекомендуется точно задавать язык для текстовых данных, однако при неуверенности можно выбрать опцию **ВСЕ**. Для текстов на японском языке опция **ВСЕ** недоступна. Опция **ВСЕ** увеличивает время извлечения, так как при ее применении сначала используется автоматическое распознавание языка с просмотром всех документов и записей для идентификации языка текстов. При включении этой опции все записи или документы на поддерживаемых и лицензированных языках читаются механизмом извлечения с использованием внутренних словарей на конкретных языках. Дополнительную информацию смотрите в разделе “Идентификатор языка” на стр. 212. Обратитесь к торговому представителю, если хотите приобрести лицензию на поддерживаемый язык, к которому в настоящее время не имеете доступа.

## Узел Анализ текстовых связей: Вкладка Модель

Вкладка Модель содержит одну опцию, влияющую на скорость и точность процесса извлечения.

**Оптимизировать для скорости скоринга.** Эта опция, которая по умолчанию выбрана, нацелена на создание компактное модели с быстрыми оценками. При отключении этой опции создается модель, оцениваемая медленнее, но гарантирующая полную согласованность типов и понятий, то есть гарантирующая, что понятию всегда будет назначаться только один тип.

## Узел Анализ текстовых связей: Вкладка Эксперт

На этом узле автоматически включается поддержка извлечения результатов паттернов анализа текстовых связей (Text Link Analysis, TLA). Вкладка Эксперт содержит некоторые дополнительные параметры, влияющие на то, как будут извлекаться и обрабатываться текстовые данные. Параметры в этом диалоговом окне управляют базовым поведением, а также некоторыми дополнительными стратегиями извлечения.

Кроме того, существует ряд лингвистических ресурсов и опций, также влияющих на результаты извлечения, которыми управляет выбранный вами шаблон ресурсов.

**Для текста на голландском, английском, французском, немецком, итальянском, португальском и испанском**

**Допускать ошибки пунктуации.** Эта опция временно нормализует текст, содержащий ошибки пунктуации (например, неверно используемые знаки препинания) во время извлечения, чтобы повысить извлекаемость понятий. Эта опция особенно полезна для коротких текстов низкого качества (например, ответы при опросе с произвольным ответом, электронная переписка, данные CRM), а также для текста, содержащего много сокращений.

**Допускать орфографические ошибки при минимальном числе символов корня [n].** Эта опция применяет метод нечеткой группировки, который помогает группировать в одну концепцию слова, которые часто пишутся с ошибками, а также вариативные написания слова. Алгоритм нечеткой группировки перед сравнением временно удаляет из извлеченных слов все гласные (кроме первой) и двойные или тройные согласные, так что тунель и тоннель попадут в одну группу. Методы нечеткой группировки, однако, не применяются, если различным терминам назначены различные типы, кроме типа <Неизвестный>.

Кроме того, можно задать минимально необходимое число символов *корня* при использовании нечеткой группировки. Число символов корня в термине рассчитывается как общее число символов минус число символов окончания; кроме того, в случае термина-словосочетания вычитаются детерминативы и предлоги. Например, в термине упражнения будет насчитано 9 символов корня “упражнени”, поскольку буква *я* на конце слова относится к окончанию множественного числа. Аналогичным образом в пакет яблок насчитывается 10 символов корня (“пакет яблок”), а в магнитола для автомобиля насчитывается 17 символов корня (“магнитол автомобил”). Этот метод подсчета используется только при проверке применимости нечеткой группировки и не используется в алгоритмах сравнения слов.

*Примечание*: Если окажется, что некоторые слова группируются неправильно, такие пары слов можно исключить из метода при помощи явного объявления в разделе **Нечеткая группировка: исключения** на вкладке Расширенные ресурсы. Дополнительную информацию смотрите в разделе “Нечеткая группировка” на стр. 206.

**Извлечь одиночные термины.** Эта опция извлекает отдельные слова (одиночные термины), если слово не входит в словосочетание и если это существительное или нераспознанная часть речи.

**Извлечь нелингвистические объекты.** Эта опция извлекает нелингвистические объекты, такие как номера телефонов, номера социального страхования, время, даты, валюты, цифры, проценты, адреса электронной почты и HTTP-адреса. Вы можете включить или исключить те или иные типы нелингвистических объектов в разделе **Нелингвистические объекты: конфигурация** на вкладке Расширенные ресурсы. Выключив ненужные объекты, вы сэкономите время обработки механизмом извлечения. Дополнительную информацию смотрите в разделе “Конфигурация” на стр. 210.

**Алгоритм верхнего регистра.** Эта опция извлекает простые и составные термины, не входящие во встроенные словари, если первая буква термина - в верхнем регистре. Это хороший способ извлечь большинство имен собственных.

**Группировать частичные и полные личные имена, где возможно.** Эта опция группирует имена, которые по-разному появляются в тексте. Эта возможность полезна, поскольку имена часто употребляются в начале текста в полной форме, а затем - в краткой. Эта опция пытается сопоставить каждый одиночный термин с типом <Неизвестный> последнему слову в любом составном термине, типизированном как <Личный>. Например, если найден терм *иванов*, получивший вначале тип <Неизвестный>, механизм извлечения поищет составные термины в типе <Личный>, содержащие *иванов* как последнее слово, например, *александр иванов*. Эта опция применяется только к фамилии, поскольку первое имя почти никогда не извлекается как одиночный термин.

**Максимум неслужебных слов при перестановке.** Эта опция задает максимально допустимое число неслужебных слов при применении метода перестановки. Этот метод перестановок группирует как близкие словосочетания, содержащие в своем составе одни и те же неслужебные слова, если игнорировать форму слова. Например, если задать ограничение в два неслужебных слова, будут обработаны такие извлеченные словосочетания, как `компания клиенту` и `клиенту от нашей компании`. В этом примере такие словосочетания будут сгруппированы в итоговом списке понятий, поскольку считаются одинаковыми, если проигнорировать слова `от нашей`.

Для текста на японском

При использовании текстовых данных на японском языке можно выбрать, какой вторичный анализатор следует применить.

**Вторичный анализ.** При извлечения базовые ключевые слова извлекаются при помощи набора типов по умолчанию. Но, если выбрать тот или иной вторичный анализатор, можно получить много дополнительных и более богатых понятий, поскольку теперь экстрактор будет учитывать частицы и вспомогательные глаголы как часть концепции. Кроме того, при анализе эмоциональной окраски можно включить большое число дополнительных типов. После выбора вторичного анализатора можно сгенерировать результаты Text Link Analysis.

**Примечание:** При вызове вторичного анализатора извлечение занимает больше времени.

- **Анализ зависимостей**. При выборе этой опции извлечение понятий производится с дополнительным учетом частиц по сравнению с извлечением базовых типов и ключевых слов. Кроме того, при анализе зависимостей можно получить более богатые результаты паттернов TLA.
- **Анализ эмоциональной окраски**. При выборе этого анализатора извлекаются дополнительные понятия и, если применимо, результаты паттернов TLA. Помимо базовых типов можно воспользоваться более чем 80 типами эмоциональной окраски . При помощи таких типов можно раскрывать в тексте понятия и паттерны, выражающие эмоции, настроения и мнения. Фокус анализа эмоциональной окраски управляется тремя опциями: **Все эмоциональные окраски**, **Только репрезентативные эмоциональные окраски** и **Только заключения**.

## Выходные данные узла TLA

После запуска узла анализа текстовых связей (Text Link Analysis, TLA) данные реструктуризуются. Важно понимать способ, которым исследование текстовых данных реструктуризует используемые данные. Если вы хотите получить иную структуру для исследования данных, для достижения этой цели можно использовать узлы на палитре Операции с полями. Например, если вы работали с данными, в которых каждая строка представляла текстовую запись, будет создано по одной строке для каждого паттерна, обнаруженного в исходных текстовых данных. Для каждой строки в выходных данных существует 15 полей:

- Шесть полей (**Номер понятия**, такое как **Понятие1**, **Понятие2**, ... и **Понятие6**) представляют все понятия, обнаруживаемые при сопоставлении паттернов.
- Шесть полей (**Номер типа**, такое как **Тип1**, **Тип2**, ... и **Тип6**) представляют тип для каждого понятия.
- **Имя правила** представляет имя правила текстовых связей, используемого для сопоставления текста и генерирования выходных данных.
- Поле, где используется имя поля ID, заданного вами на узле и представляющего ID записи или документа, встреченный в данных.
- **Совпавший текст** представляет фрагмент текстовых данных в исходной записи или документе, совпавший с паттерном TLA.

*Примечание*: Правила паттернов анализа текстовых связей для текста на японском генерируют только один или два результата паттернов слотов.

*Примечание*: Любые уже существующие потоки, содержащие узел анализа текстовых связей из выпуска до версии 5.0, могут быть выполнимы не полностью, пока вы не обновите узлы. Некоторые

усовершенствования в более поздних версиях IBM SPSS Modeler требуют замены более старых узлов более новыми версиями, которые лучше внедряются и мощнее.

Возможен также автоматической перевод, выполняемый с некоторых языков. Эта возможность позволяет исследовать документы на языке, на котором вы не можете ни говорить, ни читать. Если вы хотите использовать эту возможность перевода, у вас должен быть доступ к SDL SaaS (Software as a Service - Программное обеспечение как служба). Дополнительную информацию смотрите в разделе “Параметры перевода” на стр. 56.

## Кэширование результатов TLA

Если используется кэширование, результаты анализа текстовых связей (text link analysis, TLA) остаются в потоке. Чтобы избежать повторения извлечения результатов при каждом выполнении потока, выберите узел анализа текстовых связей и в наборе меню выберите **Изменить > Узел > Кэш > Включить**. При следующем выполнении потока выходные данные будут кэшироваться на узле. В значке узла появится крошечное изображение "документа", цвет которого при заполнении кэша изменится с белого на зеленый. Кэш сохраняется в течение сеанса. Чтобы сохранить кэш на другой день (после того, как поток будет закрыт и снова открыт), выберите узел и в наборе меню выберите **Изменить > Узел > Кэш > Сохранить кэш**. При следующем открытии потока можно будет перезагрузить кэш вместо того, чтобы запускать перевод повторно.

Другой вариант: можно сохранить кэш или включить кэширование узла, щелкнув правой кнопкой мыши по узлу и выбрав **Кэш** в контекстном меню.

## Использование узла анализа текстовых связей в потоке

Узел анализа текстовых связей используется для обращения к данным и излечения понятий в потоке. Для доступа к данным можно использовать любой исходный узел.

### Пример: Узел файла статистики с узлом анализа текстовых связей

Использование узла анализа текстовых связей показано в следующем примере.


*Рисунок 15. Пример: Узел файла статистики с узлом анализа текстовых связей*

1. **Узел Файл статистики (вкладка данных).** Во-первых, мы добавили этот узел в поток, чтобы задать, где хранится текст.

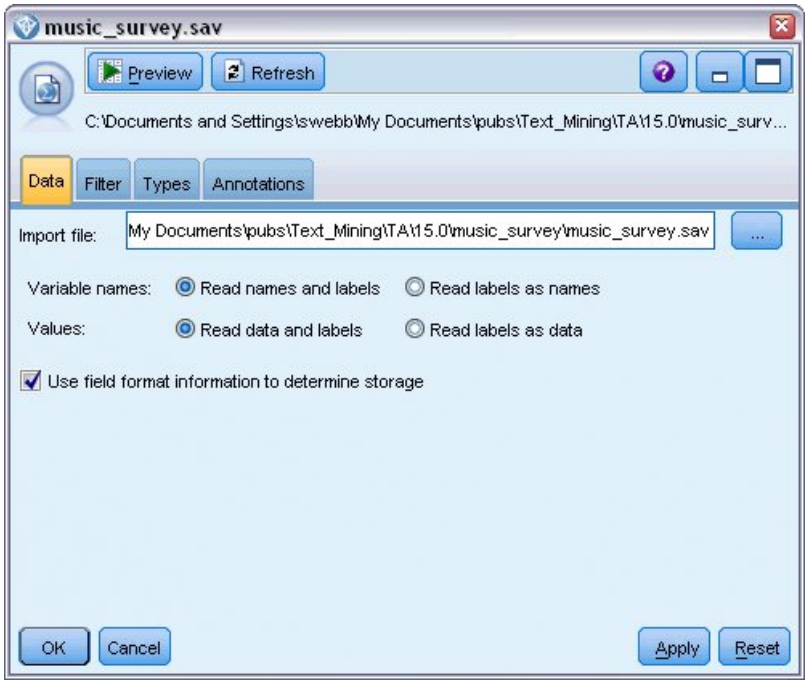



*Рисунок 16. Диалоговое окно узла Файл статистики: вкладка данных*

2. **Узел анализа текстовых связей (вкладка Поля).** Затем мы присоединяем этот узел к потоку с целью извлечения понятий для моделирования или просмотра нисходящего потока. Мы задаем поле ID и имя текстового поля, содержащего данные, а также другие параметры.

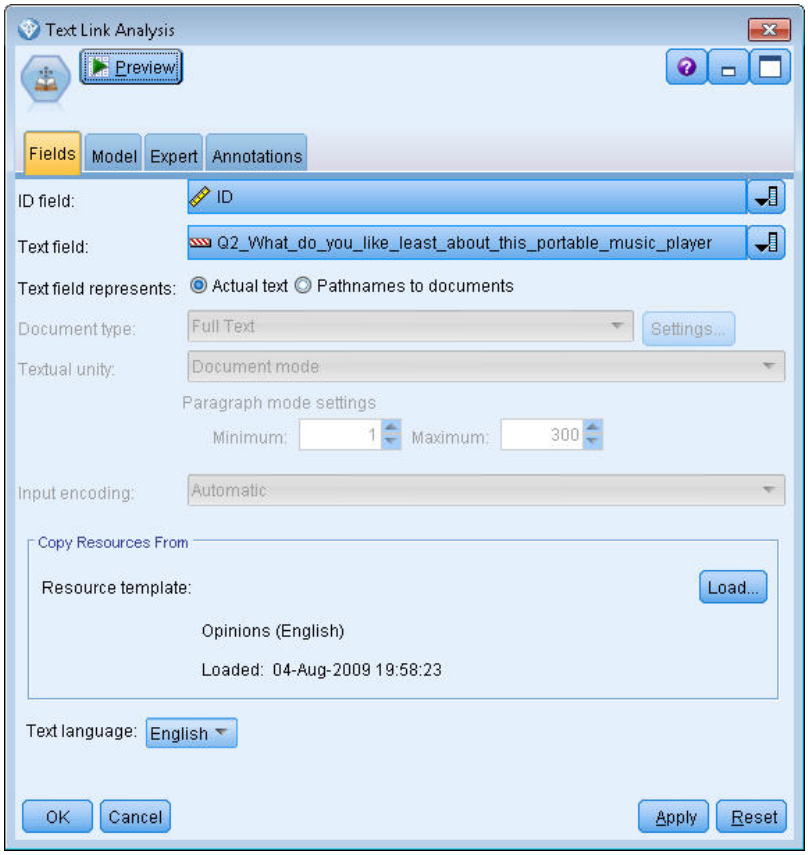
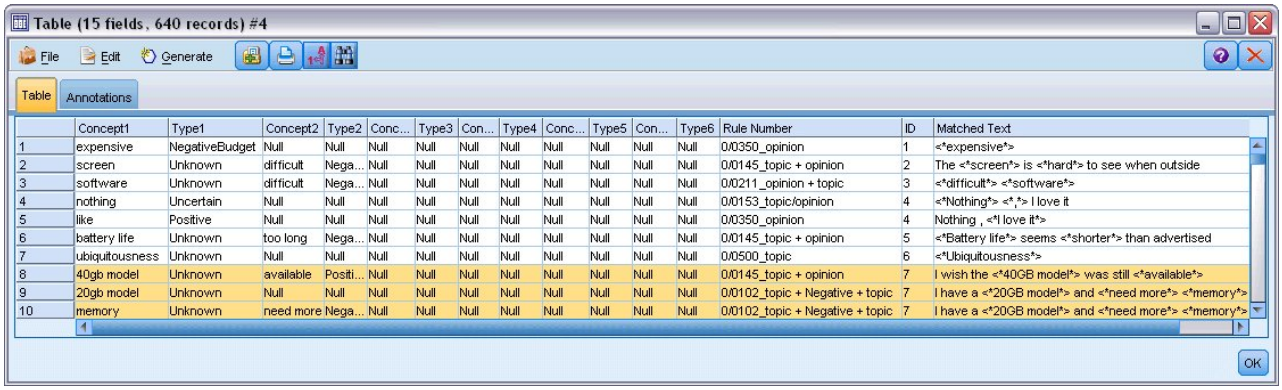



*Рисунок 17. Диалоговое окно узла анализа текстовых связей: Вкладка Поля*

3. **Узел таблицы.** И наконец, мы подключаем узел таблицы для просмотра понятий, извлеченных из наших тестовых документов. В выводимых выходных данных таблицы можно просмотреть результаты паттернов TLA, найденные в данных, после выполнения обработки этого потока с применением узла анализа текстовых связей. Некоторые результаты покажут, что было присоединено только одно понятие/тип. В других выходных данных результаты будут более сложны и содержать несколько типов и понятий. Кроме того, после прохождения данных через узел анализа текстовых связей и извлечения понятий некоторые аспекты данных будут изменены. Исходные данные в нашем примере содержат 8 полей и 405 записей. После выполнения обработки узла анализа текстовых связей будет 15 полей и 640 записей. На данный момент будет по одной строке для каждого найденного результата паттернов TLA. Например, `ID 7` превратится в три строки из исходной, поскольку будет извлечено три результата паттернов TLA. Если вы захотите слить выходные данные обратно в исходные данные, можете использовать узел слияния.




| | Concept1 | Type1 | Concept2 | Type2 | Conc... | Type3 | Con... | Type4 | Conc... | Type5 | Con... | Type6 | Rule Number | ID | Matched Text |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | expensive | NegativeBudget | Null | Null | Null | Null | Null | Null | Null | Null | Null | Null | 0/0350_opinion | 1 | <*expensive*> |
| 2 | screen | Unknown | difficult | Nega... | Null | Null | Null | Null | Null | Null | Null | Null | 0/0145_topic + opinion | 2 | The <*screen*> is <*hard*> to see when outside |
| 3 | software | Unknown | difficult | Nega... | Null | Null | Null | Null | Null | Null | Null | Null | 0/0211_opinion + topic | 3 | <*difficult*> <*software*> |
| 4 | nothing | Uncertain | Null | Null | Null | Null | Null | Null | Null | Null | Null | Null | 0/0153_topic/opinion | 4 | <*Nothing*> <*,*> I love it |
| 5 | like | Positive | Null | Null | Null | Null | Null | Null | Null | Null | Null | Null | 0/0350_opinion | 4 | Nothing , <*I love it*> |
| 6 | battery life | Unknown | too long | Nega... | Null | Null | Null | Null | Null | Null | Null | Null | 0/0145_topic + opinion | 5 | <*Battery life*> seems <*shorter*> than advertised |
| 7 | ubiquitousness | Unknown | Null | Null | Null | Null | Null | Null | Null | Null | Null | Null | 0/0500_topic | 6 | <*Ubiquitousness*> |
| 8 | 40gb model | Unknown | available | Positi... | Null | Null | Null | Null | Null | Null | Null | Null | 0/0145_topic + opinion | 7 | I wish the <*40GB model*> was still <*available*> |
| 9 | 20gb model | Unknown | Null | Null | Null | Null | Null | Null | Null | Null | Null | Null | 0/0102_topic + Negative + topic | 7 | I have a <*20GB model*> and <*need more*> <*memory*> |
| 10 | memory | Unknown | need more | Nega... | Null | Null | Null | Null | Null | Null | Null | Null | 0/0102_topic + Negative + topic | 7 | I have a <*20GB model*> and <*need more*> <*memory*> |

*Рисунок 18. Узел табличного вывода*

# Глава 5. Перевод текста для извлечения

## Узел перевода

Узел перевода может использоваться для перевода текста с поддерживаемых языков (таких как арабский, китайский и персидский) на английский или другие языки для анализа при помощи IBM SPSS Modeler Text Analytics. Это разрешает исследование документов на двухбайтных языках, поддержка которых в противном случае оказалась бы невозможной, и позволяет аналитикам извлекать понятия из документов на иностранном языке, даже если они не понимают исследуемый язык. Имейте в виду, что для использования узла перевода у вас должна быть возможность соединения с SDL SaaS (Software as a Service - Программное обеспечение как служба).

При исследовании текстовых данных на любом из этих языков просто добавьте узел перевода перед узлом моделирования Исследование текстовых данных в потоке. На узле перевода можно также включить поддержку кэширования, чтобы избежать повторения перевода при каждом выполнении обработки потока.

Этот узел можно найти на вкладке IBM SPSS Modeler Text Analytics палитры узлов в нижней части окна IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “IBM SPSS Modeler Text Analytics узлов” на стр. 8.

Кэширование перевода.. Если перевод кэшируется, переведенный текст сохраняется не во внешних файлах, а в потоке. Чтобы избежать повторения перевода при каждом выполнении обработки потока, выберите узел перевода и в наборе меню выберите **Изменить > Узел > Кэш > Включить**. При следующем выполнении потока выходные данные перевода будут кэшироваться на узле. В значке узла появится крошечное изображение "документа", цвет которого при заполнении кэша изменится с белого на зеленый. Кэш сохраняется в течение сеанса. Чтобы сохранить кэш на другой день (после того, как поток будет закрыт и снова открыт), выберите узел и в наборе меню выберите **Изменить > Узел > Кэш > Сохранить кэш**. При следующем открытии потока можно будет перезагрузить кэш вместо того, чтобы запускать перевод повторно.

Другой вариант: можно сохранить кэш или включить кэширование узла, щелкнув правой кнопкой мыши по узлу и выбрав **Кэш** в контекстном меню.

**Важно!** Если вы пытаетесь получить информацию по сети через прокси-сервер, нужно включить использование прокси-сервера в файле `net.properties` и для клиента, и для сервера IBM SPSS Modeler Text Analytics. Следуйте подробным указаниям в этом файле. Они применимы при доступе к сети через узел

Веб-фид или при получении лицензии программы как службы (Software as a Service, SaaS) для SDL, поскольку такие соединения проходят через Java. По умолчанию этот файл расположен в каталоге *C:\Program Files\IBM\SPSS\Modeler\18\jre\lib\net.properties*.

**Примечание:** Узел перевода для скоринга в конфигурации IBM SPSS Collaboration and Deployment Services - Scoring использовать нельзя.

## Узел перевода: Вкладка Перевод

**Текстовое поле** Выберите поле, содержащее подлежащий исследованию текст, имя пути документа или имя пути каталога документов. Это поле зависит от источника данных. Можно задать любое строковое поле, даже с параметром `Direction=None` или `Type=Typeless`.

**Текстовое поле представляет**. Укажите, что именно содержит текстовое поле, заданное в предыдущем параметре. Варианты выбора:

- **Фактический текст** Выберите эту опцию, если поле содержит точный текст, из которого следует извлечь понятия.
- **Имена путей к документам** Выберите эту опцию, если поле содержит один или несколько имен путей к положению внешних документов, содержащих подлежащий извлечению текст. Например, эту опцию следует выбрать в случае использования узла Список файлов для считывания списка документов. Дополнительную информацию смотрите в разделе “Узел списка файлов” на стр. 11.

**Кодировка входных данных** Выберите кодировку исходного текста. Для начала можно выбрать опцию **Автоматически**, но если будет замечено, что некоторые файлы обрабатываются неверно, мы рекомендуем выбрать фактическую кодировку из приведенного здесь списка. Опция Автоматически при работе с короткими текстовыми данными, такими как короткие записи базы данных, может определять кодировку неверно. Для выходных текстовых данных с этого узла используется кодировка `UTF-8`.

**Параметры** Задайте параметры перевода для потока.

- **Подключение языковой пары**. Выберите языковую пару, которую вы хотите использовать; доступные языковые пары появляются в этом списке автоматически после конфигурирования ссылки на службу в диалоговом окне **Параметры перевода**. Дополнительную информацию смотрите в разделе “Параметры перевода”.
- **Точка контакта**. Если предварительно были созданы *точки контактов SDL*, выберите одну из них для использования в соединении с переводом.
- **Сохранять и по возможности повторно использовать ранее переведенный текст** Указывает, что результаты перевода следует сохранять и, если при следующем выполнении обработки потока будет представлено то же самое число записей/документов, содержимое следует считать одинаковым и использовать результаты перевода повторно для экономии времени обработки. Если во время выполнения эта опция включена, а число записей не совпадает с сохраненным в прошлый раз, текст будет переведен полностью, а затем сохранен под именем метки для следующего выполнения. Эта опция доступна, только если выбран язык перевода SDL.

  **Примечание:** Если текстовые данные сохраняются в потоке, можно также включить поддержку кэширования на узле перевода. В этом случае, когда бы ни был доступен кэш, помимо повторно использования результатов перевода будет также игнорироваться все в предшествующем потоке.
- **Метка** При выборе опции **Сохранять и по возможности повторно использовать ранее переведенный текст** нужно задать имя метки для результатов. Эта метка используется для идентификации ранее переведенного текста. Если не задать метку, при выполнении обработки потока в свойства потока будет добавлено предупреждение, и повторное использование будет невозможно.

## Параметры перевода

В этом диалоговом окне можно определить подключение перевода SDL SaaS (Software as a Service - Программное обеспечение как служба) и управлять им; это подключение перевода можно использовать

повторно при каждом выполнении вами перевода. Определив здесь подключение, можно будет быстро выбрать подключение языковой пары во время перевода без необходимости заново вводить все параметры подключения.

Подключение языковой пары определяет исходный язык и язык перевода, а также подробности соединения с URL для сервера. Например, *китайский - английский* означает, что исходный текст - на китайском, а окончательный перевод будет на английском. Каждое подключение, к которому вы будете обращаться через онлайновые сервисы SDL, нужно определить вручную.

**Важно!** Если вы пытаетесь получить информацию по сети через прокси-сервер, нужно включить использование прокси-сервера в файле `net.properties` и для клиента, и для сервера IBM SPSS Modeler Text Analytics. Следуйте подробным указаниям в этом файле. Они применимы при доступе к сети через узел Веб-фид или при получении лицензии программы как службы (Software as a Service, SaaS) для SDL, поскольку такие соединения проходят через Java. По умолчанию этот файл расположен в каталоге *C:\Program Files\IBM\SPSS\Modeler\18\jre\lib\net.properties*.

**URL соединения** Введите URL для подключения SDL SaaS.

**Ключ API** Введите ключ, предоставленный вам SDL.

**ID учетной записи** Введите уникальный ID, предоставленный вам SDL.

**ID пользователя** Введите уникальный ID, предоставленный вам SDL.

**Проверить** Нажмите кнопку **Проверить**, чтобы проверить правильность конфигурации подключения и посмотреть найденные для этого подключения языковые пары.

## Использование узла перевода

Для извлечения понятий из текстовых данных на поддерживаемых языках перевода, таких как арабский, китайский или персидский, перед любым узлом исследования текстовых данных в потоке можно добавить узел перевода.

Если подлежащий переводу текст содержится в одном или нескольких внешних файлах, список их имен может быть считан при помощи узла Список файлов. В этом случае можно добавить узел перевода между узлом списка файлов и последующими узлами исследования текстовых данных, а для выходных данных использовать положение, где находится переведенный текст.

# Глава 6. Просмотр текста внешних источников

## Узел программы просмотра файлов

При исследовании собрания документов можно задать полные имена путей файлов непосредственно на узле моделирования Исследование текстовых данных и на узле перевода. Однако, если данные выводятся на узел Таблица, вы увидите только полное имя пути документа, а не тест в нем. Узел программы просмотра файлов может использоваться как аналог узла Таблица и позволяет обращаться к фактическим текстовым данным в каждом из документов без необходимости слияния их всех вместе в один файл.

Узел программы просмотра файлов может помочь лучше понять результаты извлечения текстовых данных благодаря обеспечению доступа к исходному (то есть непереведенному) тексту, из которого были извлечены понятия, поскольку в противном случае он был бы недоступен в потоке. Этот узел добавляется в поток после узла Список файлов с целью получения списка ссылок на все файлы.

Результат этого узла - окно, показывающее все элементы документов, которые были прочитаны и использованы для извлечения понятий. Щелкнув в этом окне по значку панели инструментов, можно запустить отчет во внешнем браузере, возвращающем список имен документов в виде гиперссылок. Щелкнув по ссылке, можно открыть соответствующий документ в собрании. Дополнительную информацию смотрите в разделе “Использование узла программы просмотра файлов”.

Этот узел можно найти на вкладке IBM SPSS Modeler Text Analytics палитры узлов в нижней части окна IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “IBM SPSS Modeler Text Analytics узлов” на стр. 8.

*Примечание*: При работе в режиме клиент-сервер, если в состав потока входят узлы программы представления файлов, собрания документов должны храниться на сервере, в каталоге веб-сервера. Поскольку узел вывода исследования текстовых данных генерирует список документов, хранящихся в каталоге веб-сервера, разрешениями для этих документов управляют параметры защиты этого веб-сервера.

## Параметры узла программы просмотра файлов

Для узла программы просмотра файлов можно задать следующие параметры:

**Поле документа.** Выберите поле в данных, содержащее полное имя и путь документов, которые следует вывести.

**Заголовок для сгенерированной страницы HTML.** Создайте заголовок для вывода в верхней части страницы, содержащей список документов.

## Использование узла программы просмотра файлов

Как используется узел программы просмотра файлов, показано в следующем примере.

## Пример: узел списка файлов и узел программы просмотра файлов


*Рисунок 19. Поток, иллюстрирующий использование узла программы просмотра файлов*

1. **Узел Список файлов (вкладка параметров).** Сначала мы добавляем этот узел, чтобы указать, где находятся документы.

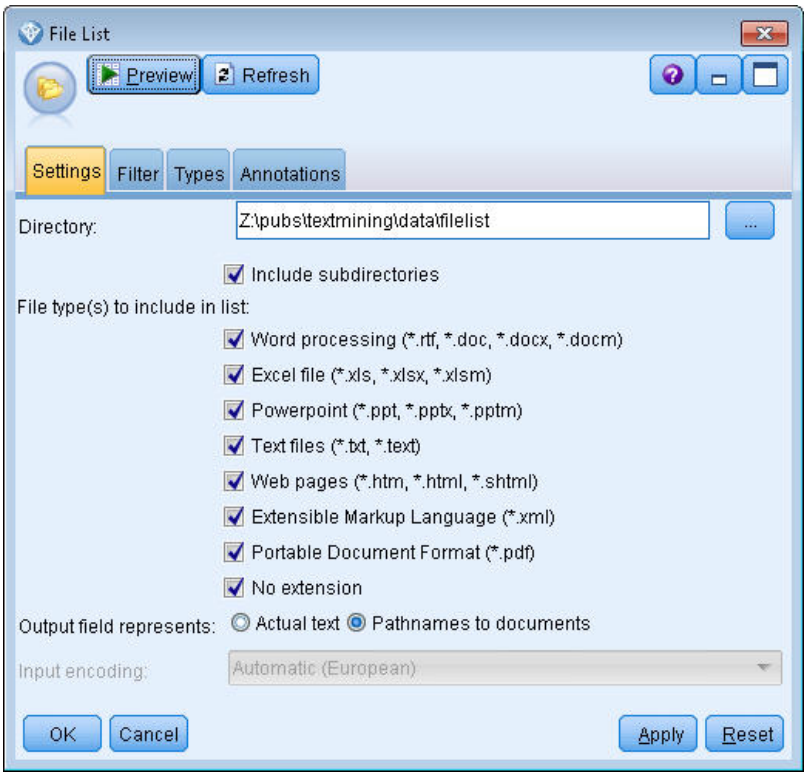



*Рисунок 20. Диалоговое окно узла списка файлов: Вкладка Параметры*

2. **Узел программы просмотра файлов (вкладка Параметры).** Далее мы присоединяем узел программы просмотра файлов для генерирования списка документов.

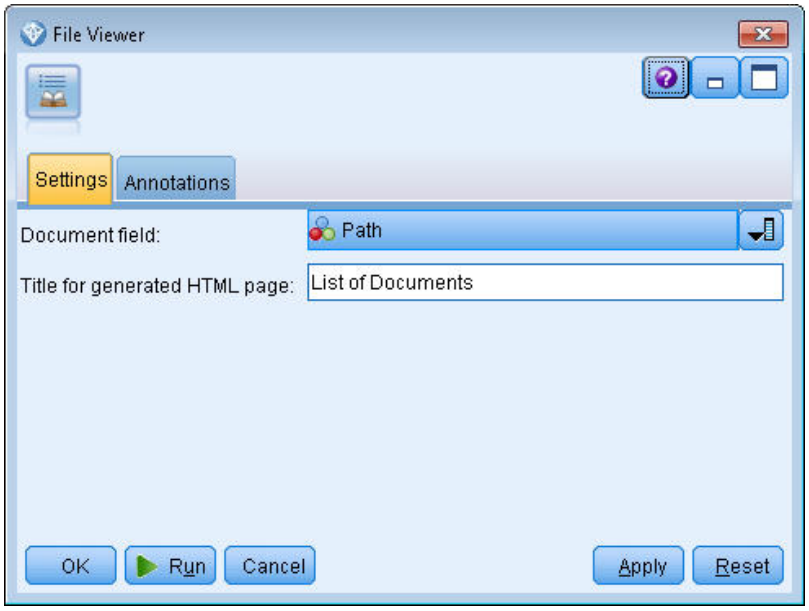



*Рисунок 21. Диалоговое окно узла программы просмотра файлов: Вкладка Параметры*

3. **Диалоговое окно вывода программы просмотра файлов.** Затем мы выполняем обработку потока, выводящего список документов в новом окне.

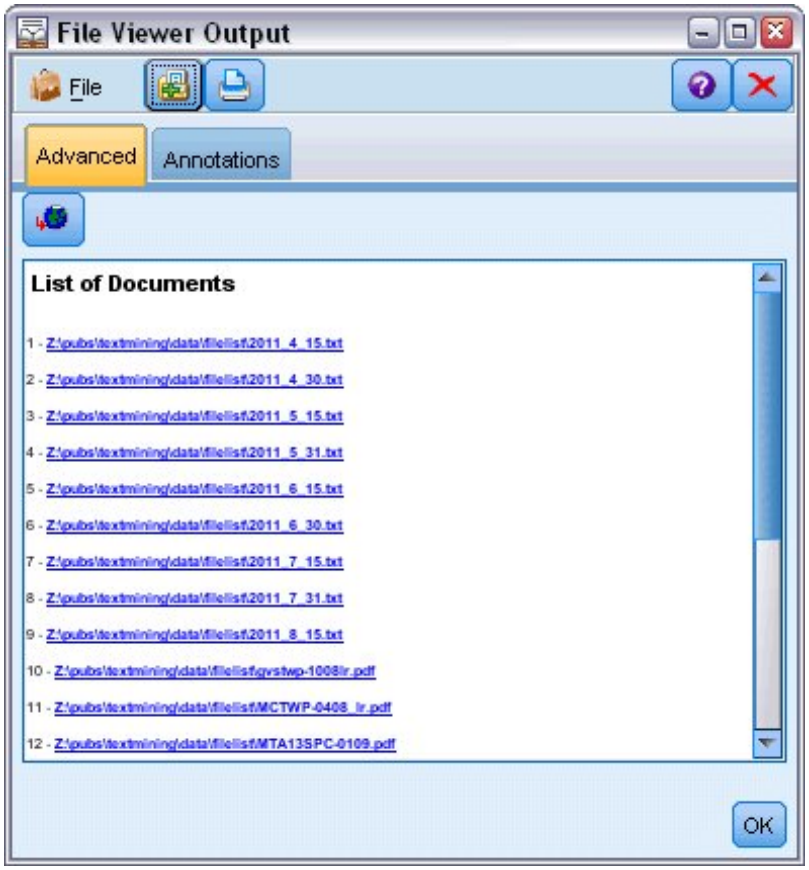



*Рисунок 22. Вывод программы просмотра файлов*

4. Чтобы просмотреть документы, мы нажимаем кнопку панели инструментов, выводящую глобус с красной стрелкой. Она открывает список гиперссылок на документы в используемом браузере.

# Глава 7. Свойства узла для сценариев

У IBM SPSS Modeler есть язык сценариев, при помощи которых можно выполнять потоки из командной строки. Здесь мы познакомимся со свойствами конкретных узлов, поставляемых с IBM SPSS Modeler Text Analytics. Дополнительную информацию о наборе стандартных узлов, поставляемых вместе с IBM SPSS Modeler, смотрите в Руководстве по сценариям и автоматизации.

## Узел Список файлов: filelistnode

Свойства в приведенной ниже таблице можно использовать в сценариях. Сам узел называется `filelistnode`.

*Таблица 7. Свойства сценариев узла Список файлов*

| Свойства сценариев | Тип переменной |
|---|---|
| `path` | *строка* |
| `recurse` | *флаг* |
| `word_processing` | *флаг* |
| `excel_file` | *флаг* |
| `powerpoint_file` | *флаг* |
| `text_file` | *флаг* |
| `web_page` | *флаг* |
| `xml_file` | *флаг* |
| `pdf_file` | *флаг* |
| `no_extension` | *флаг* |

Примечание: Параметр 'Создать список' более не доступен, и все выходные данные сценариев, содержащих эту опцию, будут автоматически преобразованы в 'файлы'.

## Узел веб-фидов: webfeednode

Свойства в приведенной ниже таблице можно использовать в сценариях. Сам узел называется `webfeednode`.

*Таблица 8. Свойства сценариев узла Веб-фид*

| Свойства сценариев | Тип переменной | Описание свойства |
|---|---|---|
| `urls` | *строка1 строка2 ...строкаn* | Каждый URL задается в списковой структуре. URL в списке разделяются символами “\n” |
| `recent_entries` | *флаг* | |
| `limit_entries` | *целое* | Число самых новых записей, читаемых из одного URL |
| `use_previous` | *флаг* | Для сохранения и повторного использования кэша веб-фидов. |
| `use_previous_label` | *строка* | Имя для сохраняемого веб-кэша. |
| `start_record` | *строка* | Открывающий тег для формата, иного чем RSS. |
| `url` *n* `.title` | *строка* | Для каждого URL в списке здесь нужно тоже определить URL. Первый URL будет `url1.title`, где номер соответствует его позиции в списке URL. Это открывающий тег, содержащий заголовок содержимого. |
| `url` *n* `.short_description` | *строка* | То же, что и для `url` *n* `.title`. |

*Таблица 8. Свойства сценариев узла Веб-фид (продолжение)*

| Свойства сценариев | Тип переменной | Описание свойства |
|---|---|---|
| `url` *n* `.description` | *строка* | То же, что и для `url` *n* `.title`. |
| `url` *n* `.authors` | *строка* | То же, что и для `url` *n* `.title`. |
| `url` *n* `.contributors` | *строка* | То же, что и для `url` *n* `.title`. |
| `url` *n* `.published_date` | *строка* | То же, что и для `url` *n* `.title`. |
| `url` *n* `.modified_date` | *строка* | То же, что и для `url` *n* `.title`. |
| `html_alg` | `Нет`<br>`HTMLCleaner` | Метод фильтрации содержимого. |
| `discard_lines` | *флаг* | Отбрасывание коротких строк. Используется со свойством `min_words` |
| `min_words` | *целое* | Минимальное число слов. |
| `discard_words` | *флаг* | Отбрасывание коротких строк. Используется со свойством `min_avg_len` |
| `min_avg_len` | *целое* | |
| `discard_scw` | *флаг* | Отбрасывание строк с множеством односимвольных слов. Используется со свойством `max_scw` |
| `max_scw` | *целое* | Максимальная процентная доля односимвольных слов в строке (0-100). |
| `discard_tags` | *флаг* | Отбрасывание строк, содержащих определенные теги. |
| `теги` | *строка* | Специальным символам должен предшествовать эскейп-символ обратной дробной черты (\). |
| `discard_spec_words` | *флаг* | Отбрасывание строк, содержащих конкретные строковые значения. |
| `words` | *строка* | Специальным символам должен предшествовать эскейп-символ обратной дробной черты (\). |

## Узел Text Mining: TextMiningWorkbench

При помощи приведенных ниже параметров можно определить или изменить узел с использованием сценариев. Сам узел называется `TextMiningWorkbench`.

**Важно!** В сценариях невозможно задать другой шаблон ресурса. Если нужен некоторый шаблон, его необходимо выбрать в диалоговом окне узла.

*Таблица 9. Свойства сценариев узла моделирования Text Mining*

| Свойства сценариев | Тип переменной | Описание свойства |
|---|---|---|
| `текст` | *field* | |
| `method` | `ReadText`<br>`ReadPath` | |
| `docType` | *целое* | Допустимы значения (0,1,2), где 0 = `Полностью текстовый`, 1 = `Структурированный текст` и 2 = `XML` |

*Таблица 9. Свойства сценариев узла моделирования Text Mining (продолжение)*

| Свойства сценариев | Тип переменной | Описание свойства |
|---|---|---|
| кодировка | Автоматически<br>"UTF-8"<br>"UTF-16"<br>"ISO-8859-1"<br>"US-ASCII"<br>"CP850"<br>"EUC-JP"<br>"SHIFT-JIS"<br>"ISO2022-JP" | Имейте в виду, что значения со специальными символами, такими как "UTF-8", должны заключаться в кавычки во избежание путаницы с математическими операциями. |
| unity | *целое* | Возможные значения - (0,1), где 0 = Абзац, а 1 = Документ |
| para_min | *целое* | |
| para_max | *целое* | |
| mtag | *строка* | Содержит все параметры mtag (из диалогового окна Параметры для файлов XML) |
| mclef | *строка* | Содержит все параметры mclef (из диалогового окна Параметры для файлов структурированного текста) |
| partition | *поле* | |
| custom_field | *флаг* | Показывает, будет ли задано поле разделения. |
| use_model_name | *флаг* | |
| model_name | *строка* | |
| use_partitioned_data | *флаг* | Если определено поле раздела, для построения модели используются только данные из раздела обучения. |
| model_output_type | Интерактивно<br>Модель | Значение Интерактивно задает создание модели категорий. Значение Модель задает создание модели понятий. |
| use_interactive_info | *флаг* | Только при интерактивном построении в сеансе инструментальной среды. |
| reuse_extraction_results | *флаг* | Только при интерактивном построении в сеансе инструментальной среды. |
| interactive_view | Категории<br>TLA<br>Кластеры | Только при интерактивном построении в сеансе инструментальной среды. |
| extract_top | *целое* | Этот параметр используется, когда model_type = Concept |
| use_check_top | *флаг* | |
| check_top | *целое* | |
| use_uncheck_top | *флаг* | |
| uncheck_top | *целое* | |

*Таблица 9. Свойства сценариев узла моделирования Text Mining (продолжение)*

| Свойства сценариев | Тип переменной | Описание свойства |
|---|---|---|
| язык | de<br>en<br>es<br>fr<br>it<br>ja<br>nl<br>pt | |
| frequency_limit | *целое* | Устарело в 14.0. |
| concept_count_limit | *целое* | Предельное число извлечений для понятий, у которых глобальная частота не меньше этого значения. недоступно для текста на японском языке |
| fix_punctuation | *флаг* | недоступно для текста на японском языке |
| fix_spelling | *флаг* | недоступно для текста на японском языке |
| spelling_limit | *целое* | недоступно для текста на японском языке |
| extract_uniterm | *флаг* | недоступно для текста на японском языке |
| extract_nonlinguistic | *флаг* | недоступно для текста на японском языке |
| upper_case | *флаг* | недоступно для текста на японском языке |
| group_names | *флаг* | недоступно для текста на японском языке |
| permutation | *целое* | Максимальное число нефункциональных перестановок слов (значение по умолчанию 3). Недоступно для текста на японском языке. |
| Заключения jp_algorithmset, только репрезентативные, только Все настроения | 0<br>1<br>2 | Только для извлечения из текстов на японском языке.<br>0 = Вторичное извлечение настроений<br>1 = Извлечение зависимостей<br>2 = Вторичный анализатор не задан. |
| jp_algorithm_sense_mode | 0<br>1<br>2 | Только для извлечения из текстов на японском языке.<br>0 = Только заключения<br>2 = Только репрезентативные<br>3 = Все настроения. |

## Слепок модели Text Mining: TMWBModelApplier

Свойства в приведенной ниже таблице можно использовать в сценариях. Сам слепок называется TMWBModelApplier.

*Таблица 10. Свойства слепка модели Text Mining*

| Свойства сценариев | Тип переменной | Описание свойства |
|---|---|---|
| scoring_mode | Поля<br>Записи | |
| field_values | Флаги<br>Количества | Эта опция недоступна в слепке модели категорий. Для поля Флаги задайте значение TRUE или FALSE |
| true_value | *строка* | Для поля Флаги определите значение, соответствующее логической истине. |
| false_value | *строка* | Для поля Флаги определите значение, соответствующее логической лжи. |

*Таблица 10. Свойства слепка модели Text Mining (продолжение)*

| Свойства сценариев | Тип переменной | Описание свойства |
|---|---|---|
| extension_concept | *строка* | Задайте расширение для имени поля. Имена полей генерируются с использованием имени понятия и этого расширения. Куда поместить это расширение, задается значением add_as. |
| extension_category | *строка* | Расширение имени поля. Можно задать использование расширяющего префикса/суффикса для имени поля или использование кодов категорий. Имена полей генерируются с использованием имени категории и этого расширения. Куда поместить это расширение, задается значением add_as. |
| add_as | Суффикс<br>Префикс | |
| fix_punctuation | *флаг* | |
| excluded_subcategories_descriptors | RollUpToParent<br>Ignore | Только для моделей категорий. Если подкатегория исключена. При помощи этой опции можно задать, как нужно обработать дескрипторы, принадлежащие тем подкатегориям, которые не были включены в скоринг. Есть две опции.<br>• Ignore. Опция Исключить дескрипторы полностью из скоринга задает, что дескрипторы тех подкатегорий, которые исключены (у которых выключены переключатели), игнорируются и не используются при скоринге.<br>• RollUpToParent. Опция Агрегировать дескрипторы с дескрипторами в родительской категории задает, что дескрпиторы тех подкатегорий, которые исключены (у которых выключены переключатели), используются как дескрипторы в родительской категории (надкатегории этой категории). Если исключено несколько уровней подкатегорий, для исключенной подкатегории низкого уровня выполняется откат до ближайшей доступной родительской категории |
| check_model | *флаг* | Устарело в версии 14 |
| текст | *field* | |
| method | ReadText<br>ReadPath | |
| docType | *целое* | Допустимы значения (0,1,2), где 0 = Полностью текстовый, 1 = Структурированный текст и 2 = XML |
| кодировка | Автоматически<br>"UTF-8"<br>"UTF-16"<br>"ISO-8859-1"<br>"US-ASCII"<br>"CP850"<br>"EUC-JP"<br>"SHIFT-JIS"<br>"ISO2022-JP" | Имейте в виду, что значения со специальными символами, такими как "UTF-8", должны заключаться в кавычки во избежание путаницы с математическими операциями. |

*Таблица 10. Свойства слепка модели Text Mining (продолжение)*

| Свойства сценариев | Тип переменной | Описание свойства |
|---|---|---|
| язык | de<br>en<br>es<br>fr<br>it<br>ja<br>nl<br>pt | |

## Узел Анализ текстовых связей: textlinkanalysis

Чтобы определить или изменить узел при помощи сценариев, можно использовать параметры в следующей таблице. Сам узел называется textlinkanalysis.

**Важно!** Задать шаблон ресурсов посредством сценариев невозможно. Чтобы выбрать шаблон, необходимо использовать диалоговое окно узла.

*Таблица 11. Свойства сценариев узла анализа текстовых связей (TLA)*

| Свойства сценариев | Тип переменной | Описание свойства |
|---|---|---|
| id_field | *поле* | |
| текст | *field* | |
| method | ReadText<br>ReadPath | |
| docType | *целое* | Допустимы значения (0,1,2), где 0 = Полностью текстовый, 1 = Структурированный текст и 2 = XML |
| кодировка | Автоматически<br>"UTF-8"<br>"UTF-16"<br>"ISO-8859-1"<br>"US-ASCII"<br>"CP850"<br>"EUC-JP"<br>"SHIFT-JIS"<br>"ISO2022-JP" | Имейте в виду, что значения со специальными символами, такими как "UTF-8", должны заключаться в кавычки во избежание путаницы с математическими операциями. |
| unity | *целое* | Возможные значения - (0,1), где 0 = Абзац, а 1 = Документ |
| para_min | *целое* | |
| para_max | *целое* | |
| mtag | *строка* | Содержит все параметры mtag (из диалогового окна Параметры для файлов XML) |
| mclef | *строка* | Содержит все параметры mclef (из диалогового окна Параметры для файлов структурированного текста) |
| язык | de<br>en<br>es<br>fr<br>it<br>ja<br>nl<br>pt | |

*Таблица 11. Свойства сценариев узла анализа текстовых связей (TLA) (продолжение)*

| Свойства сценариев | Тип переменной | Описание свойства |
|---|---|---|
| concept_count_limit | *целое* | Предельное число извлечений для понятий, у которых глобальная частота не меньше этого значения. недоступно для текста на японском языке |
| fix_punctuation | *флаг* | недоступно для текста на японском языке |
| fix_spelling | *флаг* | недоступно для текста на японском языке |
| spelling_limit | *целое* | недоступно для текста на японском языке |
| extract_uniterm | *флаг* | недоступно для текста на японском языке |
| extract_nonlinguistic | *флаг* | недоступно для текста на японском языке |
| upper_case | *флаг* | недоступно для текста на японском языке |
| group_names | *флаг* | недоступно для текста на японском языке |
| permutation | *целое* | Максимальное число нефункциональных перестановок слов (значение по умолчанию 3). Недоступно для текста на японском языке. |
| Заключения jp_algorithmset, только репрезентативные, только Все настроения | 0<br>1<br>2 | Только для извлечения из текстов на японском языке.<br>0 = Вторичное извлечение настроений<br>1 = Извлечение зависимостей<br>2 = Вторичный анализатор не задан. |
| jp_algorithm_sense_mode | 0<br>1<br>2 | Только для извлечения из текстов на японском языке.<br>0 = Только заключения<br>2 = Только репрезентативные<br>3 = Все настроения. |

## Узел перевода: translatenode

Свойства в приведенной ниже таблице можно использовать в сценариях. Сам узел называется translatenode.

*Таблица 12. Свойства узла Translate*

| Свойства сценариев | Тип переменной | Описание свойства |
|---|---|---|
| текст | *field* | |
| method | ReadText<br>ReadPath | |

*Таблица 12. Свойства узла Translate (продолжение)*

| Свойства сценариев | Тип переменной | Описание свойства |
|---|---|---|
| кодировка | Автоматически<br>"Big5", "Big5-HKSCS" ,<br>"UTF-8", "UTF-16",<br>"US-ASCII", "Latin1",<br>"CP850", "CP874",<br>"CP1250", "CP1251",<br>"CP1252", "CP1253",<br>"CP1254","CP1255",<br>"CP1256","CP1257",<br>"CP1258","GB18030",<br>"GB2312","GBK",<br>"eucJP","JIS7",<br>"SHIFT_JIS", "eucKR","TSCII",<br>"ucs2","KOI8-R",<br>"KOI8-U","ISO8859-1",<br>"ISO8859-2","ISO8859-3",<br>"ISO8859-4","ISO8859-5",<br>"ISO8859-6","ISO8859-7",<br>"ISO8859-8", "ISO8859-8-i",<br>"ISO8859-9","ISO8859-10",<br>"ISO8859-13","ISO8859-14",<br>"ISO8859-15","IBM 850",<br>"IBM 866","Apple Roman",<br>"TIS-620" | Имейте в виду, что значения со специальными символами, такими как "UTF-8", должны заключаться в кавычки во избежание путаницы с математическими операциями. |
| lw_server_type | LOC<br>WAN<br>HTTP | |
| lw_hostname | *строка* | |
| lw_port | *целое* | |
| url | *строка* | URL сервера перевода |
| apiKey | *строка* | |
| user_id | *строка* | |
| lpid | *целое* | Если задано *language_from* или *language_from_id*, это свойство не используется. |
| translate_from | Arabic, Chinese,<br>Traditional Chinese, Czech,<br>Danish,Dutch,<br>English,French,<br>German, Greek,<br>Hindi, Hungarian,<br>Italian, Japanese,<br>Korean, Persian,<br>Polish, Portuguese,<br>Romanian, Russian,<br>Spanish, Somali,<br>Шведский | |

*Таблица 12. Свойства узла Translate (продолжение)*

| Свойства сценариев | Тип переменной | Описание свойства |
|---|---|---|
| translate_from_id | ara, chi, cht, cze, dan, dut, eng, fra, ger, gre, hin, hun, ita, jpn, kor, per, pol, por, rum, rus, som, spa, swe | |
| translate_to | English | |
| translate_to_id | eng | |
| translation_accuracy | *целое* | Задает желаемый уровень точности для процесса перевода (выберите значение от 1 до 3). |
| use_previous_translation | *флаг* | Указывает, что результаты перевода уже существуют с предыдущего выполнения, и их можно использовать повторно. |
| translation_label | *строка* | Введите метку, чтобы идентифицировать результаты перевода для повторного использования. |

# Глава 8. Режим интерактивного сеанса инструментальной среды

Можно задать запуск интерактивного сеанса инструментальной среды из узла моделирования Text Mining во время выполнения потока. В инструментальной среде можно извлечь ключевые понятия из текстовых данных, построить категории, изучить паттерны и кластеры TLA и сгенерировать модели категорий. В этой главе мы обсуждаем с высокой перспективы работу с интерфейсом инструментальной среды и другими главными элементами, включая:

- **Результаты извлечения.** После извлечения вы располагаете так называемыми *понятиями*, то есть идентифицированными ключевыми словами и словосочетаниями, извлеченными из текстовых данных. Эти понятия сгруппированы в *типы*. Пользуясь понятиями и типами, можно изучать данные и создавать категории. С последними можно работать в представлении **Категория и понятия**.
- **Категории.** При помощи дескрипторов (таких как результаты извлечения, паттерны и правила) как определений можно вручную или автоматически создать набор категорий, которые назначаются тем документам и записям, в которых встречаются части определения категории. С последними можно работать в представлении **Категория и понятия**.
- **Кластеры.** *Кластер* объединяет понятия, если между ними обнаружены связи. Понятия группируются по сложному алгоритму, в котором, помимо других факторов, учитывается, сколько раз два понятия встречаются вместе, по сравнению с тем, сколько раз они встречаются по отдельности. С такими объединениями можно работать в представлении **Кластеры**. Кроме того, понятия из кластера можно добавить в категорию.
- **Паттерны TLA.** Если в ваших лингвистических ресурсах есть правила паттернов TLA (text link analysis, анализ текстовых связей), или вы используете шаблон ресурсов, в которому уже есть какие-то правила TLA, вы можете извлечь паттерны из текстовых данных. Такие паттерны могут быть полезны для обнаружения интересующих вас взаимосвязей между понятиями в ваших данных. Кроме того, эти паттерны можно использовать как дескрипторы в категориях. С ними можно работать в представлении **Text Link Analysis**. Для текста на японском нужно выбрать дополнительный анализатор и включить извлечение TLA.
- **Лингвистические ресурсы.** Процесс извлечения регулируется набором параметров и пользуется лингвистическими определениями при извлечении текста и его обработке. С наборами параметров и определений в виде шаблонов и библиотек можно работать в представлении **Редактор ресурсов**.

## Представление категорий и понятий

Интерфейс прикладной программы состоит из ряда представлений. Представление категорий и понятий - это окно, в котором можно создавать и изучать категории, а также изучать и подправлять результаты извлечения. *Категориями* называются группы тесно связанных идей и паттернов, назначаемых для документов и записей в процессе скоринга. В отличии от них *понятиями* называются результаты извлечения, относящиеся к низшему из уровней строительных блоков, так называемых дескрипторов, для категорий.

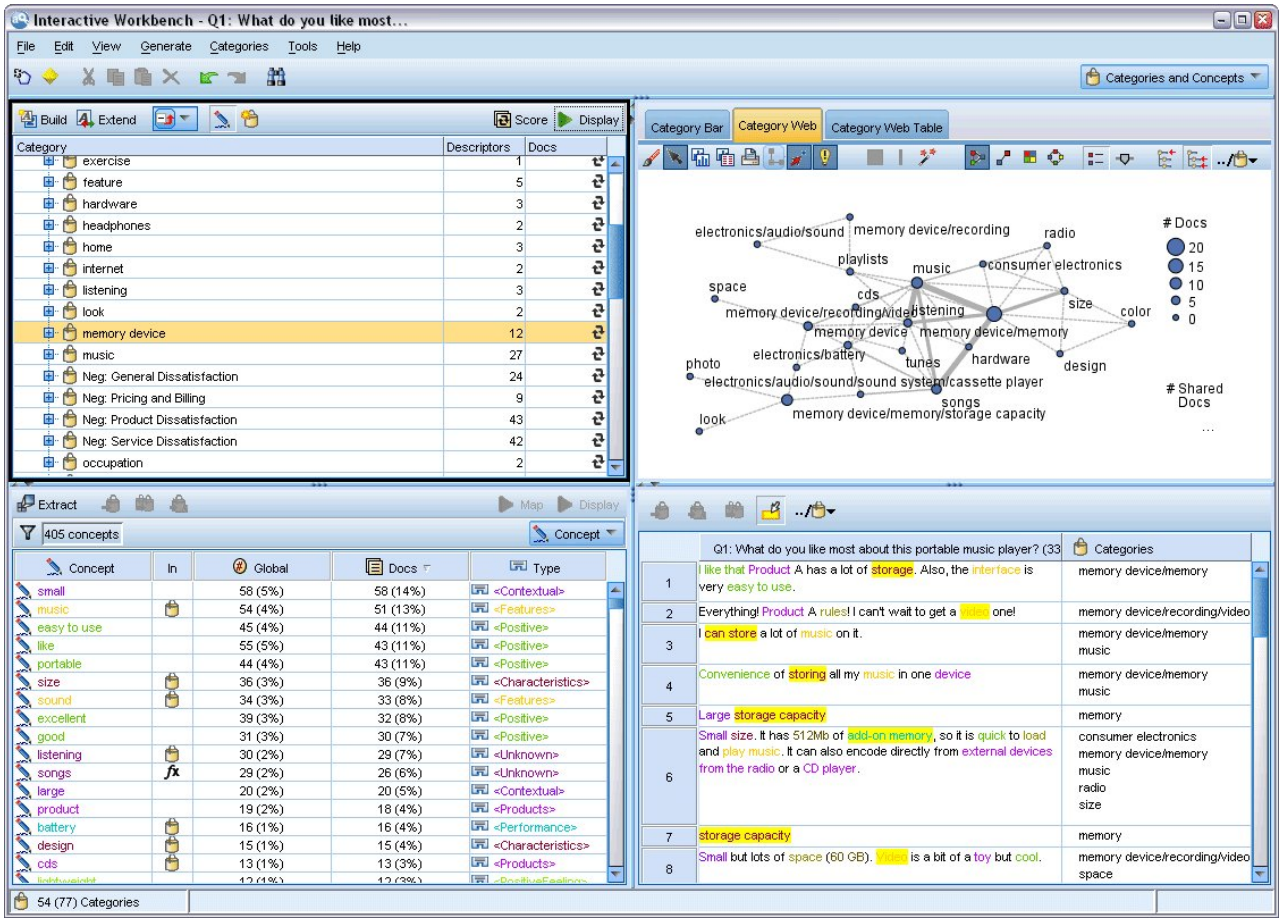



*Рисунок 23. Представление Категории и понятия*

Представление категорий и понятий разбито на четыре панели, которые можно по отдельности скрывать и выводить, выбирая имя панели в меню Вид. Дополнительную информацию смотрите в разделе Глава 10, “Категоризация текстовых данных”, на стр. 101.

## Панель категорий

Занимает верхний левый угол; эта область содержит таблицу, в которой можно работать с любыми категориями, которые вы строите. После извлечения понятий и типов из текстовых данных, можно начать строить категории при помощи методов, таких как семантические сети и включение понятий, или создавая категории вручную. При двойном щелчке по имени категории открывается диалоговое окно Определения категорий, которое содержит все дескрипторы, составляющие определение, такие как понятия, типы и правила. Дополнительную информацию смотрите в разделе Глава 10, “Категоризация текстовых данных”, на стр. 101. Не все автоматические методы доступны для всех языков.

Выбирая строки на панели, можно выводить информацию о соответствующих документах/записях или дескрипторах на панелях Данные и Визуализация.

## Панель результатов извлечения

Занимает нижний левый угол; эта область представляет результаты извлечения. При извлечении механизм извлечения читает текстовые данные, идентифицирует нужные понятия и каждому назначает тип. **Понятия** - это слова или словосочетания, извлеченные из текстовых данных. **Типы** - это объединения понятий по

смыслу, сохраненные в виде словарей типов. По завершении извлечения понятия и типы выводятся на панели результатов извлечения с цветовым кодированием. Дополнительную информацию смотрите в разделе “Результаты извлечения: Понятия и типы” на стр. 87.

Чтобы увидеть набор терминов, охваченных понятием, остановите указатель мыши на имени понятия. При этом выводится подсказка, содержащая имя понятия и одну или несколько строк терминов, сгруппированных этой понятием. К таким терминам относятся синонимы, определенные в лингвистических ресурсах (независимо от того, найдены ли они в тексте), а также извлеченные формы единственного и множественного числа, термины с перестановками, термины с нечеткой группировкой и так далее. Вы можете скопировать эти термины или просмотреть полный список терминов понятия, щелкнув правой кнопкой по имени понятия и выбрав опцию в контекстном меню.

Исследование текста - это интерактивный процесс, в котором результаты извлечения пересматриваются в контексте текстовых данных; выполняется тонкая настройка для получения новых результатов и их повторной оценки. Результаты извлечения можно уточнить, модифицировав лингвистические ресурсы. Такую тонкую настройку можно частично выполнить непосредственно на панели Результаты извлечения или на панели данных; ее также выполняют непосредственно в представлении Редактор ресурсов. Дополнительную информацию смотрите в разделе “Представление Редактор ресурсов” на стр. 80.

## Панель Визуализация

Эта область в верхнем правом углу представляет ряд перспектив для для объединений при категоризации документов/записей. Все диаграммы содержат сходную информацию, но каждая представляет ее по-своему или на своем уровне детализации. Эти диаграммы и графики полезны при анализе результатов категоризации и тонкой настройки категорий или отчетов. Например, на диаграмме можно обнаружить слишком близкие категории (скажем, более 75% их записей - общие) или слишком далекие. Содержимое диаграммы соответствует выбранному на других панелях. Дополнительную информацию смотрите в разделе “Графики и диаграммы категорий” на стр. 159.

## Панель данных

Панель данных расположена в нижнем правом углу. Эта панель представляет собой таблицу, содержащую документы или записи, соответствующие выбранному в другой области представления. В зависимости от выбранного только соответствующий текст выводится на панели данных. Выделив нужное, нажмите кнопку **Вывести**, и на панели данных появится соответствующий текст.

Если есть выделение на другой панели, соответствующие документы или записи содержат понятия, выделенные цветом для удобства идентификации в тексте. Кроме того, если остановить указатель мыши на элементе того или иного цвета, выводится подсказка с именем того понятия, под которым этот элемент был извлечен, и назначенный этому понятию тип. Дополнительную информацию смотрите в разделе “Панель Данные” на стр. 110.

## Поиск в представлении категорий и понятий

В некоторых случаях нужно быстро найти информацию в конкретном разделе. На панели инструментов Поиск можно ввести искомую строку и задать другие критерии поиска, такие как учет регистра и направление поиска. Затем можно выбрать панель, на которой нужно выполнить поиск.

**Чтобы использовать функцию Поиск**

1. В представлении категорий и понятий выберите **Правка > Поиск** в меню. Панель инструментов Поиск выводится над панелью категорий и над панелями Визуализация.
2. Введите строку искомого слова в текстовом окне. Кнопками на панели поиска можно задать учет регистра, частичное соответствие и направление поиска.
3. На панели инструментов щелкните по имени панели, на которой нужно выполнить поиск. Если соответствие найдено, текст выделяется в окне.

4. Для поиска следующего соответствия щелкните по имени панели еще раз.

## Представление кластеров

В представлении кластеров можно строить и изучать результаты поиска кластеров в ваших текстовых данных. *Кластеры* - это группировки понятий, сгенерированные алгоритмами кластеризации с учетом того, как часто встречаются понятия и как часто часто они встречаются вместе. Кластеры предназначены для того, чтобы сгруппировать те понятия, которые встречаются совместно, в то время как цель категорий - сгруппировать документы или записи с учетом того, как содержащийся в них текст соответствует дескрипторам (понятиям, правилам, паттернам) для каждого понятия.

Чем чаще понятия в кластере встречаются совместно и чем реже они встречаются с другими понятиями, тем лучше кластер отражает интересные взаимосвязи между понятиями. Два понятия считаются встреченными совместно, если они сами или их синонимы или термины встретились в одном и том же документе или записи. Дополнительную информацию смотрите в разделе Глава 11, “Анализ кластеров”, на стр. 147.

Вы можете строить кластеры и изучать их в наборе диаграмм, помогающих выявить такие взаимосвязи среди понятий, на поиск которых иначе ушло бы много времени. Хотя вы не можете добавлять свои кластеры в категории, можно при помощи диалогового окна определений кластеров добавлять понятия из кластера в категорию. Дополнительную информацию смотрите в разделе “Определения кластеров” на стр. 151.

Можно вносить изменения в параметры кластеризации, влияя на результаты. Дополнительную информацию смотрите в разделе “Построение кластеров” на стр. 148.

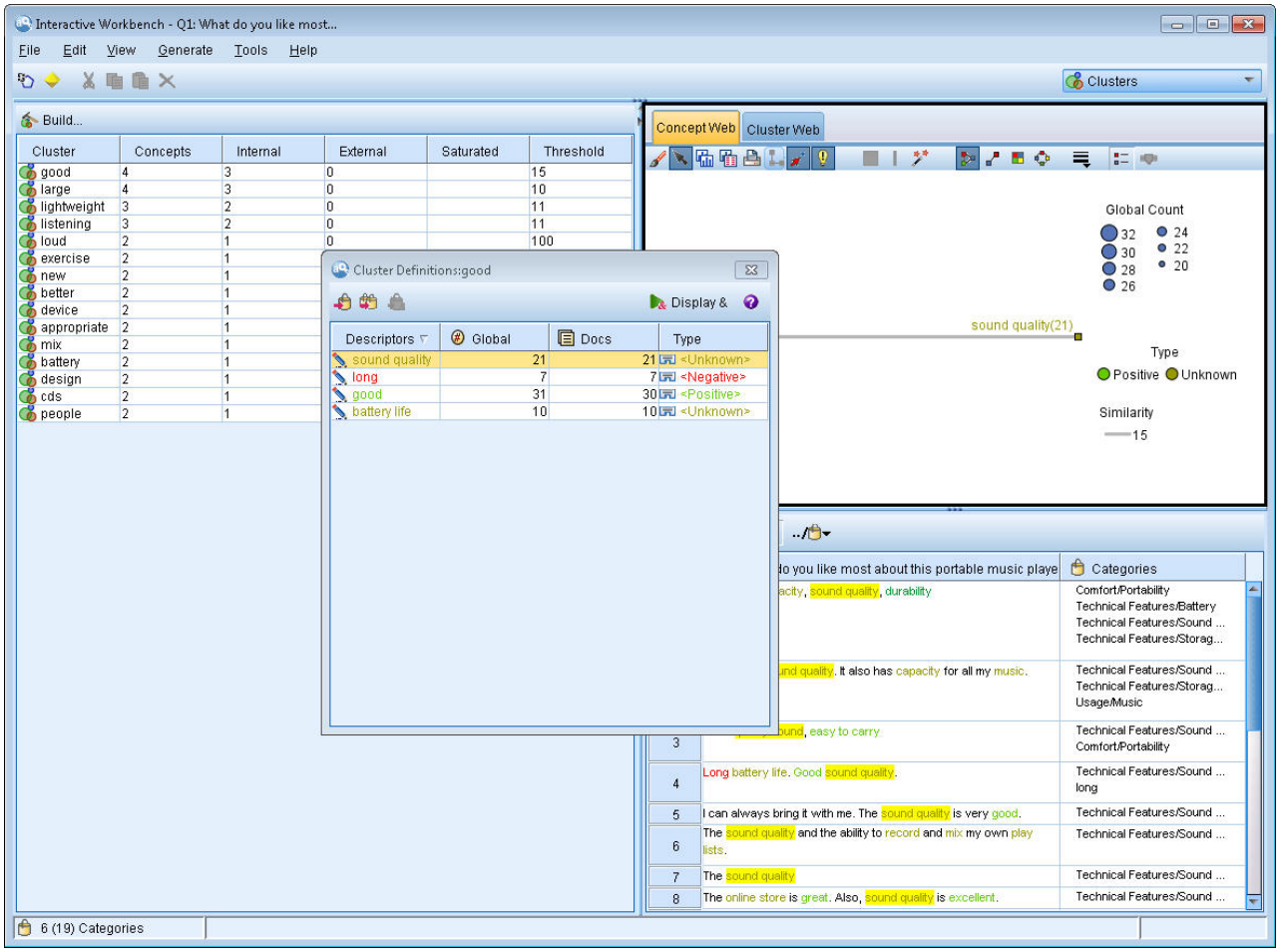



*Рисунок 24. Вид представления Кластеры*

Представление Кластеры разбито на три панели, которые можно по отдельности скрывать и выводить, выбирая имя панели в меню Вид. Обычно выводятся только две панели, Кластеры и Визуализация.

## Панель Кластеры

Расположена слева; на этой панели представлены кластеры, обнаруженные в текстовых данных. Вы можете создавать результаты кластеризации, нажимая кнопку **Построить**. Кластеры образуются по алгоритму кластеризации, который пытается идентифицировать понятия, которые часто встречаются вместе.

При новом извлечении результаты кластеризации очищаются, и для получения новейших результатов нужно построить кластеры еще раз. При построении кластеров можно изменять некоторые параметры, такие как максимальное число создаваемых кластеров, максимально допустимое число понятий в одном кластере и максимально допустимое число связей кластера с понятиями вне кластера. Дополнительную информацию смотрите в разделе “Исследование кластеров” на стр. 151.

## Панель Визуализация

Эта панель в верхнем правом углу предлагает две перспективы кластеризации: веб-диаграмма понятия и веб-диаграмма кластера. Если она не видна, ее можно открыть из меню Вид (**Вид > Визуализация**). В зависимости от выбранного на панели кластеров можно просматривать взаимодействия между кластерами или внутри кластеров. Результаты представляются в нескольких форматах:

- **Веб-диаграмма понятия.** Веб-диаграмма содержит все понятия в выбранном кластере (кластерах), а также связанные понятия вне кластера.

- **Веб-диаграмма кластера.** Веб-диаграмма содержит связи выбранного кластера (кластеров) с другими кластерами, а также любые связи между другими кластерами.

**Примечание:** Чтобы вывести веб-диаграмму кластера, нужно, чтобы были построены кластеры с внешними связями. Внешние связи - это связи между понятиями из разных кластеров (одно понятие пары в одном кластере, а другое - вне этого кластера, в другом кластере). Дополнительную информацию смотрите в разделе “Диаграммы кластеров” на стр. 161.

### Панель Данные

Панель данных расположена в нижнем правом углу и по умолчанию скрыта. Нельзя вывести результаты панели данных из панели кластеров, поскольку кластеры охватывают множество документов/записей, и результаты данных будут неинтересные. Но можно просмотреть данные, соответствующие выбранному в диалоговом окне Определения кластеров. В зависимости от выбранного в этом диалоговом окне только соответствующий текст выводится на панели данных. Выделив нужное, нажмите кнопку **Вывести &**, и на панели данных появятся документы или записи, содержащие все эти понятия одновременно.

Соответствующие документы или записи содержат понятия, выделенные цветом для удобства идентификации в тексте. Кроме того, если остановить указатель мыши на элементе того или иного цвета, выводится то понятие, под которым этот элемент был извлечен, и назначенный этому понятию тип. Панель данных может содержать несколько столбцов, но столбец текстового поля выводится всегда. Он содержит имя текстового поля, использованного при извлечении, или имя документа, если текстовые данные находятся во многих различных файлах. Доступны и другие столбцы. Дополнительную информацию смотрите в разделе “Панель Данные” на стр. 110.

## Представление Text Link Analysis

В представлении Text Link Analysis можно строить и изучать паттерны TLA, найденные в ваших текстовых данных. TLA (Text link analysis, анализ текстовых связей) - это метод сопоставления паттернов, при помощи которого можно задать правила TLA и сравнить их с фактически извлеченными понятиями и взаимосвязями, найденными в вашем тексте.

Паттерны особенно полезны при попытке обнаружить взаимосвязи между понятиями или мнениями по конкретной теме. Некоторые примеры включают в себя потребность извлекать мнения о продуктах из данных опроса, взаимосвязи в геноме из отчетов о медицинских исследованиях и взаимосвязи между людьми или местами из разведывательных данных.

После того, как были извлечены какие-то паттерны TLA, их можно изучить на панелях данных или визуализации; более того, их можно добавить в категории в представлении категорий и понятий. Чтобы получить результаты поиска паттернов TLA, нужно, чтобы какие-то правила TLA были определены в шаблоне ресурсов или в используемых библиотеках. Дополнительную информацию смотрите в разделе Глава 19, “О правилах текстовых связей”, на стр. 215.

В этом представлении представлены результаты поиска паттернов TLA, если вы выбрали извлечение паттернов. Если вы этого не делали, нужно нажать кнопку **Извлечь** и выбрать опцию включить извлечение паттернов .

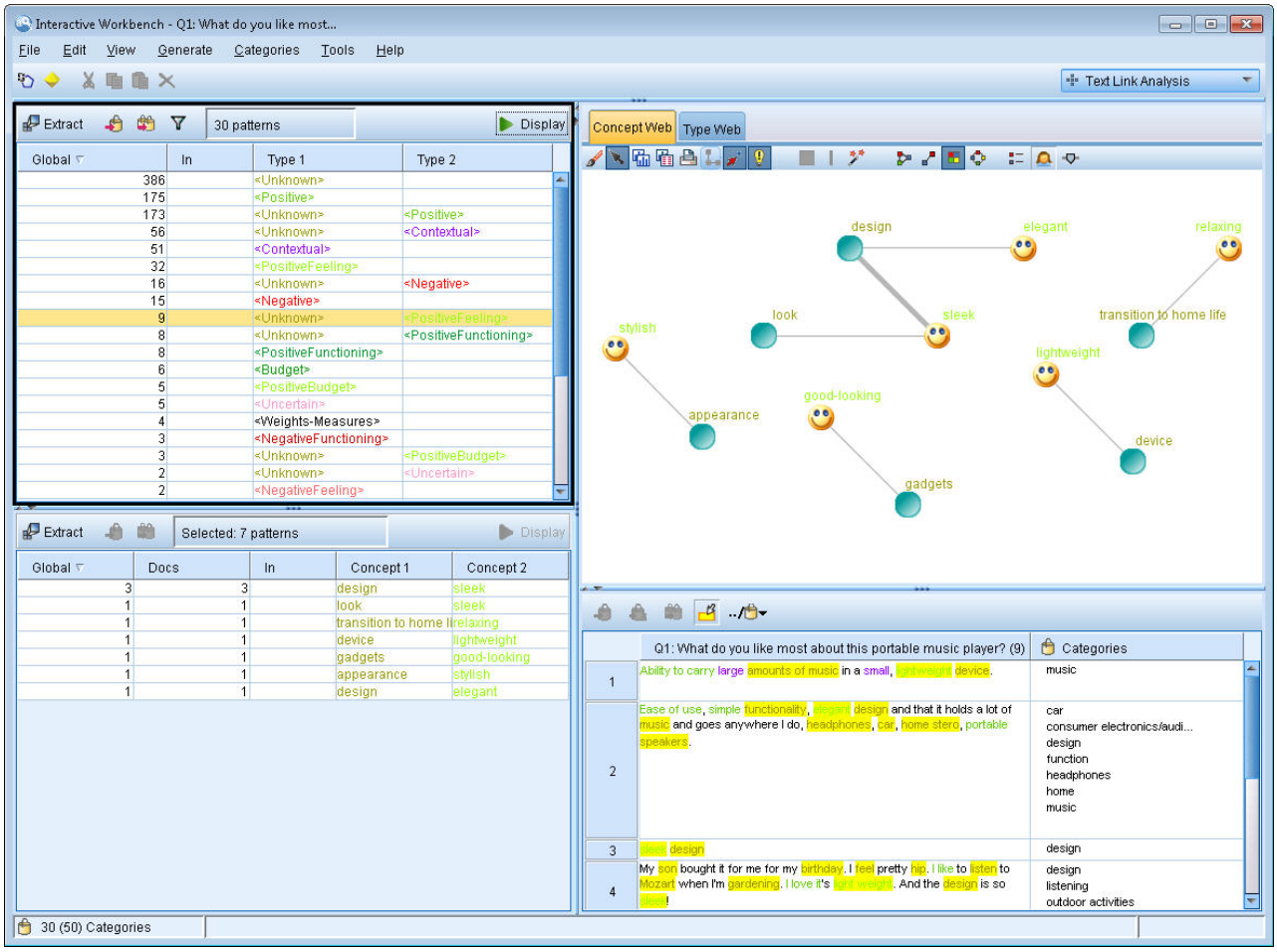



*Рисунок 25. Представление Text Link Analysis*

Представление Text Link Analysis разбито на четыре панели, которые можно по отдельности скрывать и выводить, выбирая имя панели в меню Вид. Дополнительную информацию смотрите в разделе Глава 12, “Изучаем анализ текстовых связей (Text Link Analysis, TLA)”, на стр. 153.

## Панели паттернов типа и понятия.

На взаимосвязанных панелях паттернов типа и понятия, расположенных слева, можно изучать и выбирать результаты поиска паттернов TLA. Паттерны представляют собой последовательности, которые могут содержать до шести типов или понятий. Обратите внимание на то, что для японского текста паттерны представляют собой последовательности, которые могут содержать не более одного или двух типов или понятий. Правилом паттерна TLA, определенным в лингвистических ресурсах, диктуется сложность результатов поиска паттернов. Дополнительную информацию смотрите в разделе Глава 19, “О правилах текстовых связей”, на стр. 215.

Результаты поиска паттернов сначала группируются на уровне типа, а затем делятся на паттерны понятия. Поэтому есть две различные панели результатов: Паттерны типа (вверху слева) и Паттерны понятия (внизу слева).

- **Паттерны типа.** Панель Паттерны типа представляет извлеченные паттерны, содержащие один или несколько связанных типов, соответствующих правилу паттерна TLA. Паттерны типа устроены как <Организация> + <Положение> + <Положительный>; в этом примере нужно получить положительные отзывы о конкретной организации в конкретном положении.

- **Паттерны понятия.** На панели Паттерны понятия представлены извлеченные паттерны на уровне понятия для всех типов паттернов, выбранных в данный момент на расположенной выше панели Паттерны типа. Паттерны понятия имеют структуру вида `отель + париж + чудесный`.

Здесь можно просматривать результаты, так же как результаты извлечения в представлении категорий и понятий. Если потребуется внести уточнения в типы и понятия, входящие в эти паттерны, это можно сделать на панели результатов извлечения в представлении категорий и понятий или непосредственно в редакторе ресурсов, а потом извлечь паттерны еще раз.

### Панель Визуализация

Эта панель в верхнем правом углу представления Text Link Analysis представляет веб-диаграмму выбранных паттернов как паттернов типа или паттернов понятия. Если она не видна, ее можно открыть из меню Вид (**Вид > Визуализация**). В зависимости от выбранного на других панелях можно просматривать соответствующие взаимодействия между документами/записями и паттернами.

Результаты представляются в нескольких форматах:

- **Диаграмма понятия.** Эта диаграмма представляет все понятия в выбранных паттернах. Толщина линий и размер узлов (если не выводятся значки) на диаграмме понятия показывают число глобальных вхождений в выбранной таблице.
- **Диаграмма типа.** Эта диаграмма представляет все типы в выбранных паттернах. Толщина линий и размер узлов (если не выводятся значки) на диаграмме показывают число глобальных вхождений в выбранной таблице. Узлы представлены цветом типа или значком.

Дополнительную информацию смотрите в разделе “Диаграммы TLA (Text Link Analysis, анализ ссылок в тексте)” на стр. 162.

### Панель Данные

Панель данных расположена в нижнем правом углу. Эта панель представляет собой таблицу, содержащую документы или записи, соответствующие выбранному в другой области представления. В зависимости от выбранного только соответствующий текст выводится на панели данных. Выделив нужное, нажмите кнопку **Вывести**, и на панели данных появится соответствующий текст.

Если есть выделение на другой панели, соответствующие документы или записи содержат понятия, выделенные цветом для удобства идентификации в тексте. Кроме того, если остановить указатель мыши на элементе того или иного цвета, выводится подсказка с именем того понятия, под которым этот элемент был извлечен, и назначенный этому понятию тип. Дополнительную информацию смотрите в разделе “Панель Данные” на стр. 110.

## Представление Редактор ресурсов

IBM SPSS Modeler Text Analytics быстро и точно захватывает основные понятия из текстовых данных, пользуясь надежным механизмом извлечения. В значительной мере этот механизм опирается на лингвистические ресурсы, предписывающие, каким образом следует анализировать и интерпретировать большие объемы неструктурированных текстовых данных.

Представление Редактор ресурсов служит для просмотра и тонкой настройки лингвистических ресурсов, используемых при извлечении понятий, группировки их по типам, обнаружения паттернов в текстовых данных и многого другого. IBM SPSS Modeler Text Analytics предлагает несколько заранее сконфигурированных шаблонов ресурсов. Кроме того, для некоторых языков можно использовать ресурсы в пакетах анализа текста. Дополнительную информацию смотрите в разделе “Использование пакетов анализа текста (Text Analysis Package)” на стр. 140.

Поскольку эти ресурсы не всегда идеально адаптированы к контексту ваших данных, вы можете создавать и редактировать свои ресурсы и работать с ними; это могут быть ресурсы для конкретного контекста или домена в Редактор ресурсов. Дополнительную информацию смотрите в разделе Глава 16, “Работа с библиотеками”, на стр. 179.

Чтобы упростить процедуру тонкой настройки лингвистических ресурсов, можно выполнять обычные словарные задачи непосредственно из представления Категории и понятия при помощи контекстных меню на панелях Результаты извлечения и Данные. Дополнительную информацию смотрите в разделе “Уточнение результатов извлечения” на стр. 95.

*Примечание:* Интерфейс для ресурсов, настроенный на японский текст, имеет некоторые отличия.

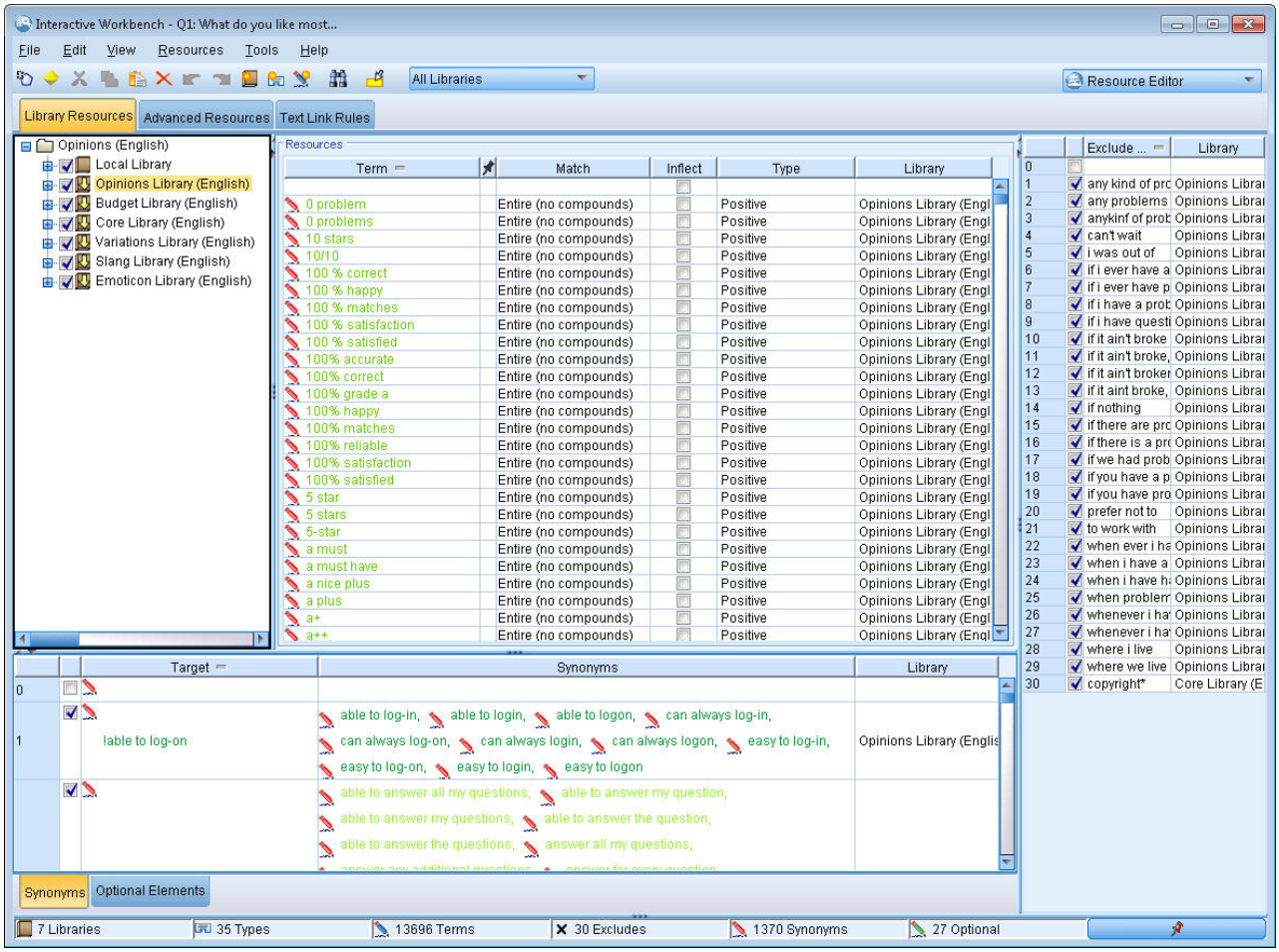



*Рисунок 26. Представление Редактор ресурсов*

Операции, выполняемые в представлении Редактор ресурсов, связаны с управлением лингвистическими ресурсами и их тонкой настройкой. Эти ресурсы хранятся в виде шаблонов и библиотек. Представление Редактор ресурсов состоит из четырех частей: панель Дерево библиотек, панель Словарь типов, панель Словарь подстановок и панель Словарь исключения.

**Примечание:** Дополнительную информацию смотрите в разделе “Интерфейс редактора” на стр. 170.

# Настройка опций

Общие опции для IBM SPSS Modeler Text Analytics можно задать в диалоговом окне Опции. В этом диалоговом окне есть следующие вкладки:

- **Сеанс.** Эта вкладка содержит общие опции и разделители.
- **Вывести.** Эта вкладка содержит опции используемых в интерфейсе цветов.
- **Звуки.** Эта вкладка содержит опции звуковых сигналов.

Чтобы изменить опции

1. В меню выберите **Сервис > Опции**. Откроется диалоговое окно Опции.
2. Выберите вкладку, содержащую информацию, которую нужно изменить.
3. Измените любую опцию.
4. Нажмите кнопку **OK**, чтобы сохранить изменения.

# Опции: вкладка Сеанс

На этой вкладке можно задать некоторые базовые параметры.

**Панель данных и Вывести диаграмму категорий.** Эти опции влияют на представление данных на панели данных и на панели Визуализация в представлении категорий и понятий.

- **Предел вывода для панели данных и веб-диаграммы категорий.** Эта опция задает максимальное число документов для вывода на панелях данных или в диаграммах в представлении категорий и понятий.
- **Показать категории для документов/записей во время вывода.** Если эта опция выбрана, документы или записи оцениваются при каждом выводе, так что все содержащие их категории можно вывести в столбце категорий на панели данных, а также в диаграммах категорий. В некоторых случаях, особенно при больших наборах данных, нужно отключить эту опцию, что существенно ускоряет вывод данных и диаграмм.

**Добавить в категорию из панели данных.** Эти опции влияют на то, какие элементы добавляются в категории при добавлении документов и записей из панели данных.

- **В представлении Категории и понятия - копировать.** Добавление документа или записи с панели данных в этом представлении копирует поверх опций **Только понятия** или **Понятия и паттерны**.
- **В представлении Анализ текстовых связей (Text Link Analysis, TLA) - копировать.** Добавление документа или записи с панели данных в этом представлении копирует поверх опций **Только паттерны** или **Понятия и паттерны**.

**Разделитель редактора ресурсов.** Выберите символ-разделитель при вводе элементов, таких как понятия, синонимы и необязательные элементы, в представлении Редактор ресурсов.

# Опции: вкладка Дисплей

На этой вкладке можно изменить опции, влияющие на оформление прикладной программы и цвета, помогающие различать элементы.

*Примечание*: Чтобы переключиться в классическое оформление продукта или оформление одного из прошлых выпусков, откройте диалоговое окно Пользовательские опции в меню Инструменты в главном окне IBM SPSS Modeler.

**Пользовательские цвета.** Отредактируйте цвета элементов на экране. Можно изменить цвет для каждого элемента в таблице. Чтобы задать пользовательский цвет, щелкните по области цвета справа от нужного элемента и выберите цвет в выпадающем списке цветов.

- **Неизвлеченный текст.** Текстовые данные, которые не были извлечены, но выводятся на панели данных.
- **Фон выделения.** Цвет фона выделения в тексте, когда выбираются элементы на панелях или текст на панели данных.

- **Фон, когда требуется извлечение.** Цвет фона панелей Результаты извлечения, Паттерны и Кластеры, показывающий, что в библиотеки внесены изменения и требуется извлечение.
- **Фон возвращенных категорий.** Цвет фон при выводе категорий после операции.
- **Тип по умолчанию.** Цвет по умолчанию для типов и понятий на панелях Данные и Результаты извлечения. Этот цвет будет применяться к любым пользовательским типам, создаваемым в редакторе ресурсов. Цвет по умолчанию можно перезадать для пользовательских словарей типов, отредактировав свойства для словарей типов в Редактор ресурсов. Дополнительную информацию смотрите в разделе “Создание типов” на стр. 191.
- **Полосатая таблица 1.** Первый из двух цветов, используемых попеременно в диалоговом окне Редактировать принудительные понятия, чтобы различать наборы строк.
- **Полосатая таблица 2.** Второй из двух цветов, используемых попеременно в диалоговом окне Редактировать принудительные понятия, чтобы различать наборы строк.

*Примечание*: Если нажать кнопку **Восстановить значения по умолчанию**, все опции в этом диалоговом окне возвращаются к значениям, заданным при установке этого продукта.

## Опции: вкладка Звуки

На этой вкладке можно изменить опции, влияющие на звуки. В разделе Звуки для событий можно указать, какой звук использовать для уведомления о событии. Доступен ряд звуков. Чтобы найти и выбрать звук, нажмите кнопку просмотра (**...**). Файлы *.wav* для создания звуков для IBM SPSS Modeler Text Analytics хранятся в подкаталоге *media* каталога установки. Если воспроизводить звуки нежелательно, можно **Отключить все звуки**. По умолчанию звуки отключены.

*Примечание*: Если нажать кнопку **Восстановить значения по умолчанию**, все опции в этом диалоговом окне возвращаются к значениям, заданным при установке этого продукта.

---

## Параметры для справки Microsoft Internet Explorer

Параметры Microsoft Internet Explorer

Большинство компонентов справки в этой прикладной программе пользуются технологией на основе Microsoft Internet Explorer. Некоторые версии Internet Explorer (включая версию, поставлявшуюся с Microsoft Windows XP, Service Pack 2) по умолчанию будут блокировать то, что они сочтут "активным содержимым" в окнах Internet Explorer на локальном компьютере. Эта настройка по умолчанию может приводить к блокированию содержимого в компонентах справки. Чтобы увидеть все содержимое справки, можно изменить работу Internet Explorer по умолчанию.

1. В меню Internet Explorer выберите:

   **Сервис > Свойства браузера...**
2. Щелкните по вкладке **Дополнительно**.
3. Прокрутите вниз до раздела **Безопасность**.
4. Выберите (включите переключатель) **Разрешать запуск активного содержимого файлов на моем компьютере**.

---

## Генерирование слепков модели и узлов моделирования

Находясь в интерактивном сеансе, можно использовать проделанную работу, чтобы сгенерировать одно из двух:

- **Узел моделирования Text Mining.** Узел моделирования, сгенерированный из интерактивного сеанса инструментальной среды, - это узел Text Mining, параметры и опции которого отражают параметры и опции, сохраненные в открытом интерактивном сеансе. Это может оказаться полезным, когда у вас больше нет исходного узла Text Mining или нужно создать новую версию. Дополнительную информацию смотрите в разделе Глава 3, “Исследование концепций и категорий”, на стр. 19.

- **Слепок модели категорий.** Слепок модели, сгенерированный из интерактивного сеанса инструментальной среды, - это слепок модели категорий. Чтобы сгенерировать слепок модели категорий, необходимо, чтобы в представлении категорий и понятий была хотя бы одна категория. Дополнительную информацию смотрите в разделе “Слепок Text Mining: модель категорий” на стр. 40.

Чтобы сгенерировать узел моделирования Text Mining

1. В меню выберите **Генерировать > Генерировать узел моделирования**. Узел моделирования Text Mining добавляется на рабочий холст с использованием всех параметров текущего интерактивного сеанса инструментальной среды. Узел получает имя согласно текстовому полю.

Чтобы сгенерировать слепок модели категорий

1. В меню выберите **Генерировать > Генерировать модель**. Слепок модели генерируется непосредственно на палитре моделей с именем по умолчанию.

## Обновление узлов моделирования и сохранение

При работе в интерактивном сеансе рекомендуется время от времени обновлять узел моделирования, чтобы сохранять изменения. Кроме того, следует обновлять узел моделирования по завершении работы в интерактивном сеансе инструментальной среды, если нужно сохранить работу. При обновлении узла моделирования содержимое сеанса инструментальной среды сохраняется на узле Text Mining, из которого был запущен интерактивный сеанс инструментальной среды. Окно вывода при этом не закрывается.

**Важно!** При обновлении не сохраняется поток. Чтобы сохранить поток, сохранение нужно выполнить в главном окне IBM SPSS Modeler после обновления узла моделирования.

Чтобы обновить узел моделирования

1. В меню выберите **Файл > Обновить узел моделирования**. В узле моделирования обновляются параметры построения и извлечения, а также любые имеющиеся опции и категории.

## Закрытие и завершение сеансов

По окончании работы в сеансе его можно покинуть тремя различными способами.

- **Сохранить.** Эта опция позволяет сначала сохранить работу обратным порядком на исходном узле моделирования для будущих сеансов, а также опубликовать любые библиотеки для их повторного использования в других сеансах. Дополнительную информацию смотрите в разделе “Совместное использование библиотек” на стр. 184. После сохранения окно сеанса закрывается, и сеанс удаляется из менеджера вывода в окне IBM SPSS Modeler.
- **Выйти.** Эта опция отбрасывает сохраненную работу, закрывает окно сеанса и удаляет сеанс из менеджера вывода в окне IBM SPSS Modeler. Для освобождения памяти мы рекомендуем сохранение всей важной работы и выход из сеанса.
- **Закрыть.** Эта опция не сохраняет и не отбрасывает никакую работу. Данная опция закрывает окно сеанса, но сеанс продолжает выполняться. Сеанс можно снова открыть, выбрав его в менеджере вывода в окне IBM SPSS Modeler.

Чтобы закрыть сеанс инструментальной среды:

1. В меню выберите **Файл > Закрыть**.

## Средства доступности через клавиатуру

Для упрощения доступа к функциональным возможностям продукта в интерфейсе интерактивной инструментальной среды предлагаются клавиши быстрого вызова. На самом простом уровне можно активировать меню окна, нажав клавишу Alt плюс нужную клавишу (например, Alt+F, чтобы открыть меню Файл), или перебирать управляющие элементы диалогового окна, нажимая клавишу Tab. В этом разделе описываются клавиши быстрого вызова для альтернативной навигации. Для интерфейса IBM SPSS Modeler существуют другие сочетания клавиш.

*Таблица 13. Общие клавиши быстрого вызова*

| **Клавиши быстрого вызова** | **Функция** |
|---|---|
| Ctrl+1 | Показать первую вкладку на панели с вкладками. |
| Ctrl+2 | Показать вторую вкладку на панели с вкладками. |
| Ctrl+A | Выбрать все элементы на панели, где установлен фокус. |
| Ctrl+C | Скопировать выделенный текст в буфер обмена. |
| Ctrl+E | Запустить извлечение в представлениях Категории и понятия и Анализ текстовых связей. |
| Ctrl+F | Показать панель инструментов Поиск в Редактор ресурсов/Редактор шаблонов, если она еще не показана, и перевести на нее фокус. |
| Ctrl+I | В представлении Категории и понятия открыть диалоговое окно Определения категорий для выбранной категории. В представлении Кластер открыть диалоговое окно Определения кластеров для выбранного кластера. |
| Ctrl+R | Открыть диалоговое окно Добавить термины в Редактор ресурсов/Редактор шаблонов. |
| Ctrl+T | Открыть диалоговое окно Свойства типов, чтобы создать новый тип в Редактор ресурсов/Редактор шаблонов. |
| Ctrl+V | Вставить содержимое буфера обмена. |
| Ctrl+X | Вырезать выбранный элемент из Редактор ресурсов/Редактор шаблонов. |
| Ctrl+Y | Повторить последнее действие в данном представлении. |
| Ctrl+Z | Отменить последнее действие в данном представлении. |
| F1 | Вывести справку или, находясь в диалоговом окне, вывести контекстную справку для элемента. |
| F2 | Включение и выключение режима изменений в ячейках таблицы. |
| F6 | Переместить фокус между главными панелями активного представления. |
| F8 | Переместить фокус в область разделителя панелей для изменения размеров. |
| F10 | Раскрыть главное меню Файл. |
| Кнопки стрелок вверх и вниз | Изменить вертикальный размер панели при фокусе в области разделителя панелей. |
| Кнопки со стрелками влево и вправо | Изменить горизонтальный размер панели при фокусе в области разделителя панелей. |
| Home, End | Изменить размер панелей до минимального или максимального при фокусе в области разделителя. |
| Клавиша Tab | Переместиться вперед по элементам в окне, на панели или в диалоговом окне. |
| Shift+F10 | Вывести контекстное меню для элемента. |
| Shift+Tab | Переход назад по элементам окна или диалогового окна. |
| Shift+стрелка | Выбрать символы в изменяемом поле в режиме изменений (F2). |
| Ctrl+Tab | Переместить фокус вперед в следующую главную область окна. |
| Shift+Ctrl+Tab | Переместить фокус назад в предыдущую главную область окна. |

## Клавиши быстрого вызова для диалоговых окон

Ряд клавиш быстрого вызова и клавиш программы чтения экрана будут полезны при работе с диалоговыми окнами. Иногда после входа в диалоговое окно нужно нажать клавишу Tab, чтобы переместить фокус на первый элемент управления и инициировать средство чтения экрана. В таблице ниже приведен полный список специальных клавиатурных сокращений и клавиш быстрого вызова программы чтения экрана.

*Таблица 14. Ярлыки в диалоговом окне*

| **Клавиши быстрого вызова** | **Функция** |
|---|---|
| Клавиша Tab | Переход вперед по элементам окна или диалогового окна. |
| Ctrl+Tab | Переход вперед из текстового поля к следующему элементу. |
| Shift+Tab | Переход назад по элементам окна или диалогового окна. |
| Shift+Ctrl+Tab | Переход назад из текстового поля к предыдущему элементу. |
| Клавиша пробела | Выбрать управляющий элемент или нажать кнопку, на которой находится фокус. |
| Esc | Отменить изменения и закрыть диалоговое окно. |
| Enter | Подтвердить изменения и закрыть диалоговое окно (эквивалент кнопки OK). Если фокус находится на текстовом поле, нужно нажать сначала клавиши Ctrl+Tab, чтобы выйти из него. |

# Глава 9. Извлечение понятий и типов

При каждом выполнении потока, запускающего интерактивную инструментальную среду, для текстовых данных в потоке выполняется автоматическое извлечение. Конечный результат этого извлечения - набор понятий, типов и (при наличии паттернов TLA в лингвистических ресурсах) паттернов. Понятия и типы можно просмотреть и работать с ними на панели Результаты извлечения. Дополнительную информацию смотрите в разделе “Как работает извлечение” на стр. 5.

Если вы хотите точно настроить результаты извлечения, можно изменить лингвистические ресурсы и выполнить извлечение повторно. Дополнительную информацию смотрите в разделе “Уточнение результатов извлечения” на стр. 95. Процесс извлечения опирается на ресурсы и всевозможные параметры в диалоговом окне Извлечь, определяющие, как извлечь и организовать результаты. с помощью результатов извлечения можно задать наиболее уместные (если не все) определения категорий.

## Результаты извлечения: Понятия и типы

В процессе извлечения все текстовые данные просматриваются, и в них выявляются релевантные понятия, которые затем извлекаются и распределяются по типам. По завершении извлечения результаты выводятся на панели результатов извлечения в нижнем левом углу представления Категории и понятия. При первом запуске сеанса эти понятия и типы извлекаются при помощи шаблона лингвистических ресурсов, выбранного в данном узле.

Извлеченные понятия, типы и паттерны TLA обобщенно называются **результатами извлечения** и служат как дескрипторы или блоки для построения категорий. Кроме того, понятиями, типами и паттернами можно пользоваться в правилах категорий. Автоматические методы также пользуются понятиями и типами при построении категорий.

Исследование текста - это интерактивный процесс, в котором результаты извлечения пересматриваются в контексте текстовых данных; выполняется тонкая настройка для получения новых результатов и их повторной оценки. После извлечения следует просмотреть результаты и внести желательные изменения в лингвистические ресурсы. Тонкую настройку ресурсов можно выполнять, в частности, непосредственно с панели результатов извлечения, с панели данных, из диалогового окна определений категорий или из диалогового окна определений кластеров. Дополнительную информацию смотрите в разделе “Уточнение результатов извлечения” на стр. 95. Кроме того, это можно сделать непосредственно в представлении Редактор ресурсов. Дополнительную информацию смотрите в разделе “Представление Редактор ресурсов” на стр. 80.

После тонкой настройки можно повторить извлечение, чтобы увидеть новые результаты. Выполнив с самого начала тонкую настройку результатов извлечения, вы убедитесь, что при всех повторных извлечениях получаются одинаковые результаты в ваших определениях категорий, идеально адаптированных в контексте данных. Таким образом, документы или записи будут попадать в ваши определения категорий точным, воспроизводимым образом.

Понятия

В процессе извлечения текстовые данные просматриваются и анализируются с целью выявить интересные или значимые слова (например, `выборы` или `мирный`) и словосочетания (например, `президентские выборы`, `выборы президента` или `мирные договоры`) в тексте. Эти слова и словосочетания вместе называются *терминами*. Благодаря лингвистическим ресурсам извлекаются соответствующие термины, а аналогичные термины собираются вместе с главным термином, называемым **понятие**.

Чтобы увидеть набор терминов, охваченных понятием, остановите указатель мыши на имени понятия. При этом выводится подсказка, содержащая имя понятия и одну или несколько строк терминов,

сгруппированных этой понятием. К таким терминам относятся синонимы, определенные в лингвистических ресурсах (независимо от того, найдены ли они в тексте), а также извлеченные формы единственного и множественного числа, термины с перестановками, термины с нечеткой группировкой и так далее. Вы можете скопировать эти термины или просмотреть полный список терминов понятия, щелкнув правой кнопкой по имени понятия и выбрав опцию в контекстном меню.

По умолчанию понятия выводятся в нижнем регистре и сортируются в порядке убывания числа документов (столбец Doc.) . При извлечении понятиям приписывается тип, что помогает группировать сходные понятия. Согласно этому типу понятия получают цветовой код. Цвета определяются в свойствах типов в Редактор ресурсов. Дополнительную информацию смотрите в разделе “Словари типов” на стр. 189.

Если понятие, тип или паттерн используется в определении категории, в сортируемом столбце **In** появляется значок .

Типы

**Типы** - это объединения понятий по смыслу. При извлечении понятиям приписывается тип, что помогает группировать сходные понятия. Различные встроенные типы поставляются вместе с IBM SPSS Modeler Text Analytics , например, <Положение>, <Организация>, <Персональные>, <Положительные>, <Отрицательные> и так далее. Например, тип <Положение> объединяет географические ключевые слова и места. Этот тип назначается таким понятиям, как чикаго, париж и токио. Для большинства языков понятия, которые не были найдены ни в одном словаре типов, но были извлечены из текста, автоматически получают тип <Неизвестный>Дополнительную информацию смотрите в разделе “Встроенные типы” на стр. 190.

По умолчанию извлеченные типы выводятся в представлении Тип в порядке убывания глобальной частоты. Кроме того, для удобства различения типы раскрашены согласно своему цветовому коду. Цвет задается как одно из свойств типа. Дополнительную информацию смотрите в разделе “Создание типов” на стр. 191. Кроме того, вы можете создавать свои типы.

Паттерны

Кроме того, из текстовых данных можно извлекать паттерны. Однако для этого нужна библиотека, содержащая некоторые правила паттернов Text Link Analysis (TLA) в Редактор ресурсов. Кроме того, нужно выбрать извлечение паттернов в параметре узла IBM SPSS Modeler Text Analytics или в диалоговом окне Извлечь при помощи опции **Включить извлечение паттернов Text Link Analysis**. Дополнительную информацию смотрите в разделе Глава 12, “Изучаем анализ текстовых связей (Text Link Analysis, TLA)”, на стр. 153.

## Извлечение данных

Во всех случаях, когда требуется извлечение, панель Результаты извлечения становится желтой, а под панелью инструментов на этой панели выводится сообщение **Для извлечения понятий нажмите кнопку Извлечь**.

Возможно, вам понадобится выполнить извлечение, если у вас еще нет результатов извлечения, вы внесли изменения в лингвистические ресурсы и нужно изменить результаты извлечения, или же вы переоткрыли сеанс , в котором не сохранили результаты извлечения (**Инструменты > Опции**).

**Примечание:** Если вы изменили исходный узел для вашего потока после кэширования результатов извлечения при помощи опции **Использовать работу сеанса...**, надо будет запустить новое извлечение сразу после запуска сеанса интерактивной инструментальной среды, если требуется получить обновленные результаты извлечения.

При выполнении извлечения появляется индикатор хода выполнения, позволяющий судить о состоянии процесса. В это время механизм извлечения считывает все текстовые данные, выявляет соответствующие термины и паттерны, извлекает их и назначает их типу. Затем механизм пытается сгруппировать

термины-синонимы под одним главным термином, который называется понятием. По завершении процесса полученные в результате понятия, типы и паттерны выводятся на панели Результаты извлечения.

Процесс извлечения приводит к созданию набора понятий и типов, например, паттернов Text Link Analysis (TLA), если они разрешены. На панели Результаты извлечения представления Категории и понятия можно просматривать эти понятия и типы и работать с ними. При извлечении паттернов TLA их можно увидеть в представлении Text Link Analysis.

*Примечание*: Существует связь между размером вашего набора данных и временем, которое требуется для завершения процесса извлечения. Всегда можно вставить узел выборки выше в потоке или оптимизировать конфигурацию вашего компьютера.

Для извлечения данных

1. В меню выберите **Инструменты > Извлечь**. Или же нажмите кнопку **Извлечь** на панели инструментов.
2. Если выбрана опция всегда выводить диалоговое окно Параметры извлечения, в появившемся окне можно будет внести нужные изменения. Дескрипторы этих параметров описаны далее в этой теме.
3. Нажмите кнопку **Извлечь**, чтобы начать процесс извлечения. Как только извлечение начнется, появится диалоговое окно хода выполнения. По завершении извлечения его результаты появятся на панели Результаты извлечения. По умолчанию понятия выводятся в нижнем регистре и сортируются в порядке убывания числа документов (столбец Документы) .

Результаты можно просматривать, используя опции панели инструментов для различной сортировки результатов, фильтрации результатов или переключения на другое представление (понятия или типы). Можно также уточнить результаты извлечения путем работы с лингвистическими ресурсами. Дополнительную информацию смотрите в разделе “Уточнение результатов извлечения” на стр. 95.

**Для текста на голландском, английском, французском, немецком, итальянском, португальском и испанском**

Диалоговое окно Параметры извлечения содержит основные опции извлечения.

**Включить извлечение паттерна Text Link Analysis.** Задает, что вам требуется извлекать паттерны TLA из ваших текстовых данных. Предполагается также, что ваши правила паттернов TLA находятся в одной из библиотек в редакторе ресурсов. Эта опция может существенно увеличить время извлечения. Дополнительную информацию смотрите в разделе Глава 12, “Изучаем анализ текстовых связей (Text Link Analysis, TLA)”, на стр. 153.

**Допускать ошибки пунктуации.** Эта опция временно нормализует текст, содержащий ошибки пунктуации (например, неверно используемые знаки препинания) во время извлечения, чтобы повысить извлекаемость понятий. Эта опция особенно полезна для коротких текстов низкого качества (например, ответы при опросе с произвольным ответом, электронная переписка, данные CRM), а также для текста, содержащего много сокращений.

**Допускать орфографические ошибки при минимальном числе символов корня [n].** Эта опция применяет метод нечеткой группировки, который помогает группировать в одну концепцию слова, которые часто пишутся с ошибками, а также вариативные написания слова. Алгоритм нечеткой группировки перед сравнением временно удаляет из извлеченных слов все гласные (кроме первой) и двойные или тройные согласные, так что `тунель` и `тоннель` попадут в одну группу. Методы нечеткой группировки, однако, не применяются, если различным терминам назначены различные типы, кроме типа `<Неизвестный>`.

Кроме того, можно задать минимально необходимое число символов *корня* при использовании нечеткой группировки. Число символов корня в термине рассчитывается как общее число символов минус число символов окончания; кроме того, в случае термина-словосочетания вычитаются детерминативы и предлоги. Например, в термине `упражнения` будет насчитано 9 символов корня “упражнени”, поскольку буква *я* на конце слова относится к окончанию множественного числа. Аналогичным образом в `пакет яблок` насчитывается 10 символов корня (“пакет яблок”), а в `магнитола для автомобиля` насчитывается 17

символов корня (“магнитол автомобил”). Этот метод подсчета используется только при проверке применимости нечеткой группировки и не используется в алгоритмах сравнения слов.

*Примечание*: Если окажется, что некоторые слова группируются неправильно, такие пары слов можно исключить из метода при помощи явного объявления в разделе **Нечеткая группировка: исключения** на вкладке Расширенные ресурсы. Дополнительную информацию смотрите в разделе “Нечеткая группировка” на стр. 206.

**Извлечь одиночные термины.** Эта опция извлекает отдельные слова (одиночные термины), если слово не входит в словосочетание и если это существительное или нераспознанная часть речи.

**Извлечь нелингвистические объекты.** Эта опция извлекает нелингвистические объекты, такие как номера телефонов, номера социального страхования, время, даты, валюты, цифры, проценты, адреса электронной почты и HTTP-адреса. Вы можете включить или исключить те или иные типы нелингвистических объектов в разделе **Нелингвистические объекты: конфигурация** на вкладке Расширенные ресурсы. Выключив ненужные объекты, вы сэкономите время обработки механизмом извлечения. Дополнительную информацию смотрите в разделе “Конфигурация” на стр. 210.

**Алгоритм верхнего регистра.** Эта опция извлекает простые и составные термины, не входящие во встроенные словари, если первая буква термина - в верхнем регистре. Это хороший способ извлечь большинство имен собственных.

**Группировать частичные и полные личные имена, где возможно.** Эта опция группирует имена, которые по-разному появляются в тексте. Эта возможность полезна, поскольку имена часто употребляются в начале текста в полной форме, а затем - в краткой. Эта опция пытается сопоставить каждый одиночный термин с типом <Неизвестный> последнему слову в любом составном термине, типизированном как <Личный>. Например, если найден терм *иванов*, получивший вначале тип <Неизвестный>, механизм извлечения поищет составные термины в типе <Личный>, содержащие *иванов* как последнее слово, например, *александр иванов*. Эта опция применяется только к фамилии, поскольку первое имя почти никогда не извлекается как одиночный термин.

**Максимум неслужебных слов при перестановке.** Эта опция задает максимально допустимое число неслужебных слов при применении метода перестановки. Этот метод перестановок группирует как близкие словосочетания, содержащие в своем составе одни и те же неслужебные слова, если игнорировать форму слова. Например, если задать ограничение в два неслужебных слова, будут обработаны такие извлеченные словосочетания, как компания клиенту и клиенту от нашей компании. В этом примере такие словосочетания будут сгруппированы в итоговом списке понятий, поскольку считаются одинаковыми, если проигнорировать слова от нашей.

**Опция индекса для карты понятий** Задает, что во время извлечения нужно построить индекс карты, чтобы впоследствии быстро построить карту понятий. Для редактирования параметров индекса нажмите кнопку **Параметры**. Дополнительную информацию смотрите в разделе “Построение индексов карты понятий” на стр. 95.

**Всегда показывать это диалоговое окно перед запуском извлечения.** Задайте, нужно ли показывать диалоговое окно Параметры извлечения при каждом извлечении, скрывать ли его всегда, кроме случаев перехода в меню Инструменты, и нужно ли спрашивать при каждом извлечении, хотите ли вы редактировать параметры извлечения.

Для текста на японском

Диалоговое окно Параметры извлечения содержит основные опции извлечения для текста на японском языке. По умолчанию параметры, выбранные в этом диалоговом окне, совпадают с параметрами, выбранными на вкладке Эксперт узла моделирования исследования текста. Для работы с японским текстом необходимо использовать этот текст в качестве ввода, а также выбрать шаблон японского языка или пакет

анализа текста на вкладке Модель узла исследования текста. Дополнительную информацию смотрите в разделе “Копирование ресурсов из шаблонов и файлов TAP” на стр. 26.

**Вторичный анализ.** При извлечения базовые ключевые слова извлекаются при помощи набора типов по умолчанию. Но, если выбрать тот или иной вторичный анализатор, можно получить много дополнительных и более богатых понятий, поскольку теперь экстрактор будет учитывать частицы и вспомогательные глаголы как часть концепции. Кроме того, при анализе эмоциональной окраски можно включить большое число дополнительных типов. После выбора вторичного анализатора можно сгенерировать результаты Text Link Analysis.

**Примечание:** При вызове вторичного анализатора извлечение занимает больше времени.

- **Анализ зависимостей**. При выборе этой опции извлечение понятий производится с дополнительным учетом частиц по сравнению с извлечением базовых типов и ключевых слов. Кроме того, при анализе зависимостей можно получить более богатые результаты паттернов TLA.
- **Анализ эмоциональной окраски**. При выборе этого анализатора извлекаются дополнительные понятия и, если применимо, результаты паттернов TLA. Помимо базовых типов можно воспользоваться более чем 80 типами эмоциональной окраски . При помощи таких типов можно раскрывать в тексте понятия и паттерны, выражающие эмоции, настроения и мнения. Фокус анализа эмоциональной окраски управляется тремя опциями: **Все эмоциональные окраски**, **Только репрезентативные эмоциональные окраски** и **Только заключения**.
- **Без вторичного анализатора**. Эта опция отключает все вторичные анализаторы. Ее нельзя выбрать при включенной опции **Разрешить извлечение паттернов Text Link Analysis**, поскольку для получения результатов TLA требуется вторичный анализатор.

**Включить извлечение паттерна Text Link Analysis.** Задает, что вам требуется извлекать паттерны TLA из ваших текстовых данных. Предполагается также, что ваши правила паттернов TLA находятся в одной из библиотек в редакторе ресурсов. Эта опция может существенно увеличить время извлечения. Кроме того, для извлечения результатов паттерна TLA необходимо выбрать вторичный анализатор. Дополнительную информацию смотрите в разделе Глава 12, “Изучаем анализ текстовых связей (Text Link Analysis, TLA)”, на стр. 153.

## Фильтрация результатов извлечений

При работе с очень большими наборами данных процесс извлечения может сгенерировать миллионы результатов. Для многих пользователей такой объем затруднит эффективный просмотр результатов. Поэтому, чтобы увеличить детализацию наиболее интересных результатов, их можно отфильтровать при помощи диалогового окна Фильтр, доступного на панели Результаты извлечения.

Имейте в виду, что все параметры в этом диалоговом окне Фильтр используются все вместе для фильтрации результатов извлечения, доступных для категорий.

**Фильтр по частоте.**Можно отфильтровать результаты с определенным значением глобальной частоты или частоты в документах.

- **Глобальная частотность** - это общее число вхождений понятия во всем наборе документов или записей; оно выводится в столбце **Глобально**.
- **Частотность документов** - это общее число документов или записей, в которых встречается понятие; оно выводится в столбце **Документы**.

Например, если понятие `nato` встречается 800 раз в 500 записях, мы можем сказать, что у этого понятия глобальная частотность составляет 800, а частотность документов - 500.

**И по типам.** Можно применить фильтр, чтобы выводились результаты, принадлежащие только к определенным типам. Можно выбрать все или только конкретные типы.

**И по соответствию тексту.** Кроме того, при помощи фильтра можно вывести только те результаты, которые соответствуют заданному здесь правилу. Введите набор символов для сопоставления в поле **Сопоставить текст**, а затем выберите условие, при котором следует применять сопоставление.

*Таблица 15. Условия соответствия тексту*

| Условие | Описание |
|---|---|
| **Содержит** | Текст сопоставляется, если строка встречается в любом месте. (выбрано по умолчанию) |
| **Начинается с** | Соответствие тексту признается, только если понятие или тип начинаются с заданного текста. |
| **Оканчивается на** | Соответствие тексту признается, только если понятие или тип оканчиваются заданным текстом. |
| **Точное совпадение** | Вся строка должна соответствовать имени понятия или типа. |

**И по рангу.** Можно также применить фильтр, чтобы выводились понятия только понятия, соответствующие самой высокой глобальной частотности (**Глобально**) или частотности документов (**Документы**), либо по возрастанию, либо по убыванию.

Результаты, выводимые на панели Результаты извлечения

Вот несколько примеров того, как могут выводиться результаты (на английском) на панели инструментов панели Результаты извлечения на основе фильтров.

*Таблица 16. Примеры обратной связи фильтра*

| Ответная реакция на фильтрацию | Описание |
|---|---|
|   | На панели инструментов показано число результатов. Поскольку никакого текста, соответствующего фильтру, не оказалось и максимум не достигнут, никакие дополнительные значчки не выводятся. |
|   | На панели инструментов показано число результатов, которые были ограничены задаваемым в фильтре максимумом, который в данном случае представлен числом 300. Фиолетовый значок, если он присутствует, означает, что достигнуто максимальное число понятий. Для дополнительной информации остановите указатель на значке. |
|   | На панели инструментов показано число результатов, которые были ограничены при помощи фильтра сопоставления текста. Об этом свидетельствует значок лупы (бинокля). |

Чтобы фильтровать результаты

1. В меню выберите **Инструменты > Фильтр**. Откроется диалоговое окно Фильтр.
2. Выберите и уточните нужные фильтры.
3. Нажмите кнопку **OK**, чтобы применить фильтры и увидеть новые результаты на панели Результаты извлечения.

## Исследование карт понятий

Создав карту понятий, можно выяснить, как понятия связаны между собой. Если выбрать одно понятие и нажать кнопку **Карта**, откроется окно карты понятий, позволяющее исследовать набор понятий, связанных с выбранным понятием. Вывод понятий можно отфильтровать, изменив соответствующие параметры типа: какие типы следует включить, поиск каких типов взаимосвязей выполнить и так далее.

**Важное замечание:** Чтобы можно было создать карту, сначала нужно сгенерировать индекс. Это может занять несколько минут. Однако после генерирования индекса его нужно будет сгенерировать снова, только когда будет выполняться повторное извлечение. Если вы хотите, чтобы индекс создавался при каждом извлечении автоматически, выберите эту опцию в параметрах извлечения. Дополнительную информацию

смотрите в разделе “Извлечение данных” на стр. 88.

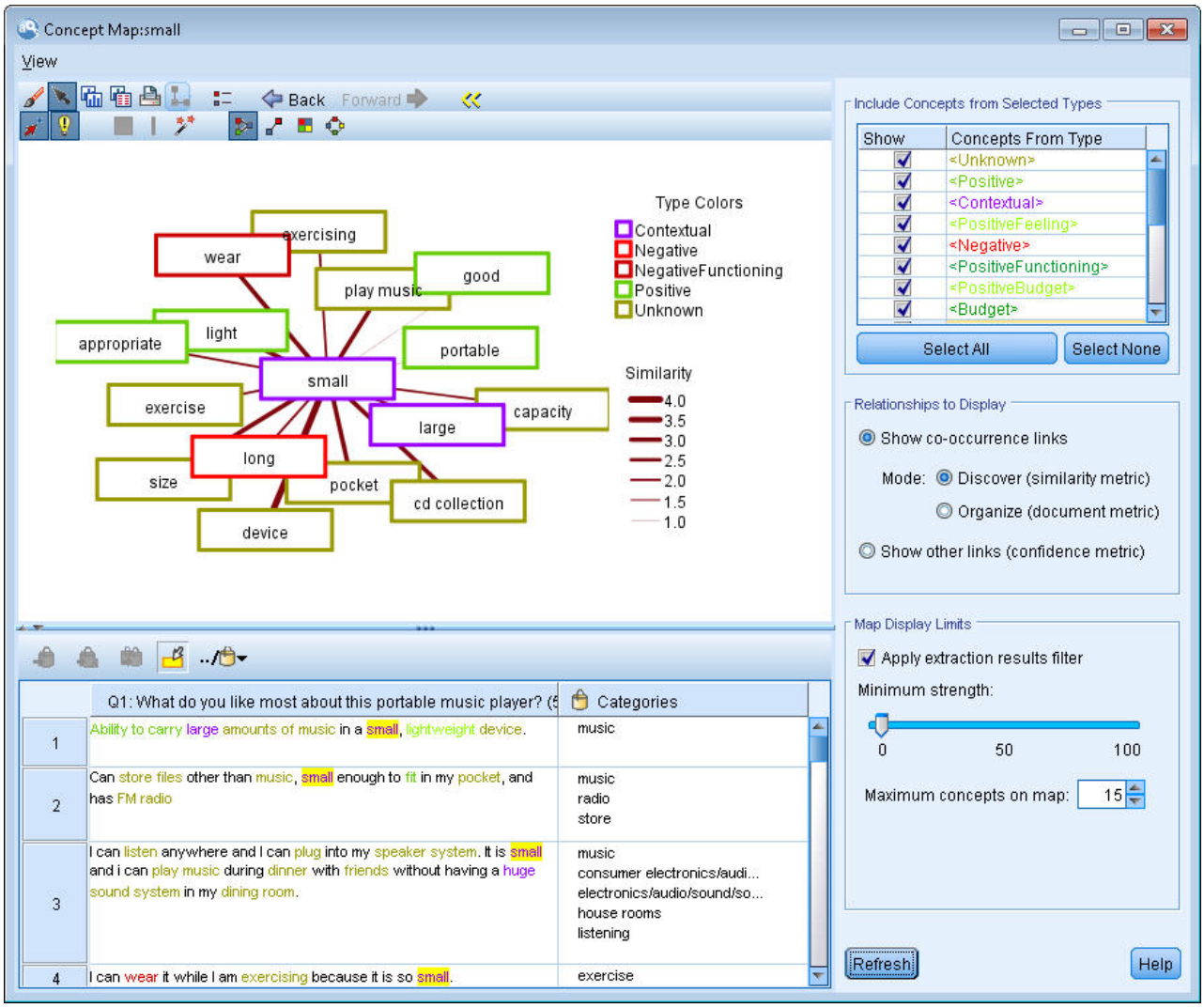



*Рисунок 27. Карта понятий для выбранного понятия*

## Чтобы просмотреть карту понятий

1. На панели Результаты извлечения выберите одно понятие.
2. На панели инструментов этой панели нажмите кнопку **Карта**. Если индекс карты уже сгенерирован, карта понятий откроется в отдельном диалоговом окне. Если индекс карты не был сгенерирован или устарел, его нужно перепостроить. Этот процесс может занять несколько минут.
3. Щелкните в области карты, чтобы ее исследовать. Если щелкнуть дважды мышью по связанному понятию, сама карта будет перерисована, и на ней появятся понятия, связанные с тем, по которому вы только что дважды щелкнули мышью.
4. На верхней панели инструментов предлагаются некоторые основные инструменты для карты, например: возврат к прежней карте, фильтрация связей в соответствии с силой взаимосвязей и открытие диалогового окна фильтра для управления типами выводимых понятий, а также разновидностями представляемых взаимосвязей. Во второй строке панели инструментов содержатся инструменты редактирования диаграмм. Дополнительную информацию смотрите в разделе “Использование палитр и панелей инструментов диаграмм” на стр. 163.
5. Если обнаруженные разновидности связей вас не устраивают, просмотрите параметры для этой карты, представленные в правой ее части.

## Параметры карты: Включить понятия из выбранных типов

На карте выводятся только понятия, принадлежащие к выбранным типам в этой таблице. Чтобы скрыть понятия определенного типа, отключите этот тип в таблице.

## Параметры карты: Вывести взаимосвязи

**Показать связи совместной встречаемости** Включите этот режим, если вы хотите вывести связи совместной встречаемости. Этот режим влияет на способ вычисления силы связи.

- *Обнаружение (показатель подобия)*. При использовании этого показателя сила связи рассчитывается при помощи более сложного вычисления, где учитывается, как часто встречаются два понятия порознь и как часто они встречаются в паре. Высокое значение силы означает, что пара понятий имеет склонность чаще встречаться совместно друг с другом, чем порознь. При помощи следующей формулы все числовые значения с плавающей запятой преобразуются в целочисленные.

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

*Рисунок 28. Формула коэффициента подобия*

В этой формуле $C_I$ - число документов или записей, в которых встречается понятие I.

$C_J$ - число документов или записей, в которых встречается понятие J.

$C_{IJ}$ - число документов или записей, в которых пара понятий I и J встречается совместно в наборе документов.

- *Организация (показатель документов)*. Сила связей при помощи этого показателя определяется посредством грубой оценки числа совместных вхождений. В общем случае, чем более часты два понятия, тем вероятнее, что время от времени они будут встречаться в паре. Высокое значение силы означает, что пара понятий часто встречается совместно друг с другом.

**Показать другие связи (показатель достоверности)**. Можно выбрать другие связи для вывода; они могут быть семантическими, производными (морфологическими) или связями включений (синтаксическими) и зависеть от числа шагов, отделяющих то или иное понятие от понятия, с которым оно связано. Эти связи могут помочь настроить ресурсы, в частности, синонимию или неоднозначность. Краткие описания каждого из этих методов группирования смотрите в разделе “Дополнительные лингвистические параметры” на стр. 113

**Примечание:** Имейте в виду, что если эти связи не были выбраны во время построения индекса или если не были найдены никакие взаимосвязи, ничего выводиться не будет. Дополнительную информацию смотрите в разделе “Построение индексов карты понятий” на стр. 95.

## Параметры карты: Пределы для вывода карты

**Применить фильтр результатов извлечения**. Если вы хотите использовать не все понятия, можно использовать фильтр на панели результатов извлечения, позволяющий ограничить то, что выводится. Затем надо будет включить эту опцию, и IBM SPSS Modeler Text Analytics будет выполнять поиск связанных понятий, применяя этот отфильтрованный набор. Дополнительную информацию смотрите в разделе “Фильтрация результатов извлечений” на стр. 91.

**Минимальная сила**. Задайте здесь минимальную силу связей. Все связанные понятия с силой взаимосвязи ниже этого предела на карте будут скрыты.

**Максимальное число понятий на карте**. Задайте максимальное число выводимых на карте взаимосвязей.

## Построение индексов карты понятий

Чтобы можно было создать карту, сначала нужно сгенерировать индекс взаимосвязей понятий. IBM SPSS Modeler Text Analytics будет обращаться к этому индексу при каждом создании карты понятий. Выбрав метод в этом диалоговом окне, можно выбрать взаимосвязи, которые следует индексировать.

**Методы группировки**. Выберите один или более методов. Краткие описания каждого из этих методов смотрите в разделе “О лингвистических методах” на стр. 116. Не все методы доступны для всех языков текстовых данных.

**Предотвращать объединение отдельных понятий в пары.** Включите этот переключатель, чтобы не объединять два понятия в группу или пару при выводе результатов. Чтобы создавать пары понятий и работать с ними, щелкните по **Работа с парами**. Дополнительную информацию смотрите в разделе “Управление парами с исключением связи” на стр. 116.

Построение индекса может занять несколько минут. Однако после генерирования индекса его нужно будет сгенерировать снова, только когда будет выполняться повторное извлечение или когда вы захотите изменить параметры, чтобы включить дополнительные взаимосвязи. Если вы хотите, чтобы индекс генерировался при каждом выполнении извлечения, эту опцию можно выбрать в параметрах извлечения. Дополнительную информацию смотрите в разделе “Извлечение данных” на стр. 88.

## Уточнение результатов извлечения

Извлечение - это итерационный процесс, с помощью которого вы можете выполнить извлечение, просмотреть результаты, внести в них изменения, а затем повторить извлечение, чтобы изменить результаты. Так как точность и непрерывность существенны для успешного исследования данных и для определения категорий, уточнение результатов извлечения с самого начала обеспечит, что при всяком повторном извлечении вы получите точно те же результаты в ваших определениях категорий. В этом случае записи и документы будут назначаться вашим категориям более точно и повторяемым образом.

Результаты извлечения служат строительными блоками для категорий. При создании категорий с использованием этих результатов извлечения записи и документы будут автоматически назначаться категориям, если они содержат текст, совпадающий с одним или несколькими дескрипторами категорий. Хотя категоризацию можно начать до выполнения уточнения лингвистических ресурсов, полезно хотя бы однажды просмотреть результаты извлечения до начала процедуры.

При пересмотре своих результатов вы можете обнаружить элементы, которые механизм извлечения должен обрабатывать по-особенному. Рассмотрим следующие примеры:

- **Нераспознанные синонимы.** Допустим, вы обнаружили несколько понятий, рассматриваемых как синонимы, таких как `smart`, `intelligent`, `bright` и `knowledgeable`, и все они появились как отдельные понятия в результатах извлечения. Вы можете создать определение синонимов, в котором понятия `intelligent`, `bright` и `knowledgeable` все сгруппированы вокруг понятия назначения `smart`. При этом все эти понятия сгруппируются вокруг `smart`, и кроме этого повысится значение глобальной частоты. Дополнительную информацию смотрите в разделе “Добавление синонимов” на стр. 96.
- **Понятия в неправильных типах.** Допустим, понятия в ваших результатах извлечения оказались в одном типе, а вы хотели бы назначить их другому типу. Другой пример. Представим, что вы обнаружили 15 понятий овощей в ваших результатах извлечения и хотите добавить все их к новому типу с названием `<Овощ>`. Для большинства языков понятия, которые не были найдены ни в одном словаре типов, но были извлечены из текста, автоматически получают тип `<Неизвестный>` Вы можете добавить понятия к типам. Дополнительную информацию смотрите в разделе “Добавление понятий к типам” на стр. 97.
- **Несущественные понятия.** Допустим, вы обнаруживаете извлеченное понятие с очень большим значением частоты появления, то есть оно находится в многих записях или документах. Однако для вас это понятие несущественно при анализе. Тогда его можно исключить из результатов извлечения. Дополнительную информацию смотрите в разделе “Исключение понятий при извлечении” на стр. 99.

- **Неправильные совпадения.** Допустим, при пересмотре записей или документов, содержащих определенное понятие, обнаруживается два неправильно сгруппированных слова, таких как `faculty` и `facility`. Это соответствие может быть связано с внутренним алгоритмом, так называемой нечеткой группировкой, в котором временно игнорируются сдвоенные или строенные согласные и гласные, чтобы сгруппировать варианты с обычными опечатками. Эти слова можно добавить в список пар слов, которые не должны группироваться. Дополнительную информацию смотрите в разделе “Нечеткая группировка” на стр. 206. В японских текстах нечеткая группировка недоступна.
- **Неизвлеченные понятия.** Допустим, вы предполагали обнаружить в результатах определенные понятия, но при пересмотре текста записи или документа оказалось, что несколько слов или словосочетаний не было извлечено. Часто эти слова - глаголы или прилагательные, которые вам не интересны. Однако в некоторых случаях может потребоваться использовать слово или фразу, которые не были извлечены как часть определения категории. Для извлечения такого понятия можно принудительно ввести термин в словарь типов. Дополнительную информацию смотрите в разделе “Принудительное включение слов в результаты извлечения” на стр. 99.

Многие из этих изменений можно выполнить непосредственно на панели Результаты извлечения или на панели Данные, в диалоговом окне Определения категорий или Определения кластеров, выбрав один или несколько элементов и щелкнув правой кнопкой мыши для доступа к контекстному меню.

После внесения изменений фоновый цвет панели изменится, указывая на необходимость повторного извлечения для просмотра изменений. Дополнительную информацию смотрите в разделе “Извлечение данных” на стр. 88. Если вы работаете с большими наборами данных, выполнение повторного извлечения окажется более эффективным после внесения нескольких изменений, а не каждый раз после одиночного изменения.

*Примечание*: Просмотреть весь набор доступных для изменения лингвистических ресурсов, используемых для получения результатов извлечения, можно в представлении Редактор ресурсов (Вид > Редактор ресурсов). В этом представлении указанные ресурсы появятся в форме библиотек и словарей. Настроить понятия и типы можно непосредственно в этих библиотеках и словарях. Дополнительную информацию смотрите в разделе Глава 16, “Работа с библиотеками”, на стр. 179.

## Добавление синонимов

*Синонимы* связывают два или более слов с одинаковым смыслом. Синонимы часто используются также для группировки слов с их сокращениями или для группировки слова в правильном написании с его ошибочными, но часто встречающимися формами. При использовании синонимов частота целевого понятия возрастает, что упрощает обнаружение сходной информации, разными способами представленной в ваших текстовых данных.

В поставляемых с продуктами библиотеках и шаблонах лингвистических ресурсов есть много предварительно определенных синонимов. Однако если обнаруживаются нераспознанные синонимы, их можно определить как синонимы, так что они будут распознаваться при следующем извлечении.

Во-первых, нужно решить, какое понятие будет главным, или целевым понятием. *Целевое понятие* - это слово или словосочетание, вокруг которого вы хотите сгруппировать все синонимические термины в окончательных результатах. При извлечении все синонимы будут группироваться вокруг этого целевого понятия. Во-вторых, нужно идентифицировать все синонимы для данного понятия. Целевое понятие подставляется для всех синонимов в результатах окончательного извлечения. Чтобы выступить как синоним, термин должен быть извлечен. Однако целевое понятие не обязательно должно быть извлечено, чтобы произошла подстановка. Например, если нужно заменять слово `intelligent` на слово `smart`, нужно определить `intelligent` синонимом, а `smart` целевым понятием.

Если создается новое определение синонима, в словарь добавляется новое целевое понятие. Затем к целевому понятию нужно добавить синонимы. При создании или изменении синонимов эти изменения записываются в словари синонимов в Редактор ресурсов. Если вы хотите просмотреть все содержимое этих

словарей синонимов или внести в них существенные изменения, предпочтительнее работать непосредственно в Редактор ресурсов. Дополнительную информацию смотрите в разделе “Словари подстановок/синонимов” на стр. 196.

Все новые синонимы будут автоматически сохраняться в первой библиотеке из дерева библиотек в представлении Редактор ресурсов, по умолчанию это *Локальная библиотека*.

**Примечание:** Если вы не можете найти определение синонима через контекстное меню или непосредственно в Редактор ресурсов, совпадение могло возникнуть при использовании внутреннего способа неявной группировки. Дополнительную информацию смотрите в разделе “Нечеткая группировка” на стр. 206.

Чтобы создать новый синоним

1. На панели Результаты извлечения или на панели Данные, в диалоговом окне Определения категорий или Определения кластеров выберите понятия, для которых вы хотите создать новый синоним.
2. Выберите пункт меню **Изменить > Добавить к синониму > Создать**. Откроется диалоговое окно Создать синоним.
3. Введите целевое понятие в текстовом поле Целевое понятие. Это понятие, вокруг которого будут группироваться все синонимы.
4. Если вы хотите добавить дополнительные синонимы, введите их в поле списка Синонимы. Для разделения синонимичных терминов используйте глобальный разделитель. Дополнительную информацию смотрите в разделе “Опции: вкладка Сеанс” на стр. 82.
5. При работе с японским текстом обозначьте тип для этих синонимов, выбрав имя типа в поле **Синонимы из типа**. Однако целевое понятие принимает тип, назначенный при извлечении. Если же целевое понятие не было извлечено как понятие, представленный в этом столбце тип будет назначен целевому понятию в результатах извлечения.
6. Нажмите кнопку **OK**, чтобы применить изменения. Диалоговое окно закроется, а у панели Результаты извлечения изменится фоновый цвет, указывая на необходимость произвести повторное извлечение, чтобы увидеть результат изменений. Если есть несколько изменений, внесите их все, прежде чем выполнять повторное извлечение.

Чтобы добавить синоним

1. На панели Результаты извлечения или на панели Данные, в диалоговом окне Определения категорий или Определения кластеров выберите понятия, которые вы хотите добавить к существующему определению синонима.
2. Выберите в меню опцию **Изменить > Добавить к синониму**. В меню будет показан набор синонимов, причем в начале списка будет самый последний созданный синоним. Выберите имя синонима, к которому вы хотите добавить выбранные понятия. Если вы нашли нужный синоним, выберите его, и все выбранные элементы будут добавлены к этому определению синонима. Если нужный синоним не показан, выберите опцию **Дополнительно**, чтобы вывести диалоговое окно Все синонимы.
3. В диалоговом окне Все синонимы можно сортировать список в порядке создания, а также по возрастанию или убыванию. Выберите имя синонима, к которому вы хотите добавить выбранные понятия, и нажмите кнопку **OK**. Диалоговое окно закроется, а понятия будут добавлены к определению синонима.

## Добавление понятий к типам

При всяком выполнении извлечения полученные понятия назначаются типам для группировки терминов, у которых есть что-то общее. IBM SPSS Modeler Text Analytics поставляется с многими встроенными типами. Дополнительную информацию смотрите в разделе “Встроенные типы” на стр. 190. Для большинства языков понятия, которые не были найдены ни в одном словаре типов, но были извлечены из текста, автоматически получают тип `<Неизвестный>`

При пересмотре результатов вы можете обнаружить некоторые понятия из одного типа, которые можно назначить другому типу, или группу слов, которые сами по себе принадлежат к новому типу. В этих случаях можно переназначить понятия другому типу или создать совершенно новый тип. Для японского текста создавать новые типы нельзя.

Допустим, например, что вы работаете над исследованием данных, относящихся к автомобилям, и хотите определить категории на основании различных частей транспортных средств. Можно создать тип с названием <Приборная панель>, чтобы сгруппировать все понятия, относящиеся к приборам и кнопкам на приборной панели автомобиля. Этому новому типу можно назначить такие понятия, как указатель уровня топлива, печка, радио и одометр.

Другой пример. Допустим, вы работаете с данными исследования, связанного с университетами и колледжами, и для извлеченного понятия Johns Hopkins (Университет Джонса Хопкинса) оказался присвоен тип <Человек>, а не тип <Организация>. В этом случае данное понятие можно добавить к типу <Организация>.

При всяком создании типа или при добавлении понятий в список терминов типа эти изменения записываются в словари типов в библиотеках ваших лингвистических ресурсов в Редактор ресурсов. Если вы хотите просмотреть содержимое этих библиотек или внести в них существенные изменения, предпочтительнее работать непосредственно в Редактор ресурсов. Дополнительную информацию смотрите в разделе “Добавление терминов” на стр. 192.

Чтобы добавить понятие к типу

1. На панели Результаты извлечения или на панели Данные, в диалоговом окне Определения категорий или Определения кластеров выберите понятия, которые вы хотите добавить к существующему типу.
2. Щелкните правой кнопкой мыши, чтобы открыть контекстное меню.
3. Выберите в меню опцию **Изменить > Добавить к типу**. В меню будет показан набор типов, причем в начале списка будут недавно созданные типы. Выберите имя типа, к которому вы хотите добавить выбранные понятия. Когда вы найдете нужное имя типа, выберите его, и выбранные понятия будут добавлены к этому типу. Если в выведенном списке нет нужного имени типа, выберите опцию **Дополнительно** для вывода диалогового окна Все типы.
4. В диалоговом окне Все типы можно сортировать список в порядке создания, а также по возрастанию или убыванию. Выберите имя типа, к которому вы хотите добавить выбранные понятия, и нажмите кнопку **OK**. Диалоговое окно закроется, а все понятия будут добавлены к типу как термины.

*Примечание*: При работе с японским текстом есть несколько случаев, когда изменение типа термина не изменяет тип, которому в конечном итоге он будет назначен в окончательном списке извлечения. Это связано с внутренними словарями, которым отдается предпочтение при извлечении некоторых базовых терминов.

Чтобы создать новый тип

1. На панели Результаты извлечения или на панели Данные, в диалоговом окне Определения категорий или Определения кластеров выберите понятия, для которых вы хотите создать новый тип.
2. Выберите пункт меню **Изменить > Добавить к типу > Создать**. Откроется диалоговое окно Свойства типа.
3. Введите новое имя для этого типа в текстовом окне Имя и внесите нужные изменения в другие поля. Дополнительную информацию смотрите в разделе “Создание типов” на стр. 191.
4. Нажмите кнопку **OK**, чтобы применить изменения. Диалоговое окно закроется, а у панели Результаты извлечения изменится фоновый цвет, указывая на необходимость произвести повторное извлечение, чтобы увидеть результат изменений. Если есть несколько изменений, внесите их все, прежде чем выполнять повторное извлечение.

## Исключение понятий при извлечении

При пересмотре результатов может обнаружиться, что в них есть понятия, которые вы не хотели бы извлекать или которые были использованы автоматическими способами построения категорий. В некоторых случаях у этих понятий может быть очень большая частота появлений, но при этом они совершенно не существенны для вашего анализа. В таких случаях можно пометить понятие, чтобы оно было исключено при окончательном извлечении. Обычно такие понятия в списке представляют собой заполняющие слова или словосочетания, служащие связности текста, но не добавляющие ничего нового, а только мешающие восприятию результатов. Добавление понятий в словарь исключения гарантирует, что они никогда не будут попадать в число извлеченных понятий.

При исключении понятий все их грамматические вариации исчезнут из списка результатов при следующем запуске извлечения. Если некоторое понятие уже содержится в категории как дескриптор, после повторного извлечения оно останется в категории, но с нулевым числом появлений.

При определении исключений эти изменения будут записаны в словаре исключения в Редактор ресурсов. Если необходимо просмотреть все определения исключений и изменить их непосредственно, предпочтительнее работать в Редактор ресурсов. Дополнительную информацию смотрите в разделе “Словари исключения” на стр. 200.

**Примечание:** В японских текстах есть примеры, когда исключение термина или типа не приводит к их исключению в конечных результатах. Это связано с внутренними словарями, которым отдается предпочтение при извлечении некоторых базовых терминов ресурсов на японском языке.

Чтобы исключить понятия

1. На панели Результаты извлечения или на панели Данные, в диалоговом окне Определения категорий или Определения кластеров выберите понятия, которые вы хотите исключать при извлечении.
2. Щелкните правой кнопкой мыши, чтобы открыть контекстное меню.
3. Выберите опцию **Исключать при извлечении**. Это понятие будет добавлено в словарь исключения в Редактор ресурсов, а у панели Результаты извлечения изменится фоновый цвет, указывая на необходимость произвести повторное извлечение, чтобы увидеть результат изменений. Если есть несколько изменений, внесите их все, прежде чем выполнять повторное извлечение.

**Примечание:** Все исключенные слова будут автоматически сохраняться в первой библиотеке из дерева библиотек в Редактор ресурсов, по умолчанию это *Локальная библиотека*.

## Принудительное включение слов в результаты извлечения

При пересмотре текстовых данных на панели Данные после извлечения может обнаружиться, что некоторые слова или словосочетания не были извлечены. Часто эти слова - глаголы или прилагательные, которые вам не интересны. Однако в некоторых случаях может потребоваться использовать слово или фразу, которые не были извлечены как часть определения категории.

Если нужно, чтобы эти слова или словосочетания извлекались, можно принудительно ввести термин в библиотеку типов. Дополнительную информацию смотрите в разделе “Принудительное назначение типов терминам” на стр. 195.

**Важно!** Обозначение термина в словаре как принудительно введенного не защищает от неправильного использования. Здесь имеется в виду, что даже при явном добавлении термина в словарь в некоторых случаях его может не оказаться на панели Результаты извлечения после повторного извлечения или этот термин появляется не точно в той форме, как вы его объявили. Хотя такое случается редко, но может произойти, когда слово или словосочетание уже было извлечено как часть более длинного словосочетания. Для предотвращения такой ситуации примените к этому типу в словаре типов опцию совпадения **Полностью (без составных элементов)**. Дополнительную информацию смотрите в разделе “Добавление терминов” на стр. 192.

# Глава 10. Категоризация текстовых данных

В представлении Категории и понятия можно создавать **категории**, по сути представляющие понятия более высокого уровня или темы, которые будут захватывать ключевые идеи, знания и отношения, выраженные в тексте.

Начиная с выпуска IBM SPSS Modeler Text Analytics 14, у категорий может также быть иерархическая структура, означающая, что они могут содержать подкатегории, а у этих подкатегорий могут быть их собственные подкатегории и так далее. Вы можете импортировать предопределенные структуры категорий (прежнее название фреймы) с иерархическими категориями, а также построить эти иерархические категории в программном продукте.

Фактически, иерархические категории позволяют построить древовидную структуру с одной или несколькими подкатегориями, чтобы сгруппировать элементы, такие как различные области понятий или тем, более точно. Простой пример можно связать с досугом; при ответе на вопрос, такой как *Чем бы вам хотелось заняться, когда будет побольше времени?*, на первом месте у вас могут быть такие категории, как *спорт*, *прикладное искусство*, *рыбалка* и так далее; под уровнем *спорт* у вас могут быть подкатегории, позволяющие понять, что это *игры с мячом*, *водные виды* и тому подобное.

Категории состоят из набора дескрипторов, таких как *понятия*, *типы*, *паттерны* и *правила категорий*. Все вместе эти дескрипторы применяются для определения, принадлежит ли документ или запись к данной категории. Просмотрев текст в документе или записи, можно выяснить, существуют ли какие-либо текстовые совпадения с дескриптором. Если совпадение найдено, документ/запись назначается данной категории. Этот процесс называется **категоризацией**.

Работать с категориями, строить их и визуально исследовать можно при помощи данных, представляемых на четырех панелях представления Категории и понятия, каждую из которых можно скрыть или показать, выбрав ее имя в меню Вид.

- **Панель Категории.** На этой панели строят категории и управляют ими. Дополнительную информацию смотрите в разделе “Панель Категории” на стр. 102.
- **Панель Результаты извлечения.** На этой панели исследуют извлеченные понятия и типы и работают с ними. Дополнительную информацию смотрите в разделе “Результаты извлечения: Понятия и типы” на стр. 87.
- **Панель Визуализация.** На этой панели визуально исследуют категории и то, как они взаимодействуют. Дополнительную информацию смотрите в разделе “Графики и диаграммы категорий” на стр. 159.
- **Панель Данные.** На этой панели исследуют и просматривают текст, содержащийся в документах и записях, соответствующих выбранным вариантам. Дополнительную информацию смотрите в разделе “Панель Данные” на стр. 110.

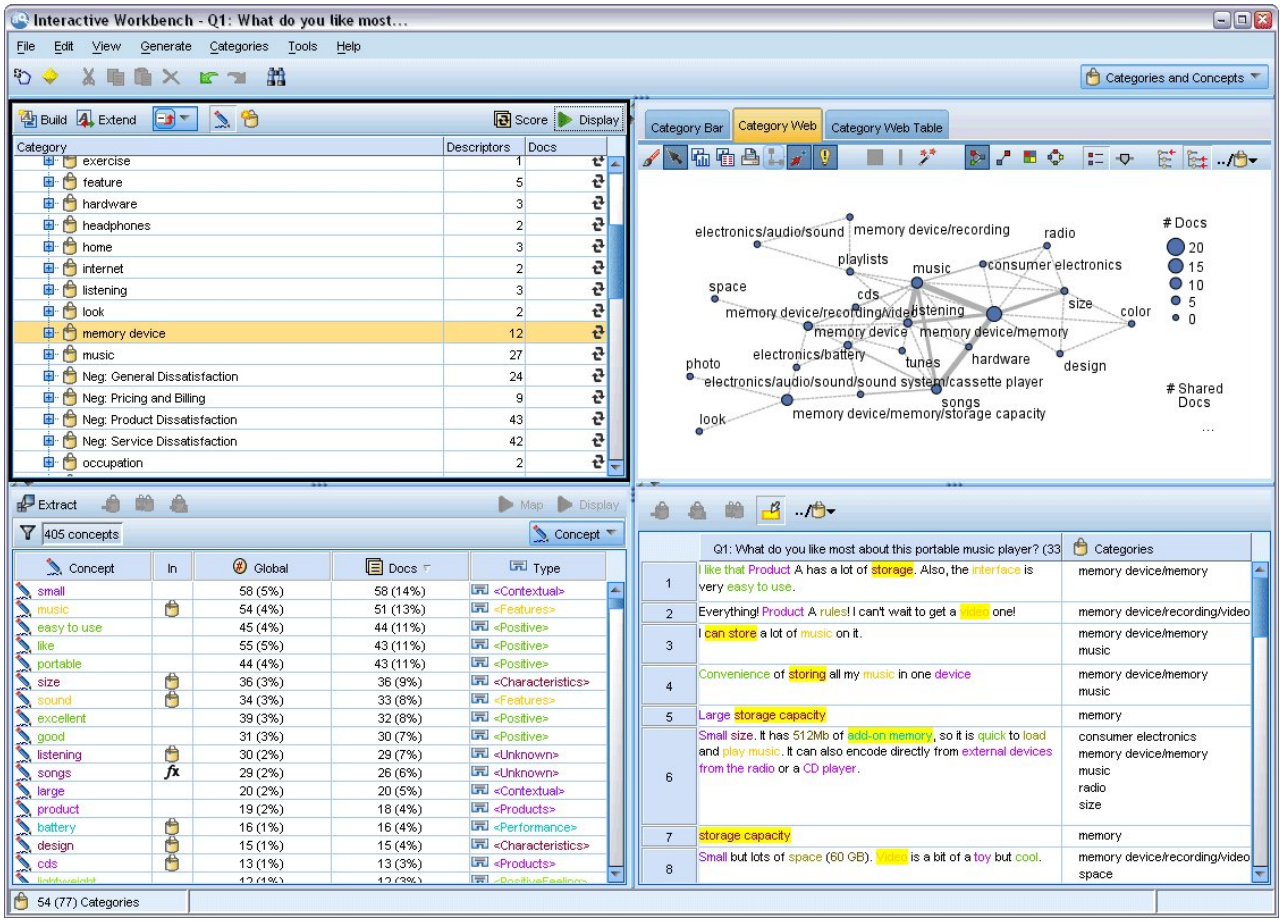



*Рисунок 29. Представление Категории и понятия*

Вы можете начать с набора категорий из пакета анализа текста (text analysis package, TAP) или импортировать файл предопределенных категорий, но может также потребоваться и создать свои собственные категории. Категории можно создать автоматически, сгенерировав их и их дескрипторы при помощи надежного набора автоматизированных методов программного продукта, где используются результаты извлечения (понятия, типы и паттерны). Категории можно также создать вручную при помощи дополнительных аналитических наработок, которые у вас могут быть. Однако создание категорий вручную и их точная настройка возможны только при помощи интерактивной инструментальной среды. Дополнительную информацию смотрите в разделе “Узел Text Mining: вкладка Модель” на стр. 24. Создать определения категорий вручную можно, перетащив в категории результаты извлечения. Эти категории или любую пустую категорию можно усилить, добавив в нее правила категорий, применив ваши собственные предопределенные категории или их сочетание.

Каждый из методов и способов хорошо подходит для определенных типов данных и ситуаций, но часто будет полезен и для объединения методов в этом же анализе с целью захвата всего диапазона документов или записей. В ходе категоризации, возможно, вы увидите другие изменения, подлежащие внесению в лингвистические ресурсы.

## Панель Категории

Панель Категории - это та область, в которой вы можете построить категории и управлять ими. Эта панель расположена в верхнем левом углу представления Категории и Понятия. После извлечения понятий и типов из ваших текстовых данных можно автоматически начать построение категорий, используя такие способы как включение понятий, совместное появление и так далее (или вручную). Дополнительную информацию смотрите в разделе “Построение категорий” на стр. 111.

При всяком создании или изменении категории документы или записи можно оценить, нажав кнопку **Скоринг**, чтобы можно было увидеть, есть ли какие-то совпадения текста с дескриптором в данной категории. Если такое совпадение обнаруживается, этот документ или запись назначается данной категории. Окончательный результат состоит в том, что большинство, если не все, из документов или записей назначаются категориям на основе дескрипторов категорий.

Таблица Дерево категорий

Таблица дерева на этой панели представляет набор категорий, подкатегорий и дескрипторов. У этого дерева есть также несколько столбцов, представляющих информацию для каждого элемента дерева. Для вывода могут быть доступны следующие столбцы:

- **Код.** Выводит значение кода для каждой категории. По умолчанию этот столбец скрыт. Этот столбец можно вывести через пункт меню **Вид > Панель Категории**.
- **Категория.** Содержит дерево категорий, показывающее имя категории и подкатегорий. Кроме этого, если нажать кнопку панели инструментов дескрипторов, появится также набор дескрипторов.
- **Дескрипторы.** Предоставляет количество дескрипторов, составляющих определение категории. Это число не включает в себя дескрипторы в подкатегориях. Число не показывается, если имя дескриптора представлено в столбце **Категории**. Сами дескрипторы можно вывести или скрыть в дереве при помощи пункта меню **Вид > Панель Категории > Все дескрипторы**.
- **Документы.** После скоринга в этом столбце приводится число документов или записей, которым присвоена данная категория и все ее подкатегории. Так, если 5 записей соответствует верхней категории на основании ее дескрипторов и 7 других записей соответствует подкатегории на основе ее дескрипторов, общим числом документов будет сумма двух значений, то есть 12 в данном случае. Однако если одна запись соответствует и верхней категории, и ее подкатегории, итогом будет 11.

Когда категорий не существует, в таблице все равно будет две строки. Верхняя строка с названием **Все документы** - это общее число документов или записей. Вторая строка с названием **Без категории** показывает количество документов/записей, которые еще не были категоризованы.

Для каждой категории на этой панели перед ее именем будет небольшой желтый значок ковша. Если дважды щелкнуть по категории или выбрать пункт меню **Вид > Определения категорий** в меню, откроется диалоговое окно Определения категорий, в котором будут представлены все элементы, называемые **дескрипторами**, которые составляют определение категории, такие как понятия, типы, паттерны и правила категорий. Дополнительную информацию смотрите в разделе “О категориях” на стр. 109. По умолчанию таблица дерева категорий не показывает дескрипторы в категориях. Если вы хотите увидеть дескрипторы непосредственно в дереве, а не в диалоговом окне Определения категорий, нажмите кнопку переключения со значком карандаша на панели инструментов. При нажатии этой кнопки переключения можно раскрыть ваше дерево и увидеть также дескрипторы.

Скоринг категорий

В столбце **Документы** таблицы дерева категорий выводится количество документов или записей, отнесенных к этой конкретной категории. Если данные значения устарели или не вычислялись, в этом столбце появится значок. Можно щелкнуть по **Оценка** на панели инструментов для данной панели, чтобы повторно вычислить количество документов. Учтите, что процесс скоринга может потребовать некоторого времени, если вы работаете с большими наборами данных.

Выбор категорий в дереве

Выполняя выбор в дереве, можно выбрать категории только одного иерархического уровня с общим родителем; то есть если выбрать категории верхнего уровня, невозможно одновременно выбрать подкатегорию. Аналогично, если выбрать две подкатегории данной категории, невозможно одновременно выбрать подкатегорию другой категории. Выбор несмежной категории приведет к потере предыдущего выбора.

Вывод на панелях Данные и Визуализация

При выборе строки в таблице можно нажать кнопку **Вывести**, чтобы обновить информацию на панелях Данные и Визуализация в соответствии с вашим выбором. Если панель не показана, она появится после нажатия кнопки **Вывести**.

Уточнение ваших категорий

С первой попытки категоризация может не дать идеальных результатов, и вполне могут присутствовать такие категории, которые вы захотите удалить или объединить с другими категориями. При пересмотре результатов извлечения может обнаружиться также, что есть некоторые полезные категории, которых вы не создавали. В таком случае в полученные результаты можно внести некоторые изменения вручную, чтобы уточнить их для конкретного контекста. Дополнительную информацию смотрите в разделе “Изменение и уточнение категорий” на стр. 143.

---

# Методы и стратегии для создания категорий

Если вы еще не выполнили извлечение или его результаты устарели, при использовании одного из методов построения или расширения категорий появляется приглашение выполнить автоматическое извлечение. После применения того или иного метода те понятия и типы, которые были сгруппированы в категорию, по-прежнему будут доступны для построения категорий другими методами. Это значит, что одно и то же понятие может участвовать во многих категориях, если вы не решите выключить повторное использование.

Чтобы научиться получать оптимальные категории, изучите следующее:

- **Методы для создания категорий**
- **Стратегии для создания категорий**
- **Советы по созданию категорий**

## Методы для создания категорий

Поскольку каждый набор данных уникален, число методов создания категорий и порядок их применения может изменяться. Кроме того, от одного набора данных к другому у вас могут изменяться цели исследования данных, и иногда необходимо поэкспериментировать с различными методами, чтобы увидеть, что дает лучшие результаты для тех или иных текстовых данных. Ни один из автоматических методов не гарантирует идеальной категоризации ваших данных; мы рекомендуем найти и применить один или несколько автоматических методов, которые хорошо работают с вашими данными.

Помимо использования пакетов анализа текста (text analysis packages, TAP, **.tap*) с заранее построенными наборами категорий, вы можете категоризовать свои ответы, используя сочетание следующих методов:

- **Методы автоматического построения.** Для категорий на основе лингвистики и частот доступен ряд опций автоматического построения. Дополнительную информацию смотрите в разделе “Построение категорий” на стр. 111.
- **Методы автоматического расширения.** Доступен ряд лингвистических методов для расширения существующих категорий путем добавления и расширения дескрипторов, чтобы захватить дополнительные записи. Дополнительную информацию смотрите в разделе “Расширение категорий” на стр. 122.
- **Ручные методы.** Имеется несколько ручных методов, таких как перетаскивание. Дополнительную информацию смотрите в разделе “Создание категорий вручную” на стр. 125.

## Стратегии создания категорий

Приведенный ниже список стратегий не исчерпывающий, но дает представление о подходах к построению категорий.

- Когда вы определяете узел Исследование текста, выберите набор категорий из пакета анализа текста (text analysis package, файл TAP), чтобы начать анализ при помощи предварительно построенных категорий.

Возможно, эти категории подойдут вам в своем исходном виде. Если же понадобится добавить дополнительные категории, вы сможете отредактировать параметры построения категорий (**Категории > Параметры построения**). Откройте диалоговое окно **Дополнительные параметры: лингвистика**, выберите опцию ввода категорий **Неиспользованные результаты извлечения** и постройте дополнительные категории.

- Когда вы определяете узел, выберите набор категорий из TAPв представлении Категории и понятия в интерактивной инструментальной среде. Затем перетащите неиспользованные понятия или паттерны в категории, как сочтете нужным. Затем расширьте существующие категории, которые отредактировали (**Категории > Расширить категории**), чтобы получить дополнительные дескрипторы, связанные с существующими дескрипторами категории.
- Построить категории автоматически при помощи дополнительных лингвистических параметров (**Категории > Построить категории**). Затем уточните категории вручную, удалив дескрипторы или объединив близкие категории, пока не будете удовлетворены полученными категориями. Кроме того, если вы первоначально построили категории **без** использования опции **Обобщать с использованием символов подстановки, где возможно**, можно попробовать упростить категории автоматически при помощи Расширенных категорий с опцией **Обобщать**.
- Импортируйте файл предопределенных категорий с максимально описательными именами категорий и/или аннотациями. Кроме того, если вы первоначально выполнили импорт **без** опции импорта или генерирования дескрипторов из имен категорий, вы можете позже в диалоговом окне Расширить категории выбрать опцию **Расширять пустые категории при помощи дескрипторов, сгенерированных из имени категории**. Затем расширьте эти категории второй раз, но теперь используйте методы группировки.
- Создайте вручную первый набор категорий, отсортировав понятия или паттерны понятий по частоте и затем перетащив наиболее интересные на панель категорий. Получив начальный набор категорий, воспользуйтесь возможностью расширения (**Категории > Расширить категории**), чтобы расширить и уточнить все выбранные категории, добавив связанные с ними дескрипторы, чтобы захватывалось больше записей.

Рекомендуется, применив эти методы, пересмотреть полученные категории и воспользоваться ручными методами для небольших уточнений, удаления ошибок классификации или добавления пропущенных записей и слов. Кроме того, поскольку при использовании различных методов могут возникать лишние категории, можно объединить или удалить ненужные категории. Дополнительную информацию смотрите в разделе “Изменение и уточнение категорий” на стр. 143.

## Подсказки по созданию категорий

Изучение некоторых подсказок может помочь принять нужное решение для создания лучших категорий.

Подсказки для соотношения категория-документ

Категории, которым назначаются документы и записи, часто невзаимоисключающие при качественном анализе текста по крайней мере по двум причинам:

- Во-первых, есть эмпирическое правило, по которому чем длиннее текст документа или записи, тем более различается выражение идей и мнений. Тем самым, существенно возрастает шанс, что документ или запись будет отнесена к нескольким категориям.
- Во-вторых, часто есть различные способы группировать и интерпретировать текстовые документы или записи, которые не разделены логически. В случае исследования с вопросом, не предполагающим вариантов ответа, относительно политических предпочтений респондентов можно создать категории *Liberal* и *Conservative* или *Republican* и *Democrat*, а также несколько более конкретных категорий, таких как *Socially Liberal*, *Fiscally Conservative* и так далее. Эти категории не должны быть взаимно исключающими или исчерпывающими.

Подсказки по количеству создаваемых категорий

Создание категорий должно происходить непосредственно при изучении данных - как только вы встречаете что-то интересное, относящееся к данным, можно создать категорию, отображающую эту информацию. В

общем случае нет рекомендованного верхнего предела для числа создаваемых категорий. Однако может возникнуть и такая ситуация, что категорий будет создано слишком много, чтобы ими можно было управлять. Применимы два принципа:

- **Частота категорий.** Чтобы категория была полезна, в ней должно содержаться минимальное число документов или записей. В одном или нескольких документах может быть что-то действительно интригующее, но если это всего один или несколько случаев из тысячи документов, эта информация будет слишком редкой для всей совокупности, чтобы ее можно было практически использовать.
- **Сложность.** Чем больше категорий вы создаете, тем больше информации нужно пересматривать и суммировать после завершения анализа. Однако при добавлении сложности в случае большого числа категорий могут не добавиться полезные подробности.

К сожалению, не существует правил, определяющих, когда количество категорий становится чрезмерно большим или каким должно быть минимальное количество записей на категорию. Вам нужно будет самостоятельно определить это на основании потребностей в конкретной ситуации.

Однако можно предложить совет, с чего начать. Хотя число категорий не должно быть чрезмерным, на начальных стадиях анализа слишком много категорий лучше, чем слишком мало. Проще сгруппировать относительно сходные категории, чем разделить наблюдения на новые категории, поэтому обычно рекомендуется стратегия, при которой работа начинается с большего числа категорий, и их число постепенно уменьшается. Учитывая итерационную основу исследования текстов и его простоту в этой программе, построение большего числа категорий приемлемо при начале работы.

## Выбор наилучших дескрипторов

Ниже представлены рекомендации по выбору или созданию наилучших дескрипторов (понятий, типов, паттернов TLA и правил категорий) для ваших категорий. Дескрипторы - это строительные блоки категорий. Когда весь текст документа или записи совпадает с дескриптором, этот документ или эта запись соответствует категории.

Пока дескриптор не будет содержать извлеченное понятие или паттерн или не будет соответствовать им, он не будет соответствовать и никаким документам или записям. Поэтому используйте понятия, типы, паттерны и правила категорий, как это описано в следующих параграфах.

Так как понятия представляют не только сами себя, но и набор подразумеваемых терминов, в том числе единственного и множественного числа, синонимов и различного правописания, только само понятие можно использовать как дескриптор или часть дескриптора. Чтобы узнать больше о подразумеваемых терминах для данного понятия, щелкните по имени понятия на панели Результаты извлечения представления Категории или Понятие. При наведении указателя мыши на имя понятия появится подсказка со всеми подразумеваемыми терминами, обнаруженными в вашем тексте при последнем извлечении. Не у всех понятий есть подразумеваемые термины. Например, если машина и автомобиль были синонимами, но слово машина было извлечено как понятие с подразумеваемым термином автомобиль, вам нужно использовать в дескрипторе только слово машина, так как дескриптор автоматически установит соответствие документов или записей со словом автомобиль.

Понятия и типы как дескрипторы

Используйте понятие как дескриптор, когда нужно найти все документы или записи, содержащие данное понятие (или любые из его подразумеваемых терминов). В этом случае использование более сложного правила категорий не требуется, так как достаточно точного имени понятия. Имейте в виду, что при использовании ресурсов, извлекающих мнения, понятия могут изменяться при извлечении паттерна TLA для извлечения более адекватного смысла предложения (смотрите пример в следующем разделе о TLA).

Например, ответ исследования, указывающий на предпочтительные фрукты покупателей как "*Лучше всего яблоки и ананасы*", может привести к извлечению понятий яблоко и ананас. При добавлении понятия яблоко в качестве дескриптора вашей категории все ответы, содержащие понятие яблоко (или его подразумеваемые термины), будут соответствовать этой категории.

Однако если вам нужно узнать, какие ответы упоминают *яблоко* любым образом, можно записать правило категории, такое как `* яблоко *`, чтобы захватывать ответы, содержащие такие понятия как `яблоко`, `свежее яблоко` или `французское яблоко столовое`.

Можно захватывать также все документы или записи, содержащие понятия, для которых был одинаково определен тип, использованный в качестве дескриптора, такой как `<Фрукт>`. Обратите внимание на то, что звездочку * нельзя использовать с типами.

Дополнительную информацию смотрите в разделе “Результаты извлечения: Понятия и типы” на стр. 87.

паттерны анализа текстовых связей (Text Link Analysis, TLA) как дескрипторы

Используйте в качестве дескриптора итоговый паттерн TLA, когда нужно захватывать более точные идеи с учетом нюансов. Когда при извлечении TLA анализируется текст, в нем поочередно обрабатывается каждое предложение или часть сложного предложения, а не сразу весь текст (документ или запись). Рассматривая все части одного предложения совместно, TLA может идентифицировать мнения, взаимосвязи между двумя элементами или, например, отрицание, чтобы понять действительный смысл. Паттерны понятий или паттерны типов можно использовать как дескрипторы. Дополнительную информацию смотрите в разделе “Паттерны типа и понятия” на стр. 155.

Например, если есть текст "*комната была не очень чистой*", можно извлечь следующие понятия: `комната` и `чистый`. Однако если в параметрах извлечения было включено извлечение TLA, алгоритм TLA может обнаружить, что `чистой` использовалось с отрицанием и фактически соответствует `не чистая`, что представляет синоним понятия `грязный`. Здесь видно, что использование самого по себе понятия `чистый` в качестве дескриптора приведет к согласованию с эти текстом, но не может захватить также другие документы или записи, упоминающие чистоту. Поэтому предпочтительнее было бы использовать паттерн понятий TLA с выходным понятием `грязный`, так как при этом возникнет и сопоставление с текстом, и более соответствующий дескриптор.

Бизнес-правила категорий как дескрипторы

Правила категорий - это операторы, автоматически классифицирующие документы или записи в категорию на основании логического выражения, использующего извлеченные понятия, типы и паттерны, а также логические операторы. Например, можно написать выражение, означающее, что следует *включить все записи, содержащие извлеченное понятие* `посольство`, *но не* понятие `аргентина`, *в эту категорию*.

Вы можете записать и использовать правила категорий как дескрипторы в ваших категориях, чтобы выразить несколько разных идей с помощью логических операторов `&`, `|` и `!()`. Подробную информацию о синтаксисе этих правил, их записи и изменениях смотрите в разделе “Использование правил категорий” на стр. 126.

- Использование правила категории с логическим оператором `&` (AND) поможет вам найти документы или записи, в которых есть два или более понятий. Два или более понятий, соединенных операторами `&`, не должны встречаться именно в одном предложении или словосочетании, но могут появиться где угодно в документе или записи, и тогда они будут рассматриваться как соответствующие категории. Например, если в качестве дескриптора создается правило категории `еда & дешево`, этот дескриптор подойдет для записи с текстом "*еда была очень дорогой, но номера стоили дешево*", хотя здесь `еда` - не то существительное, к которому относится определение `дешево`, но в тексте одновременно есть и понятие `еда`, и понятие `дешево`.
- Использование правила категории с логическим оператором `!()` (NOT) поможет вам найти документы или записи, в которых какие-то понятия встречаются, а какие-то нет. Это поможет исключить группировку информации, которая может показаться связанной на основании слов, но не контекста. Например, если в качестве дескриптора создается правило категории `<Organization> & !(ibm)`, этот дескриптор подойдет для текста *SPSS Inc. was a company founded in 1967* и не подойдет для текста *the software company was acquired by IBM.*.

- Использование правила категории с логическим оператором `|()` (OR) поможет вам найти документы или записи, в которых встречаются какие-то понятия из перечисленных понятий или типов. Например, если в качестве дескриптора создается правило категории `(personnel|staff|team|coworkers) & bad`, этот дескриптор подойдет для любых документов или записей, в которых встречаются какие-то из перечисленных существительных одновременно с понятием `bad`.
- Используйте типы в правилах категорий, чтобы сделать их более общими и более вероятно внедряемыми. Например, если вы работаете с данными гостиницы, может быть интересна информация о том, что думают посетители о персонале гостиницы. Соответствующие термины могут включать в себя такие слова как receptionist (дежурный), waiter (официант), waitress (официантка), reception desk (стойка регистрации), front desk (стойка администратора) и так далее. В этом случае вы можете создать новый тип `<HotelStaff>` и добавить к этому типу все указанные термины. Хотя правило категории можно создать для любой специализации персонала, например, `[* waitress * & nice]`, `[* desk * & friendly]`, `[* receptionist * & accommodating]`, можно создать и одно более общее правило категории, используя тип `<HotelStaff>` для захвата всех ответов с положительными мнениями о персонале гостиницы в форме `[<HotelStaff> & <Positive>]`.

*Примечание:* При включении в правила категорий паттернов TLA в этих правилах можно использовать операторы + и &. Дополнительную информацию смотрите в разделе “Использование паттернов TLA в правилах категорий” на стр. 128.

Пример разного соответствия при использовании в качестве дескрипторов понятий, TLA или правил категорий

В следующем примере показано, как использование в качестве дескриптора понятия, правила категории или паттерна TLA влияет на категоризацию документов или записей. Допустим, например, что у вас есть 5 следующих записей.

- A: *"awesome restaurant staff, excellent food and rooms comfortable and clean."* (великолепный персонал ресторана, отличная еда, комнаты удобные и чистые)
- B: *"restaurant personnel was awful, but rooms were clean."* (персонал ресторана был ужасен, но комнаты были чистыми)
- C: *"Comfortable, clean rooms."* (удобные чистые комнаты)
- D: *"My room was not that clean."* (моя комната было не очень чистой)
- E: *"Clean."* (чисто)

Так как эти записи включают слово *clean* (чисто) и вы хотите захватывать такую информацию, можно создать один из дескрипторов, указанных в следующей таблице. На основании смысла содержимого, который вы хотите захватить, можно сравнить, как использование дескрипторов разного вида может привести к разным результатам.

*Таблица 17. Как примеры записей соответствуют дескрипторам.*

| Дескриптор | A | B | C | D | E | Обоснование |
|---|---|---|---|---|---|---|
| `clean` | *соотв.* | *соотв.* | *соотв.* | *соотв.* | *соотв.* | Дескриптор - это извлеченное понятие. Каждая запись, содержащая понятие `clean`, даже запись D, так как без TLA автоматически не будет определено, что “*not clean*” означает по правилам TLA `dirty`. |
| `clean + .` | - | - | - | - | *соотв.* | Дескриптор - это паттерн TLA, представляющий само понятие `clean`. Соответствует только записи, в которой понятие `clean` при извлечении TLA получено без связанного понятия. |

*Таблица 17. Как примеры записей соответствуют дескрипторам (продолжение).*

| Дескриптор | A | B | C | D | E | Обоснование |
|---|---|---|---|---|---|---|
| [clean] | *соотв.* | *соотв.* | *соотв.* | - | *соотв.* | Дескриптор - это правило категории, которое ищет правило TLA, содержащее понятие clean само по себе или в связи с чем-то еще. Соответствуют все записи, для которых выход TLA содержит только понятие clean или понятие clean, связанное с другим понятием, таким как room, в любой позиции слота. |

## О категориях

**Категории** - это группы тесно связанных понятий, мнений или установок. Чтобы категория была полезной, ее должно быть просто описать короткой фразой или меткой, отражающей ее смысл.

Например, если вы анализируете ответы потребителей, полученные в ходе опроса о новом хозяйственном мыле, можно создать категорию с меткой *запах*, содержащую все ответы, в которых описывается запах продукта. Однако такая категория не отделяет ответы тех, кому запах показался приятным, от тех, кому он не понравился. Поскольку IBM SPSS Modeler Text Analytics способен извлекать мнения при использовании подходящих ресурсов, нужно создать две другие категории, объединяющие респондентов, которым *запах понравился* и респондентов, которым *запах не понравился*.

Создавать категории и работать с ними можно на панели Категории в верхней левой панели окна представления Категории и понятия. Каждая категория определяется одним или несколькими дескрипторами. **Дескрипторы** - это понятия, типы и паттерны, а также правила категории, использованные для ее определения.

Чтобы увидеть дескрипторы, формирующие конкретную категорию, щелкните по значку с карандашом на полосе инструментов панели Категории и затем раскройте дерево с дескрипторами. Другой вариант - выберите категорию и откройте диалоговое окно Определения категорий (**Просмотр > Определения категорий**).

При автоматическом построении категорий при помощи методов построения категорий, таких как добавление понятий, эти методы будут рассматривать понятия и типы в качестве дескрипторов для создания категорий. Если вы извлекаете паттерны TLA, можно также добавить паттерны или части этих паттернов в качестве дескрипторов категорий. Дополнительную информацию смотрите в разделе Глава 12, “Изучаем анализ текстовых связей (Text Link Analysis, TLA)”, на стр. 153. Если же вы строите кластеры, в них можно добавлять понятия, используя для этого новые или существующие категории. Наконец, можно вручную создать правила категорий, которые будут служить дескрипторами в ваших категориях. Дополнительную информацию смотрите в разделе “Использование правил категорий” на стр. 126.

## Свойства категорий

Помимо дескрипторов, у категорий есть также редактируемые свойства, при помощи которых категории можно переименовывать и снабжать метками и аннотациями .

Существуют следующие свойства:

- **Имя.** Это имя выводится в дереве по умолчанию. Если категория создается автоматическим методом, ее имя задается автоматически.
- **Метка.** Метки полезны как описания категорий при использовании их в других продуктах или в других таблицах и графиках. Если выбрать опцию вывода метки, метка используется в интерфейсе для идентификации категории.
- **Код.** Номер кода соответствует значению кода категории. .

- **Аннотация.** В этом поле для каждой категории можно добавить краткое описание. Если категория сгенерирована в диалоговом окне Построить категории, замечание в эту аннотацию добавляется автоматически. Вы также можете добавить текст выборки в аннотацию непосредственно с панели данных, выбрав текст и щелкнув в меню по **Категории > Добавить в аннотацию**.

## Панель Данные

При создании категорий бывают моменты, когда нужно просмотреть некоторые из текстовых данных, с которыми вы работаете. Например, при создании категории, в которую входит 640 документов, можно взглянуть на некоторые или все документы, чтобы посмотреть, какой текст был фактически написан. Можно просмотреть записи и документы на панели данных, расположенной в нижнем правом углу. Если она не видна по умолчанию, выберите **Вид > Панели > Данные** в меню.

На панели данные выводится по одной строке для документа или записи, соответствующей выбранному варианту на панели Категории, панели Результаты извлечения или в диалоговом окне Определения категорий до определенного предела для вывода. По умолчанию число документов или записей, выводимых на панели Данные, ограничено, что позволяет быстрее просмотреть данные. Но это можно изменить в диалоговом окне Опции. Дополнительную информацию смотрите в разделе “Опции: вкладка Сеанс” на стр. 82.

Вывод и обновление панели данных

Панель данные не обновляет свой вывод автоматически, поскольку при более крупных наборах данных выполнение автоматического обновления данных может занять некоторое время. Поэтому всякий раз, выбрав вариант на другой панели в этом представлении или диалоговом окне Определения категорий, нажмите кнопку **Показать**, чтобы обновить содержимое панели Данные.

Текстовые документы или записи

Если ваши текстовые данные имеют форму записей, и текст относительно небольшой длины, текстовое поле на панели данных содержит полный текст. Но при работе с записями и наборами данных большего размера столбец текстового поля содержит небольшую часть текста и открывает панель Предварительный просмотр текста справа, содержащую большего размера порцию или весь текст записи, выбранной в таблице. Если ваши текстовые данные имеют форму отдельных документов, панель данных содержит имя файла документа. При выборе документа открывается панель Предварительный просмотр текста с текстом выбранного документа.

Цвета и выделение

при выводе данных понятия и дескрипторы, найденные в документах или записях, выделяются цветом для удобства идентификации в тексте. Цветовые коды соответствуют типам, заданным для понятий. Кроме того, если остановить указатель мыши на элементе того или иного цвета, выводится то понятие, под которым этот элемент был извлечен, и назначенный этому понятию тип. Текст, который не был извлечен, выводится черным. Чаще всего не извлечены остаются такие слова, как соединители (*и* или *с*), местоимения (*меня* или *они*) и глаголы (*был*, *есть*, *принимать*).

Столбцы панели данных

Столбец текстового поля всегда видимый, и можно задать вывод других столбцов. Чтобы вывести другие столбцы, выберите **Вид > Панель данных** в меню и выберите столбец, который нужно вывести на панели данных. Для вывода могут быть доступны следующие столбцы:

- **"Имя текстового поля" (#)/документы.** Добавляет столбец для текстовых данных, из которого были извлечены понятия и тип. Если данные представлены в документах, столбец называется Документы и видимы только имя файла или полный путь документа. Чтобы увидеть текст этих документов, нужно заглянуть на панель Предварительный просмотр текста. Число строк на панели данных показано в скобках после имени этого столбца. Иногда показаны не все документы или записи из-за ограничения в

диалоговом окне Опции, цель которого - повысить скорость загрузки. При достижении максимума после числа добавляется **- Max**. Дополнительную информацию смотрите в разделе “Опции: вкладка Сеанс” на стр. 82.

- **Категории.** Выводит все категории, к которым принадлежит запись. Если этот столбец выведен, обновление панели данных может занять больше времени, пока собирается новейшая информация.
- **Ранг значимости.** Содержит ранг каждой записи в одной категории. Ранг показывает, насколько хорошо запись соответствует категории по сравнению с остальными записями в этой категории. Выберите категорию на панели категорий (верхняя левая панель), чтобы увидеть ранг. Дополнительную информацию смотрите в разделе “Релевантность категорий”.
- **Число категорий.** Выводит число категорий, к которым принадлежит запись.

## Релевантность категорий

Помочь построению лучших категорий может пересмотр релевантности документов или записей в каждой категории, а также всех категорий, к которым принадлежит документ или запись.

Релевантность категории для записи

При всяком появлении документов или записей на панели Данные все категории, к которым они принадлежат, перечисляются в столбце Категории. Когда документ или запись принадлежит к нескольким категориям, категории в этом столбце появляются по порядку, от наибольшей до наименьшей релевантности соответствия. Предполагается, что первая из перечисленных категорий - наилучшая, соответствующая этой записи или документу. Дополнительную информацию смотрите в разделе “Панель Данные” на стр. 110.

Релевантность записи для категории

При выборе категории можно пересмотреть релевантность каждой из ее записей в столбце Ранг релевантности на панели Данные. Этот ранг релевантности обозначает, насколько хорошо документ или запись подходит для выбранной категории по сравнению с другими записями в этой категории. Чтобы просмотреть ранг записей для одной категории, выберите эту категорию на панели Категории (верхняя левая панель), и в столбце появится ранг для документа или категории. По умолчанию этот столбец не выводится, но можно выбрать опцию его показа. Дополнительную информацию смотрите в разделе “Панель Данные” на стр. 110.

Чем меньше номер ранга записи, тем лучше соответствие и тем более релевантна эта запись выбранной категории, то есть 1 соответствует наилучшему совпадению. Если у нескольких записей релевантность одинакова, каждая из них появляется с одинаковым рангом, после которого показан знак равенства (=), указывающий на совпадение значений релевантности. Например, ранги могут быть следующими: 1=, 1=, 3, 4 и так далее, то есть может быть две записи, соответствие которых данной категории рассматривается как наилучшее.

**Подсказка**: Текст наиболее релевантной записи можно добавить в аннотацию категории, чтобы улучшить ее описание. Добавить этот текст можно непосредственно с панели Данные, выбрав текст и использовав пункт меню **Категории > Добавить в аннотацию**.

---

## Построение категорий

У вас могут быть категории из пакета текстового анализа, но категории можно также построить автоматически при помощи ряда лингвистических и частотных методов. В диалоговом окне Параметры построения категорий с помощью методов автоматизированных лингвистических и частотных методов можно построить категории на основе либо понятий, либо паттернов понятий.

В общем случае категории могут состоять из разнородных дескрипторов, таких как типы, понятия, паттерны TLA (Text Link Analysis - анализ текстовых связей) и правила категорий. При построении категорий методами автоматизированного построения категорий окончательным категориям присваивается имя,

добавляемое после понятия или паттерна понятия (в зависимости от выбранных вами входных данных); каждая категория содержит набор дескрипторов. Эти дескрипторы могут быть в форме правил или понятий категорий и содержать все связанные понятия, обнаруженные используемыми методами.

После построения о категориях можно получить информацию, посматривая их на панели Категории или анализируя на графиках и диаграммах. Затем при помощи ручных методов в них можно внести второстепенные корректировки, удалить все неправильные категории или добавить записи или слова, которые могли оказаться пропущены. После применения метода понятия, типы и паттерны, которые были сгруппированы в категорию, останутся по-прежнему доступными для других методов. Кроме того, поскольку использование различных методов может также сгенерировать излишние или неуместные категории, такие категории можно слить или удалить. Дополнительную информацию смотрите в разделе “Изменение и уточнение категорий” на стр. 143.

**Важно!** В прошлых выпусках правила совместного появления и синонимов заключались в квадратные скобки. В этом выпуске квадратные скобки обозначают результат паттерна Text Link Analysis. Правила совместного появления и синонимов берутся в круглые скобки, например, (акустические системы|динамики).

Чтобы построить категории:

1. В наборе меню выберите **Категории > Построить категории**. Появится окно сообщения (если только вы не выбрали опцию Никогда не подсказывать).
2. Выберите, хотите ли вы выполнить построение сейчас или сначала отредактировать параметры.

- Чтобы начать построение категорий при помощи текущих параметров, нажмите кнопку **Построить сейчас**. Часто, чтобы начать процесс построения категорий, выбранных по умолчанию значений параметров достаточно. Начнется процесс построения категорий и откроется диалоговое окно с ходом выполнения.
- Чтобы проверить и изменить параметры построения, нажмите кнопку **Редактировать**.

*Примечание*: Максимально можно вывести 10000 категорий. При достижении или превышении этого числа выводится предупреждение. Если такое случится, следует уменьшить число построенных категорий, изменив опции построения или расширения категорий.

Входные поля

Категории строятся из дескрипторов, получаемых либо из паттернов типов, либо из типов. В приведенной таблице можно выбрать отдельные типы или паттерны для включения в процесс построения категорий.

**Паттерны типов**. Если выбрать паттерны типов, категории будут построены на основе паттернов, а не из типов и понятий самостоятельно. Этим способом будут построены категории все записей или документов, содержащих паттерн понятий, принадлежащий к выбранному типу паттернов. Так, если в таблице выбрать паттерн типов <Бюджет> и <Положительные>, могут быть сгенерированы такие категории, как стоимость & <Положительные> или ставки & отличные.

Если для автоматизированного построения категорий в качестве входных данных используются паттерны типов, бывают случаи, когда методы определяют несколько способов формирования структуры категорий. Формально нет единственного правильного способа генерирования категорий, однако, возможно, вы поймете, что одна структура подходит для анализа лучше, чем другая. В этом случае для настройки вывода может оказаться полезным назначить в качестве предпочтительного фокуса тип. Все генерируемые категории верхнего уровня будут поступать из понятия типа, который вы выберете здесь (и никакого другого типа). Каждая подкатегория будет содержать паттерн текстовых связей из этого типа. Выберите этот тип в поле **Структурировать категории по типу паттернов:**, и таблица будет обновлена так, чтобы выводились только применимые паттерны, содержащие выбранный тип. Довольно часто будет предварительно выбрана категория <Неизвестные>. Вследствие этого будут выбраны все паттерны, содержащие тип <Неизвестные> (для текста не на японском языке). . В таблице выводятся типы в нисходящем порядке, начиная с типа с самым большим числом записей или документов (**Doc.**).

**Типы**. Если выбрать типы, категории будут построены из понятий, принадлежащих выбранным типам. Так, если выбрать тип <Бюджет>, могут быть сгенерированы такие категории, как стоимость или цена, поскольку стоимость и цена - это понятия, назначаемые типу <Бюджет> .

По умолчанию будут выбраны только типы, захватывающие большинство записей выбираемых записей или документов. Этот предварительный выбор позволяет быстро сфокусировать внимание на наиболее интересных типах и избежать построения не имеющих значения категорий. В таблице выводятся типы в нисходящем порядке, начиная с типа с самым большим числом записей или документов (**Doc.**). По умолчанию типы из библиотеки Opinions в таблице типов не выбраны.

Выбираемые входные данные влияют на категории, которые вы получите. Если выбрать использовать в качестве входных данных Типы, можно будет подробнее рассмотреть явно связанные понятия. Например, если построить категории, применив в качестве входных данных Типы, можно получить категорию Фрукты с такими понятиями, как яблоки, груши, цитрусовые, апельсины и так далее. Выбрав в качестве входных данных тип паттерны и выбрав, например, паттерн <Неизвестные> + <Положительные>, можно получить категорию фрукты + <Положительные> с фруктами одним из двух видов, таких как фрукты + вкусные и яблоки + хорошие. Этот второй результат содержит только два паттерна понятий, поскольку остальные вхождения фруктов необязательно будут специфицированы как положительные. И хотя это и может оказаться достаточно полезным для текущих текстовых данных, в продольных исследованиях, где будут использоваться различные наборы документов, может потребоваться встроить вручную другие дескрипторы, такие как цитрусовые + положительные, или использовать типы. Использование типов в качестве входных данных автономно поможет найти все возможные фрукты.

Методы

Поскольку набор данных уникален, число методов и порядок их применения с течением времени может изменяться. Поскольку от одного набора данных к другому ваши цели исследования данных могут меняться, может потребоваться опробовать различные методы, чтобы понять, какой из них генерирует наилучшие результаты для настоящих текстовых данных.

Вовсе необязательно быть специалистом по этим параметрам, чтобы их использовать. По умолчанию часто используемые и средние значения параметров уже выбраны. Поэтому диалоговые окна дополнительных параметров можно обойти и сразу приступить к построению категорий. Таким же образом, при внесении здесь изменений не придется возвращаться в диалоговое окно параметров каждый раз, поскольку всегда сохраняются последние заданные значения параметров.

Выберите либо лингвистические, либо частотные методы и нажмите кнопку Дополнительные параметры, чтобы вывести параметры для выбранных методов. Ни один из автоматических методов не гарантирует идеальной категоризации ваших данных; мы рекомендуем найти и применить один или несколько автоматических методов, которые хорошо работают с вашими данными.Выполнять построение, используя одновременно и лингвистические, и частотные методы, нельзя.

- **Расширенные лингвистические методы**. Дополнительную информацию смотрите по адресу “Дополнительные лингвистические параметры”.
- **Расширенные частотные методы**. Дополнительную информацию смотрите по адресу “Дополнительные частотные параметры” на стр. 121.

## Дополнительные лингвистические параметры

При построении категорий доступен ряд вариантов выбора расширенных лингвистических методов построения категорий, включая *вывод корня понятия* (для японского языка недоступен), *включение понятий*, *семантические сети* (только для текста на английском) и *правила совместного появления*. Эти методы используются для создания категорий по отдельности или в сочетании друг с другом.

Имейте в виду, что поскольку каждый набор данных уникален, число методов и порядок их применения с течением времени могут меняться. Поскольку ваши цели исследования данных могут изменяться от одного набора данных к другому, может потребоваться попробовать различные методы, чтобы понять, какой из

них генерирует наилучшие результаты для реальных текстовых данных. Ни один из автоматических методов не гарантирует идеальной категоризации ваших данных; мы рекомендуем найти и применить один или несколько автоматических методов, которые хорошо работают с вашими данными.

В диалоговом окне Дополнительные параметры: Лингвистика доступны следующие области и поля:

Вход и выход

**Входные данные категорий**. Выберите, из чего будут построены категории:

- **Неиспользованные результаты извлечения**. Эта опция разрешает строить категории по результатам извлечения, не использованным в существующих категориях. Этим минимизируется тенденция сопоставлять записи нескольким категориям и ограничивается число созданных категорий.
- **Все результаты извлечения**. Эта опция разрешает строить категории по любым результатам извлечения. Это особенно полезно, когда категорий еще мало или совсем нет.

**Выходные данные категорий.** Выберите общую структуру для категорий, которые будут построены:

- **Иерархические с подкатегориями**. Эта опция включает поддержку создания подкатегорий на нескольких уровнях иерархии. Можно задать глубину иерархии категорий, выбрав максимальное число уровней (в поле **Максимальное число создаваемых уровней**), которые можно будет создать. Если выбрать 3, категории смогут содержать подкатегории, и у этих подкатегорий тоже будут подкатегории.
- **Плоские категории (только один уровень)**. Эта опция включает поддержку построения только одного уровня категорий, означающую, что никакие подкатегории сгенерированы не будут.

Методы группирования

Каждый из доступных методов хорошо подходит для определенных типов данных и ситуаций, но часто для получения полного набора документов или записей бывает полезно сочетать методы в ходе одного анализа. Вы можете увидеть понятие в нескольких категориях или найти лишние категории.

**Вывод корня понятия**. Этот метод создает категории, исходя из некоторого понятия, путем поиска связанных с ним других понятий, когда соответствующие компоненты морфологически родственны или имеют общие корни. Этот метод очень полезен для выявления синонимичных понятий, выраженных сочетаниями слов, поскольку понятия в каждой созданной категории являются синонимами или близки по значению. Он работает с данными различной длины и создает небольшое число компактных категорий. Например, понятие `возможности развиваться` могло бы объединиться с понятиями `возможности для развития` и `возможное развитие`. Дополнительную информацию смотрите в разделе “Вывод корня понятия” на стр. 117. Для текста на японском эта опция недоступна.

**Семантическая сеть**. Этот метод начинает обработку с выявления возможных направлений обхода каждого понятия в его расширенном индексе взаимосвязей слов, а затем создает категории, группируя связанные понятия. Этот метод дает наилучшие результаты, когда понятия известны семантической сети и не слишком неоднозначны. Он менее полезен, когда текст содержит специальную терминологию или неизвестные сети жаргонизмы. Например, понятие `яблоки Гренни Смит` может быть объединено с понятиями `яблоки Гала` и `яблоки Вайнсеп`, поскольку все это представители одного класса объектов (сорта яблок). В другом примере понятие `животное` можно объединить с понятиями `кошка` и `кенгуру`, поскольку это гипонимы для понятия `животное`. Этот метод в данном выпуске применим только к английским текстам. Дополнительную информацию смотрите в разделе “Семантические сети” на стр. 119.

**Вложенные понятия**. Этот метод строит категории путем группировки составных терминов (словосочетаний) на основе наличия в них слов, являющихся поднаборами других или, наоборот, включающих в себя другие слова как поднаборы. Например, понятие `сидение` будет объединено с понятиями `сидение с ремнем безопасности`, `ремень безопасности` и `пряжка ремня безопасности`. Дополнительную информацию смотрите в разделе “Включение понятий” на стр. 118.

**Совместная встречаемость**. Этот метод создает категории, исходя из совместной встречаемости понятий в тексте. Идея состоит в том, что когда понятия или шаблоны понятий часто встречаются вместе в документах и записях, эта совместная встречаемость отражает некую базовую взаимосвязь, которая может оказаться полезной в ваших определениях категорий. Когда слова имеют существенную тенденцию к совместной встречаемости, создается правило совместной встречаемости, которое может выступать в качестве дескриптора для новой подкатегории. Например, если во многих записях встречаются слова цена и доступность (но по отдельности эти слова встречаются редко), соответствующие понятия можно сгруппировать, создав правило совместной встречаемости (цена & доступная) и включить их, например, в подкатегорию категории цена. Дополнительную информацию смотрите в разделе “Правила совместного появления” на стр. 120.

**Минимальное число документов.** Чтобы определить, насколько интересными могут быть совместные вхождения, задайте минимальное число документов или записей, в которых термины встречаются совместно, чтобы использовать эти случаи в качестве дескриптора категории.

**Максимальное расстояние поиска.** Выберите, на каком удалении должны выполнить поиск методы, прежде чем создать категории. Чем меньше это значение, тем меньше будет результатов поиска; вместе с тем в таких результатах будет меньше шума, и они с большей вероятностью окажутся зависимы друг от друга. Чем выше это значение, тем больше будет результатов; однако такие результаты могут оказаться ненадежны или не соответствовать цели поиска. Хотя эта опция глобально применяется ко всем методам, она наиболее эффективна для сетей совместного появления и семантических сетей.

**Предотвращать объединение отдельных понятий в пары.** Включите этот переключатель, чтобы не объединять два понятия в группу или пару при выводе результатов. Чтобы создавать пары понятий и работать с ними, щелкните по **Работа с парами...**. Дополнительную информацию смотрите в разделе “Управление парами с исключением связи” на стр. 116.

**По возможности обобщать с помощью символов подстановки.** Выберите эту опцию, чтобы разрешить программному продукту генерировать в категориях общие правила, применяя в качестве символа подстановки звездочку. Например, вместо создания нескольких дескрипторов, таких как [яблочный пирог + .] и [яблочное пюре + .] с помощью символов подстановки может быть сгенерировано: [яблочн* + .]. В случае обобщения при помощи символов подстановки часто будет возвращаться в точности такое же число записей или документов, что и ранее. Однако у этой опции есть преимущество, позволяющее уменьшить число дескрипторов категорий и упростить эти дескрипторы. Дополнительно эта опция повышает способность категоризации большего количества записей или документов благодаря использованию этих категорий для новых текстовых данных (например, в лонгитюдных/волновых исследованиях).

Другие опции для построения категорий

В дополнение к выбору вариантов применяемых методов группирования можно отредактировать несколько других опций построения следующим образом:

**Максимальное число создаваемых категорий** верхнего уровня. Эта опция позволяет ограничить число категорий, которые можно будет сгенерировать при нажатии соседней с ней кнопки Построить категории. В некоторых случаях можно будет получить лучшие результаты, если задать это значение высоким, а затем удалить все неинтересные категории.

**Минимальное число дескрипторов и/или подкатегорий для одной категории.** С помощью этой опции задается минимальное число дескрипторов и/или подкатегорий, которые должна содержать создаваемая категория. Эта опция помогает ограничить создание категорий, не захватывающих существенного числа записей или документов.

**Разрешить присутствие дескрипторов в нескольких категориях.** Если эта опция выбрана, она разрешает использование дескрипторов в нескольких категориях, которые будут построены в дальнейшем. Как правило, эта опция будет выбрана, поскольку элементы обычно или "естественным образом" попадают в несколько категорий, и разрешение такого их поведения приводит к построению категорий более высокого

качества. Если не включить эту опцию, будет уменьшено перекрытие записей в нескольких категориях, а в зависимости от типа используемых вами данных это может оказаться желательным. Однако для большинства типов данных ограничение числа дескрипторов для одной категории обычно приводит к потере качества или охватываемого категориями диапазона. Например, предположим, что у вас было понятие производитель автомобильных сидений. Если включить эту опцию, настоящее понятие сможет присутствовать в одной категории на основе текста автомобильное сиденье и в другой категории на основе текста производитель. Но если эта опция не включена, хотя вы по-прежнему можете получить обе эти категории, понятие производитель автомобильных сидений будет присутствовать в качестве дескриптора в категории, которой оно лучше всего соответствует на основе нескольких факторов, включая число записей, в которых встречается и автомобильное сиденье, и производитель.

**Разрешить повторяющиеся имена категорий по**. Выберите, как обрабатывать любые новые категории или подкатегории, имена которых совпадут с именами уже существующих категорий. Можно либо выполнять слияние новых категорий (и их дескрипторов) с существующими категориями с тем же именем, либо выбрать вариант пропуска создания всяческих категорий, если дубликаты их имен будут найдены в существующих категориях.

### Управление парами с исключением связи

При построении категорий, кластеризации и отображении понятий внутренние алгоритмы группируют слова по известным ассоциациям. Чтобы исключить образование пар понятий (их связывание), можно включить эту возможность в диалоговом окне **Дополнительные параметры построения категорий**, диалоговом окне **Построение кластеров** и в диалоговом окне **Параметры индексации отображения понятий** и нажать кнопку **Управление парами**.

В появившемся диалоговом окне **Управление исключением связей** можно добавить, изменить или удалить пары понятий. Вводите по одной паре на строку. Введенные здесь пары не будут появляться при построении или расширении категорий, кластеризации и отображении понятий. Точно вводите формы слова, как они вам требуются, например, версия слова с проставленным ударением будет отличаться от версии без ударения.

Например, чтобы задать, что понятия hot dog и dog не должны группироваться, надо добавить эту пару как отдельную строку в таблице.

## О лингвистических методах

При построении или расширении категорий доступен ряд вариантов выбора расширенных лингвистических методов построения категорий, включая *вывод корня понятия* (для японского языка недоступен), *включение понятий*, *семантические сети* (только для английского языка) и *правила совместного появления*. Эти методы используются для создания категорий по отдельности или в сочетании друг с другом.

Вовсе необязательно быть специалистом по этим параметрам, чтобы их использовать. По умолчанию часто используемые и средние значения параметров уже выбраны. Если вы хотите, это диалоговое окно дополнительных параметров можно обойти и сразу приступить к построению или расширению категорий. Таким же образом, при внесении здесь изменений не придется возвращаться в диалоговое окно параметров каждый раз, поскольку то, что вы последний раз использовали, запоминается.

Однако имейте в виду, что поскольку каждый набор данных уникален, число методов и порядок их применения с течением времени может меняться. Поскольку ваши цели исследования данных могут изменяться от одного набора данных к другому, может потребоваться попробовать различные методы, чтобы понять, какой из них генерирует наилучшие результаты для реальных текстовых данных. Ни один из автоматических методов не гарантирует идеальной категоризации ваших данных; мы рекомендуем найти и применить один или несколько автоматических методов, которые хорошо работают с вашими данными.

Вот основные автоматизированные лингвистические методы для построения категорий:

- **Вывод корня понятия.** Этот метод создает категории посредством принятия понятия и нахождения других связанных с ним понятий путем анализа, выясняющего, связаны ли с ним морфологически другие компоненты понятия. Дополнительную информацию смотрите в разделе “Вывод корня понятия”. Для текста на японском эта опция недоступна.
- **Включение понятий.** Этот метод создает категории посредством принятия понятия и нахождения других понятий, в которые оно входит. Дополнительную информацию смотрите в разделе “Включение понятий” на стр. 118.
- **Семантическая сеть.** Этот метод начинает обработку с выявления возможных направлений обхода каждого понятия в его расширенном индексе взаимосвязей слов, а затем создает категории, группируя связанные понятия. Дополнительную информацию смотрите в разделе “Семантические сети” на стр. 119. Эта опция доступна только для текста на английском.
- **Совместное появление.** В этом способе создаются правила совместного появления, которые можно использовать для создания новой категории, расширения категории или в качестве вводных данных для другого способа работы с категориями. Дополнительную информацию смотрите в разделе “Правила совместного появления” на стр. 120.

## Вывод корня понятия

*Примечание:* Этот способ недоступен для японского текста.

Способ вывода корня понятия создает категории, принимая понятие, находя другие понятия, которые связаны с первым и анализируя эти понятия на наличие морфологически связанных компонентов. Компонент - это слово. Этот способ пытается сгруппировать понятия, просматривая окончания (суффиксы) всех компонентов понятия и находя другие понятия, которые могли бы произойти от них. Идея состоит в том, что если если одна словоформа происходит от другой, их понятия могут совпадать или быть близкими по смыслу. Чтобы идентифицировать окончания, используются внутренние правила, относящиеся к конкретному языку. Например, понятие `возможности развиваться` могло бы объединиться с понятиями `возможности для развития` и `возможное развитие`.

Способ вывода корня понятия можно использовать для любого вида текста. Этот способ сам по себе создает довольно мало категорий, и каждая категория может содержать не так много понятий. Понятия в каждой категории - это или синонимы, или ситуативно связанные формы. Использование этого алгоритма может оказаться полезным даже при построении категорий вручную; находимые им синонимы могут быть синонимами тех понятий, которые вы в настоящее время используете.

*Примечание*: группировку понятий можно предотвратить, задавая их непосредственно. Дополнительную информацию смотрите в разделе “Управление парами с исключением связи” на стр. 116.

Разделение терминов на компоненты и изменение их грамматических форм

При применении алгоритмов вывода корня понятий или включения понятий термины сначала разбиваются на компоненты (слова), а затем компоненты приводятся к основной грамматической форме. При применении этого способа понятия и связанные с ними термины загружаются и расщепляются на компоненты с помощью разделителей, таких как пробелы, дефисы или апострофы. Например, термин `системный администратор` разделяется на компоненты {`администратор, системный`}.

Однако некоторые части исходного слова не могут использоваться, их называют стоп-словами. В английском языке примерами таких игнорируемых компонентов могут быть `a`, `and`, `as`, `by`, `for`, `from`, `in`, `of`, `on`, `or`, `the`, `to` и `with`.

Например, у термина `examination of the data` есть набор компонентов {`data, examination`}, а к игнорируемым стоп-словам относятся `of` и `the`. Кроме этого, порядок компонентов для их набора не определяется. Таким образом, следующие три термина могут быть эквивалентными: `cough relief for child`, `child relief from a cough` и `relief of child cough`, так как у них одинаковый набор компонентов {`child, cough, relief`}. Всякий раз, когда пара терминов определяется как эквивалентная, соответствующие понятия сливаются и образовывают новое понятие, содержащее все эти термины.

Кроме этого, так как компоненты термина могут отличаться окончанием (склоняться или спрягаться), внутренне применяются относящиеся к конкретному языку правила для идентификации эквивалентных терминов независимо от изменчивости формы окончания, например, для формы множественного числа. Таким образом, термины `level of support` и `support levels` могут идентифицироваться как эквивалентные, очищенной от окончаний формой единственного числа может быть `level`.

Как работает вывод корня понятия

После разделения терминов на компоненты и приведения их к основной грамматической форме (смотрите предыдущий раздел) алгоритм вывода корня понятия анализирует окончания компонентов (или суффиксы), чтобы найти корень компонента, а затем группирует понятия с другими понятиями, у которых схожие корни. Окончания идентифицируются с использованием набора лингвистических правил отклонения, специфичных для языка текста. Например, для английского языка существует грамматическое правило, по которому компонент понятия, оканчивающийся на суффикс `ical`, может быть произведен из понятия с тем же основным корнем и с суффиксом `ic`. Используя это правило (и избавление от окончаний), этот алгоритм мог бы сгруппировать такие понятия как `epidemiologic study` и `epidemiological studies`.

После разделения термина на компоненты и определения игнорируемых компонентов (например, `in` и `of`), алгоритм вывода корня понятия мог бы сгруппировать также понятия `studies in epidemiology` и `epidemiological studies`.

Был выбран набор правил порождения компонентов, чтобы большинство сгруппированных этим алгоритмом понятий стали синонимами: понятия `epidemiologic studies`, `epidemiological studies`, `studies in epidemiology` - это эквивалентные термины. Для полноты картины можно использовать некоторые порождающие правила, позволяющие алгоритму группировать ситуативно связанные понятия. Например, этот алгоритм может сгруппировать понятия `empire builder` и `empire building`.

## Включение понятий

Способ включения понятий строит категории, принимая понятие и, используя алгоритмы лексических рядов, идентифицируя понятия, включенные в другие понятия. Идея состоит в том, что для случая, когда слова в понятии представляют собой подмножество слов другого понятия, отображается внутренняя семантическая взаимосвязь. Включение - это мощный прием, который можно использовать с любым типом текста.

Этот способ хорошо работает с семантическими сетями, но может использоваться и отдельно. Включение понятий может дать лучшие результаты также в тех случаях, когда документы или записи содержат много специфичной для некоторой области терминологии или жаргона. Это особенно справедливо, если вы заранее настроили словари, так что специальные термины вынесены и нужным образом сгруппированы (с синонимами).

Как работает включение понятий

Прежде чем применять алгоритм включения понятий, термины должны быть разложены на компоненты и приведены к основной грамматической форме. Дополнительную информацию смотрите в разделе “Вывод корня понятия” на стр. 117. После этого алгоритм включения понятий анализирует наборы компонентов. Для каждого набора компонентов алгоритм ищет другой набор, представляющий из себя подмножество исходного набора компонентов.

Например, если у вас есть понятие `continental breakfast` с набором компонентов {`breakfast, continental`}, а также понятие `breakfast` с набором компонентов {`breakfast`}, алгоритм заключит, что `continental breakfast` - это тип `breakfast` и сгруппирует эти понятия вместе.

В более крупном примере, если на панели Результаты извлечения у вас есть понятие `seat` и применяется этот алгоритм, такие понятия, как `safety seat`, `leather seat`, `seat belt`, `seat belt buckle`, `infant seat carrier` и `car seat laws` будут также сгруппированы в одну категорию.

После разделения термина на компоненты и определения игнорируемых компонентов (например, in и of), алгоритм включения понятий может распознать, что понятие advanced spanish course включает в себя понятие course in spanish.

*Примечание*: группировку понятий можно предотвратить, задавая их непосредственно. Дополнительную информацию смотрите в разделе “Управление парами с исключением связи” на стр. 116.

## Семантические сети

В этом выпуске использование семантических сетей доступно только для англоязычного текста.

Этот метод строит категории, используя встроенную сеть взаимосвязей слов. Из-за этого использование данного метода может привести к очень хорошим результатам, когда термины конкретны и не слишком многозначны. Однако не следует ожидать, что использование семантических сетей приведет к нахождению большого числа связей между техническими или очень специализированными понятиями. При работе с такими понятиями может оказаться, что предпочтительнее способы включения понятий и вывода корня понятий.

Как работает семантическая сеть

Смысл метода семантических сетей состоит в усилении известных взаимосвязей слов для создания категорий синонимов или гипонимов. **Гипоним** - это вторичное, более частное понятие по сравнению с другим понятием, то есть более конкретный элемент иерархической взаимосвязи (другое название - взаимосвязь ISA). Например, если понятие - это animal, cat и kangaroo - это гипонимы animal, так как они представляют собой частные виды животных.

Кроме взаимосвязей гипонимов и синонимов метод семантических сетей проверяет также частичные и полные связи между понятиями типа <Положение>. Например, понятия normandy, provence и france будут группироваться в одну категорию, так как Нормандия и Прованс - это части Франции.

Работа семантических сетей начинается с идентификации возможных смыслов каждого понятия в семантической сети. Когда понятия идентифицируются как синонимы и гипонимы, они группируются в одну категорию. Например, этим способом можно создать одну категорию, содержащую следующие три понятия: eating apple (столовое яблоко), dessert apple (десертное яблоко) и granny smith (сорт Гранни Смит), так как в семантической сети есть информация о том, что: 1) dessert apple - это синоним eating apple, и 2) granny smith - это сорт eating apple (то есть гипоним eating apple).

Взятые по отдельности, многие понятия, особенно одиночные термины, неоднозначны. Например, понятие буфет может означать или способ подачи пищи, или элемент мебели. Если в наборе понятий есть понятия еда, мебель и буфет, алгоритм должен будет выбирать между группировкой буфет с едой или с мебелью. Учтите, что в некоторых случаях выбор алгоритма может не походить для контекста конкретного набора записей или документов.

Метод семантических сетей обычно превосходит результаты включения понятий для определенных типов данных. Хотя и семантическая сеть, и включение понятий распознают, что apple pie - это частный вид понятия pie, только семантическая сеть распознает, что tart - это тоже вид pie.

Семантические сети могут работать в связи с другими способами. Допустим, например, что вы выбрали использование и семантической сети, и включения понятий, и семантическая сеть сгруппировала понятие teacher с понятием tutor (так как tutor - это частный вид понятия teacher). Алгоритм включения понятий может сгруппировать понятия graduate tutor и tutor, то есть в результате два алгоритма создадут выходную категорию, содержащую все три понятия: tutor, graduate tutor и teacher.

Опции семантической сети

Есть несколько дополнительных параметров, которые могут быть полезны для этого метода.

- Измените **Максимальное расстояние поиска**. Выберите, на каком удалении должны выполнить поиск методы, прежде чем создать категории. Чем меньше это значение, тем меньше будет результатов поиска; вместе с тем в таких результатах будет меньше шума, и они с большей вероятностью окажутся зависимы друг от друга. Чем выше это значение, тем больше может быть результатов; однако такие результаты могут оказаться ненадежны или не соответствовать цели поиска.

Например, в зависимости от расстояния поиска алгоритм, начиная с `Danish pastry` ищет понятия до `coffee roll` (родительский элемент), затем до `bun` (родительский элемент еще уровнем выше) и далее по иерархии до `bread`.

При меньшем расстоянии поиска этот алгоритм создаст меньше категорий, с которыми может быть проще работать, если вы чувствуете, что создается очень большая категория или очень много элементов группируется вместе.

**Важно!** Кроме этого, мы не рекомендуем не применять опцию **Согласовать грамматические ошибки для минимального предела символов корня** (определенную на вкладке Эксперт узла или в диалоговом окне Извлечение) для нечеткой группировки при использовании этого метода, так как несколько ложных группировок могут крайне отрицательно сказаться на результатах.

## Правила совместного появления

Правила совместного появления позволяют обнаружить и сгруппировать понятия, тесно связанные с набором документов или записей. Основной смысл состоит в том, что если понятия часто встречаются в документах или записях вместе, такое совместное появление отображает взаимосвязь, которая может быть существенна для определений категорий. В этом способе создаются правила совместного появления, которые можно использовать для создания новой категории, расширения категории или в качестве вводных данных для другого способа работы с категориями. Два понятия считаются строго совместными, если они часто появляются вместе в наборе записей и редко по отдельности во всех других записях. Этот способ может привести к хорошим результатам при работе с большими наборами данных, в которые входит по крайней мере несколько сотен документов или записей.

Например, если во многих записях есть слова `price` и `availability`, эти понятия можно сгруппировать в правило совместного появления `(price & available)`. Другой пример. Если понятия `масло`, `джем`, `сэндвич` чаще появляются вместе, чем по отдельности, их можно сгруппировать в правило совместного появления `(масло & джем & сэндвич)`.

**Важно!** В прошлых выпусках правила совместного появления и синонимов заключались в квадратные скобки. В этом выпуске квадратные скобки обозначают результат паттерна Text Link Analysis. Правила совместного появления и синонимов берутся в круглые скобки, например, `(акустические системы|динамики)`.

Как работаю правила совместного появления

При использовании этого способа документы или записи просматриваются на наличие двух или более понятий, у которых есть тенденция совместного появления. Два или более понятий считаются строго совместными, если они часто появляются в наборе документов или записей вместе и редко - по отдельности в любых других документах или записях.

При обнаружении совместно появляющихся понятий формируется правило категории. Такие правила состоят из двух или более понятий, соединенных логическим оператором &. Эти правила - это логические операторы, которые будут автоматически классифицировать документы или записи как относящиеся к категории, если весь набор понятий из правила присутствует в документе или записи.

Опции для правил совместного появления

Если вы работаете с правилами совместного появления, можно точно подстроить несколько параметров, влияющих на итоговые правила:

- Измените параметр **Максимальное расстояние поиска**. Выберите, как далеко друг от друга будут искаться понятия из правила совместного появления. При увеличении расстояния поиска минимальное значение сходства, требуемое для каждого совместного появления, понижается; в результате можно создать много правил совместного появления, но те из них, у которых будет малое значение сходства, чаще всего окажутся и малозначимыми. При уменьшении расстояния поиска минимальное требуемое значение сходства увеличится; в результате будет создано меньше правил совместного появления, но они с большей вероятностью окажутся значимыми (более строгими).
- **Минимальное число документов.** Минимальное число записей или документов, которые должны содержать данную пару понятий, чтобы они рассматривались как понятия совместного появления; чем меньше значение этой опции, тем легче обнаружить факты совместного появления. Увеличение этого значения приведет к меньшему количеству результатов совместного появления, но они будут более значимыми. Допустим, например, что понятия "яблоко" и "груша" найдены вместе в 2 записях (и ни разу по отдельности). Если для параметра **Минимальное число документов** задано значение 2 (значение по умолчанию), способ совместного появления создаст правило категории (яблоко & груша). Если это значение увеличить до 3, правило создаваться не будет.

*Примечание*: При малых наборах данных (< 1000 ответов) вы не сможете обнаружить совместного появления при использовании параметров по умолчанию. В таких случаях попробуйте увеличить значение расстояния поиска.

*Примечание*: группировку понятий можно предотвратить, задавая их непосредственно. Дополнительную информацию смотрите в разделе “Управление парами с исключением связи” на стр. 116.

## Дополнительные частотные параметры

Вы можете построить категории на основе метода прямой и механической частотности. Этим методом можно построить по одой категории для каждого элемента (типа, понятия или паттерна), который будет найден свыше заданного числа записей или документов. Дополнительно можно построить одну категорию для всех элементов, встречающихся реже. При подсчете предметом нашего обращения является число записей или документов, содержащих извлеченное понятие (и любые его синонимы), тип или паттерн, а не общее число вхождений во всем тексте.

Сгруппировав часто встречающиеся элементы, можно получить интересные результаты, поскольку они могут указывать на часто встречающийся или значимый ответ. Этот метод весьма полезен для неиспользованных результатов извлечения после того, как были применены другие методы. Еще одно применение - запустить этот метод сразу же после извлечения, если никаких других категорий не существует, отредактировать результаты, удалив неинтересные категории, а затем развернуть эти категории так, чтобы они соответствовали еще большему количеству записей или документов. Дополнительную информацию смотрите в разделе “Расширение категорий” на стр. 122.

Вместо использования этого метода можно отсортировать понятия или паттерны понятий по возрастанию числа записей или документов на панели Результаты извлечения, а затем перетащить самые верхние на панель Категории, чтобы создать соответствующие категории.

В диалоговом окне Дополнительные параметры: Частотность доступны следующие поля:

**Сгенерировать дескрипторы категорий как**. Выберите вид входных данных для дескрипторов. Дополнительную информацию смотрите в разделе “Построение категорий” на стр. 111.

- **Уровень понятий**. Выбор этой опции означает, что будет использоваться частотность понятий или паттернов понятий. Понятия будут использоваться, если в качестве входных данных для построения категорий были выбраны типы, а паттерны понятий будут использоваться, если были выбраны паттерны типов. В общем случае при применении этого метода к уровню понятий будут сгенерированы более конкретные результаты, поскольку понятия и паттерны понятий представляют более низкий уровень измерений.
- **Уровень типов**. Выбор этой опции означает, что будет использоваться частотность типов или паттернов типов. Типы будут использоваться, если в качестве входных данных для построения категорий были

выбраны типы, а паттерны типов будут использоваться, если были выбраны паттерны типов. Применение этого метода к уровню типов позволяет получить быстрое представление для типа представления информации (если имеется).

**Минимальное число документов для элементов, у которых должна быть своя собственная категория.** Эта опция позволяет построить категории из чаще всего встречающихся элементов. Эта опция ограничивает вывод только категориями, содержащими дескриптор, встреченный по крайней мере в X записях или документах, где X - значение, которое следует ввести для этой опции.

**Сгруппировать все остальные элементы в категорию с именем**. Эта опция позволяет сгруппировать все часто встречающиеся понятия или типы в одну 'обобщающую' категорию с именем по вашему выбору. По умолчанию эта категория называется *Другое*.

**Входные данные категорий**. Выберите группу, к которой следует применить метод:

- **Неиспользованные результаты извлечения**. Эта опция разрешает строить категории по результатам извлечения, не использованным в существующих категориях. Этим минимизируется тенденция сопоставлять записи нескольким категориям и ограничивается число созданных категорий.
- **Все результаты извлечения**. Эта опция разрешает строить категории по любым результатам извлечения. Это особенно полезно, когда категорий еще мало или совсем нет.

**Разрешить повторяющиеся имена категорий по**. Выберите, как обрабатывать любые новые категории или подкатегории, имена которых совпадут с именами уже существующих категорий. Можно либо выполнять слияние новых категорий (и их дескрипторов) с существующими категориями с тем же именем, либо выбрать вариант пропустить создание любой категории, если в существующих категориях окажется дубликат ее имени.

## Расширение категорий

Расширение - это процесс автоматического добавления или улучшения дескрипторов для 'развития' существующих категорий. Его цель - создать улучшенную категорию, захватывающую связанные записи или документы, которые не были исходно назначены в эту категорию.

Выбираемые вами методы автоматического группирования пытаются выявить понятия, паттерны TLA TLA (Text Link Analysis - анализ текстовых связей) и правила категорий, связанные с существующими дескрипторами категорий. Затем эти новые понятия, паттерны и правила категорий добавляются в качестве новых дескрипторов в существующие дескрипторы. В состав методов группирования для расширения входят *вывод корня понятия* (для японского языка недоступен), *включение понятий*, *семантические сети* (только для английского языка) и *правила совместного появления*. Метод **расширения пустых категорий при помощи дескрипторов, сгенерированных из имени категории** генерирует дескрипторы при помощи слов в именах категорий, поэтому, чем описательней имена категорий, тем лучше результаты.

*Примечание*: Частотные методы при расширении категорий недоступны.

Расширение - это замечательный способ интерактивного улучшения используемых категорий. Вот несколько примеров, когда может быть расширена категория:

- После перетаскивания паттернов понятий для создания категорий на панели Категории.
- После создания категорий ручным способом и добавления простых правил категорий и дескрипторов.
- После импорта файла предварительно заданных категорий , в котором у категорий весьма описательные имена.
- После уточнения категорий, поступивших из выбранного вами TAP (text analysis package - пакет анализа текста).

Категорию можно расширять неоднократно. Например, если вы импортировали файл предопределенных категорий с весьма описательными именами, можно выполнить расширение при помощи опции **Расширить**

**пустые категории при помощи дескрипторов, сгенерированных из имени категории**, чтобы получить первый набор дескрипторов, а затем снова выполнить расширение этих категорий. Однако в других случаях неоднократное расширение может привести к построению слишком общей категории, если дескрипторы будут становиться все шире и шире. Поскольку в методах построения и расширения групп используются схожие базовые алгоритмы, расширение категорий непосредственно после их построения вряд ли позволит получить более интересные результаты.

**Советы.**

- Если вы пытаетесь выполнить расширение и не хотите использовать результаты, операцию можно всегда отменить сразу же после выполнения расширения (**Изменить > Откат**).
- Расширение может сгенерировать несколько правил категорий в категории, точно соответствующей тому же набору документов, поскольку во время выполнения процесса правила строятся независимо. По желанию можно просмотреть категории и удалить излишние, отредактировав описание категории вручную. Дополнительную информацию смотрите в разделе “Изменение дескрипторов категорий” на стр. 144.

Чтобы выполнить расширение категорий:

1. На панели Категории выберите категории, которые вы хотите расширить.
2. В наборе меню выберите **Категории > Расширить категории**. Появится окно сообщения (если только не было выбрана опция Никогда не подсказывать).
3. Выберите, хотите ли вы выполнить построение сейчас или сначала отредактировать параметры.

- Чтобы начать расширение категорий при помощи текущих параметров, нажмите кнопку **Расширить сейчас**. Начнется процесс расширения и откроется диалоговое окно с ходом выполнения.
- Чтобы проверить и изменить значения параметров, нажмите кнопку **Редактировать**.

По завершении попытки расширения все категории, для которых были найдены новые дескрипторы, будут помечены на панели Категории словом **Расширено**, чтобы их можно было быстро идентифицировать. Текст Расширено остается, пока вы либо еще раз не расширите категории, не отредактируете их другим способом, либо не очистите их через контекстное меню.

*Примечание*: Максимально можно вывести 10000 категорий. При достижении или превышении этого числа выводится предупреждение. Если такое случится, следует уменьшить число построенных категорий, изменив опции построения или расширения категорий.

Каждый из методов, доступных при построении или расширении категорий, хорошо подходит к определенным типам данных и ситуаций, но часто будет полезен и для объединения методов в этом же анализе с целью захвата всего диапазона документов или записей. В интерактивной инструментальной среде понятия и типы, которые были сгруппированы в категорию, будут по-прежнему доступны для следующего построения категорий. Это означает, что вы сможете увидеть понятие в нескольких категориях или найти излишние категории.

В диалоговом окне Расширение категорий: Параметры доступны следующие области и поля:

**Расширить с помощью.** Выберите, какие входные данные будут использоваться для расширения категорий:

- **Неиспользованные результаты извлечения**. Эта опция разрешает строить категории по результатам извлечения, не использованным в существующих категориях. Этим минимизируется тенденция сопоставлять записи нескольким категориям и ограничивается число созданных категорий.
- **Все результаты извлечения**. Эта опция разрешает строить категории по любым результатам извлечения. Это особенно полезно, когда категорий еще мало или совсем нет.

Методы группирования

Краткие описания каждого из этих методов смотрите в разделе “Дополнительные лингвистические параметры” на стр. 113. В состав этих методов входят:

- **Вывод корня понятия** *(для японского языка недоступен)*
- **Семантическая сеть** (*только для текста на английском; если выбрана только опция Обобщать, этот метод не используется*)
- **Включение понятий**
- Подопция **Совместное появление** и **Минимальное число документов**

Ряд типов исключен из метода семантических сетей на постоянной основе, поскольку эти типы не генерируют нужных результатов. В состав этих типов входят <Положительные>, <Отрицательные>, <IP>, прочие нелингвистические типы и так далее.

**Максимальное расстояние поиска.** Выберите, на каком удалении должны выполнить поиск методы, прежде чем создать категории. Чем меньше это значение, тем меньше будет результатов поиска; вместе с тем в таких результатах будет меньше шума, и они с большей вероятностью окажутся зависимы друг от друга. Чем выше это значение, тем больше будет результатов; однако такие результаты могут оказаться ненадежны или не соответствовать цели поиска. Хотя эта опция глобально применяется ко всем методам, она наиболее эффективна для сетей совместного появления и семантических сетей.

**Предотвращать объединение отдельных понятий в пары.** Включите этот переключатель, чтобы не объединять два понятия в группу или пару при выводе результатов. Чтобы создавать пары понятий и работать с ними, щелкните по **Работа с парами...**. Дополнительную информацию смотрите в разделе “Управление парами с исключением связи” на стр. 116.

**По возможности:** Выберите, упрощать ли расширение или/и обобщать ли дескрипторы при помощи символов подстановки.

- **Расширение и обобщение.** Эта опция расширяет выбранные категории, а затем обобщает дескрипторы. Если выбрать обобщение, программный продукт создаст в категориях общие правила категорий, применив в качестве символа подстановки звездочку. Например, вместо создания нескольких дескрипторов, таких как [яблочный пирог + .] и [яблочное пюре + .] с помощью символов подстановки может быть сгенерировано: [яблочн* + .]. В случае обобщения при помощи символов подстановки часто будет возвращаться в точности такое же число записей или документов, что и ранее. Однако у этой опции есть преимущество, позволяющее уменьшить число дескрипторов категорий и упростить эти дескрипторы. Дополнительно эта опция повышает способность категоризации большего количества записей или документов благодаря использованию этих категорий для новых текстовых данных (например, в лонгитюдных/волновых исследованиях).
- **Только расширение**. Эта опция расширяет категории без обобщения. Она может оказаться полезной для первого выбора опции **Только расширение** для категорий, созданных вручную, и последующего расширения этих же категорий повторно при помощи опции **Расширение и обобщение**.
- **Только обобщение.** Эта опция обобщает дескрипторы, не расширяя используемые категории никаким другим способом.

  *Примечание:* Выбор этой опции отключает поддержку опции **Семантическая сеть**; это происходит потому, что опция **Семантическая сеть** доступна, только если расширению подлежит описание.

Другие опции для расширения категорий

В дополнение к выбору вариантов применяемых методов можно отредактировать любые из следующих опций:

**Максимальное число элементов для расширения дескриптора по**. При расширении дескриптора с помощью элементов (понятий, типов и прочих выражений) определите максимальное число элементов, которые можно будет добавить в один дескриптор. Если в качестве этого предела задать 10, в существующий дескриптор можно будет добавить не более 10 дополнительных элементов. Если добавлению подлежит более 10 элементов, метод останавливает добавление новых элементов после добавления десятого из них. Это позволяет сократить список дескрипторов, но не гарантирует, что сначала будут использоваться наиболее интересные элементы. Возможно, вы предпочтете сократить размер расширения без заметного снижения

качества при помощи опции **Обобщать по возможности с помощью символов подстановки**. Эта опция применяет только дескрипторы, содержащие логические выражения & (AND) или ! (NOT).

**Расширить также подкатегории.** Эта опция расширяет также и все подкатегории выбранных категорий.

**Расширить пустые категории дескрипторами, сгенерированными из имени категории.** Этот метод применяется только к пустым категориям, у которых 0 дескрипторов. Если категория уже содержит дескрипторы, она не будет расширена этим способом. Эта опция пытается создать дескрипторы для каждой категории автоматически на основе слов, из которых состоит имя категории. Имя категории просматривают, чтобы выяснить, не совпадают ли слова в нем с какими-либо извлеченными понятиями. Если распознается понятие, оно используется для нахождения соответствующих паттернов понятий, и с применением их всех формируются дескрипторы для категории. Эта опция генерирует лучшие результаты, если имена категорий одновременно и длинные, и описательные. Это быстрый метод генерирования дескрипторов категорий, которые, в свою очередь, включают для категории поддержку захвата записей, содержащих эти дескрипторы. Эта опция наиболее полезна, если вы импортируете категории из другого места или создаете категории вручную с длинными описательными именами.

**Сгенерировать дескрипторы как.** Эта опция применяется, только если выбрана предшествующая опция.

- **Понятия.** Выберите эту опцию, чтобы сгенерировать окончательные дескрипторы в форме понятий независимо от того, были ли они извлечены из исходного текста или нет.
- **Паттерны.** Выберите эту опцию, чтобы сгенерировать окончательные дескрипторы в форме паттернов независимо от того, были ли извлечены окончательные дескрипторы или какие-либо паттерны.

## Создание категорий вручную

В дополнение к использованию автоматизированных способов построения категорий и редактора правил создавать категории можно также вручную. Для этого существуют следующие способы:

- Создание пустой категории, в которую по одному будут добавляться новые элементы. Дополнительную информацию смотрите в разделе “Создание новых категорий и переименование категорий”.
- Перетаскивание терминов, типов и паттернов на панель категорий. Дополнительную информацию смотрите в разделе “Создание категорий перетаскиванием” на стр. 126.

## Создание новых категорий и переименование категорий

Вы можете создать пустые категории для добавления в них понятий и типов. Можно также переименовывать категории.

Чтобы создать новую пустую категорию

1. Перейдите на панель Категории.
2. Выберите в меню **Категории > Создать пустую категорию**. Откроется диалоговое окно Свойства категории.
3. В поле Имя введите имя для этой категории.
4. Нажмите кнопку **OK**, чтобы принять имя и закрыть диалоговое окно. Диалоговое окно закроется, и на панели появится имя новой категории.

Теперь можно начать добавлять объекты в эту категорию. Дополнительную информацию смотрите в разделе “Добавление дескрипторов к категориям” на стр. 143.

Чтобы переименовать категорию

1. Выберите категорию и выберите в меню **Категории > Переименовать категорию**. Откроется диалоговое окно Свойства категории.
2. В поле Имя введите новое имя для этой категории.
3. Нажмите кнопку **OK**, чтобы принять имя и закрыть диалоговое окно. Диалоговое окно закроется, и на панели появится имя новой категории.

## Создание категорий перетаскиванием

Ручной способ перетаскивания не основан на алгоритмах. Можно создать категории на панели Категории, перетаскивая следующие объекты:

- Извлеченные понятия, типы или паттерны с панели Результаты извлечения на панель Категории.
- Извлеченные понятия с панели Данные на панель Категории.
- Целые строки с панели Данные на панель Категории. При этом будет создана категория, состоящая из всех извлеченных понятий и паттернов, которые содержатся в данной строке.

*Примечание*: На панели Результаты извлечения поддерживается выбор нескольких элементов для упрощения их совместного перетаскивания.

**Важно!** С панели Данные нельзя перетаскивать понятия, которые не были извлечены из текста. Если вы хотите принудительно извлечь понятие, обнаруженное в данных, необходимо добавить это понятие к типу. Затем снова запустите извлечение. Новые результаты извлечения будут содержать только что добавленное понятие. Затем вы сможете использовать его в своей категории. Дополнительную информацию смотрите в разделе “Добавление понятий к типам” на стр. 97.

**Чтобы создать категории с помощью перетаскивания:**

1. На панели Результаты извлечения или на панели Данные выберите одно или несколько понятий, паттернов, типов, записей или частичных записей.
2. Удерживая нажатой кнопку мыши, перетащите нужный элемент в существующую категорию или в область панели, чтобы создать новую категорию.
3. При достижении области, где вы хотели бы оставить элемент, отпустите кнопку мыши. Элемент будет добавлен на панель Категории. Измененные категории будут отмечены специальным фоновым цветом. Этот цвет называется **category feedback background** (фон обратной связи категорий). Дополнительную информацию смотрите в разделе “Настройка опций” на стр. 82.

*Примечание*: Итоговой категории автоматически будет присвоено имя. При желании это имя можно изменить. Дополнительную информацию смотрите в разделе “Создание новых категорий и переименование категорий” на стр. 125.

Если нужно просмотреть, какие записи присвоены категории, выберите эту категорию на панели Категории. Панель данных будет автоматически обновлена и покажет все записи этой категории.

---

## Использование правил категорий

Категории можно создать разными способами. Один из них - определить правила категорий для выражения идей. Правила категорий - это операторы, автоматически классифицирующие документы или записи в категорию на основании логического выражения, использующего извлеченные понятия, типы и паттерны, а также логические операторы. Например, можно написать выражение, означающее, что следует *включить все записи, содержащие извлеченное понятие* `посольство`, *но не* понятие `аргентина`, *в эту категорию*.

Хотя некоторые правила категорий создаются автоматически при использовании таких способов группировки, как *совместное появление* и *вывод корня понятия* (**Категории > Параметры построения > Дополнительные понятия: лингвистика**), правила категорий можно создать и вручную в редакторе правил, понимая смысл данных в категории и контекст. Каждое правило присоединяется к одной категории, так что каждое соответствие документов или записей правилу приведет к их учету для этой категории.

Правила категорий помогают повысить качество и продуктивность результатов исследования текста и улучшают последующий количественный анализ, обеспечивая более конкретную категоризацию ответов. Ваш опыт и знание бизнеса обеспечат более точное понимание данных и контекста. Это понимание можно усилить, переводя знания в правила категорий, чтобы категоризовать ваши документы или записи еще точнее и эффективнее, комбинируя извлеченные элементы с логическими операторами.

Возможность создания этих правил повышает точность кодирования и продуктивность, позволяя объединить ваши бизнес-знания с технологиями извлечения этого продукта.

*Примечание*: Примеры соответствия правил тексту смотрите в разделе “Примеры правил категорий” на стр. 132

## Синтаксис правил категорий

Хотя некоторые правила категорий создаются автоматически при использовании таких способов группировки, как *совместное появление* и *вывод корня понятия* (**Категории > Параметры построения > Дополнительные понятия: лингвистика**), правила категорий можно создать и вручную в редакторе правил. Каждое правило - это дескриптор одной категории, поэтому каждое соответствие документов или записей правилу приведет к их учету для этой категории.

*Примечание*: Примеры соответствия правил тексту смотрите в разделе “Примеры правил категорий” на стр. 132

При создании или изменении правила оно должно быть открыто в редакторе правил. Для расширения условий совпадения можно добавить понятия, типы или паттерны, а также символы подстановки. При использовании извлеченных понятий, типов и паттернов можно выиграть, находя и все связанные понятия.

**Важно!** Для предотвращения распространенных ошибок мы рекомендуем перетаскивать понятия непосредственно с панели Результаты извлечения, с панелей Анализ текстовых связей и с панели Данные в редактор текстовых связей или, где это возможно, добавлять их через контекстное меню.

После распознавания понятий, типов и паттернов рядом с текстом появится значок.

*Таблица 18. Значки извлечения*

| Значок | Описание |
|---|---|
|  | Извлеченное понятие |
|  | Извлеченный тип |
|  | Извлеченный паттерн |

Операторы и синтаксис связей

В следующей таблице представлены символы, с помощью которых можно определить синтаксис правил. Используйте эти символы с понятиями, типами и паттернами для создания правила.

*Таблица 19. Поддерживаемый синтаксис*

| Символ | Описание |
|---|---|
| `&` | Логическое "и". Например, `a & b` содержит `a` *и* `b`, например:<br>`- вторжение & соединенные штаты`<br>`- 2016 & олимпиада`<br>`- хороший & фрукт` |
| `\|` | Логическое "или" - включающий логический оператор, то есть соответствие достигается при обнаружении любого из элементов. Например, `a \| b` содержит или `a`, *или* `b`, например:<br>`- атака \| франция`<br>`- кондоминиум \| апартаменты` |
| `!()` | Логическое "не". Например, `!(a)` не содержит `a`:<br>`!(хороший & отель)`, `теракт & !(австрия)` или `!(золото) & !(медь)` |

*Таблица 19. Поддерживаемый синтаксис (продолжение)*

| Символ | Описание |
|---|---|
| * | В зависимости от использования символ подстановки может представлять что угодно, от одного символа до целого слова. Дополнительную информацию смотрите в разделе “Использование символов подстановки в правилах категорий” на стр. 130. |
| () | Разделитель выражений. Все выражения в скобках оцениваются первыми. |
| + | Соединитель в паттерне, используемый для образования паттерна с конкретным порядком. При наличии этого символа должны использоваться квадратные скобки. Дополнительную информацию смотрите в разделе “Использование паттернов TLA в правилах категорий”. |
| [] | Разделитель в паттерне, который требуется при поиске сопоставления на основе извлеченного паттерна TLA внутри правила категорий. Содержимое в квадратных скобках относится к паттернам TLA и никогда не будет сопоставлять понятия или типы на основании простого совместного появления. Если вы не извлекли этот паттерн TLA, сопоставление будет невозможно. Дополнительную информацию смотрите в разделе “Использование паттернов TLA в правилах категорий”. Не используйте квадратные скобки, если вы ищете совпадение понятий или типов, а не паттернов.<br>*Примечание*: В старых версиях правила совместного появления и синонимов, генерируемые способами построения категорий, обычно заключались в квадратные скобки. Теперь во всех новых версиях квадратные скобки указывают на наличие паттерна TLA. Вместо них в правилах, создаваемых способом совместного появления и синонимами, используются обычные скобки (акустическая система\|колонки). |

Операторы & и | коммутативны, то есть a & b = b & a и a | b = b | a.

Буквальное использование символа с помощью обратной дробной черты

Если в понятии есть символ, который можно понимать и как символ синтаксиса, перед ним нужно поместить обратную дробную черту, чтобы правило интерпретировалось правильно. Символ обратной дробной черты (\) используется для указания буквального использования символов, которые без него понимались бы как специальные значения. При перетаскивании объектов в редактор символы обратной дробной черты расставляются автоматически.

Обратную дробную черту нужно использовать перед следующими символами, если их нужно рассматривать как обычные символы, а не элементы синтаксиса правил:
& ! | + < > ( ) [ ] *

Например, так как понятие r&d содержит оператор "и" (&), при использовании редактора правил нужно добавить обратную дробную черту, то есть использовать запись r\&d.

## Использование паттернов TLA в правилах категорий

Паттерны анализа текстовых связей (Text Link Fnalysis, TLA) можно явно определить в правилах категорий, что позволит получать более конкретные и связанные с контекстом результаты. При определении паттерна в правиле категорий вы можете обойти более простые результаты извлечения понятий, и в результатах появятся только документы и записи, согласованные на основе паттернов TLA.

**Важно!** Чтобы сопоставлять документы в ваших правилах категорий при помощи паттернов TLA, нужно запускать извлечение при включенной возможности анализа текстовых связей. Правило категорий будет искать сопоставления, найденные в этом процессе. Если вы не выбрали исследование результатов TLA на вкладке Модель узла исследования текста, можно включить извлечение TLA в параметрах извлечения в интерактивном сеансе, а затем повторить извлечение. Дополнительную информацию смотрите в разделе “Извлечение данных” на стр. 88.

**Ограничение квадратными скобками.** Паттерн TLA нужно окружить квадратными скобками [ ], если он используется внутри правила категорий. Ограничитель паттерна требуется, если вы ищете сопоставление на

основе извлеченного паттерна TLA. Так как правила категорий могут содержать типы, понятия и паттерны, использование квадратных скобок показывает правилу, что содержимое внутри них относится к извлеченному паттерну TLA. Если вы не извлекли этот паттерн TLA, сопоставление будет невозможно. Если вы встретите паттерн без квадратных скобок, например, фрукт + хороший, на панели Категории, это скорее всего будет означать, что этот паттерн был добавлен непосредственно в категорию вне редактора правил. Если вы добавите паттерн понятий непосредственно в категорию из представления анализа текстовых связей, он появится без квадратных скобок. Однако при использовании паттерна в правиле необходимо заключить его в квадратные скобки внутри правила категорий, например, [банан + !(хороший)].

**Использования знака + в паттернах.** В IBM SPSS Modeler Text Analytics может присутствовать паттерн из 6 частей, или слот. Если нужно указать на важность порядка следования, используйте для соединения каждого из элементов знак +, например, [компания1 + приобрела + компания2]. Здесь порядок важен, так как он определяет смысл, какая именно компания приобрела другую. Порядок определяется не последовательной структурой предложения, а тем, как структурирован выход паттерна TLA. Например, если у вас есть текст "*Я люблю Париж*" и нужно извлечь это содержание, паттерн TLA более вероятно будет выглядеть как [париж + любить] или [<Положение> + <Положительное>], а не [<Положительное> + <Положение>], так как по умолчанию в общем случае ресурсы мнений обычно размещают мнения на второй позиции в паттернах из двух частей. Поэтому для исключения проблем предпочтительнее использовать паттерн непосредственно как дескриптор в вашей категории. Однако если нужно использовать паттерн как часть более сложного оператора, обратите особенное внимание на порядок элементов в паттернах, представленных анализом текстовых связей, так как порядок играет большую роль для самой возможности найти соответствие.

Допустим, у вас есть следующие два текстовых выражения: "Я *люблю ананас*" и "Я ненавижу *ананас*, но *люблю* клубнику". Выражение люблю & ананас будет соответствовать обоим текстам, так как это выражение понятия, а не правила текстового связывания (не заключено в квадратные скобки). Выражение ананас + люблю соответствует только предложению "Я *люблю ананас*", так как во втором тексте слово *люблю* теперь связано с *клубника*.

**Группировка при помощи паттернов.** Правила можно упростить с помощью ваших собственных паттернов. Пусть вы хотите захватить следующие три выражения: кайенский перец + люблю, жгучий перец + люблю и перец + люблю. Их можно сгруппировать в одно правило категорий [* перец & люблю]. Если у вас есть еще одно выражение, острый перец + хорошо, можно сгруппировать все четыре выражение правилом [* перец + <Положительное>].

**Порядок в паттернах.** Чтобы лучше организовать выход, предоставленные в установленных с продуктом паттернах правила анализа текстовой связи пытаются вывести основные патерны в порядке, не зависящем от порядка слов в предложении. Допустим, например, что у вас есть запись с текстом "*Хорошие презентации.*" и другая запись, содержащая текст "*презентации были хорошими*", оба текста будут сопоставляться по одному правилу и выводиться в результатах паттерна понятий в том же порядке, что и презентация + хороший, а не презентация + хороший и также хороший + презентация. И в двухслотовом паттерне, как в этом примере, понятия, назначенные типам в библиотеке Opinions (Мнения), по умолчанию будут представлены в выходе последними, например, яблоко + плохой.

*Таблица 20. Синтаксис паттернов и использование логических выражений*

| Выражение | Сопоставляет документ или запись, которая |
|---|---|
| [] | Содержит любой паттерн TLA. *В правилах категорий* требуется ограничитель паттерна, если вы ищете сопоставление на основе извлеченного паттерна TLA. Содержимое внутри квадратных скобок относится к паттернам TLA, а не просто к понятиям и типам. Если вы не извлекли этот паттерн TLA, сопоставление будет невозможно.<br><br>Если вы хотели создать правило без паттернов, можно использовать !([]). |
| [a] | Содержит паттерн, в котором по крайней мере один элемент - это a, независимо от его положения в паттерне. Например, [древесина] может соответствовать [древесина + хороший] или просто [древесина + .] |

*Таблица 20. Синтаксис паттернов и использование логических выражений (продолжение)*

| Выражение | Сопоставляет документ или запись, которая |
|---|---|
| [a + b] | Содержит паттерн понятий. Например, [древесина + хороший].<br>*Примечание:* Если вы хотите захватывать только этот паттерн, не добавляя другие элементы, рекомендуется добавлять паттерн непосредственно в категорию, а не создавать правило с ним. |
| [a + b + c] | Содержит паттерн понятий. Знак + указывает на важность порядка элементов сопоставления. Например, [компания1 + приобрела + компания2]. |
| [<A> + <B>] | Содержит любой паттерн с типом <A> в первом слоте и с типом <B> во втором, и существует ровно два слота. Знак + указывает на важность порядка элементов сопоставления. Например, [<Бюджет> + <Отрицательное>].<br>*Примечание:* Если вы хотите захватывать только этот паттерн, не добавляя другие элементы, рекомендуется добавлять паттерн непосредственно в категорию, а не создавать правило с ним. |
| [<A> & <B>] | Содержит паттерн любого типа с типом <A> и с типом <B>. Например, [<Бюджет> & <Отрицательное>]. Этот паттерн TLA никогда не будет извлечен, однако при записи в таком виде он на самом деле эквивалентен [<Бюджет> + <Отрицательное>]\|[<Отрицательное> + <Бюджет>]. Порядок элементов сопоставления не важен. Кроме этого, в паттерне могут быть и другие элементы, но по крайней мере должны быть <Бюджет> и <Отрицательное>. |
| [a + .] | Содержит паттерн, в котором a - это единственное понятие, и в других слотах нет ничего для этого паттерна. Например,<br>[древесина + .] соответствует паттерну понятий, для которого единственный выход - это понятие древесина. Если вы добавили понятие древесина как дескриптор категории, будут получены все записи, где древесина - это понятие, включающее положительные утверждения о древесине. Однако при использовании [древесина + .] будут согласованы результаты паттернов только таких записей, где представлена древесина и никакие другие взаимосвязи или мнения, например, не будет соответствия с древесина + фантастика.<br>*Примечание:* Если вы хотите захватывать только этот паттерн, не добавляя другие элементы, рекомендуется добавлять паттерн непосредственно в категорию, а не создавать правило с ним. |
| [<A> + <>] | Содержит паттерн, в котором <A> - это единственный тип. Например, [<Бюджет> + <>] соответствует паттерну, в котором единственный выход - это понятие типа <Бюджет>.<br>*Примечание:* Можно использовать обозначение <> для пустого типа, только помещая его после знака + паттерна в паттерне типов, как в [<Бюджет> + <>], но нельзя писать [цена + <>].<br>*Примечание:* Если вы хотите захватывать только этот паттерн, не добавляя другие элементы, рекомендуется добавлять паттерн непосредственно в категорию, а не создавать правило с ним. |
| [a + !(b)] | Содержит по крайней мере один паттерн, содержащий понятие a, но не содержащий понятия b. Должна включать по крайней мере один паттерн.<br><br>Например, [цена + !(высокий)]<br><br>или для типов, [!(<Фрукт>\|<Овощ>) + <Положительное>] |
| !([<A> & <B>]) | Не содержит конкретного паттерна. Например, !([<Бюджет> & <Отрицательное>]). |

*Примечание*: Примеры соответствия правил тексту смотрите в разделе “Примеры правил категорий” на стр. 132

## Использование символов подстановки в правилах категорий

К понятиям в правилах можно добавлять символы подстановки для расширения возможностей совпадения. Символ подстановки звездочка * может располагаться перед словом или после него для указания, как можно сопоставлять понятия. Есть два типа использования символов подстановки:

- **Аффиксные символы подстановки**. Эти символы подстановки используются как префиксы или суффиксы без пробелов между звездочкой и строкой. Например, operat* может соответствовать *operat*, *operate*, *operates*, *operations*, *operational* и так далее.

- **Символы подстановки как слова**. Эти символы подстановки располагаются перед понятием или после него с пробелом между понятием и звездочкой. Например, `* operation` может соответствовать *operation*, *surgical operation*, *post operation* и так далее. Кроме этого, символ подстановки как слово может использоваться вместе с аффиксным символом подстановки, как в `* operat* *`, что будет соответствовать *operation*, *surgical operation*, *telephone operator*, *operatic aria* и так далее. Как видно из последнего примера, символы подстановки рекомендуется использовать осторожно, чтобы не слишком широко забрасывать сеть и не захватывать нежелательные соответствия.

**Исключения!**

- Символ подстановки никогда не может стоять сам по себе. Например, форма `(яблоко | * )` неприемлема.
- Символ подстановки нельзя использовать для сопоставления имен типов. `<Отрицательное*>` не будет совпадать вообще ни с каким типом.
- Невозможно отфильтровать определенные типы от сопоставления с понятиями, найденными с помощью символов подстановки. Тип, которому назначено понятие, используется автоматически.
- Символ подстановки не может располагаться в середине последовательности слов, в том числе в начале или в конце слова (открыт* счет) или как отдельный элемент (открыт * счет). Нельзя использовать символы подстановки и в именах типов. Например, `слово* слово`, как в `apple* recipe`, никогда не будет сопоставлено с applesauce recipe или с чем-либо еще. Однако `apple* *` даст совпадение с *applesauce recipe*, *apple pie*, *apple* и так далее. В другом примере `слово * слово`, таком как `apple * toast`, не будет совпадения с *apple cinnamon toast* или с чем-либо еще, так как символ подстановки помещается между двумя другими словами. Однако `apple *` даст совпадение с *apple cinnamon toast*, *apple*, *apple pie* и так далее.

*Таблица 21. Использование символов подстановки*

| Выражение | Сопоставляет документ или запись, которая |
|---|---|
| `*apple` | Содержит понятие, оканчивающееся представленными буквами, но в виде префикса может быть любое количество букв. Например: `*apple` заканчивается на *apple*, но могут быть и префиксы, такие как:<br>- `apple`<br>- `pineapple`<br>- `crabapple` |
| `apple*` | Содержит понятие, начинающееся представленными буквами, но в виде суффикса может быть любое количество букв. Например: `apple*` начинается буками *apple*, но может присутствовать или нет суффикс:<br>- `apple`<br>- `applesauce`<br>- `applejack`<br><br>Например, `apple* & !(pear* \| quince)` содержит понятие, начинающееся с букв `apple`, но не понятие с начальными буквами *pear* и не понятие `quince`: это НЕ совпадает с `apple & quince`,<br><br>но может совпасть с<br>- `applesauce`<br>- `apple & orange` |
| `*product*` | Содержит понятие, внутри которого есть буквы `product`, а также в виде префикса и/или суффикса может быть любое количество букв.<br><br>Например, `*продукт*` может совпадать с:<br>- `продукт`<br>- `субпродукт`<br>- `непродуктивно` |

*Таблица 21. Использование символов подстановки (продолжение)*

| Выражение | Сопоставляет документ или запись, которая |
|---|---|
| `* loan` | Содержит понятие со словом `loan`, но может быть составным объектом с другим словом спереди. Например, `* loan` может соответствовать:<br>- `loan`<br>- `car loan`<br>- `home equity loan`<br><br>Например, `[* доставка + <Отрицательное>]` содержит понятие, оканчивающееся на слово `доставка` в первой позиции и тип `<Негативное>` во второй позиции и может соответствовать следующим паттернам понятий:<br>- `пакетная доставка + медленно`<br>- `суточная доставка + опоздание` |
| `event *` | Содержит понятие со словом `event`, но может быть составным объектом с другим словом сзади. Например, `event *` может соответствовать:<br>- `event`<br>- `event location`<br>- `event planning committee` |
| `* apple *` | Содержит понятие, которое может начинаться с любого слова, далее содержит слово `apple` и, возможно, еще какое-то слово. `*` означает отсутствие или наличие слова, поэтому понятие `apple` тоже подходит. Например, `* apple *` может соответствовать:<br>- `gala applesauce`<br>- `granny smith apple crumble`<br>- `famous apple pie`<br>- `apple`<br><br>Например, конструкция `[* резервирован* * + <положительное>]`, содержащее понятие со словом `резервирован` (в любой части понятия) в первой позиции и тип `<Положительное>` во второй позиции, соответствует паттернам понятий:<br>- `система резервирования + хорошо`<br>- `резервирование онлайн + хорошо` |

*Примечание*: Примеры соответствия правил тексту смотрите в разделе “Примеры правил категорий”

## Примеры правил категорий

Следующий пример демонстрирует, как правила по-разному соответствуют записям в зависимости от используемого для их выражения синтаксиса.

Примеры записей

Допустим, у вас есть две записи:

- **Запись A**: *“when I checked my wallet, I saw I was missing 5 dollars.”* (проверяя кошелек, я увидели, что не хватает 5 долларов)
- **Запись B**: *“$5 was found at the picnic area, but the blanket was missing.”* (5 долларов нашли на месте пикника, но одеяла не было)

В следующих двух таблицах показано, что можно извлечь отсюда для понятий и типов, а также для паттернов понятий и типов.

Извлеченные из примера понятия и типы

*Таблица 22. Пример извлеченных понятий и типов*

| Извлеченное понятие | Определенный тип понятий |
|---|---|
| wallet | <Нет данных > |
| missing | <Отрицательные> |
| USD5 | <Денежная единица> |
| blanket | <Нет данных > |
| picnic area | <Нет данных > |

Извлеченные из примера паттерны анализа связей текста (Text Link Analysis, TLA)

*Таблица 23. Вывод извлеченных из примера паттернов анализа TLA*

| Извлеченные паттерны понятий | Извлеченные паттерны типов | Из записи |
|---|---|---|
| picnic area + . | <Неизвестно> + <> | Запись B |
| wallet + . | <Неизвестно> + <> | Запись A |
| blanket + missing | <Неизвестно> + <Отрицательное> | Запись B |
| USD5 + . | <Валюта> + <> | Запись B |
| USD5 + missing | <Валюта> + <Отрицательное> | Запись A |

Как соотносятся возможные правила категорий

В следующей таблицы представлены примеры синтаксиса, которые можно ввести в редактор правил категорий. Здесь не все правила работают и не все соответствуют той же записи. Посмотрите, как различный синтаксис влияет на сопоставление записей.

*Таблица 24. Примеры правил*

| Синтаксис правила | Результат |
|---|---|
| USD5 & missing | Подходят обе записи A и B, так как обе они содержат извлеченное понятие missing и извлеченное понятие USD5. Это эквивалентно следующему: (USD5 & missing) |
| missing & USD5 | Подходят обе записи A и B, так как обе они содержат извлеченное понятие missing и извлеченное понятие USD5. Это эквивалентно следующему: (missing & USD5) |
| missing & <Валюта> | Подходят обе записи A и B, так как обе они содержат извлеченное понятие missing и понятие, соответствующее типу <Валюта>. Это эквивалентно следующему: (missing & <Валюта>) |
| <Валюта> & missing | Подходят обе записи A и B, так как обе они содержат извлеченное понятие missing и понятие, соответствующее типу <Валюта>. Это эквивалентно следующему: (<Валюта> & missing) |
| [USD5 + missing] | Подходит запись A, но **не** подходит B, так как запись B не создает паттерна выхода TLA, содержащего USD5 + missing (смотрите предыдущую таблицу). Это эквивалентно выходу паттерна TLA: USD5 + missing |
| [missing + USD5] | Не подходит ни запись A, ни запись B, так как ни один извлеченный паттерн TLA (смотрите предыдущую таблицу) не соответствует представленному здесь порядку с понятием missing на первом месте. Это эквивалентно выходу паттерна TLA: USD5 + missing |

*Таблица 24. Примеры правил (продолжение)*

| Синтаксис правила | Результат |
|---|---|
| `[missing & USD5]` | Подходит запись A, но **не** подходит B, так как такой паттерн TLA не был извлечен из записи B. Использование символа `&` означает, что порядок при сопоставлении не важен, поэтому это правило ищет сопоставления или с `[missing + USD5]`, или с `[USD5 + missing]`. Совпадение есть только для формы `[USD5 + missing]` из записи A. |
| `[missing + <Валюта>]` | Не подходит ни запись A, ни запись B, так как ни один извлеченный паттерн TLA не соответствует этому порядку. Эквивалента нет, так как выход TLA основан только на терминах `(USD5 + missing)` или типах `(<Валюта> + <Отрицательное>)`, но не на смешанных понятиях и типах. |
| `[<Валюта> + <Отрицательное>]` | Подходит запись A, но не подходит B, так как такой паттерн TLA не был извлечен из записи B. Это эквивалентно выходу паттерна TLA: `<Валюта> + <Отрицательное>` |
| `[<Отрицательное> + <Валюта>]` | Не подходит ни запись A, ни запись B, так как ни один извлеченный паттерн TLA не соответствует этому порядку. По умолчанию в шаблоне `Мнения`, когда найдено *тема* и *мнение*, *тема* (`<Валюта>`) занимает первую позицию слота, а *мнение* (`<Отрицательное>`) - вторую. |

## Создание правил категории

Создавая или редактируя правило, вы должны открывать его в редакторе правил. Вы можете добавлять понятия, типы и паттерны, а также использовать символы подстановки для расширения соответствий. При использовании распознанных понятий, типов и паттернов выигрыш достигается благодаря нахождению всех связанных понятий. Например, когда вы используете концепцию, то правилу соответствуют также все связанные термины, формы множественного числа и синонимы. Аналогичным образом при использовании типа все его понятия также захватываются правилом.

Открыть редактор правил можно, выбрав редактирование существующего правила или щелчком правой кнопкой по имени категории с выбором **Создать правило**.

Можно пользоваться контекстными меню, перетаскиванием и ручным вводом понятий, типов и паттернов в редакторе. Из них можно составлять выражения правил, используя как связки логические операции `&`, `!()`, `|`. Чтобы избегать обычных ошибок, рекомендуется перетаскивать понятия в редактор правил непосредственно с панели результатов извлечения или с панели данных. Будьте внимательны к синтаксису правил, чтобы избегать ошибок. Дополнительную информацию смотрите в разделе “Синтаксис правил категорий” на стр. 127.

*Примечание*: Примеры того, как правила выполняют сопоставление с текстом, смотрите в разделе “Примеры правил категорий” на стр. 132.

Чтобы создать правило

1. Если вы еще не извлекли данные или ваше извлечение устарело, выполните его сейчас. Дополнительную информацию смотрите в разделе “Извлечение данных” на стр. 88.

   *Примечание*: Если отфильтровать извлечение так, что не останется видимых понятий, при попытке создать или отредактировать правило категории выводится сообщение об ошибке. Чтобы предотвратить его, измените фильтр извлечения, сделав какие-то понятия доступными.
2. На панели категорий выберите категорию, в которой нужно добавить правило.
3. В меню выберите **Категории > Создать правило**. В окне откроется панель редактора правил категории.
4. В поле Имя правила введите имя для вашего правила. Если не ввести имя, в качестве автоматического имени будет использоваться выражение правила. В дальнейшем имя правила можно изменить.
5. В текстовом поле большего размера для ввода выражения можно:

- Непосредственно ввести текст в поле или перетащить его с другой панели. Используйте только извлеченные понятия, типы и паттерны. Например, если введено слово `коты`, а на панели результатов извлечения выведена только форма единственного числа, `кот`, редактор не распознает слово `коты`. В последнем случае форма единственного числа могла бы автоматически включить форму множественного числа; в тех случаях, когда этого не происходит, можно использовать символ подстановки. Дополнительную информацию смотрите в разделе “Синтаксис правил категорий” на стр. 127.
- Выберите понятия, типы или паттерны, которые нужно добавить в правила, и используйте меню.
- Добавьте логические операции как связки между элементами в правиле. Нажимайте кнопки на панели инструментов, чтобы добавить в правило логические операции "и" **&**, "или" **|**, "не" **!()**, а также скобки **()** и прямые скобки для паттернов **[ ]**.

6. Нажмите кнопку **Тестировать правило**, чтобы проверить, что правило правильно сформатировано. Дополнительную информацию смотрите в разделе “Синтаксис правил категорий” на стр. 127. Число найденных документов или записей выводится в круглых скобках рядом с текстом **Результат теста**. Справа от этого текста выводятся распознанные элементы правила и сообщения об ошибках. Если диаграмма рядом с типом, паттерном или понятием выведена с красным восклицательным знаком, это показывает, что элемент не соответствует ни одному известному извлечению. Если таких соответствий нет, правило не найдет ни одной записи.
7. Чтобы протестировать часть правила, выберите эту часть и щелкните по **Тестировать выбранное**.
8. Если найдены ошибки, внесите необходимые изменения и еще раз протестируйте правило.
9. Завершив работу, щелкните по **Сохранить и закрыть**, чтобы еще раз сохранить правило и закрыть редактор. Новое правило появится в папке категории.

## Редактирование и удаление правил

Созданное и сохраненное правило можно в любое время отредактировать. Дополнительную информацию смотрите в разделе “Синтаксис правил категорий” на стр. 127.

Если правило больше не нужно, его можно удалить.

Чтобы отредактировать правила

1. Выберите правило в таблице дескрипторов в диалоговом окне определений категорий.
2. В меню выберите **Категории > Редактировать правило** или щелкните дважды по имени правила. Откроется редактор с выбранным правилом.
3. Внесите изменения в правило, пользуясь результатами извлечения и кнопками на панели инструментов.
4. Снова протестируйте правило, чтобы убедиться, что оно возвращает ожидаемые результаты.
5. Выберите **Сохранить и закрыть**, чтобы снова сохранить правило и закрыть редактор.

Чтобы удалить правило

1. Выберите правило в таблице дескрипторов в диалоговом окне определений категорий.
2. В меню выберите **Изменить > Удалить**. Правило будет удалено из категории.

## Импорт и экспорт предопределенных категорий

Если вы сохранили свои категории в файле Microsoft Excel (*.*xls*, *.*xlsx*), их можно импортировать в IBM SPSS Modeler Text Analytics .

Вы можете также экспортировать категории, содержащиеся в открытом сеанса интерактивной инструментальной среды в файлы Microsoft Excel (*.*xls*, *.*xlsx*). При экспорте категорий можно по своему выбору добавить или исключить дополнительную информацию, например, дескрипторы и оценки. Дополнительную информацию смотрите в разделе “Экспорт категорий” на стр. 140.

Если у ваших предопределенных категорий нет кодов или вам нужны новые коды, можно автоматически сгенерировать новый набор кодов для набора категорий в панели категорий, выбрав в меню **Категории > Управление категориями > Автогенерация кодов**. При этом будут удалены все существующие коды и выполнена автоматическая перенумерация кодов.

## Импорт предопределенных категорий

Вы можете импортировать свои предопределенные категории в IBM SPSS Modeler Text Analytics . Перед импортом убедитесь, что предопределенные категории находятся в файле Microsoft Excel (**.xls*, **.xlsx*) и структурированы в одном из поддерживаемых форматов. Можно также выбрать автоматическое определение формата продуктом. Поддерживаются следующие файлы:

- **Формат плоского списка**: Дополнительную информацию смотрите в разделе “Формат плоского списка” на стр. 137.
- **Компактный формат**: Дополнительную информацию смотрите в разделе “Компактный формат” на стр. 138.
- **Формат с отступами**: Дополнительную информацию смотрите в разделе “Формат с отступами” на стр. 138.

Для импорта предопределенных категорий

1. В меню интерактивной инструментальной среды выберите **Категории > Управление категориями > Импортировать предопределенные категории**. Откроется мастер по импорту предопределенных категорий.
2. В выпадающем списке Папка выберите диск, на котором находится файл.
3. Выберите файл из списка. Имя файла появляется в текстовом поле Имя файла.
4. Выберите рабочий лист, содержащий предопределенные категории из списка. Имя рабочего листа появляется в поле Рабочий лист.
5. Чтобы начать выбирать формат данных, щелкните по **Далее**.
6. Выберите формат для файла или выберите опцию, позволяющую продукту определять формат автоматически. Автообнаружение работает лучше всего с наиболее распространенными форматами.
   - **Формат плоского списка**: Дополнительную информацию смотрите в разделе “Формат плоского списка” на стр. 137.
   - **Компактный формат**: Дополнительную информацию смотрите в разделе “Компактный формат” на стр. 138.
   - **Формат с отступами**: Дополнительную информацию смотрите в разделе “Формат с отступами” на стр. 138.
7. Чтобы задать дополнительные опции импорта, нажмите кнопку **Далее**. Если было выбрано автоматическое определение формата, вы будете перенаправлены к заключительному шагу.
8. Если одна или несколько строк содержат заголовки столбцов или другую постороннюю информацию, выберите номер строки, с которой хотите начать импорт, при помощи опции **Начать импорт со строки**. Например, если имена ваших категорий начинаются со строки 7, надо ввести для этой опции номер 7 для правильного импорта файла.
9. Если файл содержит коды категорий, выберите опцию **Содержит коды категорий**. Это поможет мастеру правильно распознать ваши данные.
10. Сравните цветные ячейки с пояснениями, чтобы убедиться в правильности распознавания данных. Ошибки, обнаруженные в файле, будут выделены красным цветом и указаны под таблицей предпросмотра формата. Если выбран неверный формат, вернитесь назад и выберите другой формат. Если в файл нужно внести исправления, сделайте это и перезапустите мастер, выбрав этот файл повторно. Для нормального завершения работы мастера необходимо исправить все ошибки.
11. Чтобы просмотреть набор импортируемых категорий и подкатегорий и задать, как будут создаваться дескрипторы для этих категорий, нажмите кнопку **Далее**.

12. Просмотрите в таблице набор категорий, которые будут импортироваться. Если вы не видите ключевых слов, которые ожидали увидеть в качестве дескрипторов, возможно, что они не были распознаны во время импорта. Убедитесь, что они должным образом снабжены префиксами и выводятся в нужной ячейке.
13. Выберите, как следует обрабатывать уже существующие категории в сеансе вашего .
    - **Заменить все существующие категории**. Эта опция очищает все существующие категории, вместо которых далее используются только импортированные категории.
    - **Добавить к существующим категориям**. Эта опция импортирует категории и объединяет все импортированные категории с уже существующими одноименными. При добавлении категорий к существующим категориям надо определить, как будут обрабатываться дубликаты. Одна возможность (опция **Слияние**) - объединять импортируемые категории с существующими, если имена этих категорий совпадают. Другая возможность (опция **Исключить из импорта**) - запретить импорт категории, если категория с такими именем уже существует.
14. **Импортировать ключевые слова в качестве дескрипторов** - опция импорта ключевых слов, найденных в ваших данных, в качестве дескрипторов для связанной категории.
15. **Расширить категории путем создания дескрипторов** - эта опция генерирует дескрипторы из слов, представляющих имя категории или подкатегории, а также слов, составляющих аннотацию. Если эти слова совпадают с извлеченными результатами, они добавляются в качестве дескрипторов в категорию. Эта опция дает наилучшие результаты, когда имена или аннотации категорий достаточно длинны и информативны. Это быстрый метод создания дескрипторов для категории, благодаря которым категория может захватывать записи, содержащие эти дескрипторы.
    - Поле **Из** позволяет выбрать текст, на основе которого будут создаваться дескрипторы, и указать имена категорий и подкатегорий, слова из аннотаций или оба эти источника.
    - Поле **Как** позволяет выбрать, создавать ли эти дескрипторы в форме понятий или паттернов TLA. Если извлечение TLA не выполнялось, опции **паттернов** в этом мастере будут отключены.
16. Для импорта предопределенных категорий в панель Категории нажмите кнопку **Готово**.

## Формат плоского списка

В формате плоского списка есть только один верхний уровень категорий без всякой иерархии, то есть не содержащий подкатегорий или подсетей. Имена категорий находятся в одном столбце.

Следующая информация может содержаться в файле этого формата:

- Необязательный столбец **коды** содержит числовые значения, которые уникальным образом определяют каждую категорию. Если указать, что файл данных содержит коды (опция **Содержит коды категорий** на шаге **Параметры содержимого**), в ячейке непосредственно слева от названия категории должен быть столбец с уникальными кодами для каждой категории. Если ваши данные не содержат кодов, но вы хотели бы создать коды позже, это можно будет сделать в любое время (**Категории > Управление категориями > Автогенерация кодов**).
- *Обязательный* столбец **имена категорий** содержит все имена категорий. Этот столбец обязателен для импорта в данном формате.
- Необязательные **аннотации** в ячейке непосредственно справа от имени категории. Эти аннотации представляют собой текст, описывающий ваши категории/подкатегории.
- Необязательные **ключевые слова** можно импортировать в качестве дескрипторов для категорий. Для их распознавания эти ключевые слова должны находиться в ячейке непосредственно под именем связанной с ними категории/подкатегории, а перед списком ключевых слов должен стоять символ подчеркивания (_), например, `_огнестрельное оружие, оружие / стрелковое оружие`. Ячейка ключевого слова может содержать одно или несколько слов для описания каждой категории. Эти слова будут импортированы как дескрипторы или проигнорированы в зависимости от опции, выбранной для последнего шага мастера. Позже эти дескрипторы будут сопоставляться с извлеченными из текста результатами. При обнаружении совпадения соответствующая запись или документ будут присвоены категории, содержащей данный дескриптор.

*Таблица 25. Формат плоского списка с кодами, ключевыми словами и аннотациями*

| **Столбец A** | **Столбец B** | **Столбец C** |
|---|---|---|
| Код категории (*необязательный*) | Имя категории | Аннотация |
| | _Список дескрипторов/ключевых слов (*необязательно*) | |

## Компактный формат

Компактный формат структурирован подобно формату плоского списка, но используется с иерархическими категориями. Поэтому для него обязателен столбец уровня кода, позволяющий задать иерархический уровень каждой категории и подкатегории.

Следующая информация может содержаться в файле этого формата:

- *Обязательный* столбец **уровня кода** содержит числа, указывающие иерархическое положение последующей информации в этой строке. Например, если заданы значения 1, 2 и 3, и у вас есть и категории, и подкатегории, тогда 1 будет соответствовать категориям, 2 - подкатегориям, а 3 - подподкатегориям. Если у вас есть только категории и подкатегории первого уровня, 1 будет соответствовать категориям, а 2 - подкатегориям. И так далее, до желаемой глубины вложения категорий.
- Необязательный столбец **коды** содержит значения, которые уникальным образом определяют каждую категорию. Если указать, что файл данных содержит коды (опция **Содержит коды категорий** на шаге **Параметры содержимого**), в ячейке непосредственно слева от названия категории должен быть столбец с уникальными кодами для каждой категории. Если ваши данные не содержат кодов, но вы хотели бы создать коды позже, это можно будет сделать в любое время (**Категории > Управление категориями > Автогенерация кодов**).
- *Обязательный* столбец **имена категорий** содержит все имена категорий и подкатегорий. Этот столбец обязателен для импорта в данном формате.
- Необязательные **аннотации** в ячейке непосредственно справа от имени категории. Эти аннотации представляют собой текст, описывающий ваши категории/подкатегории.
- Необязательные **ключевые слова** можно импортировать в качестве дескрипторов для категорий. Для их распознавания эти ключевые слова должны находиться в ячейке непосредственно под именем связанной с ними категории/подкатегории, а перед списком ключевых слов должен стоять символ подчеркивания (_), например, `_огнестрельное оружие, оружие / стрелковое оружие`. Ячейка ключевого слова может содержать одно или несколько слов для описания каждой категории. Эти слова будут импортированы как дескрипторы или проигнорированы в зависимости от опции, выбранной для последнего шага мастера. Позже эти дескрипторы будут сопоставляться с извлеченными из текста результатами. При обнаружении совпадения соответствующая запись или документ будут присвоены категории, содержащей данный дескриптор.

*Таблица 26. Пример компактного формата с кодами*

| **Столбец A** | **Столбец B** | **Столбец C** |
|---|---|---|
| Иерархический уровень кода | Код категории (*необязательный*) | Имя категории |
| Иерархический уровень кода | Код подкатегории (*необязательный*) | Имя подкатегории |

*Таблица 27. Пример компактного формата без кодов*

| **Столбец A** | **Столбец B** |
|---|---|
| Иерархический уровень кода | Имя категории |
| Иерархический уровень кода | Имя подкатегории |

## Формат с отступами

В формате файла с отступами содержимое является иерархическим, что означает, что файл содержит категории и один или несколько уровней подкатегорий. Кроме того, его структура содержит отступы для

обозначения этой иерархии. Каждая строка в файле содержит категорию или подкатегорию, но подкатегории отделены отступами от категорий, подподкатегории отделены отступами от категорий и так далее. Можно создать эту структуру вручную в Microsoft Excel или воспользоваться структурой, проэкспортированной и другого продукта и сохраненной в формате Microsoft Excel.

- **Коды и имена категорий верхнего уровня** занимают, соответственно, столбцы A и B. Если коды отсутствуют, то имена категорий занимают столбец A.
- **Коды и имена подкатегорий** занимают, соответственно, столбцы B и C. Если коды отсутствуют, то имена подкатегорий занимают столбец B. Подкатегория является элементом категории. При отсутствии категорий верхнего уровня не может быть подкатегорий.

*Таблица 28. Структура с отступами с кодами*

| **Столбец A** | **Столбец B** | **Столбец C** | **Столбец D** |
|---|---|---|---|
| Код категории (*необязательный*) | Имя категории | | |
| | Код подкатегории (*необязательный*) | Имя подкатегории | |
| | | Код подподкатегории (*необязательный*) | Имя подподкатегории |

*Таблица 29. Структура с отступами без кодов*

| **Столбец A** | **Столбец B** | **Столбец C** |
|---|---|---|
| Имя категории | | |
| | Имя подкатегории | |
| | | Имя подподкатегории |

Следующая информация может содержаться в файле этого формата:

- Необязательные **коды** должны быть значениями, которые уникальным образом определяют каждую категорию или подкатегорию. Если указать, что файл данных содержит коды (опция **Содержит коды категорий** на шаге **Параметры содержимого**), в ячейке непосредственно слева от имени категории/подкатегории должен быть уникальный код для каждой категории или подкатегории. Если ваши данные не содержат кодов, но вы хотели бы создать коды позже, это можно будет сделать в любое время (**Категории > Управление категориями > Автогенерация кодов**).
- *Обязательное* **имя** для каждой категории и подкатегории. Подкатегории должны быть отделены отступами от категорий на одну ячейку справа в отдельной строке.
- Необязательные **аннотации** в ячейке непосредственно справа от имени категории. Эти аннотации представляют собой текст, описывающий ваши категории/подкатегории.
- Необязательные **ключевые слова** можно импортировать в качестве дескрипторов для категорий. Для их распознавания эти ключевые слова должны находиться в ячейке непосредственно под именем связанной с ними категории/подкатегории, а перед списком ключевых слов должен стоять символ подчеркивания (_), например, `_огнестрельное оружие, оружие / стрелковое оружие`. Ячейка ключевого слова может содержать одно или несколько слов для описания каждой категории. Эти слова будут импортированы как дескрипторы или проигнорированы в зависимости от опции, выбранной для последнего шага мастера. Позже эти дескрипторы будут сопоставляться с извлеченными из текста результатами. При обнаружении совпадения соответствующая запись или документ будут присвоены категории, содержащей данный дескриптор.

**Важно!** Если на одном из уровней используется код, необходимо добавить код для каждой категории и подкатегории. В противном случае процесс импорта завершится неудачно.

## Экспорт категорий

Вы можете также экспортировать категории, содержащиеся в открытом сеанса интерактивной инструментальной среды в файлы формата Microsoft Excel (*.*xls*, *.*xlsx*). Данные, которые будут экспортироваться, поступают в основном из текущего содержимого панели Категории или из свойств категорий. Поэтому если вы планируете экспортировать также значение оценки для **документов**, рекомендуется выполнить оценку повторно.

*Таблица 30. Опции экспорта категорий*

| **Всегда экспортируется...** | **Экспортируется необязательно...** |
|---|---|
| • Коды категории, если есть<br>• Имена категории (и подкатегории)<br>• Уровни кода, если есть (*Плоский*/*Компактный* формат)<br>• Заголовки столбцов (*Плоский*/*Компактный* формат) | • **Документы.** Значения<br>• Аннотации к категориям<br>• Имена дескрипторов<br>• Количество дескрипторов |

**Важно!** При экспорте дескрипторов они преобразуются в текстовые строки и получают префикс - символ подчеркивания. Если вы повторно выполняете импорт в этот продукт, возможность отличать дескрипторы - паттерны от дескрипторов - правил категорий и дескрипторов - простых понятий будет утрачена. Если вы намереваетесь снова использовать эти категории в данном продукте, мы настоятельно рекомендуем вместо этого создать файл пакета анализа текста (text analysis package, TAP), поскольку формат TAP сохранит все дескрипторы в их текущем виде, как и все ваши категории, коды, а также используемые лингвистические ресурсы. Файлы TAP можно использовать в обоих продуктах, IBM SPSS Modeler Text Analytics и IBM SPSS Text Analytics for Surveys . Дополнительную информацию смотрите в разделе “Использование пакетов анализа текста (Text Analysis Package)”.

Для экспорта предопределенных категорий

1. В меню интерактивной инструментальной среды выберите **Категории > Управление категориями > Экспортировать категории**. Откроется мастер по экспорту категорий.
2. Выберите положение и введите имя экспортируемого файла.
3. В текстовом поле Имя файла введите имя для выходного файла.
4. Для выбора формата экспорта ваших данных категорий нажмите кнопку **Далее**.
5. Выберите один из следующих форматов:
   - **Формат плоского или компактного списка**: Дополнительную информацию смотрите в разделе “Формат плоского списка” на стр. 137. Плоский список не содержит подкатегорий. Дополнительную информацию смотрите в разделе “Компактный формат” на стр. 138. Формат компактного списка содержит иерархические категории.
   - **Формат с отступами**: Дополнительную информацию смотрите в разделе “Формат с отступами” на стр. 138.
6. Чтобы приступить к выбору экспортируемого содержимого и просмотру предлагаемых данных, нажмите кнопку **Далее**.
7. Просмотрите содержимое экспортируемого файла.
8. Выберите или отмените выбор дополнительных опций экспортируемого содержимого, например, **Аннотаций** или **Имен дескрипторов**.
9. Чтобы экспортировать категории, нажмите кнопку **Готово**.

# Использование пакетов анализа текста (Text Analysis Package)

Пакет анализа текста (text analysis package, TAP) служит шаблоном для категоризации текстовых ответов. Использование TAP - это удобный способ категоризовать текстовые данные при минимальном вмешательстве, поскольку содержит предварительно построенные наборы категорий и лингвистические ресурсы, чтобы закодировать большое число записей быстро и автоматически. Для извлечения ключевых понятий текстовые данные анализируются и исследуются при помощи лингвистических ресурсов. На основе

найденных в тексте ключевых понятий и паттернов можно разбить записи по набору категорий, выбранному в TAP. Можно создать свой пакет TAP или изменить существующий пакет.

TAP состоит из следующих элементов:

- **Наборы категорий**. Набор категорий по существу составлен из предварительно заданных категорий, кодов категорий, дескрпипторов для каждой категории и, наконец, имени для всего набора категорий. Дескрипторы - это лингвистические элементы (понятия, типы, паттерны и правила), например, термин *дешево* или паттерн *хорошая цена*. Дескрипторы служат для определения категории, чтобы помещать в нее документ или запись, когда текст соответствует любому из дескрипторов категории.
- **Лингвистические ресурсы**. Лингвистические ресурсы - это набор библиотек и расширенных ресурсов, которые настраиваются для извлечения ключевых понятий и паттернов. Извлеченные понятия и паттерны, в свою очередь, служат дескрипторами, по которым записи можно поместить в категорию из набора категорий.

Можно создать свой пакет TAP, изменить существующий или загрузить пакеты анализа текста.

Выбрав TAP и набор категорий , IBM SPSS Modeler Text Analytics может извлекать и категоризовать ваши записи.

*Примечание*: пакеты TAP можно создавать и использовать взаимозаменяемым образом для IBM SPSS Text Analytics for Surveys и IBM SPSS Modeler Text Analytics.

## Создание пакетов анализа текста

Всякий раз, когда у вас есть сеанс по крайней мере с одной категорией и несколькими ресурсами, можно создать пакет анализа текста (text analysis package, TAP) из содержимого открытого сеанса интерактивной инструментальной среды. Набор категорий и дескрипторов (понятий, типов, правил или выходов паттернов TLA) можно добавить в TAP вместе со всеми лингвистическими ресурсами, открытыми в редакторе ресурсов.

Вы сможете увидеть, для какого языка были созданы эти ресурсы. Язык задается на вкладке Расширенные ресурсы Редактор шаблонов или Редактор ресурсов.

Чтобы создать пакет анализа текста

1. Выберите пункт меню **Файл > Пакеты анализа текста > Создать пакет**. Появится диалоговое окно Создать пакет.
2. Перейдите в каталог, где вы хотите сохранить TAP. По умолчанию пакеты TAP сохраняются в подкаталог `\TAP` каталога установки продукта.
3. Введите имя TAP в поле **Имя файла**.
4. Введите метку в поле **Метка пакета**. При вводе имени файла это имя автоматически появляется как метка, но ее можно изменить.
5. Чтобы исключить набор категорий из TAP, выключите переключатель **Включить в состав**. При этом набор категорий не будет добавлен в пакет. По умолчанию в TAP включается по одному набору категорий на вопрос. В TAP всегда должен быть хотя бы один набор категорий.
6. Переименуйте любые наборы категорий. В столбце **Новый набор категорий** представлены собственные имена по умолчанию, которые генерируются добавлением префикса `Cat_` к имени текстовой переменной. После щелчка по ячейке имя становится доступным для изменения. Щелчок мышью в любом другом месте или нажатие клавиши Enter применяет переименование. При переименовании набора категорий его имя изменяется только в TAP, а имя переменной в открытом сеансе не изменяется.
7. При необходимости измените порядок наборов категорий, используя стрелки справа от таблицы наборов категорий.
8. Нажмите кнопку **Сохранить**, чтобы создать пакет анализа текста. Диалоговое окно закроется.

## Загрузка пакетов анализа текста

При конфигурировании узла моделирования исследования текста необходимо указать ресурсы, которые будут использоваться при извлечении. Вместо шаблона ресурсов можно выбрать пакет анализа текста (text analysis package, TAP), чтобы скопировать в узел не только его ресурсы, но и набор категорий.

TAP наиболее привлекательны при интерактивном создании моделей категорий, так как набор категорий можно использовать как стартовую точку для категоризации. При выполнении потока запускается сеанс интерактивной инструментальной среды, и этот набор категорий появляется на панели Категории. При этом вы сразу оцениваете документы и записи с использованием этих категорий и продолжаете уточнять, строить и расширять их, пока не будут удовлетворены существующие требования. Дополнительную информацию смотрите в разделе “Методы и стратегии для создания категорий” на стр. 104.

Начиная с версии 14, щелкнув по **Загрузка** и выбрав TAP, можно увидеть также, на каком языке были определены ресурсы в этом пакете TAP.

Чтобы загрузить пакет анализа текста

1. Перейдите к узлу моделирования исследования текста.
2. На вкладке Модели выберите *Пакет анализа текста* в разделе **Копировать ресурсы из**.
3. Нажмите кнопку **Загрузить**. Откроется диалоговое окно Загрузить пакет анализа текста.
4. Перейдите в положение пакета TAP, содержащего ресурсы и набор категорий, которые вы хотите скопировать в узел. По умолчанию пакеты TAP сохраняются в подкаталог `\TAP` каталога установки продукта.
5. Введите имя TAP в поле **Имя файла**. Метка будет показана автоматически.
6. Выберите категорию, которую вы хотите использовать. Это набор категорий, который появится в сеансе интерактивной рабочей среды. Эти категории затем можно настроить и улучшить вручную или с использованием опций Построить категории или Расширить категории.
7. Щелкните по **Загрузить**, чтобы скопировать содержимое пакета анализа текста в узел. Закроется диалоговое окно. После загрузки TAP копия TAP появится в узле, поэтому все изменения ресурсов и категорий не будут отображаться в TAP, пока вы в явном виде не измените и повторно не загрузите его.

## Изменение пакетов анализа текста

При усовершенствовании набора категорий и лингвистических ресурсов или при создании полностью нового набора категорий можно изменить пакет анализа текста (text analysis package, TAP), чтобы упростить повторное использование этих усовершенствований позднее. Для этого вы должны работать в открытом сеансе, содержащем информацию, которую нужно поместить в TAP. При внесении изменений можно выбрать присоединение наборов категорий, замену ресурсов, изменение метки пакета, переименование или изменение порядка наборов категорий.

Чтобы изменить пакет анализа текста

1. Выберите пункт меню **Файл > Пакеты анализа текста > Изменить пакет**. Появится диалоговое окно Изменить пакет.
2. Перейдите к каталогу, содержащему пакет анализа текста, который вы хотите изменить.
3. Введите имя TAP в поле **Имя файла**.
4. Чтобы заменить лингвистические ресурсы в TAP на ресурсы из текущего сеанса, выберите опцию **Заменить ресурсы из этого пакета на ресурсы открытого сеанса**. Обычно есть смысл менять лингвистические ресурсы, если они использовались для извлечения ключевых понятий и паттернов, использованных для создания определений категорий. Наличие самых последних лингвистических ресурсов позволяет получить наилучшие результаты для категоризации ваших записей. Если не выбрать эту опцию, уже находящиеся в пакете лингвистические ресурсы останутся неизменными.
5. Чтобы изменить только лингвистические ресурсы, убедитесь, что выбрана опция **Заменить ресурсы из этого пакета на ресурсы открытого сеанса** и выбраны только текущие наборы категорий, которые уже были в TAP.

6. Чтобы включить новый набор из открытого сеанса в TAP, включите переключатель для каждого набора категорий, который нужно добавить. Можно добавить один или несколько наборов категорий или не добавлять их вовсе.
7. Чтобы удалить наборы категорий из TAP, выключите соответствующий переключатель **Включить в состав**. Можно выбрать удаление набора категорий, уже входящего в TAP, так как вы добавляете улучшенный набор. Для этого выключите переключатель **Включить в состав** у соответствующего набора категорий в столбце Текущий набор категорий. В TAP всегда должен быть хотя бы один набор категорий.
8. При необходимости переименуйте наборы категорий. После щелчка по ячейке имя становится доступным для изменения. Щелчок мышью в любом другом месте или нажатие клавиши Enter применяет переименование. При переименовании набора категорий его имя изменяется только в TAP, а имя переменной в открытом сеансе не изменяется. Если у двух наборов категорий имена совпадают, они будут выделяться красным цветом, пока вы не исправите дублирование.
9. Чтобы создать новый пакет, для которого содержимое сеанса будет объединено с содержимым выбранного TAP, нажмите кнопку **Сохранить как новый**. Откроется диалоговое окно Сохранить как пакет анализа текста. Следуйте инструкциям ниже.
10. Нажмите кнопку **Изменить**, чтобы сохранить изменения, внесенные в выбранный TAP.

Чтобы сохранить пакет анализа текста

1. Перейдите в каталог, где вы хотите сохранить файл TAP. По умолчанию файлы TAP сохраняются в подкаталог TAP каталога установки.
2. Введите имя TAP в поле Имя файла.
3. Введите метку в поле Метка пакета. При вводе имени файла это имя автоматически используется как метка. Однако можно переименовать эту метку. Метка должна присутствовать обязательно.
4. Нажмите кнопку **Сохранить**, чтобы создать новый пакет.

## Изменение и уточнение категорий

После создания нескольких категорий вам обязательно понадобится проверить и уточнить их. Кроме уточнения лингвистических ресурсов, вы должны проверить категории, изучив возможности их объединения или очистки их определений, а также проверив некоторые категоризованные документы или записи. Вы можете пересмотреть также документы или записи в категории и внести необходимые поправки, чтобы эти категории были определены с учетом возможных нюансов и отличий.

Для создания своих категорий вы можете использовать встроенные автоматизированные способы построения категорий; однако вам может потребоваться некоторая окончательная настройка этих категорий. После использования одного или нескольких способов в окне может появиться несколько категорий. Можно просмотреть данные в категориях и выполнить некоторые корректировки, чтобы их определения оказались для вас удобными. Дополнительную информацию смотрите в разделе “О категориях” на стр. 109.

Здесь представлены некоторые опции уточнения ваших категорий, большинство из которых описано далее:

## Добавление дескрипторов к категориям

После использования автоматических способов у вас могут все еще остаться результаты извлечения, не использованные ни в каких из определений категорий. Необходимо пересмотреть этот список на панели Результаты извлечения. Если вы обнаружите элементы, которые хотелось бы переместить в категорию, можно добавить их в существующую или новую категорию.

Чтобы добавить понятие или тип в категорию

1. На панелях Результаты извлечения и Данные выберите элементы, которые вы хотите добавить в новую или существующую категорию.

2. В меню выберите **Категории > Добавить к категории**. В диалоговом окне Все категории появится набор категорий. Выберите категорию, в которую вы хотите добавить выбранные элементы. Если вы хотите добавить эти элементы в новую категорию, выберите опцию **Новая категория**. На панели Категории появится новая категория с именем первого выбранного элемента.

## Изменение дескрипторов категорий

После создания нескольких категорий можно открыть каждую из них и просмотреть все дескрипторы, составляющие ее описание. В диалоговом окне Определения категорий можно внести несколько изменений в ваши дескрипторы категорий. Кроме того, если категории показаны деревом категорий, можно работать с ними и в этом представлении.

Чтобы изменить категорию

1. На панели Категории выберите категорию, которую вы хотите изменить.
2. В меню выберите **Просмотр > Определения категорий**. Откроется диалоговое окно Определения категорий.
3. Выберите дескриптор, который вы хотите изменить, и нажмите соответствующую кнопку на панели инструментов.

В следующей таблице описаны все кнопки панели инструментов, которые можно использовать для изменения определений категорий.

*Таблица 31. Кнопки и описания в панели инструментов.*

| Значки | Описание |
|---|---|
|  | Удаляет выбранные дескрипторы из категории. |
|  | Перемещает выбранные дескрипторы в новую или существующую категорию. |
|  | Перемещает выбранные дескрипторы из формы правила категорий с операторами & в категорию. Дополнительную информацию смотрите в разделе “Использование правил категорий” на стр. 126. |
|  | Перемещает каждый из выбранных дескрипторов как свою собственную новую категорию |
|  **Вывод** | Изменяет контент, выводимый на панели Данные и на панели Визуализация в соответствии с выбранными дескрипторами |

## Перемещение категорий

При необходимости вы можете поместить категорию в другую существующую категорию или переместить дескрипторы в другую категорию.

Чтобы переместить категорию

1. На панели Категории выберите категории, которые вы хотите переместить в другую категорию.
2. В меню выберите **Категории > Переместить в категорию**. В меню представлен набор категорий, причем в верхней части списка находится последняя созданная категория. Выберите имя категории, в которую вы хотите переместить выбранные понятия.

- Если вы нашли нужное имя категории, выберите его, и все выбранные элементы будут добавлены в эту категорию.
- Если нужное имя не показано, выберите опцию **Еще** для вывода диалогового окна Все категории, после чего выберите нужную категорию из списка.

## Сведение категорий

Если у вас есть иерархическая структура категорий с категориями и подкатегориями, ее можно укрупнить сведением. При сведении категории все дескрипторы из ее подкатегорий переносятся в выбранную категорию, а все пустые после этого подкатегории удаляются. При этом все документы, использованные для согласования подкатегорий, будут относиться к выбранной категории.

Чтобы выполнить сведение для категорию

1. На панели Категории выберите категорию (верхнего уровня или подкатегорию), для которой вы хотите выполнить сведение.
2. В меню выберите **Категории > Выполнить сведение категорий**. Подкатегории будут удалены, а их дескрипторы перейдут к выбранной категории.

## Слияние или объединение категорий

Если вы хотите объединить две или более существующих категорий в новую категорию, их можно объединить слиянием. При слиянии категорий создается новая категория с общим именем. Все понятия, типы и паттерны, использованные в дескрипторах категорий, будут перенесены в эту новую категорию. Позже вы сможете переименовать эту категорию, изменив ее свойства.

Чтобы слить категорию или часть категории

1. На панели Категории выберите элементы, которые вы хотели бы слить совместно.
2. В меню выберите **Категории > Слить категории**. Откроется диалоговое окно Свойства категорий, в котором можно ввести имя для вновь создаваемой категории. Выбранные категории сливаются в новую категорию как подкатегории.

## Удаление категорий

Если категория больше не требуется, ее можно удалить.

Чтобы удалить категорию

1. На панели Категории выберите категорию или категории, которые вы хотите удалить.
2. В меню выберите **Изменить > Удалить**.

# Глава 11. Анализ кластеров

В представлении Кластеры (**Вид > Кластеры**) можно построить и исследовать кластеры понятий. **Кластер** - это группировка связанных понятий, генерируемая алгоритмами кластеризации на основе того, как часто эти понятия встречаются в наборе документов/записей и ка часто они встречаются совместно друг с другом в одном и том же документе (другое название - **встречаемость**). Каждое понятие в кластере встречается совместно хотя бы еще с одним понятием в кластере. Кластеры предназначены для того, чтобы сгруппировать те понятия, которые встречаются совместно, в то время как цель категорий - сгруппировать документы или записи с учетом того, как содержащийся в них текст соответствует дескрипторам (понятиям, правилам, паттернам) для каждого понятия.

Хороший кластер - это кластер с понятиями, прочно связанными и встречающимися часто, и с несколькими связями с понятиями в других кластерах. При работе с большими базами данных этот метод может привести к значительному увеличению времени обработки.

*Примечание*: Чтобы выполнить построение с применением только поднабора всех документов или записей, используйте опцию **Максимальное число документов, используемых для вычисления кластеров**.

Кластеризация - это процесс, начинающийся анализом набора понятий и поиском часто встречающихся совместно понятий в документах. Два понятия, которые встречаются в документе совместно, считаются парой понятий. Далее процесс кластеризации оценивает **значение подобия** каждой пары понятий путем сравнения числа документов, в которых эта пара встречается совместно, с числом документов, в которых встречается каждое из этих понятий. Дополнительную информацию смотрите в разделе “Вычисление значений связей подобия” на стр. 150.

И наконец, процесс кластеризации группирует схожие понятия в кластеры посредством агрегирования и учитывает их значения связей и параметры, задаваемые в диалоговом окне Построить кластеры. Под агрегацией мы подразумеваем добавление понятий или слияние более мелких кластеров в более крупные кластеры, пока кластер не становится насыщенным. Кластер становится **насыщенным**, когда дополнительное слияние понятий или более мелких кластеров приводит к превышению этим кластером заданных значений параметров в диалоговом окне Построить кластеры (числа понятий, внутренних связей или внешних связей). Кластер принимает имя понятия в кластере, у которого самое большое общее число связей с другими понятиями в этом кластере.

В конечном счете не все пары понятий оказываются в одном и том же кластере вместе, поскольку может существовать более прочная связь в другом кластере либо насыщенность может воспрепятствовать слиянию кластеров, в которых они встречаются. По этой причине существуют и внутренние, и внешние связи.

- **Внутренние связи** - это связи между парами понятий в кластере. Не все понятия в кластере связаны друг с другом. Однако каждое понятие связано хотя бы с одним другим понятием в кластере.
- **Внешние связи** - это связи между парами понятий в раздельных кластерах (понятием в одном кластере и внешним понятием, находящимся в другом кластере).

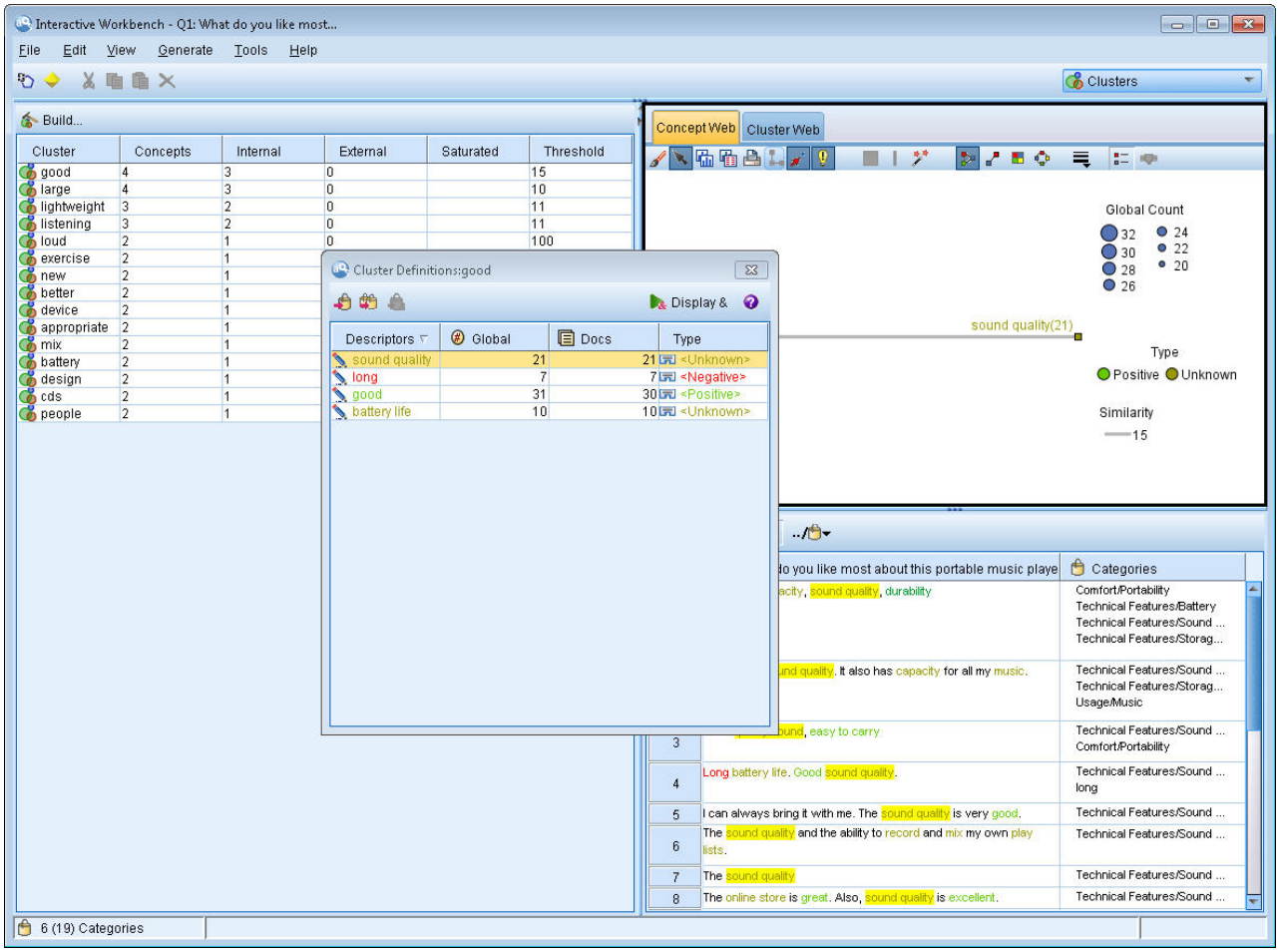



*Рисунок 30. Вид представления Кластеры*

Представление Кластеры организовано в трёх панелях, каждую из которых можно скрыть или показать, выбрав ее имя в меню Вид:

- **Панель Кластеры.** На этой панели можно построить кластеры и управлять ими. Дополнительную информацию смотрите в разделе “Исследование кластеров” на стр. 151.
- **Панель Визуализация.** На этой панели можно визуально исследовать кластеры и их взаимодействия. Дополнительную информацию смотрите в разделе “Диаграммы кластеров” на стр. 161.
- **Панель Данные.** На этой панели можно исследовать и просмотреть текст, содержащийся в документах и записях, соответствующих выбранным вариантам в диалоговом окне Определения кластеров. Дополнительную информацию смотрите в разделе “Определения кластеров” на стр. 151.

## Построение кластеров

При первом обращении к представлению Кластеры никакие кластеры не выводятся. Кластеры можно построить, перейдя в меню (**Инструменты > Построить кластеры**) или нажав на панели инструментов кнопку **Построить...**. Это действие открывает диалоговое окно Построить кластеры, в котором можно задать параметры и определить пределы для построения кластеров.

*Примечание*: Всякий раз, когда результаты извлечения перестают соответствовать ресурсам, эта панель становится желтой, ровно как и панель Результаты извлечения. Выполнив излечение повторно, можно получить его самые последние результаты, и желтый цвет исчезнет. Однако при каждом выполнении извлечения панель Кластеры очищается, и требуется перепостроение ресурсов. Подобным образом, кластеры не сохраняются от одного сеанса к другому.

В диалоговом окне Построить кластеры доступны следующие области и поля:

Входные поля

Таблица **Входные поля**. Кластеры строятся из дескрипторов, получаемых из определенных типов. В этой таблице можно выбрать типы для включения в процесс построения. По умолчанию предварительно будут выбраны типы, захватывающие большинство записей или документов.

**Понятия для кластеризации:** Выберите метод выбора понятий, который вы хотите использовать для кластеризации. Сократив число понятий, можно ускорить процесс кластеризации. Для кластеризации можно использовать ряд понятий верхнего уровня, процентную долю понятий верхнего уровня или все понятия.

- **Число на основе подсчета документов.** При выборе опции **Максимальное число понятий** введите число понятий, которые следует рассматривать для кластеризации. Понятия выбираются на основе числа понятий с наивысшим значением подсчета документов. Подсчет документов - это число документов или записей, в которых встречается понятие.
- **Процент на основе подсчета документов.** При выборе опции **Максимальный процент понятий** введите процентную долю понятий, которые следует рассматривать для кластеризации. Понятия выбираются на основе процентной доли понятий с наивысшим значением подсчета документов.

**Максимальное число документов, которые следует использовать для вычисления кластеров.** По умолчанию значения связей вычисляются с применением всего набора документов или записей. Однако в некоторых случаях может потребоваться ускорить процесс кластеризации, ограничив число документов или записей, используемых для вычисления связей. Ограничение числа документов может привести к снижению качества кластеров. Для использования этой опции включите переключатель слева и введите максимальное число документов или записей, которые следует применить.

Пределы для выходных полей

**Максимальное число создаваемых кластеров**. Это значение представляет собой максимальное число кластеров, генерируемых и выводимых на панели Кластеры. Во время процесса кластеризации насыщенные кластеры выводятся перед ненасыщенными, и поэтому многие кластеры в результатах будут насыщенными. Чтобы увидеть больше ненасыщенных кластеров, этот параметр можно изменить, указав значение, превышающее число насыщенных кластеров.

**Максимальное число понятий в кластере**. Это значение представляет собой максимальное число понятий, которые может содержать кластер.

**Минимальное число понятий в кластере**. Это значение представляет собой минимальное число понятий, которые должны быть связаны, чтобы был создан кластер.

**Максимальное число внутренних связей**. Это значение представляет собой максимальное число внутренних связей, которые может содержать кластер. Внутренние связи - это связи между парами понятий в кластере.

**Максимальное число внешних связей**. Это значение представляет собой максимальное число связей с понятиями вне кластера. Внешние связи - это связи между парами понятий в раздельных кластерах.

**Минимальное значение связи**. Это значение представляет собой наименьшее значение связи, принимаемое для пары понятий, чтобы она рассматривалась при кластеризации. Значение связи вычисляется по формуле преобразования подобия. Дополнительную информацию смотрите в разделе “Вычисление значений связей подобия” на стр. 150.

**Предотвращать объединение отдельных понятий в пары.** Включите этот переключатель, чтобы не объединять два понятия в группу или пару при выводе результатов. Чтобы создавать пары понятий и работать с ними, щелкните по **Работа с парами**. Дополнительную информацию смотрите в разделе “Управление парами с исключением связи” на стр. 116.

## Вычисление значений связей подобия

Знание только числа документов, в котором встречается пара понятий, само по себе не скажет вам, насколько эти два понятия подобны друг другу. В таких случаях может оказаться полезным значение подобия. Значение связи подобия измеряется путем сравнения числа документов совместной встречаемости с числом отдельных документов для каждого понятия в этой взаимосвязи. При вычислении подобия единица измерения - это число документов (подсчет документов), в которых находится понятие или пара понятий. Понятие или пара понятий "находятся" в документе, если они встречается в этом документе *хотя бы* один раз. Можно выбрать вариант, чтобы толщина линий на графике Понятие представляла значение связи подобия на графиках.

Алгоритм выявляет самые прочные взаимосвязи, а это означает, что тенденция совместной встречаемости понятий в текстовых данных будет намного выше, чем тенденция их встречаемости независимо друг от друга. Внутренним образом алгоритм выдает коэффициент подобия от 0 до 1, где значение 1 означает, что два понятия всегда встречаются вместе и никогда не встречаются по отдельности. Затем получившийся коэффициент подобия умножается на 100 и округляется до ближайшего целого числа. Коэффициент подобия вычисляется по формуле, приведенной на следующем рисунке.

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

*Рисунок 31. Формула коэффициента подобия*

Здесь:

- $C_I$ - число документов или записей, в которых встречается понятие I.
- $C_J$ - число документов или записей, в которых встречается понятие J.
- $C_{IJ}$ - число документов или записей, в которых пара понятий I и J совместно встречается в наборе документов.

Для примера допустим, что у вас есть 5000 документов. Пусть I и J будут извлекаемыми понятиями, а IJ - вхождением пары понятий I и J. В следующей таблице предлагаются два сценария, демонстрирующих, как вычисляются коэффициент и значение связи.

*Таблица 32. Пример частотности понятий*

| **Понятие/пара** | **Сценарий A** | **Сценарий B** |
|---|---|---|
| Понятие: I | Встречается в 20 документах | Встречается в 30 документах |
| Понятие: J | Встречается в 20 документах | Встречается в 60 документах |
| Пара понятий: IJ | Совместно встречается в 20 документах | Совместно встречается в 20 документах |
| Коэффициент подобия | 1 | 0.22222 |
| Значение связи подобия | 100 | 22 |

В сценарии A понятия I и J, а также пара понятий IJ встречаются в 20 документах, что дает коэффициент подобия 1, означающий, что эти понятия всегда встречаются вместе. Значение связи подобия для этой пары будет 100.

В сценарии B понятие I встречается в 30 документах, а понятие J - в 60 документах, но пара IJ встречается только в 20 документах. В результате коэффициент подобия будет равен 0,22222. Значение связи подобия для этой пары округляется с понижением до 22.

## Исследование кластеров

После построения кластеров на панели Кластеры можно увидеть набор результатов. Для каждого кластера в таблице доступна следующая информация:

- **Кластер.** Это имя кластера. Кластерам присваиваются имена, добавляемые после понятия с наивысшим числом внутренних связей.
- **Понятия.** Это число понятий в кластере. Дополнительную информацию смотрите в разделе “Определения кластеров”.
- **Внутренние.** Это число внутренних связей в кластере. Внутренние связи - это связи между парами понятий в кластере.
- **Внешние.** Это число внешних связей в кластере. Внешние связи - это связи между парами понятий, если одно понятие находится в одном кластере, а другое понятие - в другом кластере.
- **Насыщенные.** Если присутствует обозначение, оно указывает, что этот кластер мог бы быть крупнее, но тогда будет превышен один или несколько пределов, и поэтому процесс кластеризации для этого кластера закончен, и кластер считается *насыщенным*. В конце процесса кластеризации насыщенные кластеры выводятся перед ненасыщенными, и поэтому многие кластеры в результате будут насыщенными. Чтобы увидеть больше ненасыщенных кластеров, можно изменить параметр **Максимальное число создаваемых кластеров**, указав значение, превышающее число насыщенных кластеров, или уменьшить **Минимальное значение связи**. Дополнительную информацию смотрите в разделе “Построение кластеров” на стр. 148.
- **Порог.** Для всех совместно встречающихся пар понятий в кластере это самое низкое значение связи подобия из всех в кластере. Дополнительную информацию смотрите в разделе “Вычисление значений связей подобия” на стр. 150. Кластер с высоким значением порога означает, что у понятий в этом кластере более высокое общее подобие и они более тесно связаны, чем понятия в кластере, значение порога которого ниже.

Чтобы узнать больше о данном кластере, его можно выбрать, и на панели визуализации справа появятся два графика, помогающие исследовать кластеры. Дополнительную информацию смотрите в разделе “Диаграммы кластеров” на стр. 161. Содержимое таблицы можно также вырезать и вставить в другую прикладную программу.

Всякий раз, когда результаты извлечения перестают соответствовать ресурсам, цвет этой панели становится желтым, как и цвет панели Результаты извлечения. Выполнив излечение повторно, можно получить его самые последние результаты, и желтый цвет исчезнет. Однако каждый раз, когда выполняется извлечение, панель Кластеры очищается, и требуется повторное построение ресурсов. Более того, кластеры не сохраняются от одного сеанса к другому.

## Определения кластеров

Все понятия в кластере можно посмотреть, выбрав его на панели Кластеры и открыв диалоговое окно (**Вид > Определения кластеров**).

Все понятия в выбранном кластере появятся в диалоговом окне Определения кластеров. Если в диалоговом окне Определения кластеров выбрать одно или несколько понятий и нажать кнопку **Вывести &**, на панели Данные появятся все записи или документы, в которых *все выбранные понятия встречаются вместе*. Однако если на панели кластеры выбран кластер, никакие текстовые записи и документы на панели Данные не выводятся. Общую информацию о панели Данные смотрите в разделе “Панель Данные” на стр. 110.

При выборе понятий в этом диалоговом окне изменится также и веб-график понятий. Дополнительную информацию смотрите в разделе “Диаграммы кластеров” на стр. 161. Таким же образом, если выбрать одно или несколько понятий в диалоговом окне Определения кластеров, на панели Визуализация появятся все внешние и внутренние связи от этих понятий.

Описания столбцов

Выводящиеся значки позволяет легко идентифицировать каждый дескриптор.

*Таблица 33. Столбцы и значки дескрипторов*

| Столбцы | Описание |
|---|---|
| **Дескрипторы** | Имя понятия. |
|  **Глобальный** | Показывает, сколько раз этот дескриптор встречается во всем наборе данных (другое название - глобальная частотность). |
|  **Число документов** | Показывает число документов или записей, в которых встречается этот дескриптор (другое название - частотность документов). |
| **Тип** | Показывает типы, к которым принадлежит этот дескриптор. Если дескриптор представляет собой правило категории, никакие имена типов в этом столбце не выводятся. |

Действия панели инструментов

В этом диалоговом окне можно также выбрать одно или несколько понятий для использования в категории. Есть несколько способов сделать это, но самый интересный - это выбрать понятия, совместно встречающиеся в кластере, и добавить их в качестве правила категории. Дополнительную информацию смотрите в разделе “Правила совместного появления” на стр. 120. При помощи кнопок панели инструментов понятия можно добавить в категории.

*Таблица 34. Кнопки панели инструментов для добавления понятий в категории*

| Значки | Описание |
|---|---|
|  | Добавление понятий в новую или существующую категорию |
|  | Добавление понятий в форме правила категории & в новую или существующую категорию. Дополнительную информацию смотрите в разделе “Использование правил категорий” на стр. 126. |
|  | Добавление каждого из выбранных понятий в качестве его собственной новой категории |
|  | Изменяет контент, выводимый на панели Данные и на панели Визуализация в соответствии с выбранными дескрипторами |

*Примечание*: Контекстные меню позволяют также добавить понятия в тип, как синонимы или в качестве исключаемых элементов.

# Глава 12. Изучаем анализ текстовых связей (Text Link Analysis, TLA)

В представлении TLA (Text Link Analysis, анализ текстовых связей) можно изучить результаты извлечения паттернов TLA. TLA - это метод сопоставления паттернов, при помощи которого можно задать правила паттернов и сравнить их с фактически извлеченными понятиями и взаимосвязями, найденными в вашем тексте.

Например, иногда недостаточно извлекать идеи о той или иной организации. При помощи TLA можно также узнать о связях этой организацией с другими организациями или о людях, входящих в эту организацию. Кроме того, при помощи TLA можно извлекать мнения о товарах или, для некоторых языков, взаимосвязи между генами.

Получив некоторые результаты извлечения паттернов TLA, можно просмотреть их на панелях паттернов типа и понятия в представлении TLA. Дополнительную информацию смотрите в разделе “Паттерны типа и понятия” на стр. 155. Можно дополнительно изучить их на панелях данных или визуализации в этом представлении. И самое, возможно, существенное то, что их можно добавить в категории.

Если вы еще не сделали этого, можете выбрать **Извлечь** и **Включить извлечение паттернов TLA** в диалоговом окне Параметры извлечения. Дополнительную информацию смотрите в разделе “Извлечение результатов паттернов TLA” на стр. 154.

Чтобы получить результаты извлечения паттернов TLA, нужно, чтобы какие-то правила паттернов TLA были определены в шаблоне ресурсов или в используемых библиотеках. Можно использовать паттерны TLA в некоторых шаблонах ресурсов, поставляемых вместе с IBM SPSS Modeler Text Analytics. Какого рода взаимосвязи можно извлечь, целиком зависит от правил TLA, определенных в ваших ресурсах. Вы можете определять свои правила TLA для всех языков текста, *кроме* японского. Паттерны содержат макросы, списки слов и промежутки между словами; паттерн представляет собой логический запрос, или правило, которое сравнивается со входным текстом. Дополнительную информацию смотрите в разделе Глава 19, “О правилах текстовых связей”, на стр. 215.

Когда правило паттерна TLA соответствует тексту, этот текст можно извлечь как паттерн и реструктурировать как выходные данные. Затем результаты появляются на панелях представления TLA. Панели можно по отдельности скрывать и выводить, выбирая имя панели в меню Вид:

- **Панели паттернов типа и понятия.** Вы можете строить и изучать свои паттерны на этих двух панелях. Дополнительную информацию смотрите в разделе “Паттерны типа и понятия” на стр. 155.
- **Панель Визуализация.** На этой панели можно в наглядном виде изучать взаимодействие между понятиями и типами в ваших паттернах. Дополнительную информацию смотрите в разделе “Диаграммы TLA (Text Link Analysis, анализ ссылок в тексте)” на стр. 162.
- **Панель Данные.** Можно изучать и просматривать текст, содержащийся в документах и записях, который соответствует выбранному на другой панели. Дополнительную информацию смотрите в разделе “Панель Данные” на стр. 157.

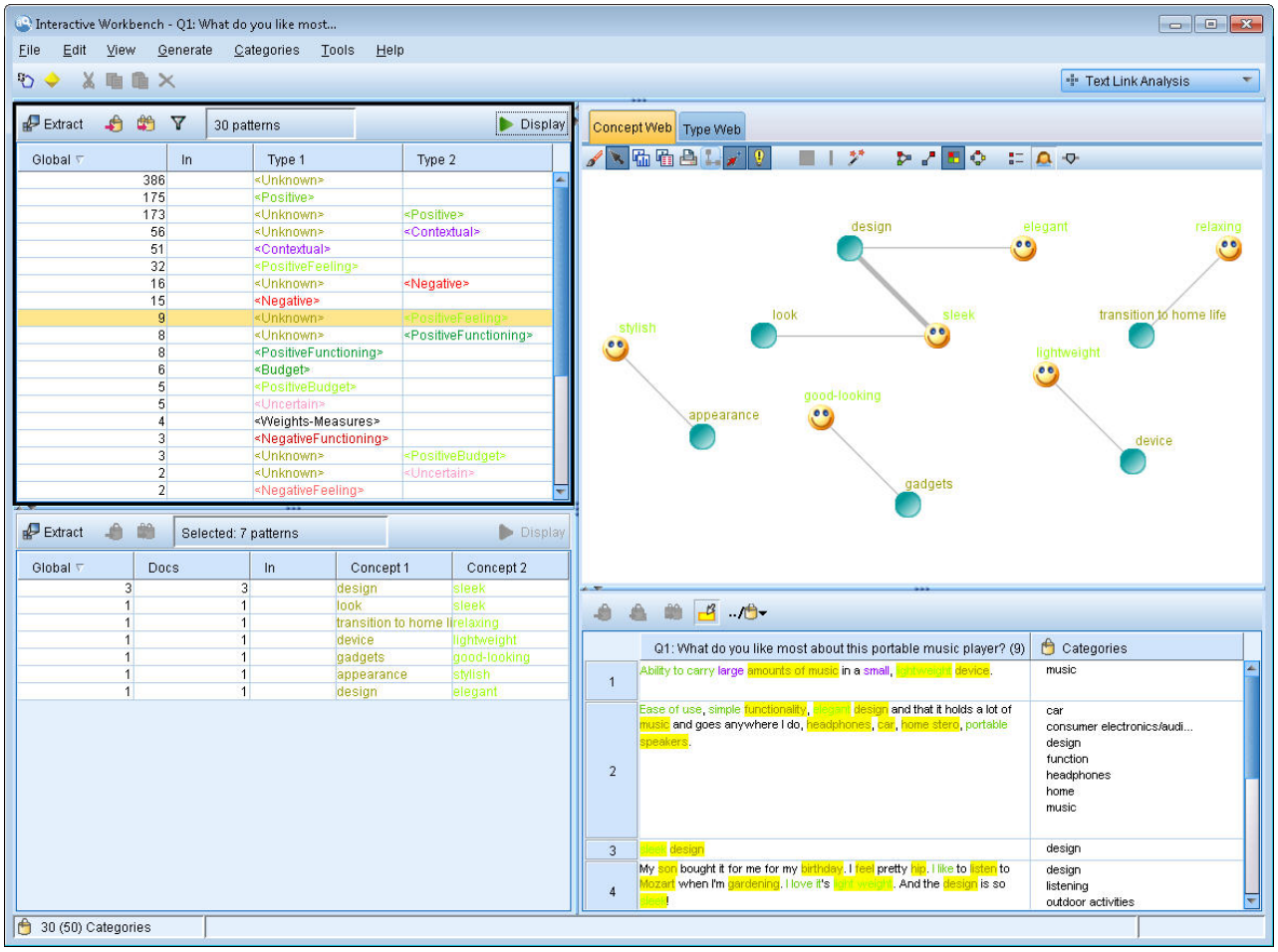



*Рисунок 32. Представление Text Link Analysis*

## Извлечение результатов паттернов TLA

Процесс извлечения приводит к созданию набора понятий и типов, например, паттернов Text Link Analysis (TLA), если они разрешены. Если извлечены паттерны TLA, их можно видеть в представлении анализа текстовых связей (Text Link Analysis, TLA). Если результаты извлечения рассинхронизировались с ресурсами, панели паттернов становятся желтыми, показывая, что при новом извлечении возможны другие результаты.

Нужно выбрать извлечение паттернов в параметре узла или в диалоговом окне Извлечь при помощи опции **Включить извлечение паттернов Text Link Analysis**. Дополнительную информацию смотрите в разделе “Извлечение данных” на стр. 88.

*Примечание*: Существует связь между размером вашего набора данных и временем, которое требуется для завершения процесса извлечения. Информация о статистике производительности и рекомендации приводятся в указаниях по установке. Можно добавить узел выборки выше данного узла или оптимизировать конфигурацию компьютера.

Для извлечения данных

1. В меню выберите **Инструменты > Извлечь**. Или же нажмите кнопку **Извлечь** на панели инструментов.

2. Внесите изменения в нужные опции. Следует иметь в виду, что для появления результатов извлечения паттернов TLA необходимо, чтобы на этой вкладке была выбрана опция **Включить извлечение паттернов Text Link Analysis**, а также чтобы в вашем паттерне были правила TLA. Дополнительную информацию смотрите в разделе “Извлечение данных” на стр. 88.
3. Нажмите кнопку **Извлечь**, чтобы начать процесс извлечения.

Как только извлечение начнется, появится диалоговое окно хода выполнения. Если нужно отменить извлечение, нажмите кнопку **Отмена**. По завершении извлечения это диалоговое окно закрывается, и результаты выводятся на этой панели. Дополнительную информацию смотрите в разделе “Паттерны типа и понятия”.

## Паттерны типа и понятия

Паттерны состоят из двух частей, сочетая понятия и типы. Паттерны особенно полезны при попытке обнаружить мнения по некоторой теме или взаимосвязи между понятиями. Например, иногда недостаточно просто извлечь имя продукта вашего конкурента. В таком случае вы можете поискать в извлеченных паттернах примеры, когда документ или запись содержит текст, в котором продукт сочли хорошим, плохим или дорогостоящим.

Паттерны могут содержать до шести типов или понятий. По этой причине строки на обоих панелях паттернов могут содержать до шести слотов, или позиций. Каждый слот соответствует определенной позиции элемента в правиле паттерна TLA, определенном в лингвистических ресурсах. В интерактивном сеансе инструментальной среды, если слот не содержит значения, он не выводится в таблице. Например, если самый длинный результат извлечения паттерна содержит только четыре слота, последние два слота не выводятся. Дополнительную информацию смотрите в разделе Глава 19, “О правилах текстовых связей”, на стр. 215.

При извлечении паттернов они сначала группируются на уровне типа, а затем делятся на паттерны понятия. Поэтому есть две различные панели результатов: **Паттерны типа** (вверху слева) и **Паттерны понятия** (внизу слева). Чтобы увидеть все возвращенные паттерны понятия, выберите все паттерны типа. Тогда на нижней панели, с паттернами понятия, будут выведены все паттерны понятия до значения максимума по рангу (заданного в диалоговом окне Фильтр).

**Паттерны типа.** Эта панель представляет результаты извлечения паттернов, содержащих один или несколько связанных типов, соответствующих правилу паттерна TLA. Паттерны типа устроены как <Организация> + <Положение> + <Положительный>; в этом примере нужно получить положительные отзывы о конкретной организации в конкретном положении. Соответствующий синтаксис следующий:

```
<Тип1> + <Тип2> + <Тип3> + <Тип4> + <Тип5> + <Тип6>
```

**Паттерны понятия.** На этой панели представлены результаты извлечения паттернов на уровне понятия для всех типов паттернов, выбранных в данный момент на расположенной выше панели Паттерны типа. Паттерны понятия имеют структуру вида отель + париж + чудесный. Соответствующий синтаксис следующий:

```
понятие1 + понятие2 + понятие3 + понятие4 + понятие5 + понятие6
```

Если результаты извлечения паттерна используют меньше допустимых шести слотов, выводится только необходимое число слотов (или столбцов). Любые пустые слоты, найденные между двумя заполненными слотами, отбрасываются, так что паттерн <Тип1>+<>+<Тип2>+<>+<>+<> может быть представлен как <Тип1>+<Тип3>. В случае паттерна понятия это может быть понятие1+.+понятие2 (где . представляет пустое значение).

Здесь можно просматривать результаты, так же как результаты извлечения в представлении категорий и понятий. Если потребуется внести уточнения в типы и понятия, входящие в эти паттерны, это можно сделать на панели результатов извлечения в представлении категорий и понятий или непосредственно в редакторе

ресурсов, а потом извлечь паттерны еще раз. Если понятие, тип или паттерн используются в определении категории, сами по себе или в составе правила, значок категории или правила выводится в столбце **В** в таблице паттерна или результатов извлечения.

## Фильтрация результатов TLA

При работе с очень большими наборами данных процесс извлечения может сгенерировать миллионы результатов. Для многих пользователей такой объем затруднит эффективный просмотр результатов. Но можно отфильтровать эти результаты, сосредоточившись на самых интересных. Эти параметры можно изменять в диалоговом окне Фильтр, ограничивая показ паттернов. Все эти параметры используются совместно.

В представлении TLA диалоговое окно Фильтр содержит следующие области и поля.

**Фильтр по частоте.**Можно отфильтровать результаты с определенным значением глобальной частоты или частоты в документах.

- **Глобальная частота** - это число вхождений паттерна во всем наборе документов или записей; выводится в столбце **Глобальная**.
- **Частота в документах** - это число документов или записей, в которых встречается паттерн; выводится в столбце **В документах**.

Например, если паттерн встречается 300 раз в 500 записях, считается, что его глобальная частота 300, а частота в документах - 500.

**И по соответствию тексту.** Кроме того, при помощи фильтра можно вывести только те результаты, которые соответствуют заданному здесь правилу. Введите набор символов, соответствие с которым будет искаться, в поле **Соответствие тексту** и выберите поиск этого текста в пределах имен понятий или типов, указав номер слота или все сразу. Затем выберите условие, в котором нужно применить соответствие (заключать имя типа в угловые скобки не требуется). Выберите **И** или **ИЛИ** в выпадающем списке, чтобы правило требовало соответствия обоим операторам или хотя бы одному из них, и определите второй оператор соответствия тексту, аналогичным образом.

*Таблица 35. Условия соответствия тексту*

| Условие | Описание |
|---|---|
| Содержит | Соответствие тексту признается, если строка встречается в любом месте. (выбрано по умолчанию) |
| Начинается с | Соответствие тексту признается, только если понятие или тип начинаются с заданного текста. |
| Оканчивается на | Соответствие тексту признается, только если понятие или тип оканчиваются заданным текстом. |
| Точное совпадение | Вся строка должна соответствовать имени понятия или типа. |

**И по рангу.** Кроме того, при помощи фильтра можно вывести только заданное число первых паттернов согласно их глобальной частоте (**Глобальная**) или частоте в документах (**В документах**) в возрастающем или убывающем порядке. Такое значение максимума по рангу ограничивает общее число выводимых паттернов.

Если применяется фильтр, продукт добавляет паттерны типов, пока не будет превышено максимальное число паттернов понятий (максимум по рангу). Сначала ищутся паттерны типов максимального ранга, затем вычисляется суммарное число соответствующих паттернов понятий. Если суммарное число не превышает максимума по рангу, паттерны выводятся в этом представлении. Затем в сумму добавляется число паттернов понятий для следующего паттерна типа. Если это число плюс общее число всех паттернов понятий в предыдущем паттерне типа меньше максимума по рангу, эти паттерны также выводятся в этом представлении. Так продолжается до вывода максимально возможного числа паттернов, без превышения максимума по рангу.

Результаты, выведенные на панели паттернов

Пусть у вас английская версия программного обеспечения; вот несколько примеров, как можно вывести результаты на панели паттернов с учетом фильтров на панели инструментов.


*Рисунок 33. Фильтрация результатов - пример 1*

В этом примере на панели инструментов показано, что число выведенных паттернов было ограничено максимумом по рангу, заданным в фильтре. Если присутствует фиолетовый значок, это значит, что было достигнуто максимальное число паттернов. Для дополнительной информации остановите указатель на значке. Смотрите предыдущее объяснение к фильтру **И по рангу**.


*Рисунок 34. Фильтрация результатов - пример 2*

В этом примере на панели инструментов значок лупы показывает, что число результатов было ограничено при помощи фильтра соответствия тексту. Чтобы увидеть сопоставляемый текст, остановите указатель на значке.

Чтобы фильтровать результаты

1. В меню выберите **Сервис > Фильтр**. Откроется диалоговое окно Фильтр.
2. Выберите и уточните нужные фильтры.
3. Нажмите кнопку **OK**, чтобы применить фильтры и увидеть новые результаты.

## Панель Данные

При извлечении и изучении паттернов TLA иногда нужно пересмотреть некоторые данные, с которыми вы работаете. Например, иногда нужно посмотреть фактические записи, в которых обнаружена группа паттернов. Можно просмотреть записи и документы на панели данных, расположенной в нижнем правом углу. Если она не видна по умолчанию, выберите **Вид > Панели > Данные** в меню.

Панель данных содержит по одной строке для каждого документа или записи, выбранных в представлении, в заданном пределе вывода. По умолчанию число документов или записей, показанных на панели данных, ограничено для скорейшего вывода данных. Но это можно изменить в диалоговом окне Опции. Дополнительную информацию смотрите в разделе “Опции: вкладка Сеанс” на стр. 82.

Вывод и обновление панели данных

Панель данных не обновляется автоматически, поскольку при больших наборах данных автоматическое обновление занимает некоторое время. Поэтому при выборе паттерна типа или понятия в этом представлении можно выбрать **Вывести**, чтобы обновить содержимое панели данных.

Текстовые документы или записи

Если ваши текстовые данные имеют форму записей, и текст относительно небольшой длины, текстовое поле на панели данных содержит полный текст. Но при работе с записями и наборами данных большего размера столбец текстового поля содержит небольшую часть текста и открывает панель Предварительный просмотр текста справа, содержащую большего размера порцию или весь текст записи, выбранной в таблице. Если ваши текстовые данные имеют форму отдельных документов, панель данных содержит имя файла документа. При выборе документа открывается панель Предварительный просмотр текста с текстом выбранного документа.

Цвета и выделение

при выводе данных понятия и дескрипторы, найденные в документах или записях, выделяются цветом для удобства идентификации в тексте. Цветовые коды соответствуют типам, заданным для понятий. Кроме того, если остановить указатель мыши на элементе того или иного цвета, выводится то понятие, под которым этот элемент был извлечен, и назначенный этому понятию тип. Текст, который не был извлечен, выводится черным. Чаще всего не извлечены остаются такие слова, как соединители (*и* или *с*), местоимения (*меня* или *они*) и глаголы (*был*, *есть*, *принимать*).

Столбцы панели данных

Столбец текстового поля всегда видимый, и можно задать вывод других столбцов. Чтобы вывести другие столбцы, выберите **Вид > Панель данных** в меню и выберите столбец, который нужно вывести на панели данных. Для вывода могут быть доступны следующие столбцы:

- **"Имя текстового поля" (#)/документы.** Добавляет столбец для текстовых данных, из которого были извлечены понятия и тип. Если данные представлены в документах, столбец называется Документы и видимы только имя файла или полный путь документа. Чтобы увидеть текст этих документов, нужно заглянуть на панель Предварительный просмотр текста. Число строк на панели данных показано в скобках после имени этого столбца. Иногда показаны не все документы или записи из-за ограничения в диалоговом окне Опции, цель которого - повысить скорость загрузки. При достижении максимума после числа добавляется **- Max**. Дополнительную информацию смотрите в разделе “Опции: вкладка Сеанс” на стр. 82.
- **Категории.** Выводит все категории, к которым принадлежит запись. Если этот столбец выведен, обновление панели данных может занять больше времени, пока собирается новейшая информация.
- **Ранг значимости.** Содержит ранг каждой записи в одной категории. Ранг показывает, насколько хорошо запись соответствует категории по сравнению с остальными записями в этой категории. Выберите категорию на панели категорий (верхняя левая панель), чтобы увидеть ранг. Дополнительную информацию смотрите в разделе “Релевантность категорий” на стр. 111.
- **Число категорий.** Выводит число категорий, к которым принадлежит запись.

# Глава 13. Диаграммы визуализации

В представлении категорий и понятий, в представлении кластеров и в представлении Text Link Analysis есть панель визуализации в верхнем правом углу окна. При помощи этой панели можно изучать данные. Доступны следующие диаграммы.

- **Представление категорий и понятий.** В этом представлении есть три диаграммы: *Столбчатая диаграмма категории*, *Веб-диаграмма категории* и *Веб-таблица категории*. В этом представлении диаграммы обновляются только при нажатии кнопки **Вывести**. Дополнительную информацию смотрите в разделе “Графики и диаграммы категорий”.
- **Представление кластеров.** В этом представлении есть две веб-диаграммы: *Веб-диаграмма понятия* и *Веб-диаграмма кластера*. Дополнительную информацию смотрите в разделе “Диаграммы кластеров” на стр. 161.
- **Представление Text Link Analysis.** В этом представлении есть две веб-диаграммы: *Веб-диаграмма понятия* и *Веб-диаграмма типа*. Дополнительную информацию смотрите в разделе “Диаграммы TLA (Text Link Analysis, анализ ссылок в тексте)” на стр. 162.

Дополнительную информацию обо всех общих панелях инструментов и палитрах, служащих для редактирования диаграмм, смотрите в разделах о редактировании диаграмм в электронной справке или в файле *modeler_nodes_general_book.pdf*, доступном в папке *\Documentation\en* в IBM SPSS Modeler DVD.

## Графики и диаграммы категорий

При построении категорий важно уделить время на то, чтобы пересмотреть определения категорий, содержащиеся в них документы или записи и перекрытие этих категорий. Панель визуализации поддерживает несколько перспектив для ваших категорий. Панель визуализации расположена в правом верхнем углу представления Категории и понятия . Если она не видна, ее можно открыть из меню Вид (**Вид > Панели > Визуализация**).

В этом представлении панель визуализации поддерживает три перспективы для объединений при категоризации документов или записей . Диаграммы и графики на этой панели полезны при анализе результатов категоризации и тонкой настройки категорий или отчетов. При тонкой настройке категорий на этой панели можно просматривать определения категорий в поисках слишком близких категорий (например, совместно использующих более 75% своих документов или записей ) или слишком определенных. Если две категории очень похожи, иногда удобнее объединить их в одну. Другой вариант - уточнить определения категорий, удалив часть дескрипторов из той или иной категории.

В зависимости от выбранного на панели результатов извлечения, на панели категорий или в диалоговом окне определений категорий можно видеть те или иные взаимодействия между документами/записями и категориями на каждой вкладке этой панели. Все они представляют сходные сведения, но в разной форме или с разным уровнем подробности. Однако для обновления диаграммы по текущему выбранному нужно щелкнуть по значку **Вывести** на панели инструментов или в диалоговом окне, где сделан выбор.

Панель визуализации в представлении категорий и понятий поддерживает следующие графики и диаграммы:

- **Столбчатая диаграмма категорий.** Таблица и столбчатая диаграмма представляют перекрытие между документами/записями , соответствующими выбранному и связанным с ним категориям. Кроме того, столбчатая диаграмма представляет отношения числа документов/записей в категориях к общему числу документов/записей. Дополнительную информацию смотрите в разделе “Столбчатая диаграмма категорий” на стр. 160.
- **Веб-диаграмма категорий.** На этой диаграмме представлено перекрытие документов/записей для категорий, которым принадлежат документы/записи согласно выбранному на других панелях. Дополнительную информацию смотрите в разделе “Диаграмма сети категорий” на стр. 160.

- **Веб-таблица категорий.** В этой таблице представлена та же информация, что на вкладке веб-категорий, но в табличном формате. Таблица содержит три столбца, которые можно сортировать щелчком по заголовку столбца. Дополнительную информацию смотрите в разделе “Таблица Сеть категорий”.

Дополнительную информацию смотрите в разделе Глава 10, “Категоризация текстовых данных”, на стр. 101.

## Столбчатая диаграмма категорий

На этой вкладке выводится таблица и столбчатая диаграмма, показывающая перекрытие документов/записей, соответствующих вашему выбору и связанным категориям. Столбчатая диаграмма представляет также отношение числа документов/записей в категориях к полному числу документов или записей. Макет этой диаграммы изменить невозможно. Однако вы можете отсортировать столбцы, щелкнув по их заголовкам.

Диаграмма содержит следующие столбцы:

- **Категория.** В этом столбце представлено имя категории из вашего выбора. По умолчанию первой представлена самая общая категория из выбранных.
- **Столбец.** В этом столбце визуально представлено отношение числа документов или записей в данной категории к общему числу документов или записей.
- **% выбора.** В этом столбце представлена процентное выражение для отношения общего числа документов или записей для категории к общему числу выбранных документов или записей.
- **Документы.** В этом столбце представлено число выбранных документов или записей в данной категории.

## Диаграмма сети категорий

На этой вкладке показана диаграмма сети категорий. Эта сеть представляет перекрытие документов или записей для категорий, к которым принадлежат документы или записи в соответствии с выбором на других панелях. Если у категории есть метки, эти метки появятся на диаграмме. Можно выбрать макет диаграммы (сетевая, круглая, направленная или решеточная), используя кнопки панели инструментов на этой панели.

В сети каждый узел представляет категорию. С помощью мыши можно выбирать и перемещать узлы по панели. Размер узла представляет относительный размер на основе числа выбранных вами документов или записей для данной категории. Толщина и цвет линии между двумя категориями обозначает число их общих документов или записей. Если навести указатель мыши на узел в режиме Исследование, появится подсказка с именем (или меткой) категории и общим числом документов или записей в этой категории.

*Примечание*: Режим исследования по умолчанию включается для диаграмм, в которых вы можете перемещать узлы. Однако вы можете переключиться на режим Изменение, чтобы изменить макеты диаграмм, в том числе цвета, шрифты, подписи и др. Дополнительную информацию смотрите в разделе “Использование палитр и панелей инструментов диаграмм” на стр. 163.

## Таблица Сеть категорий

На этой вкладке выводится та же информация, что и на вкладке Сеть категорий, но в табличном формате. В этой таблице три столбца, которые можно сортировать, щелкнув по заголовку:

- **Частота.** В этом столбце представлено число совместно используемых, то есть общих документов или записей для двух категорий.
- **Категория 1.** В этом столбце показано имя первой категории и в скобках после него общее число содержащихся в этой категории документов или записей.
- **Категория 2.** В этом столбце показано имя второй категории и в скобках после него общее число содержащихся в этой категории документов или записей.

## Диаграммы кластеров

После построения кластеров их можно исследовать визуально на веб-диаграммах на панели Визуализация. На панели Визуализация предлагаются две точки зрения на кластеризацию: веб-диаграмма понятий и веб-диаграмма кластеров. Веб-диаграммы на этой панели могут использоваться для анализа результатов кластеризации и помогают выявить некоторые понятия и правила, которые, возможно, вы захотите добавить в используемые категории. Панель Визуализация находится в верхнем правом углу представления Кластеры. Если она не видна, ее можно открыть из меню Вид (**Вид > Панели > Визуализация**). Выбрав кластер на панели Кластеры, можно автоматически вывести соответствующие диаграммы на панели Визуализация.

*Примечание*: По умолчанию диаграммы выводятся в интерактивном режиме или в режиме выделения, в котором узлы можно двигать. Однако в режиме редактирования можно отредактировать макеты диаграмм, включая цвет, шрифты, пояснения и прочее. Дополнительную информацию смотрите в разделе “Использование палитр и панелей инструментов диаграмм” на стр. 163.

В представлении Кластеры есть две веб-диаграммы.

- **Веб-диаграмма понятий**. Эта диаграмма представляет все понятия в одном или нескольких выбранных кластерах, а также связанные понятия вне выбранного кластера. Эта диаграмма может помочь посмотреть, как связаны понятия внутри кластера, и все внешние связи. Дополнительную информацию смотрите в разделе “Веб-диаграмма понятия”.
- **Веб-диаграмма кластеров**. Эта диаграмма представляет выбранные кластеры со всеми внешними связями между выбранными кластерами, показанными пунктирными линиями. Дополнительную информацию смотрите в разделе “Веб-диаграмма Кластеры”.

Дополнительную информацию смотрите в разделе Глава 11, “Анализ кластеров”, на стр. 147.

## Веб-диаграмма понятия

На этой вкладке выводится веб-диаграмма, показывающая все понятия в одном или нескольких выбранных кластерах, а также связанные понятия вне выбранного кластера. Эта диаграмма может помочь посмотреть, как связаны понятия внутри кластера, и все внешние связи. Каждое понятие в кластере представляется в виде узла, маркированного цветом, соответствующим цвету типа. Дополнительную информацию смотрите в разделе “Создание типов” на стр. 191.

Внутренние связи между понятиями в кластере представлены линиями, толщина каждой из которых прямолинейно соответствует либо числу документов для совместного вхождения каждой пары понятий, либо значению связи подобия в зависимости от вашего варианта выбора на панели инструментов диаграмм. Показаны также внешние связи между понятиями кластера и понятиями вне кластера.

Если в диалоговом окне Определения кластеров выбрать понятия, на веб-диаграмме понятий будут выведены эти понятия и все соответствующие внутренние и внешние связи с этими понятиями. Все связи между другими понятиями, не включающие в себя ни одно из выбранных понятий, на этой диаграмме не выводятся.

*Примечание*: По умолчанию диаграммы выводятся в интерактивном режиме или в режиме выделения, в котором узлы можно двигать. Но в режиме Правка можно отредактировать макеты диаграмм, включая цвета, шрифты, пояснения и другие элементы. Дополнительную информацию смотрите в разделе “Использование палитр и панелей инструментов диаграмм” на стр. 163.

## Веб-диаграмма Кластеры

На этой вкладке выводится веб-диаграмма, показывающая выбранные кластеры. Внешние связи между выбранными кластерами, а также любые связи между другими кластерами выводятся пунктирными линиями. На веб-диаграмме кластеров каждый узел представляет целый кластер, а толщина линий, проведенных между узлами представляет число внешних связей между двумя кластерами.

**Важно!** Чтобы вывести веб-диаграмму кластера, нужно, чтобы были построены кластеры с внешними связями. Внешние связи - это связи между понятиями из разных кластеров (одно понятие пары в одном кластере, а другое - вне этого кластера, в другом кластере).

К примеру скажем, что у нас два кластера. `Кластер A` содержит три понятия: `A1`, `A2` и `A3`. `Кластер B` содержит два понятия: `B1` и `B2`. Связаны следующие понятия: `A1-A2`, `A1-A3`, `A2-B1` (внешние), `A2-B2` (внешние), `A1-B2` (внешние) и `B1-B2`. Это означает, что на веб-диаграмме кластеров будут представлены три внешние связи.

*Примечание*: По умолчанию диаграммы выводятся в интерактивном режиме или в режиме выделения, в котором узлы можно двигать. Но в режиме Правка можно отредактировать макеты диаграмм, включая цвета, шрифты, пояснения и другие элементы. Дополнительную информацию смотрите в разделе “Использование палитр и панелей инструментов диаграмм” на стр. 163.

## Диаграммы TLA (Text Link Analysis, анализ ссылок в тексте)

После извлечения паттернов TLA (Text Link Analysis, анализ текстовых связей) можно изучить их в наглядном виде при помощи веб-диаграмм на панели Визуализация. Панель визуализации предлагает две перспективы для паттернов TLA: веб-диаграмму (паттернов) понятия и веб-диаграмму (паттернов) типа. При помощи веб-диаграмм на этой панели можно представить паттерны в наглядном виде. Панель визуализации расположена в правом верхнем углу представления TLA. Если она не видна, ее можно открыть из меню Вид (**Вид > Панели > Визуализация**). Если ничего не выбрано, область диаграммы пуста.

*Примечание*: По умолчанию диаграммы выводятся в интерактивном режиме или в режиме выделения, в котором узлы можно двигать. Но в режиме Правка можно отредактировать макеты диаграмм, включая цвета, шрифты, пояснения и другие элементы. Дополнительную информацию смотрите в разделе “Использование палитр и панелей инструментов диаграмм” на стр. 163.

Представление TLA содержит две веб-диаграммы.

- **Веб-диаграмма понятия**. Эта диаграмма представляет все понятия в выбранных паттернах. Толщина линий и размер узлов (если не выводятся значки) на диаграмме понятия показывают число глобальных вхождений в выбранной таблице. Дополнительную информацию смотрите в разделе “Веб-диаграмма понятия”.
- **Веб-диаграмма типа**. Эта диаграмма представляет все типы в выбранных паттернах. Толщина линий и размер узлов (если не выводятся значки) на диаграмме показывают число глобальных вхождений в выбранной таблице. Узлы представлены цветом типа или значком. Дополнительную информацию смотрите в разделе “Веб-диаграмма типа”.

Дополнительную информацию смотрите в разделе Глава 12, “Изучаем анализ текстовых связей (Text Link Analysis, TLA)”, на стр. 153.

## Веб-диаграмма понятия

Эта веб-диаграмма представляет все понятия, охваченные текущим выделением. Например, если выделить паттерн типа, которому соответствует три паттерна понятия, на данной диаграмме будет показано три набора связанных понятий. Толщина линий и размер узлов на диаграмме понятия представляют глобальные частоты. Диаграмма в наглядном виде представляет ту же информацию, которая содержится в выбранном на панелях паттернов. Типы каждого понятия представлены цветом или значком в зависимости от выбора на панели инструментов диаграммы. Дополнительную информацию смотрите в разделе “Использование палитр и панелей инструментов диаграмм” на стр. 163.

## Веб-диаграмма типа

Эта веб-диаграмма представляет все паттерны типа для текущего выделения. Например, если выделить два паттерна понятия, на данной диаграмме будет показано по одному узлу на каждый тип в выбранных паттернах и связи между ними, найденные в одном и том же паттерне. Толщина линий и размер узлов представляют глобальные частоты для этого набора. Диаграмма в наглядном виде представляет ту же

информацию, которая содержится в выбранном на панелях паттернов. Помимо имен типов, которые выводятся на диаграмме, типы идентифицируются цветом или значком типа в зависимости от выбора на панели инструментов диаграммы. Дополнительную информацию смотрите в разделе “Использование палитр и панелей инструментов диаграмм”.

## Использование палитр и панелей инструментов диаграмм

У каждой диаграммы есть панель инструментов, предоставляющая быстрый доступ к некоторым общим палитрам, с помощью которых можно выполнять несколько действий с вашими диаграммами. Панели в разных представлениях (Категории и Понятия, Кластеры и Анализ текстовых связей) немного отличаются. Можно выбирать между режимами представлений *Исследование* и *Изменение*.

В то время как режим исследования позволяет аналитически исследовать данные и значения, представленные визуализацией, режим редактирования позволяет менять макет и внешний вид визуализации. Например, можно изменить шрифты и цвета, чтобы они соответствовали руководству по стилю оформления вашей организации. Чтобы войти в этот режим, выберите в меню **Вид > Панель визуализации > Режим редактирования** (или щелкните по значку на панели инструментов).

В режиме редактирования имеется несколько панелей инструментов, которые влияют на разные аспекты внешнего вида визуализации. Можно скрыть неиспользуемые панели, чтобы увеличить пространство диалогового окна, в котором показана диаграмма. Чтобы показать или скрыть панели инструментов, щелкните по соответствующему названию панели инструментов или палитры в меню Вид.

Дополнительную информацию о всех общих панелях инструментов и палитрах, используемых для изменения диаграмм, смотрите в разделе Изменение диаграмм в оперативной справке или в файле *modeler_nodes_general_book.pdf*, находящемся в папке *\Documentation\en* в IBM SPSS Modeler DVD.

*Таблица 36. Кнопки панели инструментов Text Analytics.*

| Кнопка/список | Описание |
|---|---|
|  | Включает режим Изменение. Переключитесь в режим Изменение для изменения вида диаграмм, например, для увеличения размера шрифта, для изменения цветов, чтобы они соответствовали руководству по вашему корпоративному стилю, или для удаления меток и подписей. |
|  | Включает режим Исследование. По умолчанию включен режим Исследование, означающий, что вы можете перемещаться между узлами и перетаскивать их по диаграмме, а также наводить указатель мыши на объекты диаграммы для получения дополнительной информации подсказок. |
|  | Выберите тип вывода сети для диаграмм в представлении Категории и Понятия, а также в представлении Анализ текстовых связей.<br>• **Круговой макет.** Общий макет, который можно применить для всех диаграмм. В этом макете диаграмма строится в предположении, что все связи ненаправленные, а узлы равноценные. Узлы размещаются только по периметру круга.<br>• **Сетевой макет.** Общий макет, который можно применить к любой диаграмме. В этом макете диаграмма строится в предположении, что все связи ненаправленные, а узлы равноценные. На этом макете узлы располагаются свободно.<br>• **Направленный макет.** Макет, который нужно использовать только для направленных диаграмм. Этот макет создает структуры типа дерева, начиная от корневых узлов и распространяясь до конечных элементов, и использует организацию элементов по цвету. Этот макет хорошо работает для вывода иерархических данных.<br>• **Макет с сеткой.** Общий макет, который можно применить к любой диаграмме. В этом макете диаграмма строится в предположении, что все связи ненаправленные, а узлы равноценные. Во всем пространстве узлы размещаются только по сетке. |

*Таблица 36. Кнопки панели инструментов Text Analytics (продолжение).*

| Кнопка/список | Описание |
|---|---|
|  | Представление силы связи. Выберите, линии какой толщины будут представлены на диаграмме. Это применимо только для представления Кластеры. Сетевая диаграмма Кластеры показывает только число внешних связей между кластерами. Можно выбрать следующие варианты:<br>• **Сходство.** Толщина обозначает количество внешних связей между двумя кластерами<br>• **Совместное появление.** Толщина обозначает количество документов, для которых имеет место совместное появление дескрипторов. |
|  | Кнопка переключателя, при нажатии выводится подпись. Если эта кнопка не нажата, подпись не выводится. |
|  | Кнопка переключателя, при нажатии на диаграмме выводятся значки типа вместо цветов типа. Это применимо только в представлении Анализ текстовых связей. |
|  | Кнопка переключателя, при нажатии под диаграммой выводится ползунок Связи. Результаты можно фильтровать, перемещая стрелку ползунка. |
|  | Диаграмма будет выведена для самого верхнего уровня выбранных категорий без учета их подкатегорий. |
|  | Диаграмма будет выведена для самого низкого уровня выбранных категорий. |
|  | Эта опция управляет тем, как в результатах будут выводиться имена подкатегорий.<br>• **Полный путь категории.** При выборе этой опции имена категорий будут выводиться с полным путем родительских категорий, если они есть, а для разделения имен категорий и подкатегорий будет использоваться дробная черта.<br>• **Короткий путь категории.** При выборе этой опции будут выводиться только имена категорий, но при наличии родительских категорий будет вставляться многоточие.<br>• **Категории нижнего уровня.** При выборе этой опции будет выводиться только имя категория без полного пути и без обозначения наличия родительских категорий. |

# Глава 14. Редактор ресурсов сеанса

IBM SPSS Modeler Text Analytics быстро и точно находит и извлекает ключевые понятия из текстовых данных. Этот процесс извлечения в большой степени полагается на лингвистические ресурсы, чтобы определять, как необходимо извлекать информацию из текстовых данных. По умолчанию эти ресурсы поступают из шаблонов ресурсов.

IBM SPSS Modeler Text Analytics поставляется с набором специализированных **шаблонов ресурсов**, которые содержит ряд лингвистических и нелингвистических ресурсов, в форме библиотек и расширенных ресурсов, чтобы помочь определить, как будут обрабатываться и извлекаться ваши данные. Дополнительную информацию смотрите в разделе Глава 15, “Шаблоны и ресурсы”, на стр. 169.

В диалоговом окне узла можно загрузить в узел копию ресурсов шаблона. Если вы находись в интерактивном сеансе инструментальной среды, можно при желании настроить эти ресурсы специально для данных этого узла. Во время интерактивного сеанса инструментальной среды можно работать с ресурсами в представлении Редактор ресурсов. Когда запускается интерактивный сеанс, выполняется извлечение с помощью ресурсов, загруженных в диалоговом окне узла, если вы не кэшировали свои данные и результаты извлечения в вашем узле.

## Редактирование ресурсов в редакторе ресурсов

Редактор Редактор ресурсов обеспечивает доступ к набору ресурсов, которые используются для получения результатов извлечения (понятия, типы и паттерны) для интерактивного сеанса инструментальной среды. Этот редактор очень похож на редактор Редактор шаблонов, но в редакторе Редактор ресурсов вы редактируете ресурсы для данного сеанса. Когда вы закончили работу над вашими ресурсами и любую другую работу, которую вы делали, можно обновить узел моделирования, чтобы сохранить эту работу, так что ее можно будет восстановить в последующем интерактивном сеансе инструментальной среды. Дополнительную информацию смотрите в разделе “Обновление узлов моделирования и сохранение” на стр. 84.

Если вы хотите работать непосредственно над шаблонами, используемыми для загрузки ресурсов в узлы, мы рекомендуем использовать редактор Редактор шаблонов. Многие задачи, которые можно выполнять в редакторе Редактор ресурсов, выполняются точно так же, как если бы они находились в редакторе Редактор шаблонов, а именно:

- **Работа с библиотеками**. Дополнительную информацию смотрите в разделе Глава 16, “Работа с библиотеками”, на стр. 179.
- **Создание словарей типов**. Дополнительную информацию смотрите в разделе “Создание типов” на стр. 191.
- **Добавление терминов в словари**. Дополнительную информацию смотрите в разделе “Добавление терминов” на стр. 192.
- **Создание синонимов**. Дополнительную информацию смотрите в разделе “Определение синонимов” на стр. 197.
- **Импорт и экспорт шаблонов**. Дополнительную информацию смотрите в разделе “Импорт и экспорт шаблонов” на стр. 177.
- **Публикация библиотек**. Дополнительную информацию смотрите в разделе “Публикация библиотек” на стр. 185.

**Для текста на голландском, английском, французском, немецком, итальянском, португальском и испанском**

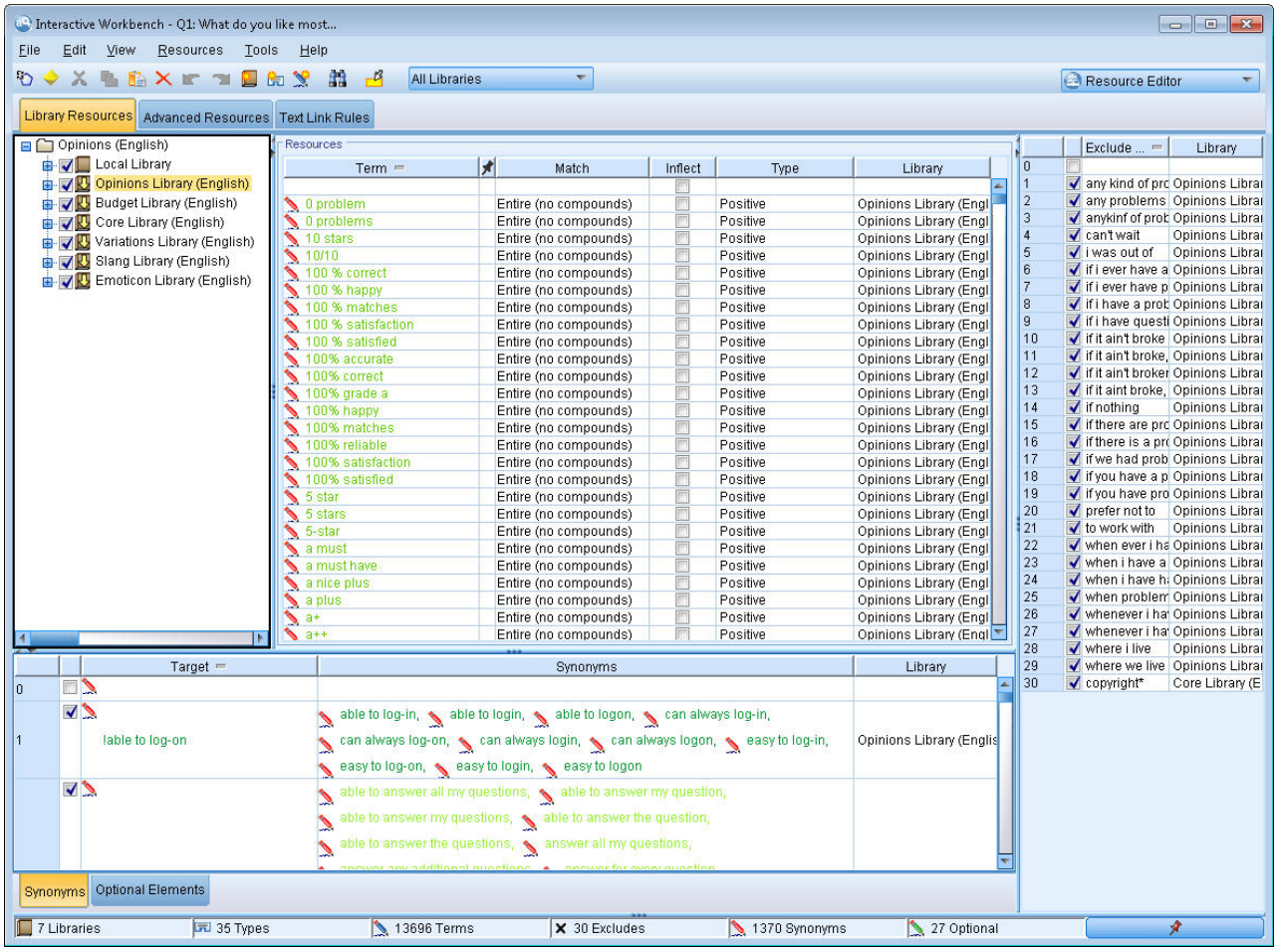



*Рисунок 35. Представление Редактора ресурсов для всех языков, кроме японского*

**Для текста на японском**

Интерфейс редактора для японского текстового языка отличается от интерфейса для других текстовых языков.

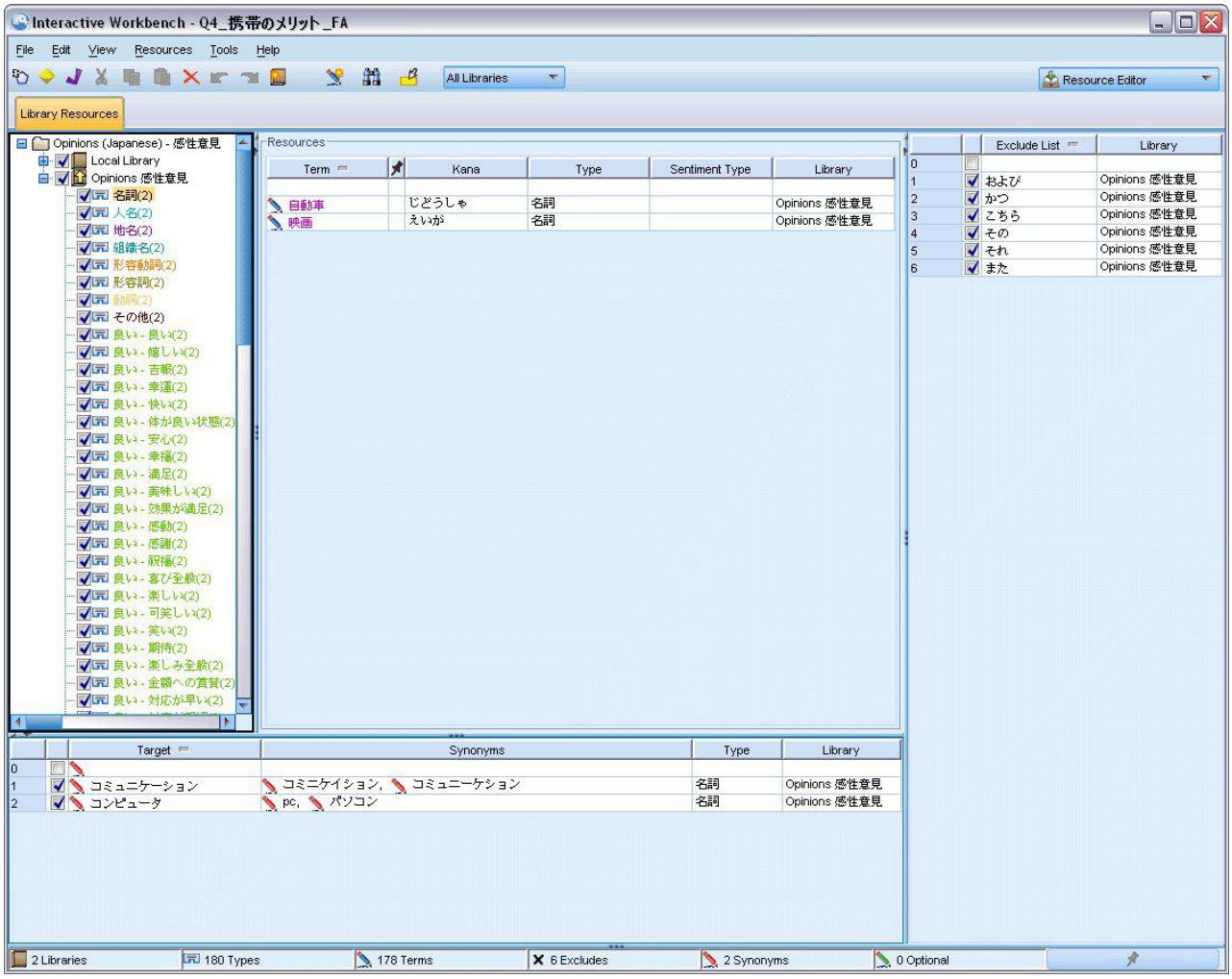



*Рисунок 36. Представление Редактора ресурсов для японского текста*

## Создание и изменение шаблонов

При всяком изменении ресурсов, если вы хотите повторно использовать их впоследствии, можно сохранить эти ресурсы как шаблон. При этом можно выбрать сохранение или с существующим, или с новым именем шаблона. Тогда, если впоследствии этот шаблон будет загружен, вы сможете получить те же ресурсы. Дополнительную информацию смотрите в разделе “Копирование ресурсов из шаблонов и файлов TAP” на стр. 26.

*Примечание*: Ваши библиотеки можно также опубликовать и сделать доступными для совместного использования. Дополнительную информацию смотрите в разделе “Совместное использование библиотек” на стр. 184.

Чтобы создать (или изменить) шаблон

1. В меню представления Редактор ресурсов выберите опцию **Ресурсы > Создать шаблон ресурсов**. Откроется диалоговое окно Создать шаблон ресурсов.
2. Введите новое имя в поле Имя шаблона, если вы хотите создать новый шаблон. Выберите шаблон в таблице, если вы хотите переопределить существующий шаблон на текущие загруженные ресурсы.
3. Нажмите кнопку **Сохранить**, чтобы создать шаблон.

**Важно!** Так как шаблоны загружаются при их выборе на узле, а не при выполнении потока, для получения самых последних изменений убедитесь, что этот шаблон ресурсов загружен на всех других узлах, где он используется. Дополнительную информацию смотрите в разделе “Изменение ресурсов узла после загрузки” на стр. 175.

## Переключение между шаблонами ресурсов

Если вы хотите заменить текущие загруженные ресурсы в сеансе на копию ресурсов из другого шаблона, можно переключиться между ресурсами. При этом будут перезаписаны все текущие загруженные ресурсы в сеансе. Если переключение нужно для доступа к некоторым предопределенным правилам паттернов анализа текстовых связей (Text Link Analysis, TLA), убедитесь, что выбирается паттерн, для которого они отмечены в столбце TLA.

**Важно!** Невозможно переключиться с японского шаблона на шаблон с другим языком и наоборот.

Переключение между ресурсами в частности полезно в тех ситуациях, когда вы хотите и восстановить работу сеанса (категории, паттерны и ресурсы), и загрузить измененную копию ресурсов из паттерна, не теряя результаты другой работы в сеансе. Можно выбрать шаблон, контент которого нужно скопировать в Редактор ресурсов, и нажать кнопку **OK.** При этом будут замещены ресурсы, которые вы используете в этом сеансе. Не забудьте изменить узел моделирования в конце сеанса, если нужно сохранить эти изменения для следующего запуска сеанса интерактивной инструментальной среды.

*Примечание*: Если в течение интерактивного сеанса переключиться на контент другого шаблона, именем шаблона, перечисленным на узле, будет по-прежнему имя последнего загруженного и скопированного шаблона. Чтобы получить выигрыш от этих ресурсов или другой работы в сеансе, измените узел моделирования, прежде чем закрывать сеанс, и выберите на узле опцию **Использовать работу сеанса**. Дополнительную информацию смотрите в разделе “Обновление узлов моделирования и сохранение” на стр. 84.

Чтобы переключиться между ресурсами

1. В меню представления Редактор ресурсов выберите опцию **Ресурсы > Переключиться между шаблонами ресурсов**. Откроется диалоговое окно Переключить ресурсы.
2. Выберите из показанных в этой таблице шаблонов тот, который вы хотите использовать.
3. Нажмите кнопку **OK**, чтобы отказаться от текущих загруженных ресурсов и загрузить на их место копию ресурсов выбранного шаблона. Если вы внесли изменения в ваши ресурсы и хотите сохранить библиотеки для дальнейшего использования, можно опубликовать, изменить или назначить их для совместного использования до переключения. Дополнительную информацию смотрите в разделе “Совместное использование библиотек” на стр. 184.

# Глава 15. Шаблоны и ресурсы

IBM SPSS Modeler Text Analytics быстро и точно находит и извлекает ключевые понятия из текстовых данных. Этот процесс извлечения в большой степени полагается на лингвистические ресурсы, чтобы определять, как необходимо извлекать информацию из текстовых данных. Дополнительную информацию смотрите в разделе “Как работает извлечение” на стр. 5. Можно подстроить эти ресурсы в представлении Редактор ресурсов.

При установке программного обеспечения вы также получаете ряд специализированных ресурсов. Эти поставляемые ресурсы позволяют пользоваться преимуществами многолетних исследований и тонкой настройки для определенных языков и определенных областей применения. Поскольку поставленные ресурсы, возможно, не всегда идеально адаптированы к контексту ваших данных, можно редактировать эти шаблоны ресурсов или даже создавать и использовать пользовательские библиотеки, настроенные на данные вашей организации. Эти ресурсы поступают в различных формах, и каждый из них можно использовать в вашем сеансе. Ресурсы можно найти в следующих областях:

- **Шаблоны ресурсов**. Шаблоны составлены из ряда библиотек, типов и некоторых расширенных ресурсов, вместе формирующих специализированный набор ресурсов, адаптированных к конкретной области или контексту, например, мнения о товарах.
- **Пакеты анализа текста (TAP)**. В дополнение к ресурсам, хранящимся в шаблоне, TAP также объединяют один или несколько специализированных наборов категорий, созданных с использованием этих ресурсов, так, что и категории, и ресурсы хранятся вместе и являются повторно используемыми. Дополнительную информацию смотрите в разделе “Использование пакетов анализа текста (Text Analysis Package)” на стр. 140.
- **Библиотеки**. Библиотеки служат строительными блоками как для TAP, так и для шаблонов. Их можно также добавлять по отдельности к ресурсам в ваш сеанс. Каждая библиотека состоит из нескольких словарей, используемых, чтобы определять и управлять типами, синонимами и списками исключения. Хотя библиотеки также поставляются по отдельности, они предварительно упакованы вместе в шаблонах и в TAP. Дополнительную информацию смотрите в разделе Глава 16, “Работа с библиотеками”, на стр. 179.

*Примечание*: Во время извлечения также используются некоторые скомпилированные внутренние ресурсы. Эти скомпилированные ресурсы содержат большое количество определений, дополняющих типы в библиотеке Core. Эти скомпилированные ресурсы нельзя редактировать.

Редактор Редактор ресурсов обеспечивает доступ к набору ресурсов, которые используются для получения результатов извлечения (понятия, типы и паттерны). Существует много задач, которые можно выполнять в редакторе Редактор ресурсов, в частности:

- **Работа с библиотеками**. Дополнительную информацию смотрите в разделе Глава 16, “Работа с библиотеками”, на стр. 179.
- **Создание словарей типов**. Дополнительную информацию смотрите в разделе “Создание типов” на стр. 191.
- **Добавление терминов в словари**. Дополнительную информацию смотрите в разделе “Добавление терминов” на стр. 192.
- **Создание синонимов**. Дополнительную информацию смотрите в разделе “Определение синонимов” на стр. 197.
- **Обновление ресурсов в TAP**. Дополнительную информацию смотрите в разделе “Изменение пакетов анализа текста” на стр. 142.
- **Создание шаблонов**. Дополнительную информацию смотрите в разделе “Создание и изменение шаблонов” на стр. 167.

- **Импорт и экспорт шаблонов**. Дополнительную информацию смотрите в разделе “Импорт и экспорт шаблонов” на стр. 177.
- **Публикация библиотек**. Дополнительную информацию смотрите в разделе “Публикация библиотек” на стр. 185.

## Сравнение Редактора шаблонов с Редактором ресурсов

Есть два основных метода для работы с вашими шаблонами, библиотеками и ресурсами и для их редактирования. Можно работать с лингвистическими ресурсами в редакторе Редактор шаблонов или в Редактор ресурсов.

Редактор шаблонов

Редактор Редактор шаблонов позволяет создавать и редактировать шаблоны ресурсов без интерактивного сеанса инструментальной среды и независимо от определенного узла или потока. Можно использовать этот редактор, чтобы создавать или редактировать шаблоны ресурсов, прежде чем загрузить их в узел Анализа текстовых связей и в узел моделирования исследования текста.

Доступ к редактору Редактор шаблонов возможен через главную панель инструментов IBM SPSS Modeler из меню **Инструменты > Редактор шаблонов текстовой аналитики**.

Редактор ресурсов

Редактор Редактор ресурсов, доступ к которому возможен из интерактивного сеанса инструментальной среды, позволяет работать с ресурсами в контексте определенного узла и набора данных. Когда вы добавляете к потоку узел моделирования исследования текста, можно загрузить копию содержимого шаблона ресурса или копию пакета анализа текста (наборы категорий *и* ресурсов), чтобы управлять способом извлечения текста для исследования текста. При запуске интерактивного сеанса инструментальной среды, в дополнение к созданию категорий, извлечению паттернов анализа текстовых связей и созданию моделей категорий, можно также подстроить ресурсы для данных этого сеанса в интегрированном представлении Редактор ресурсов. Дополнительную информацию смотрите в разделе “Редактирование ресурсов в редакторе ресурсов” на стр. 165.

Каждый раз, когда вы работаете над ресурсами в интерактивном сеансе инструментальной среды, эти изменения применяются только к данному сеансу. Если вы хотите сохранить свою работу (ресурсы, категории, паттерны и т.д.), чтобы можно было продолжить работу в последующем сеансе, надо обновить узел моделирования. Дополнительную информацию смотрите в разделе “Обновление узлов моделирования и сохранение” на стр. 84.

Если вы хотите сохранить внесенные вами изменения в исходном шаблоне, содержимое которого было скопировано в узел моделирования, чтобы этот обновленный шаблон можно было загружать в другие узлы, вы можете сделать шаблон из ресурсов. Дополнительную информацию смотрите в разделе “Создание и изменение шаблонов” на стр. 167.

## Интерфейс редактора

Операции, которые вы выполняете в Редакторе шаблонов или в редакторе Редактор ресурсов, относятся к управлению лингвистическими ресурсами и их тонкой настройке. Эти ресурсы хранятся в виде шаблонов и библиотек. Дополнительную информацию смотрите в разделе “Словари типов” на стр. 189.

Вкладка Ресурсы библиотек

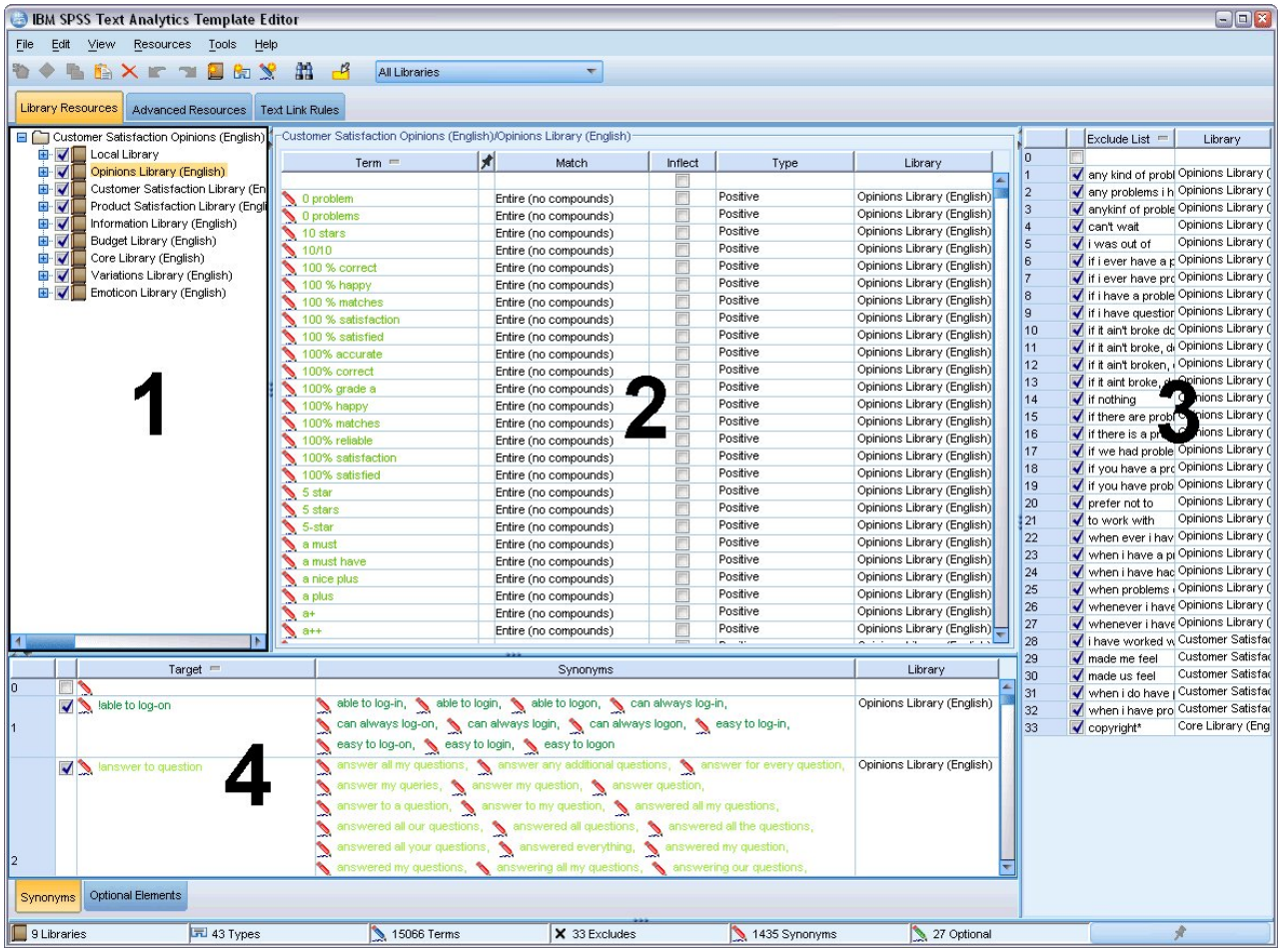



*Рисунок 37. Редактор шаблонов исследования текста*

Интерфейс состоит из следующих четырех частей:

**1. Панель Дерево библиотек.** Занимает верхний левый угол; содержит дерево библиотек. В этом дереве можно включать и выключать библиотеки, а также фильтровать представления на других панелях, выбирая библиотеку в дереве. Многие операции в этом дереве можно выполнять при помощи контекстных меню. Если раскрыть библиотеку в дереве, выводится набор типов, которые в ней содержатся. Кроме того, этот список можно фильтровать при помощи меню **Вид**, если нужно сфокусироваться на конкретной библиотеке.

**2. Списки терминов с панели Словари типов.** Эта панель, расположенная справа от дерева библиотек, содержит списки терминов из словарей типов для библиотек, выбранных в дереве. **Словарь типов** - это собрание терминов, которые нужно группировать под одним именем метки или типа. После того, как механизм извлечения прочел ваши текстовые данные, он сравнивает найденные в текстах слова с терминами в словарях типов. Если извлеченная понятие присутствует как термин в словаре типов, ей назначается это имя типа. Словарь типов можно считать словарем различных терминов, у которых есть что-то общее. Например, тип `<Расположение>` в библиотеке ядра содержит такие понятия, как `новый орлеан`, `великобритания` и `нью-йорк`. Все эти термины представляют географические положения. Библиотека может представлять один или несколько словарей типов. Дополнительную информацию смотрите в разделе “Словари типов” на стр. 189.

**3. Панель Словарь исключения.** Эта панель, расположенная справа, содержит собрание терминов, которые будут исключены из окончательных результатов извлечения. Термины, содержащиеся в этом словаре исключения, не выводятся на панели результатов извлечения. Исключенные термины можно сохранить в

библиотеке по вашему выбору. Но панель Словарь исключения содержит все исключенные термины для всех библиотек, видимых в дереве библиотек. Дополнительную информацию смотрите в разделе “Словари исключения” на стр. 200.

**4. Панель Словарь подстановок.** Эта панель, занимающая нижний левый угол, содержит на отдельных вкладках синонимы и необязательные элементы. Синонимы и необязательные элементы помогают группировать близкие термины в одно сводное целевое понятие в окончательных результатах извлечения. Этот словарь может содержать известные синонимы и пользовательские синонимы и элементы, а также частые ошибочные написания в паре с правильными. Определения синонимов и необязательных элементов можно сохранить в библиотеке по вашему выбору. Но панель Словарь подстановок содержит все содержимое для всех библиотек, видимых в дереве библиотек. Хотя эта панель содержит все синонимы или необязательные элементы из всех библиотек, на ней выводятся совместно подстановки для всех библиотек в дереве. Библиотека может содержать только один словарь подстановок. Дополнительную информацию смотрите в разделе “Словари подстановок/синонимов” на стр. 196. Обратите внимание на то, что вкладка Необязательные элементы неприменима к японоязычным ресурсам.

*Примечания*:

- Если нужна фильтрация, чтобы выводить только информацию, относящуюся к одной библиотеке, можно изменить представление библиотек при помощи выпадающего списка на панели инструментов. Он содержит объект верхнего уровня **Все библиотеки**, а также по дополнительному объекту для каждой отдельной библиотеки. Дополнительную информацию смотрите в разделе “Просмотр библиотек” на стр. 182.
- Интерфейс редактора для японского языка отличается от интерфейса для других языков.

Вкладка Расширенные ресурсы

Расширенные ресурсы доступны на второй вкладке представления редактора. На этой вкладке можно просматривать и редактировать расширенные ресурсы. Дополнительную информацию смотрите в разделе Глава 18, “О расширенных ресурсах”, на стр. 203.

**Важно!** Эта вкладка недоступна для ресурсов, настроенных под японский текст.

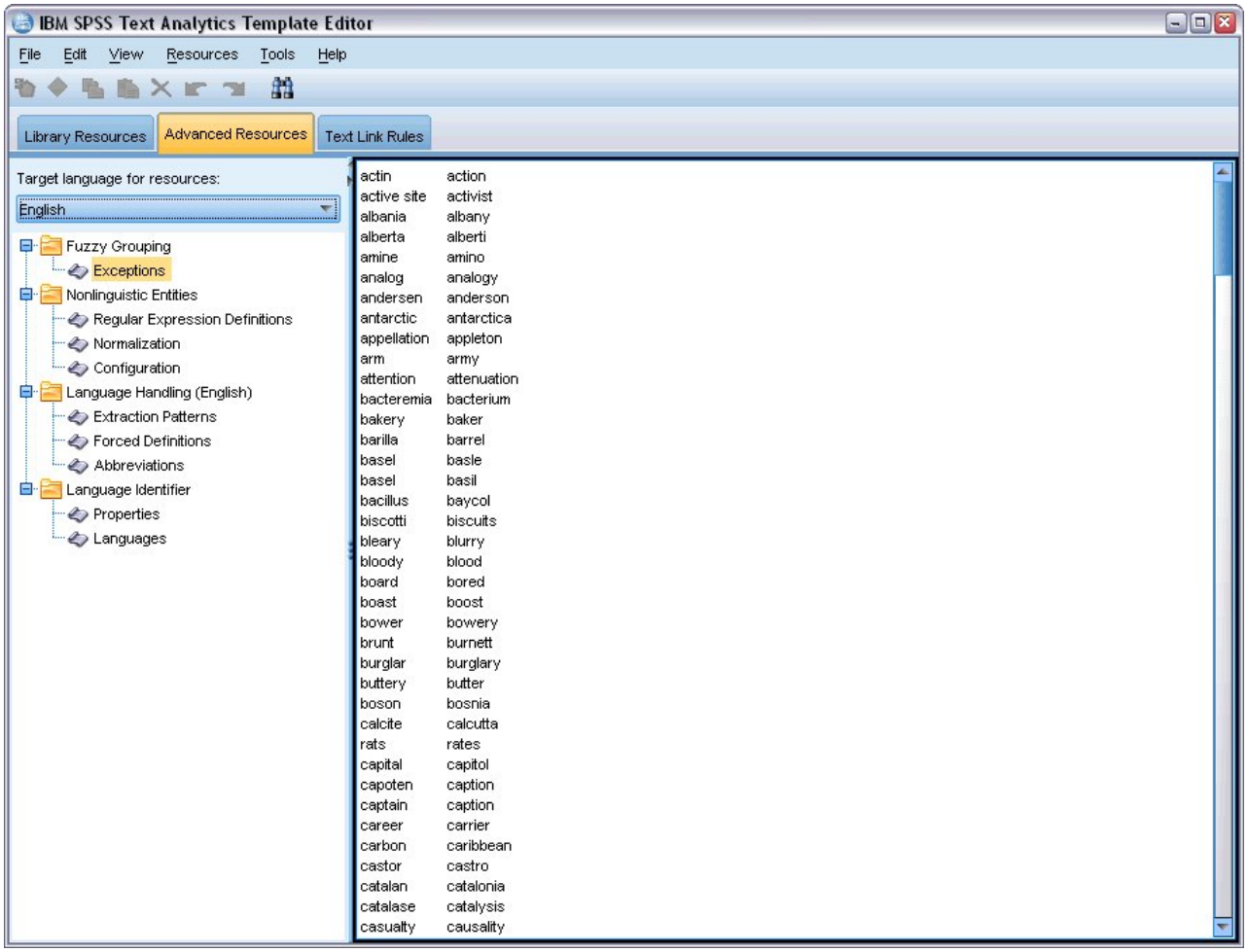



*Рисунок 38. Редактор шаблонов исследования текста - вкладка Расширенные ресурсы*

Вкладка Правила текстовых связей

Начиная с версии 14, правила анализа текстовых связей доступны для редактирования на собственной вкладке представления редактора. Можно работать в редакторе правил, создать собственные правила, и даже запускать имитации, чтобы видеть, как правила влияют на результаты анализа текстовых связей. Дополнительную информацию смотрите в разделе Глава 19, “О правилах текстовых связей”, на стр. 215.

**Важно!** Эта вкладка недоступна для ресурсов, настроенных под японский текст.

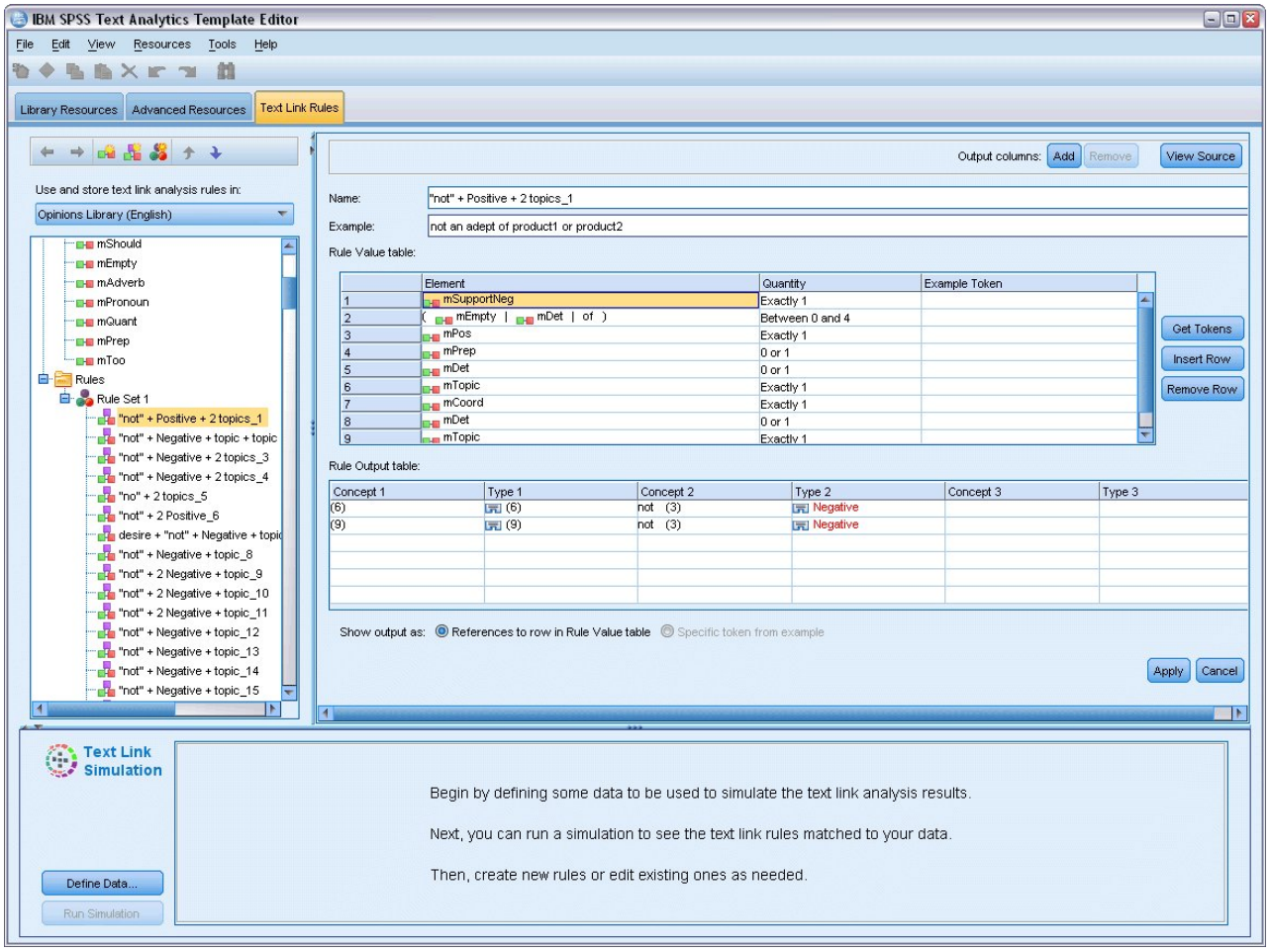



*Рисунок 39. Редактор шаблонов исследования текста - вкладка Правила текстовых связей*

## Открытие шаблонов

Когда вы запускаете редактор Редактор шаблонов, вам предлагают открыть шаблон. Можно также открыть шаблон из меню Файл. Если вам требуется шаблон, содержащий некоторые правила Анализа текстовых связей (Text Link Analysis, TLA), убедитесь, что выбран шаблон, у которого в столбце TLA есть значок. Язык, для которого был создан шаблон, показан в столбце Язык.

Если вы хотите импортировать шаблон, не показанный в таблице, или если вы хотите экспортировать шаблон, можно использовать кнопки в диалоговом окне Открыть шаблон. Дополнительную информацию смотрите в разделе “Импорт и экспорт шаблонов” на стр. 177.

Чтобы открыть шаблон

1. В меню редактора Редактор шаблонов, выберите **Файл > Открыть шаблон ресурсов**. Откроется диалоговое окно Открыть шаблон ресурсов.
2. Выберите из показанных в этой таблице шаблонов тот, который вы хотите использовать.
3. Нажмите кнопку **OK**, чтобы открыть этот шаблон. Если у вас в настоящее время был открыт другой шаблон в редакторе, при нажатии кнопки OK он будет оставлен, и будет показан шаблон, выбранный вами. Если вы внесли изменения в свои ресурсы и хотите сохранить ваши библиотеки для будущего

использования, можно опубликовать, обновить и совместно использовать их, прежде чем открыть другие. Дополнительную информацию смотрите в разделе “Совместное использование библиотек” на стр. 184.

## Сохранение шаблонов

В Редактор шаблонов можно сохранять изменения, внесенные в шаблон. При этом можно выбрать сохранение или под существующим, или под новым именем шаблона.

Если вы внесли изменения в шаблон, который ранее уже был загружен на узел, потребуется повторно загрузить содержимое этого шаблона, чтобы получить последние изменения. Дополнительную информацию смотрите в разделе “Копирование ресурсов из шаблонов и файлов TAP” на стр. 26.

Есть и другой вариант: когда вы используете опцию **Использовать сохраненную интерактивную работу** на вкладке Модель узла Исследование текстов, означающую использование ресурсов из предыдущего сеанса интерактивной инструментальной среды, нужно переключиться на ресурсы этого шаблона из сеанса интерактивной инструментальной среды. Дополнительную информацию смотрите в разделе “Переключение между шаблонами ресурсов” на стр. 168.

*Примечание*: Ваши библиотеки можно также опубликовать и сделать доступными для совместного использования. Дополнительную информацию смотрите в разделе “Совместное использование библиотек” на стр. 184.

Чтобы сохранить шаблон:

1. В меню Редактор шаблонов выберите опцию **Файл > Сохранить шаблон ресурсов**. Откроется диалоговое окно Сохранить шаблон ресурсов.
2. Введите новое имя в поле Шаблон, если вы хотите сохранить этот шаблон как новый. Выберите шаблон в таблице, если вы хотите переопределить существующий шаблон на текущие загруженные ресурсы.
3. При желании введите описание, чтобы показывать комментарий или аннотацию в таблице.
4. Нажмите кнопку **Сохранить**, чтобы сохранить шаблон.

**Важно!** Так как ресурсы из шаблонов или TAP загружаются/копируются на узел, необходимо изменить ресурсы, повторно загрузив их, если в шаблон внесены изменения и вы хотите воспользоваться преимуществами этих изменений в существующем потоке. Дополнительную информацию смотрите в разделе “Изменение ресурсов узла после загрузки”.

## Изменение ресурсов узла после загрузки

По умолчанию при добавлении узла в поток набор ресурсов из шаблона по умолчанию загружается и встраивается в ваш узел. Если вы изменяете ресурсы или используете TAP при их загрузке, копия этих ресурсов затем перезапишет существующие ресурсы. Так как шаблоны и TAP непосредственно с узлом не связаны, любые изменения шаблона или TAP не становятся автоматически доступными на уже существующем узле. Чтобы использовать преимущества этих изменений, необходимо изменить ресурсы на этом узле. Ресурсы можно изменить одним из двух способов.

Способ 1: повторная загрузка ресурсов на вкладке Модель

Если вы хотите изменить ресурсы на узле, используя новый или измененный шаблон или TAP, можно повторно загрузить их на вкладке Модель этого узла. В результате повторной загрузки вы замените копию ресурсов узла на более современную копию. Для удобства измененные дата и время появятся на вкладке Модель вместе с именем исходного шаблона. Дополнительную информацию смотрите в разделе “Копирование ресурсов из шаблонов и файлов TAP” на стр. 26.

Однако при работе с данными интерактивного сеанса на узле моделирования исследования текстов и при выборе опции **Использовать работу сеанса** на вкладке Модель будет использоваться сохраненная работа

сеанса и сохраненные ресурсы, а кнопка **Загрузить** будет отключена. Это отключение связано с тем, что в какой-то момент сеанса интерактивной инструментальной среды вы выбрали опцию **Изменить узел моделирования** и сохранили категории, ресурсы и другую работу сеанса. В этом случае при желании заменить или обновить эти ресурсы вы можете обратиться к другому способу переключения ресурсов в Редактор ресурсов.

Способ 2: переключение ресурсов в Редактор ресурсов

Всякий раз, когда в течение интерактивного сеанса вы хотите использовать другие ресурсы, можно поменять их при помощи диалогового окна Переключить ресурсы. Это особенно важно при желании повторно использовать существующую работу с категориями, но при этом заменить ресурсы. В таком случае можно выбрать опцию **Использовать работу сеанса** на вкладке Модель узла моделирования исследования текстов. При этом будет отключена возможность повторной загрузки шаблона через диалоговое окно узла, и вместо этого будут сохранены параметры и изменения, сделанные в течение вашего сеанса. После этого вы можете запустить сеанс интерактивной инструментальной среды, выполнив поток, и переключить ресурсы в Редактор ресурсов. Дополнительную информацию смотрите в разделе “Переключение между шаблонами ресурсов” на стр. 168.

Чтобы сохранить работу сеанса для последующих сеансов, в том числе ресурсы, нужно изменить узел моделирования из сеанса интерактивной инструментальной среды, чтобы ресурсы (и другие данные) были снова сохранены на узле. Дополнительную информацию смотрите в разделе “Обновление узлов моделирования и сохранение” на стр. 84.

*Примечание*: Если в течение интерактивного сеанса переключиться на контент другого шаблона, именем шаблона, перечисленным на узле, будет по-прежнему имя последнего загруженного и скопированного шаблона. Чтобы получить выигрыш от этих ресурсов или другой работы в сеансе, измените узел моделирования, прежде чем закрывать сеанс.

## Управление шаблонами

Существуют также некоторые основные управленческие задачи, которые вы можете хотеть время от времени выполнять на ваших шаблонах, такие как переименование шаблонов, импорт и экспорт шаблонов или удаление устаревших шаблонов. Эти задачи выполняются в диалоговом окне Управление шаблонами. Импорт и экспорт шаблонов позволяют использовать шаблоны совместно с другими пользователями. Дополнительную информацию смотрите в разделе “Импорт и экспорт шаблонов” на стр. 177.

*Примечание*: Вы не можете переименовать или удалить шаблоны, установленные (или поставленные) с этим программным продуктом. Вместо этого, если вы хотите переименовать такой шаблон, можно открыть установленный шаблон и создать новый шаблон с именем по вашему выбору. Можно удалять свои пользовательские шаблоны; однако при попытке удалить поставленный шаблон он будет возвращен к первоначально установленной версии.

Чтобы переименовать шаблон

1. В меню выберите **Ресурсы > Управление шаблонами ресурсов**. Откроется диалоговое окно Управление шаблонами.
2. Выберите шаблон, который вы хотите переименовать, и нажмите кнопку **Переименовать**. Поле имени становится доступным для редактирования в таблице.
3. Введите новое имя и нажмите клавишу Enter. Откроется диалоговое окно подтверждения.
4. Если вас устраивает изменение имени, щелкните по **Да**. Если нет, щелкните по **Нет**.

Чтобы удалить шаблон

1. В меню выберите **Ресурсы > Управление шаблонами ресурсов**. Откроется диалоговое окно Управление шаблонами.
2. В диалоговом окне Управлять шаблонами выберите шаблон, который вы хотите удалить.

3. Щелкните по **Удалить** . Откроется диалоговое окно подтверждения.
4. Нажмите кнопку **Да**, чтобы удалить шаблон, или нажмите кнопку **Нет**, чтобы отменить требование. Если вы нажали кнопку **Да**, шаблон удаляется.

## Импорт и экспорт шаблонов

Можно использовать шаблоны совместно с другими пользователями или компьютерами путем их импорта и экспорта. Шаблоны хранятся во внутренней базе данных, но их можно экспортировать как файлы **.lrt* на ваш жесткий диск.

Поскольку существуют обстоятельства, при которых вы можете захотеть импортировать или экспортировать шаблоны, существует несколько диалоговых окон, предлагающих эти возможности.

- Диалоговое окно Открыть шаблон в редакторе Редактор шаблонов
- Диалоговое окно Загрузить ресурсы в узле моделирования исследования текста и в узле Анализ текстовых связей.
- Диалоговое окно Управление шаблонами в редакторе Редактор шаблонов и в Редактор ресурсов.

Чтобы импортировать шаблон

1. В диалоговом окне щелкните по **Импорт**. Откроется диалоговое окно Импорт шаблона.
2. Выберите файл шаблона ресурсов (**.lrt*), который хотите импортировать, и нажмите кнопку **Импорт**. Можно сохранить импортируемый шаблон под другим именем или перезаписать существующий шаблон. Диалоговое окно закрывается, и шаблон появляется в таблице.

Экспортировать шаблон

1. В диалоговом окне выберите шаблон, который вы хотите экспортировать, и нажмите кнопку **Экспорт**. Откроется диалоговое окно Выбор каталога.
2. Выберите каталог, в который вы хотите экспортировать библиотеку, и нажмите кнопку **Экспорт**. Это диалоговое окно закрывается, и шаблон экспортируется с расширением имени файла (**.lrt*)

## Выход из редактора Редактор шаблонов

После завершения работы в редакторе Редактор шаблонов можно сохранить свою работу и выйти из редактора.

Чтобы выйти из редактора Редактор шаблонов

1. В меню выберите **Файл > Закрыть**. Откроется диалоговое окно Сохранить и Закрыть.
2. Выберите **Сохранить изменения в шаблоне**, чтобы сохранить открытый шаблон, прежде чем закрыть редактор.
3. Выберите **Опубликовать библиотеки**, если вы хотите опубликовать какую-либо из библиотек в открытом шаблоне, прежде чем закрыть редактор. При выборе этой опции вам будет предложено выбрать библиотеки для публикации. Дополнительную информацию смотрите в разделе “Публикация библиотек” на стр. 185.

## Резервное копирование ресурсов

Возможно, вы хотите время от времени выполнять резервное копирование своих ресурсов в качестве меры безопасности.

**Важно!** При восстановлении все содержимое ваших ресурсов будет полностью стерто, и в продукте станет доступно только содержимое файла резервной копии. Это относится и к любой открытой работе.

**Примечание:** Выполнять восстановление резервной копии можно только в той же самой основной версии вашего программного обеспечения. Например, если вы выполняете резервное копирование в версии 15, нельзя восстановить эту резервную копию в версии 16.

Чтобы создать резервную копию ресурсов

1. В меню выберите **Ресурсы > Инструменты резервного копирования > Создать резервную копию ресурсов**. Откроется диалоговое окно Резервное копирование.
2. Введите имя для файла резервной копии и нажмите кнопку **Сохранить**. Диалоговое окно закрывается, и создается файл резервной копии.

Чтобы восстановить ресурсы

1. В меню выберите **Ресурсы > Инструменты резервного копирования > Восстановить ресурсы**. Оповещение предупреждает, что при восстановлении будет перезаписано перезапишет текущее содержимое вашей базы данных.
2. Для продолжения нажмите кнопку **Да**. Откроется диалоговое окно.
3. Выберите файл, который вы хотите восстановить, и нажмите кнопку **Открыть**. Диалоговое окно закрывается, и ресурсы восстанавливаются в прикладной программе.

## Импорт файлов ресурсов

Если вы внесли изменения непосредственно в файлы ресурсов вне этого продукта, их можно импортировать в выбранную библиотеку, выбрав библиотеку и затем импорт. Кроме того, при импорте каталога можно импортировать все поддерживаемые файлы в конкретную открытую библиотеку. Импортировать можно только файлы **.txt*.

**Важно!** Для файлов японского языка необходимо, чтобы кодировка импортируемых файлов .txt была UTF8. Иначе импорт списков исключения для японского языка невозможен.

Каждый импортируемый файл должен содержать только одно значение на строке, а если содержимое структурировано одним из следующих способов:

- Список слов или словосочетаний (по одному на строку). Файл импортируется как список терминов для словаря типов, и словарь типов принимает имя файла без расширения.
- Список значений, например, *термин_1* <TAB> *термин_2*; такой импортируется как список синонимов, где *термин_1* - набор составных терминов, а *термин_2* - целевой термин.

Чтобы импортировать отдельный файл ресурсов

1. В меню выберите **Ресурсы > Импорт файлов > Импорт отдельного файла**. Откроется диалоговое окно Импорт файлов.
2. Выберите файл, который вы хотите импортировать, и нажмите кнопку **Импорт**. Содержимое файла преобразуется во внутренний формат и добавляется в вашу библиотеку.

Чтобы импортировать все файлы в каталоге

1. В меню выберите **Ресурсы > Импорт файлов > Импорт всего каталога**. Откроется диалоговое окно Импорт каталога.
2. Выберите библиотеку, в которую нужно импортировать все файлы ресурсов из списка **Импорт**. Если выбрать опцию **По умолчанию**, будет создана новая библиотека, в качестве имени которой будет использовано имя каталога.
3. Выберите библиотеку, из которой нужно импортировать файлы. Подкаталоги не считываются.
4. Щелкните по **Импорт**. Диалоговое окно закроется, и содержимое импортированных файлов ресурсов появится в редакторе в виде словарей и файлов расширенных ресурсов.

# Глава 16. Работа с библиотеками

Ресурсы, используемые механизмом извлечения для извлечения и группирования терминов из текстовых данных, всегда содержат одну или несколько библиотек. Набор библиотек можно увидеть в дереве библиотек, находящемся в верхней левой части Редактор шаблонов и Редактор ресурсов. В состав библиотек входят словари трех разновидностей: типов, подстановок и исключения. Дополнительную информацию смотрите в разделе Глава 17, “О словарях библиотек”, на стр. 189.

Шаблон ресурсов или ресурсы из TAP, выбираемые вами, содержат несколько библиотек, позволяющих немедленно начать извлечение понятий из текстовых данных. Однако вы можете создать свои собственные библиотеки, а также опубликовать их, чтобы эти библиотеки можно было использовать повторно. Дополнительную информацию смотрите в разделе “Публикация библиотек” на стр. 185.

К примеру допустим, что вы часто работаете с текстовыми данными, связанными с автомобильной промышленностью. После анализа данных вы решаете, что надо бы создать некоторые пользовательские ресурсы для обработки относящегося конкретно к этой промышленности словаря или сборника жаргонизмов. При помощи Редактор шаблонов можно создать новый шаблон, а в нем - библиотеку для извлечения и группирования автомобильных терминов. Поскольку вам снова потребуется информация в этой библиотеке, вы публикуете библиотеку в определенном репозитории, доступном в диалоговом окне **Управление библиотеками**, чтобы использовать ее повторно независимо в различных сеансах потока.

Предположим, что вы также заинтересованы в группировании терминов, относящихся конкретно к определенным подотраслям (таким как электронные устройства, механизмы, системы охлаждения) или даже к конкретному изготовителю или рынку. Вы можете создать библиотеку для каждой группы, а затем опубликовать созданные библиотеки, чтобы их можно было использовать повторно с несколькими наборами текстовых данных. Таким образом можно добавить библиотеки с наилучшим соответствием контексту используемых текстовых данных.

*Примечание*: На вкладке Дополнительные ресурсы можно сконфигурировать дополнительные ресурсы и управлять ими. Некоторые из них применяются ко всем библиотекам и управляют нелингвистическими объектами, нечетким группированием исключений и так далее. Дополнительно можно также можно отредактировать правила паттернов анализа текстовых связей (относящиеся к конкретным библиотекам) на вкладке Правила текстовых связей. Дополнительную информацию смотрите в разделе Глава 18, “О расширенных ресурсах”, на стр. 203.

## Поставляемые библиотеки

По умолчанию с IBM SPSS Modeler Text Analytics устанавливается несколько библиотек. Эти предварительно отформатированные библиотеки можно использовать для доступа к тысячам предопределенных терминов и синонимов, а также к многим различным типам. Эти поставляемые библиотеки точно настроены на несколько разных доменов и доступны на нескольких языках.

Существует много библиотек, но чаще всего используются следующие:

- **Локальная библиотека.** Используется для хранения определенных пользователем словарей. Это пустая библиотека, по умолчанию добавляемая ко всем ресурсам. Она содержит также пустой словарь типов. Эта библиотека наиболее полезна при внесении изменений или уточнений непосредственно в ресурсы (например, при добавлении слова или типа) из представлений Категории и понятия, Кластеры или Анализ текстовых связей. При этом такие изменения и уточнения автоматически сохраняются в первой из библиотек, перечисленных в дереве библиотек в Редактор ресурсов; по умолчанию это *Локальная библиотека*. Опубликовать эту библиотеку невозможно, так как она специфична для данных сеанса. Если вы хотите опубликовать содержимое этой библиотеки, ее сначала нужно переименовать.

- **Базовая библиотека.** Используется для большинства случаев, так как состоит из пяти встроенных базовых типов, представляющих людей, положения, организации, продукты и тип Неизвестный. Хотя вы можете увидеть только несколько терминов, перечисленных в одном из словарей типов, представленные в базовой библиотеке типы на самом деле дополняют устойчивые типы, находящиеся во внутренних скомпилированных ресурсах, поставляемых с вашим продуктом исследования текстов. Эти внутренние скомпилированные ресурсы содержат тысячи терминов для каждого типа. Из-за этого, хотя вы можете не увидеть какой-то термин в списке терминов словаря типов, он все равно будет извлекаться и ему будет присваиваться базовый тип. Это объясняет, как может извлекаться имя *George* и получать тип <Человек>, хотя в базовой библиотеке в словаре для типа <Человек> есть только имя *John*. Аналогично, даже не подключая базовую библиотеку, вы можете увидеть эти типы в результатах извлечения, так как механизм извлечения все же будет использовать скомпилированные ресурсы, содержащие эти типы.
- **Библиотека Мнения.** Чаще всего используется для извлечения из текстовых данных мнений и настроений. Эта библиотека включает в себя тысячи слов, представляющих позиции, спецификаторы и предпочтения, которые в случае совместного использования с другими терминами обозначают мнение о предмете обсуждения. В эту библиотеку входит большое количество встроенных типов, синонимов и исключений. Сюда же входит обширный набор правил паттернов, используемых для анализа текстовых связей. Чтобы воспользоваться преимуществами правил анализа текстовых связей из этой библиотеки и создаваемыми ими результатами паттернов, необходимо задать эту библиотеку на вкладке Правила текстовых связей. Дополнительную информацию смотрите в разделе Глава 19, “О правилах текстовых связей”, на стр. 215.
- **Библиотека Бюджет.** Используется для извлечения терминов, относящихся к стоимости чего-либо. Эта библиотека содержит множество слов и словосочетаний, представляющих прилагательные, спецификаторы и суждения относительно цены или качества чего-либо.
- **Библиотека Вариации.** Используется для учета тех случаев, когда некоторым языковым вариациям требуются определения синонимов для их правильной группировки. В эту библиотеку входят только определения синонимов.

Хотя некоторые из библиотек, поставляемые извне шаблонов, напоминают по содержанию некоторые шаблоны, эти шаблоны были специально настроены для конкретных прикладных программ и содержат дополнительные расширенные ресурсы. Рекомендуется использовать шаблон, разработанный специально для того типа текстовых данных, с которым вы работаете, и вносить свои изменения в эти ресурсы, а не просто добавлять отдельные библиотеки к более общему шаблону.

С IBM SPSS Modeler Text Analytics поставляются также скомпилированные ресурсы. Они всегда используются в процессе извлечения и содержат большое число определений, дополняющих встроенные словари типов в библиотеках по умолчанию. Так как эти ресурсы уже скомпилированы, их нельзя просмотреть или изменить. Однако вы можете принудительно ввести термин, тип которого определен этими скомпилированными ресурсами, в любую другую библиотеку. Дополнительную информацию смотрите в разделе “Принудительное назначение типов терминам” на стр. 195.

## Создание библиотек

Можно создать любое количество библиотек. После добавления новой библиотеки можно начать создавать словари типов в ней и вводить термины, синонимы и исключения.

Создание библиотеки

1. В меню выберите **Ресурсы > Создать библиотеку**. Откроется диалоговое окно Свойства библиотеки.
2. В текстовом поле Имя введите имя для библиотеки.
3. При необходимости введите комментарий в текстовое поле Аннотация.
4. Нажмите кнопку **Опубликовать**, если эту библиотеку нужно опубликовать до ввода в нее каких-либо данных. Дополнительную информацию смотрите в разделе “Совместное использование библиотек” на стр. 184. Библиотеку можно опубликовать и в любое время позднее.
5. Нажмите кнопку **OK**, чтобы создать библиотеку. Диалоговое окно закроется, и библиотека появится в представлении дерева. Если раскрыть библиотеки в дереве, вы увидите, что к библиотеке был

автоматически добавлен пустой словарь типов. В этом словаре можно сразу начать ввод терминов. Дополнительную информацию смотрите в разделе “Добавление терминов” на стр. 192.

## Добавление общедоступных библиотек

При желании повторно использовать библиотеку из данных другого сеанса можно добавить ее к вашим ресурсам, если только это общедоступная библиотека. **Общедоступная библиотека** - это библиотека, которая была опубликована. Дополнительную информацию смотрите в разделе “Публикация библиотек” на стр. 185.

**Важно!** Невозможно добавить библиотеку на японском языке к неяпонским ресурсам и наоборот.

При добавлении общедоступной библиотеки ее **локальная** копия встраивается в данные вашего сеанса. В эту библиотеку можно вносить изменения; однако необходимо повторно опубликовать эту версию библиотеки, если вы хотите сделать изменения доступными для совместного использования.

При добавлении общедоступной библиотеки может открыться диалоговое окно Разрешение конфликтов, если обнаружены какие-то конфликты между терминами и типами этой библиотеки и других локальных библиотек. Для завершения данной операции нужно разрешить эти конфликты или принять предлагаемые разрешения. Дополнительную информацию смотрите в разделе “Устранение конфликтов” на стр. 186.

*Примечание*: Если вы всегда обновляете свои библиотеки при запуске интерактивного сеанса инструментальной среды или публикации во время его закрытия , вы реже будете сталкиваться с рассинхронизацией библиотек. Дополнительную информацию смотрите в разделе “Совместное использование библиотек” на стр. 184.

Чтобы добавить библиотеку:

1. Выберите опцию меню **Ресурсы > Добавить библиотеку**. Откроется диалоговое окно Добавить библиотеку.
2. Выберите одну или несколько библиотек в списке.
3. Нажмите кнопку **Добавить** . Если возникнут какие-то конфликты между добавляемой и ранее существовавшими библиотеками, появится запрос на верификацию разрешений конфликтов или на их изменение до завершения операции. Дополнительную информацию смотрите в разделе “Устранение конфликтов” на стр. 186.

## Поиск терминов и типов

Поиск на различных панелях можно выполнить в редакторе при помощи возможности Найти. В меню редактора можно выбрать **Изменить > Найти**, и появится панель инструментов возможности Найти. С помощью этой панели инструментов за один раз можно найти одно вхождение. Повторным нажатием кнопки **Найти** можно найти последующие вхождения искомого термина.

При поиске редактор его выполняет только в библиотеках, выводящихся в выпадающем списке панели инструментов возможности Найти. Если выбрать опцию **Все библиотеки**, программа выполнит поиск во всём, что есть в редакторе.

При запуске поиска он начинается в области, на которую наведен фокус. Поиск продолжается по всем разделам, замыкая свой цикл возвратом в активную ячейку. При помощи стрелок с направлениями можно изменить порядок поиска. Можно также выбрать, учитывать ли при поиске регистр.

Чтобы найти строки в представлении:

1. В меню выберите **Изменить > Найти**. Откроется панель инструментов Поиск.
2. Введите строку, которую вы хотите искать.
3. Нажмите кнопку **Найти** для запуска поиска. Будет выделено следующее вхождение термина или типа.

4. Нажмите кнопку снова для перехода от одного вхождения к следующему.

## Просмотр библиотек

Можно вывести содержимое отдельной библиотеки или всех библиотек. Это может оказаться полезным, если вы работаете с множеством библиотек или хотите просмотреть содержимое библиотеки перед её публикацией. Изменение этого представления влияет только на то, что выводится на вкладке Ресурсы библиотек, но но не отключает поддержку использования никаких библиотек во время извлечения. Дополнительную информацию смотрите в разделе “Отключение локальных библиотек”.

Представление по умолчанию - **Все библиотеки**, где выводятся все библиотеки в дереве и их содержимое, представленное на других панелях. Этот вариант выбора можно изменить при помощи выпадающего списка на панели инструментов или посредством выбора меню (**Вид > Библиотеки**). При просмотре одной библиотеки в представлении все элементы в других библиотеках исчезают, но при извлечении они по-прежнему будут читаться.

Чтобы изменить представление библиотек:

1. В меню на вкладке Ресурсы библиотек выберите **Вид > Библиотеки**. Откроется меню со всеми локальными библиотеками.
2. Выберите библиотеку, которую вы хотите увидеть, или опцию **Все библиотеки**, чтобы увидеть содержимое всех библиотек. Содержимое представления будет отфильтровано в соответствии с вашим выбором.

## Управление локальными библиотеками

В отличие от общедоступных библиотек, локальные библиотеки - это библиотеки в вашем сеансе интерактивной инструментальной среды или в шаблоне. Дополнительную информацию смотрите в разделе “Управление общедоступными библиотеками” на стр. 183. Существует также несколько базовых задач управления локальными библиотеками, выполнение которых может потребоваться, в том числе переименование, отключение и удаление локальной библиотеки.

## Переименование локальных библиотек

Локальные библиотеки можно переименовать. Если переименовать локальную библиотеку, она больше не будет связана с общедоступной версией, если такая существует. Это означает, что последующие изменения больше не будут доступны для совместного использования через общедоступную версию. Можно повторно опубликовать эту локальную библиотеку с новым именем. Это означает также, что вы не сможете внести изменения, выполненные в локальной версии, в исходную версию общедоступной библиотеки.

*Примечание*: переименовать общедоступную библиотеку невозможно.

1. В меню выберите **Изменить > Свойства библиотеки**. Откроется диалоговое окно Свойства библиотеки.

Чтобы переименовать локальную библиотеку:

1. В представлении дерева выберите библиотеку, которую вы хотите переименовать.
2. В текстовом поле Имя введите новое имя для библиотеки.
3. Нажмите кнопку **OK**, чтобы принять новое имя для библиотеки. Диалоговое окно закроется, и имя библиотеки в представлении дерева изменится.

## Отключение локальных библиотек

Если нужно временно исключить библиотеку из процесса извлечения, можно выключить переключатель слева от имени библиотеки в представлении дерева. Это будет означать, что вы хотите сохранить библиотеку, но ее контент будет игнорироваться при проверке конфликтов и в процессе извлечения.

Чтобы отключить библиотеку:

1. На панели дерева библиотек выберите библиотеку, которую вы хотите отключить.
2. Нажмите клавишу пробела. Переключатель слева от имени библиотеки будет выключен.

## Удаление локальных библиотек

Можно удалить библиотеку, не удаляя ее общедоступную версию, и наоборот. При удалении локальной библиотеки она и все ее содержимое будут удалены только из сеанса. Удаление локальной версии библиотеки не приводит к удалению ее из других сеансов или к удалению ее общедоступной версии. Дополнительную информацию смотрите в разделе “Управление общедоступными библиотеками”.

Чтобы удалить локальную библиотеку:

1. В представлении дерева выберите библиотеку, которую вы хотите удалить.
2. В меню выберите **Изменить > Удалить**, чтобы удалить библиотеку. Библиотека будет удалена.
3. Если вы никогда прежде не публиковали эту библиотеку, появится сообщение с запросом, удалить или сохранить эту библиотеку. Нажмите кнопку **Удалить**, чтобы продолжить удаление, или кнопку **Сохранить**, если нужно сохранить эту библиотеку.

*Примечание*: одна библиотека должна всегда оставаться.

---

# Управление общедоступными библиотеками

Чтобы повторно использовать локальные библиотеки, их можно опубликовать, и после этого с этими библиотеками можно работать, а просмотреть их можно в диалоговом окне Управление библиотеками (**Ресурсы > Управление библиотеками**). Дополнительную информацию смотрите в разделе “Совместное использование библиотек” на стр. 184. Некоторые основные задачи управления общедоступными библиотеками, выполнение которых может потребоваться, включают в себя импорт, экспорт и удаление общедоступных библиотек. Переименовать общедоступную библиотеку невозможно.

Импорт общедоступных библиотек

1. В диалоговом окне Управление библиотеками нажмите кнопку **Импорт...**. Откроется диалоговое окно Импорт библиотеки.
2. Выберите файл библиотеки (**.lib*), которую вы хотите импортировать, а если нужно добавить ее и в число локальных библиотек, выберите опцию **Добавить библиотеку в текущий проект**.
3. Щелкните по **Импорт**. Закроется диалоговое окно. Если общедоступная библиотека с таким именем уже существует, появится запрос на переименование импортируемой библиотеки или на перезапись существующей общедоступной библиотеки.

Экспорт общедоступных библиотек

Общедоступные библиотеки можно экспортировать в формат *.lib*, чтобы они были доступны для совместного использования.

1. В диалоговом окне Управление библиотеками выберите в списке библиотеку, которую вы хотите экспортировать.
2. Щелкните по **Экспорт**. Откроется диалоговое окно Выбор каталога.
3. Выберите каталог, в который вы хотите экспортировать библиотеку, и нажмите кнопку **Экспорт**. Диалоговое окно закроется, и файл библиотеки (**.lib*) будет экспортирован.

Удаление общедоступных библиотек

Локальную библиотеку можно удалить, не удаляя общедоступную версию этой библиотеки, и наоборот. Однако если удалить библиотеку в этом диалоговом окне, ее больше нельзя будет добавить к каким-либо ресурсам сеанса, пока локальная версия не будет снова опубликована.

При удалении библиотеки, установленной с продуктом, будет восстановлена исходная установленная версия.

1. В диалоговом окне Управление библиотеками выберите библиотеку, которую вы хотите удалить. Список библиотек можно отсортировать, щелкая по соответствующим заголовкам.
2. Нажмите кнопку **Удалить**, чтобы удалить библиотеку. IBM SPSS Modeler Text Analytics верифицирует совпадение локальной версии с общедоступной библиотекой. Если это так, библиотека удаляется без оповещения. Если версии библиотек различаются, откроется оповещение с вопросом, что делать с общедоступной версией - сохранять или удалять.

## Совместное использование библиотек

Библиотеки позволяют работать с ресурсами таким способом, который легко использовать совместно в нескольких сеансах интерактивной инструментальной среды. Библиотеки могут существовать в двух состояниях (или версиях). Библиотеки, которые представляют собой часть сеанса интерактивной инструментальной среды и доступны для изменений, называются **локальными библиотеками**. При работе в сеансе интерактивной инструментальной среды вы можете внести множество изменений, например, в библиотеку *Vegetables* (Овощи). Если ваши изменения могут оказаться полезными при работе с другими данными, их можно сделать доступными, создав версию **общедоступной библиотеки** для вашей библиотеки *Овощи*. Как следует из ее названия, общедоступная библиотека может использоваться в любых других ресурсах любого сеанса интерактивной инструментальной среды.

Список общедоступных библиотек можно просмотреть в диалоговом окне Управление библиотеками. Если общедоступная версия некоторой библиотеки существует, ее можно добавить к ресурсам в другом контексте, так чтобы эти пользовательские лингвистические ресурсы можно было использовать совместно.

Изначально поставляемые библиотеки - это общедоступные библиотеки. Можно изменить ресурсы в таких библиотеках и затем создать новую общедоступную версию. Позднее к этим новым версиям можно обратиться в других сеансах интерактивной инструментальной среды.

По ходу работы с вашими библиотеками и внесения изменений версии библиотек становятся рассинхронизированными. В некоторых случаях локальная версия может оказаться более новой, чем общедоступная, а в других случаях наоборот. Может оказаться также, что и в локальной, и в общедоступной версии есть изменения, которых нет в в другой версии, если общедоступная версия была изменена из другого сеанса интерактивной инструментальной среды. Если ваши версии библиотек окажутся рассинхронизированными, их можно синхронизировать снова. Синхронизация версий библиотек состоит из повторной публикации и/или изменения локальных библиотек.

При всяком запуске или закрытии сеанса интерактивной инструментальной среды вы получите предложение синхронизировать все библиотеки, для чего требуется их изменение или повторная публикация. Кроме этого, состояние синхронизации вашей локальной библиотеки легко определить по значку, появляющемуся у имени библиотеки в представлении дерева, или просмотрев диалоговое окно Свойства библиотек. Синхронизацию можно выполнить и в любое другое время, выбрав соответствующую опцию меню. В следующей таблице представлено пять различных состояний синхронизации и связанные с ними значки.

*Таблица 37. Состояния синхронизации локальной библиотеки.*

| Значок | Описание состояния локальной библиотеки |
|---|---|
|  | Не опубликовано - локальная библиотека никогда не была опубликована. |
|  | Синхронизировано - локальная и общедоступная версии библиотеки идентичны. Это применимо также к *Локальной библиотеке*, которую нельзя опубликовать, так как она предназначена для содержания только тех ресурсов, которые относятся к конкретному сеансу. |
|  | Устарело - общедоступная версия библиотеки более новая, чем локальная версия. Вы можете обновить свою локальную версию новыми изменениями. |
|  | Более новая - локальная версия библиотеки более новая, чем общедоступная версия. Вы можете повторно опубликовать свою локальную версию в общедоступную версию. |

*Таблица 37. Состояния синхронизации локальной библиотеки (продолжение).*

| Значок | Описание состояния локальной библиотеки |
|---|---|
|  | Рассинхронизировано - и в локальной, и в общедоступной версии есть изменения, которых нет в другой версии. Необходимо решить, что именно делать - изменять свою локальную версию, или повторно ее публиковать. В случае изменения вы потеряете все внесенные правки с момента последнего изменения или публикации. Если выбрать публикацию, будут потеряны изменения, внесенные в общедоступную версию. |

*Примечание*: если вы всегда обновляете свои библиотеки при запуске сеанса интерактивной инструментальной среды или публикуете их при закрытии сеанса, появление рассинхронизированных библиотек менее вероятно.

Повторно опубликовать библиотеку можно всякий раз, когда вы сочтете, что изменения в библиотеке окажутся полезны для других потоков, которые могут также содержать данную библиотеку. После этого, если есть выигрыш от этих изменений для других потоков, можно изменить локальные версии в этих потоках. Таким образом можно создать потоки для каждого контекста или домена, применимые к вашим данным, создавая новые библиотеки и/или добавляя любое количество общедоступных библиотек к вашим ресурсам.

Если общедоступная версия библиотеки используется совместно, увеличивается вероятность появления различий между локальной и общедоступной версиями. При всяком запуске или закрытия сеанса интерактивной инструментальной среды, при публикации из этого сеанса или при открытии/закрытии шаблона из Редактор шаблонов появится сообщение, позволяющее опубликовать и/или изменить любые библиотеки, версии которых не синхронизированы с версиями в диалоговом окне Управление библиотеками. Если общедоступная версия библиотеки более новая, чем локальная, появится диалоговое окно с вопросом, не хотите ли вы выполнить обновление. Можно выбрать опцию сохранения локальной версии, как есть, вместо замены на общедоступную версию или слияния изменений в локальную версию.

## Публикация библиотек

Если вы никогда не публиковали какую-то конкретную библиотеку, ее публикация повлечет за собой создание общедоступной копии вашей локальной библиотеки в базе данных. Если вы публикуете библиотеку повторно, содержимое локальной библиотеки заменит содержимое существующей общедоступной версии. После повторной публикации эту библиотеку можно изменить в любых других сеансах потока, чтобы их локальные версии были синхронизированы с общедоступной версией. Хотя вы можете опубликовать библиотеку, ее локальная версия всегда будет сохраняться в сеансе.

**Важно!** Если вы внесли изменения в свою локальную библиотеку и со временем ее общедоступная версия также была изменена, эта библиотека будет рассматриваться как рассинхронизированная. Рекомендуется начинать работу с обновления локальной версии с учетом изменений общедоступной версии, затем вносить необходимые изменения, а в заключение снова опубликовать локальную версию, чтобы обе версии были идентичными. Если сначала внести изменения и выполнить публикацию, все изменения в общедоступной версии будут перезаписаны.

Чтобы опубликовать локальные библиотеки в базе данных:

1. В меню выберите **Ресурсы > Опубликовать библиотеки**. Откроется диалоговое окно Опубликовать библиотеки с выбранными по умолчанию библиотеками, которые нужно опубликовать.
2. Включите переключатель слева от каждой библиотеки, которую нужно опубликовать или повторно опубликовать.
3. Нажмите кнопку **Опубликовать**, чтобы опубликовать эти библиотеки в базе данных управления библиотеками.

## Обновление библиотек

При запуске или закрытии интерактивного сеанса инструментальной среды можно обновить или опубликовать любые библиотеки, которые утратили синхронность с опубликованными версиями. Если

опубликованная версия библиотеки новее локальной версии, откроется диалоговое окно с вопросом, желательно ли обновить открываемую библиотеку. Можно оставить локальную версию, не обновляя ее до опубликованной, или заменить локальную версию на опубликованную. Если опубликованная версия библиотеки новее локальной, можно обновить локальную версию, чтобы синхронизировать ее содержимое с содержимым опубликованной версии. Обновление означает внесение изменений, найденных в опубликованной версии, в локальную версию.

*Примечание*: Если вы всегда обновляете свои библиотеки при запуске интерактивного сеанса инструментальной среды или публикации во время его закрытия , вы реже будете сталкиваться с рассинхронизацией библиотек. Дополнительную информацию смотрите в разделе “Совместное использование библиотек” на стр. 184.

Чтобы обновить локальные библиотеки

1. В меню выберите **Ресурсы > Изменить библиотеки**. Откроется диалоговое окно Обновить библиотеки, в котором все библиотеки, которые нуждаются в обновлении, по умолчанию выбраны.
2. Включите переключатель слева от каждой библиотеки, которую нужно опубликовать или повторно опубликовать.
3. Нажмите кнопку **Обновить**, чтобы обновить локальные библиотеки.

## Устранение конфликтов

Конфликты локальных библиотек с опубликованными

При запуске сеанса потока IBM SPSS Modeler Text Analytics выполняет сравнение локальных библиотек с библиотеками, входящими в список в диалоговом окне Работа с библиотеками. Если локальные библиотеки в вашем сеансе не синхронизированы с опубликованными версиями, откроется диалоговое окно Предупреждение о синхронизации библиотек. Доступны следующие варианты для выбора нужных версий:

- **Все библиотеки, локальные по отношению к файлу.** Эта опция оставляет все локальные библиотеки как есть. Их можно заново опубликовать или обновить позже.
- **Все опубликованные библиотеки на этом компьютере.** Эта опция заменит выведенные локальные библиотеки на версии, найденные в базе данных.
- **Все новейшие библиотеки.** Эта опция заменит старые локальные библиотеки на более новые версии библиотек, опубликованные в базе данных.
- **Другое.** При помощи этой опции можно вручную выбрать нужные версии в таблице.

Конфликты между принудительными терминами

При добавлении опубликованной библиотеки или обновлении локальной библиотеки могут обнаружиться конфликты и дублирующие значения с участием, с одной стороны, терминов и типов в этой библиотеке и, с другой стороны, терминов и типов в других библиотеках в ваших ресурсах. В таких случаях вас просят подтвердить или изменить предлагаемые способы устранения конфликтов, после чего операция будет выполнена диалоговом окне Редактировать принудительно заданные термины. Дополнительную информацию смотрите в разделе “Принудительное назначение типов терминам” на стр. 195.

Диалоговое окно Редактировать принудительно заданные термины содержит все пары конфликтующих терминов или типов. Для наглядности конфликтующие пары отделены друг от друга чередующимися цветами фона. Цвета можно изменить в диалоговом окне Опции. Дополнительную информацию смотрите в разделе “Опции: вкладка Дисплей” на стр. 82. Диалоговое окно Редактировать принудительно заданные термины содержит две вкладки:

- **Дубликаты.** Эта вкладка содержит совпадающие термины, найденные в библиотеках. Если рядом с термином выведен значок канцелярской кнопки, это значит, что данное вхождение термина - принудительное. Если выведен значок черное X, это значит, что данное вхождение термина будет проигнорировано при извлечении, поскольку он был принудительно задан в другом месте.

- **Определяемое пользователем.** Эта вкладка содержит список всех терминов, заданных принудительно вручную на панели терминов словаря типов, а не при конфликтах.

*Примечание*: Диалоговое окно Редактировать принудительно заданные термины открывается после того, как вы добавили или обновили библиотеку. Если закрыть это окно командой отмены, вы не отмените добавление или обновление в библиотеке.

Чтобы устранить конфликты

1. В диалоговом окне Редактировать принудительно заданные термины в столбце Использовать включите радиокнопку для того термина, который хотите задать принудительно.
2. Завершив работу, нажмите кнопку **OK**, чтобы применить принудительно заданные термины и закрыть диалоговое окно. Если нажать **Отмена**, будут отменены все изменения, внесенные в этом диалоговом окне.

# Глава 17. О словарях библиотек

Ресурсы, используемые для извлечения текстовых данных, хранятся в форме шаблонов и библиотек. В состав библиотеки может входить три словаря.

- **Словарь типов** содержит собрание терминов, сгруппированных под одной меткой или именем типа. Когда механизм извлечения читает ваши текстовые данные, он сравнивает найденные в тексте слова с терминами, определенными в словарях типов. Во время извлечения слова - начальные формы терминов и синонимов типа - группируются вокруг целевого термина, называемого понятием. Извлеченные понятия назначаются словарю типа, в котором они появляются как термины. Вы можете управлять словарями типов на верхней левой и центральной панелях редактора - на панели дерева библиотек и на панели терминов. Дополнительную информацию смотрите в разделе “Словари типов”.
- **Словарь подстановок** содержит собрание слов, определенных как синонимы или как необязательные элементы, которые используются для группировки подобных терминов вокруг одного целевого термина, называемого понятием, в конечных результатах извлечения. Управлять словарями подстановок можно на нижней левой панели редактора при помощи вкладки Синонимы и вкладки Дополнительно. Дополнительную информацию смотрите в разделе “Словари подстановок/синонимов” на стр. 196.
- **Словарь исключения** содержит собрание терминов и типов, которые будут удалены из конечных результатов извлечения. Словарями исключения можно управлять на самой правой панели редактора. Дополнительную информацию смотрите в разделе “Словари исключения” на стр. 200.

Дополнительную информацию смотрите в разделе Глава 16, “Работа с библиотеками”, на стр. 179.

## Словари типов

В состав **словаря типов** входит имя типа (или метка) и список терминов. Элементы управления словарями типов находятся на верхней левой и центральной панелях вкладки Ресурсы библиотек в редакторе. Вы можете вызвать это представление, выбрав **Вид > Редактор ресурсов** в меню, если у вас запущен интерактивный сеанс инструментальной среды. Другой вариант - редактировать словари для определенного шаблона в Редактор шаблонов.

Когда механизм извлечения считывает текстовые данные, он сравнивает найденные в тексте слова с терминами, определенными в ваших словарях типов. Термины - это слова или словосочетания в словарях типов в ваших лингвистических ресурсах.

При совпадении слова с термином ему назначается имя типа для этого термина. При считывании ресурсов во время извлечения термины, которые были найдены в тексте, проходят несколько этапов обработки перед тем, как они станут понятиями на панели Результаты извлечения. Если механизм извлечения определяет несколько терминов, принадлежащих к одному и тому же словарю типов, претендующих на роль синонимов, они группируются под чаще всего встречающимся термином, и на панели Результаты извлечения они называются **понятиями**. Например, если термины `question` и `query` могут появиться под понятием с именем `question` в конце.

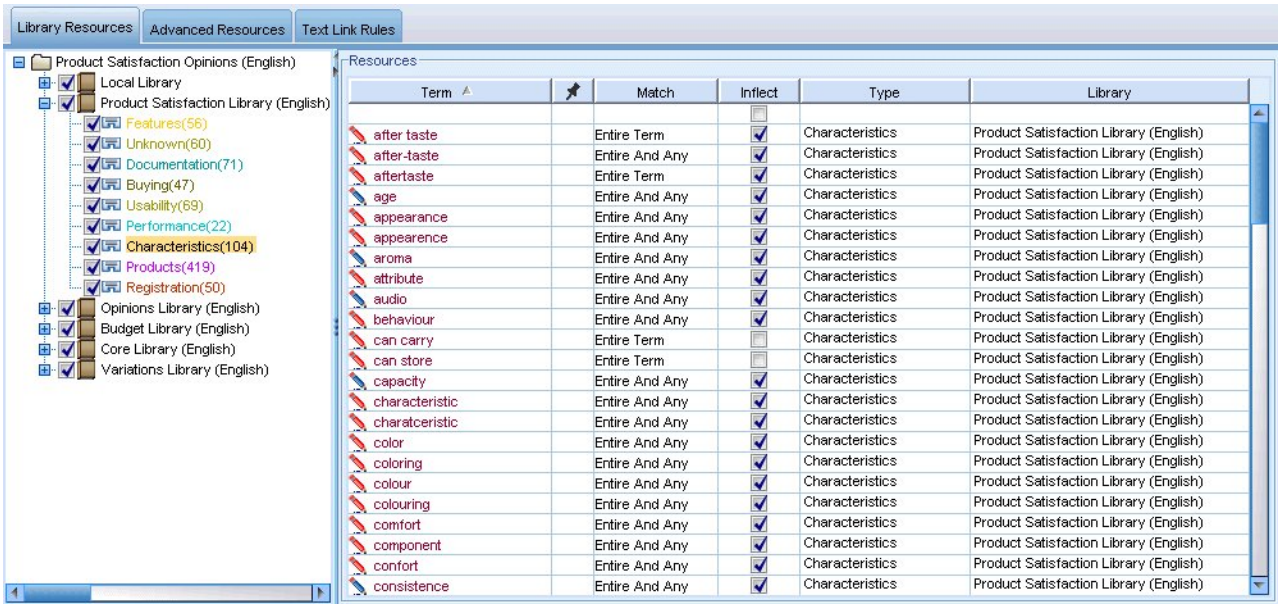



*Рисунок 40. Дерево библиотек и панель терминов*

Список словарей типов выводится на панели дерева библиотек в левой ее части. Содержимое каждого словаря типов выводится на центральной панели. В состав словарей типов входит не только список терминов. Способ, которым слова и словосочетания в текстовых данных сопоставляются с терминами определенными в словарях типов, определяется задаваемой опцией сопоставления. **Опция сопоставления** указывает, как термин привязывается по отношению к слову или словосочетанию - кандидату в текстовых данных. Дополнительную информацию смотрите в разделе “Добавление терминов” на стр. 192.

*Примечание*: К текстовым данным на японском языке применяются не все опции (из таких как опция сопоставления и флективные формы).

Кроме того, термины в словаре типов можно расширить, указав, хотите ли вы генерировать и добавлялись в словарь типов флективные формы автоматически. Путем генерирования флективных форм вы автоматически добавляете в словарь типов для терминов в форме единственного числа формы множественного числа и имена прилагательные. Дополнительную информацию смотрите в разделе “Добавление терминов” на стр. 192.

*Примечание*: Для большинства языков понятия, которые не были найдены ни в одном словаре типов, но были извлечены из текста, автоматически получают тип `<Неизвестный>`

## Встроенные типы

IBM SPSS Modeler Text Analytics поставляется с набором лингвистических ресурсов в форме пересылаемых библиотек и скомпилированных ресурсов. Поставляемые библиотеки содержат набор встроенных словарей типов, таких как `<Location>`, `<Organization>`, `<Person>` и `<Product>`.

*Примечание*: Набор встроенных типов по умолчанию для текста на японском языке отличается.

Эти словари типов используются механизмом извлечения для назначения типов (таких как тип `<Location>`, назначаемый понятию `paris`) извлекаемым им понятиям. Хотя во встроенных словарях типов и определено большое число терминов, но они не охватывают всех возможностей. Поэтому к ним можно добавить термины или создать свои собственные. Описание содержимого конкретного поставляемого словаря смотрите в аннотации в диалоговом окне Свойства типов. Выберите тип в дереве и выберите **Изменить > Свойства** в контекстном меню.

*Примечание*: Помимо поставляемых библиотек скомпилированные ресурсы (также используемые механизмом извлечения) содержат большое число определений, дополняющих встроенные библиотеки типов, но содержимое их в продукте невидимо. Однако термин, типизированный скомпилированными словарями, можно вставить в любой другой словарь. Дополнительную информацию смотрите в разделе “Принудительное назначение типов терминам” на стр. 195.

## Создание типов

Вы можете создать словари типов, помогающие сгруппировать схожие термины. Если во время процесса извлечения будут обнаружены термины, содержащиеся в таком словаре, им будет назначено имя содержащегося в словаре типа, и они будут извлечены под именем соответствующего понятия. При каждом создании библиотеки в ее состав всегда включается пустая библиотека типов, чтобы можно было сразу же приступить к вводу терминов.

**Важно!**: Для ресурсов на японском языке создать новые типы невозможно.

Если вы анализируете текстовые данные о еде и хотите сгруппировать термины, относящиеся к овощам, можно создать свой собственный словарь типов `<Овощи>`. Затем туда можно добавить такие термины, как `морковь`, `брокколи` и `шпинат`, если вы считаете, что это важные термины, которые будут встречаться в текстовых данных. Тогда во время извлечения в случае нахождения каких-либо из этих терминов они будут извлечены как понятия, и им будет присвоен тип `<Овощи>`.

Определять каждую форму слова или выражения не требуется, поскольку можно выбрать генерирование флективных форм терминов. Если выбрать эту опцию, механизм извлечения будет автоматически распознавать формы единственного или множественного числа терминов в числе других форм, принадлежащих к этому типу. Эта опция особенно полезна, если тип содержит в основном существительные, поскольку флективные формы глаголов или прилагательных вряд ли будут нужны.

Откроется диалоговое окно Свойства типов, содержащее следующие поля.

**Имя.** Имя, присваиваемое создаваемому вами словарю типов. Мы рекомендуем вам не использовать в именах типов пробелы, особенно если несколько имен типов начинаются с одного и того же слова.

*Примечание*: На имена типов и специальные символы накладываются некоторые ограничения. Например, не используйте специальные символы, такие как "@" или "!", в самом имени.

**Сопоставление по умолчанию.** Атрибут сопоставления по умолчанию указывает механизму извлечения, как сопоставлять данный термин с текстовыми данными. Каждому добавлению термина в этот словарь сопутствует автоматическое добавление в него атрибута сопоставления. Выбранный вариант сопоставления всегда можно изменить в списке терминов вручную. В состав опций входят: **Весь термин**,**В начале**, **В конце**, **В любом месте**, **В начале и В конце**, **Весь и В начале**, **Весь и В конце**, **Весь и (В начале и В конце)** и **Весь (не составной)**. Дополнительную информацию смотрите в разделе “Добавление терминов” на стр. 192. К ресурсам на японском языке эта опция не применяется.

**Добавить в.** В этом поле указывается библиотека, в которой будет создан ваш новый словарь типов.

**Генерировать флективные формы по умолчанию.** Эта опция указывает механизму извлечения использовать грамматическую морфологию для захвата и группирования добавляемых вами в этот словарь схожих форм терминов, таких как формы единственного или множественного числа. Эта опция особенно полезна, если тип содержит в основном существительные. При выборе этой опции всем добавляемым в этот тип новым терминам она будет присваиваться автоматически, хотя в этом списке ее можно изменить вручную. К ресурсам на японском языке эта опция не применяется.

**Цвет шрифта.** Это поле позволяет отличить результаты этого типа от других в интерфейсе. Кроме того, если выбрать **Использовать родительский цвет**, для этого словаря типов будет использоваться цвет типа по

умолчанию. Этот цвет по умолчанию задается в диалоговом окне опций. Дополнительную информацию смотрите в разделе “Опции: вкладка Дисплей” на стр. 82. Если выбрать **Пользовательский**, надо будет выбрать цвет в выпадающем списке.

**Аннотация.** Это поле - необязательное; оно может использоваться для любых комментариев или описаний.

Чтобы создать словарь типов:

1. Выберите библиотеку, в которой вам хотелось бы создать новый словарь типов.
2. В меню выберите **Сервис > Создать тип**. Откроется диалоговое окно Свойства типа.
3. Введите имя словаря типов в текстовом поле **Имя** и выберите нужные опции.
4. Нажмите кнопку **OK**, чтобы создать словарь типов. Новый словарь типов станет видим на панели дерева библиотек и появится на центральной панели. Можно сразу же приступить к добавлению типов. Дополнительную информацию смотрите в разделе “Добавление терминов”.

*Примечание*: Эти инструкции показывают, как вносить изменения в представление Редактор ресурсов или в Редактор шаблонов. Имейте в виду, что вы можете выполнять такую тонкую настройку непосредственно из панели Результаты извлечения, панели Данные, панели Категории или диалогового окна Определения кластеров в прочих представлениях. Дополнительную информацию смотрите в разделе “Уточнение результатов извлечения” на стр. 95.

## Добавление терминов

На панели дерева библиотек выводятся библиотеки, и ее можно развернуть, чтобы вывести словари типов, которые содержат библиотеки. На центральной панели в списке терминов выводятся термины в выбранной библиотеке или словаре типов в зависимости варианта, выбранного в дереве.

**Важно!** Для ресурсов на японском языке термины определяются по-другому.

В Редактор ресурсов можно добавить термины в словарь типов непосредственно на панели терминов или при помощи диалогового окна Добавить новые термины. Добавляемые термины могут быть как отдельными, так и составными словами. В начале списка всегда находится пустая строка, позволяющая добавить новый термин.

*Примечание*: Эти инструкции показывают, как вносить изменения в представление Редактор ресурсов или в Редактор шаблонов. Имейте в виду, что вы можете выполнять такую тонкую настройку непосредственно из панели Результаты извлечения, панели Данные, панели Категории или диалогового окна Определения кластеров в прочих представлениях. Дополнительную информацию смотрите в разделе “Уточнение результатов извлечения” на стр. 95.

Столбец Термин

В этом столбце введите в ячейку отдельные или составные слова. Цвет, которым будет представлен термин, зависит от цвета для типа, в котором хранится или куда вставляется этот термин. Цвет типов можно изменить в диалоговом окне Свойства типа. Дополнительную информацию смотрите в разделе “Создание типов” на стр. 191.

Столбец Принудительно

В этом столбце, если щелкнуть по значку канцелярской кнопки и поместить его в эту ячейку, механизм извлечения будет знать, что все остальные вхождения этого же термина в других библиотеках следует игнорировать. Дополнительную информацию смотрите в разделе “Принудительное назначение типов терминам” на стр. 195.

Столбец Сопоставление

В этом столбце выберите опцию сопоставления, чтобы указать механизму извлечения, как сопоставлять данный термин с текстовыми данными. Примеры смотрите в таблице. Значение по умолчанию можно изменить, отредактировав свойства типа. Дополнительную информацию смотрите в разделе “Создание типов” на стр. 191. В меню выберите **Изменить > Изменить соответствие**. Приведённые ниже опции сопоставления - базовые, поскольку возможны также и их сочетания:

- **Начало**. Этот тип назначается, если термин в словаре совпадает с первым словом в понятии, извлекаемом из текстовых данных. Например, если ввести `apple` (яблоки), ему будет соответствовать `apple tart` (яблочный пирог).
- **В конце**. Этот тип назначается, если термин в словаре совпадает с последним словом в понятии, извлекаемом из текстовых данных. Например, если ввести `apple` (яблоки), ему будет соответствовать `cider apple` (яблоки на сидр).
- **В любом месте**. Этот тип назначается, если термин в словаре совпадает с любым словом понятия, извлекаемого из текстовых данных. Например, если ввести `apple` (яблоки), опция **В любом месте** типизирует `apple tart` (яблочный пирог), `cider apple` (яблоки на сидр) и `cider apple tart` (пирог из яблок на сидр) одним и тем же способом.
- **Весь термин**. Этот тип назначается, если всё понятие, извлеченное из текстовых данных, совпадает с точным термином в словаре. Если добавить термин с применением опций **Весь термин**, **Весь и В начале**, **Весь и В конце**, **Весь и В любом месте** или **Весь (не составной)**, термин будет извлечен.

  Кроме того, поскольку тип `<Person>` извлекает только двухчастные имена, такие как *edith piaf* (Эдит Пиаф) или *mohandas gandhi* (Махатма Ганди), возможно вы захотите добавить в этот словарь типов только имена явно, если вам надо будет извлечь имя, упоминаемое без фамилии. Например, если вы хотите отловить все экземпляры *edith* в качестве имени, имя `edith` нужно добавить в тип `<Person>` при помощи опций **Весь термин** или **Весь и В начале**.
- **Весь (не составной)**. Этот тип назначается, если всё понятие, извлеченное из текстовых данных, совпадает с точным термином в словаре, а процесс извлечения останавливается, чтобы воспрепятствовать сопоставлению в нем термина с более длинными составными словами. Например, если ввести `apple`, опция **Весь (не составной)** типизирует `apple` (яблоки), но не извлечет составное слово `apple sauce` (яблочное пюре), если только оно не применяется где-нибудь в другом месте.

В следующей таблице предполагается, что термин `apple` (яблоки) содержится в словаре типов. В зависимости от опции сопоставления эта таблица показывает, какие понятия будут извлечены и типизированы, если они будут найдены в текстовых данных.

*Таблица 38. Примеры соответствий.*

| **Опции сопоставления для термина:**  **`apple` (яблоки)** | **Извлекаемые понятия** | | | |
|---|---|---|---|---|
| | **`apple` (яблоки)** | **`apple` *tart* (яблочный пирог)** | ***ripe*** **`apple` (спелые яблоки)** | ***homemade*** **(домашний) `apple` *tart* (яблочный пирог)** |
| **Весь термин** | ✔ | | | |
| **Начать с** | | ✔ | | |
| **В конце** | | | ✔ | |
| **В начале или В конце** | | ✔ | ✔ | |
| **Весь и В начале** | ✔ | ✔ | | |

*Таблица 38. Примеры соответствий (продолжение).*

| **Опции сопоставления для термина:**  **apple (яблоки)** | **Извлекаемые понятия** | | | |
|---|---|---|---|---|
| | **apple (яблоки)** | **apple** *tart* **(яблочный пирог)** | *ripe* **apple (спелые яблоки)** | *homemade* **(домашний) apple** *tart* **(яблочный пирог)** |
| **Весь и В конце** | ✔ | | ✔ | |
| **Весь и (В начале или В конце)** | ✔ | ✔ | ✔ | |
| **Любой** | | ✔ | ✔ | ✔ |
| **Весь и В любом месте** | ✔ | ✔ | ✔ | ✔ |
| **Весь (несоставной)** | ✔ | *никогда не извлекалось* | *никогда не извлекалось* | *никогда не извлекалось* |

Столбец Флективные

В этом столбце выберите, должен ли механизм извлечения генерировать флективные формы данного термина во время извлечения, чтоб они все были сгруппированы вместе друг с другом. Значение по умолчанию для этого столбца определяется в диалоговом окне Свойства типов, но эту опцию можно изменить для каждого случая отдельно, непосредственно в этом столбце. В меню выберите **Изменить > Изменить инфлексию**.

Столбец Тип

В этом столбце выберите тип в выпадающем списке. Список типов будет отфильтрован в соответствии с вариантом, выбранным вами на панели дерева библиотек. Первый тип в списке - это всегда тип по умолчанию, выбранный на панели дерева библиотек. В меню выберите **Изменить > Изменить тип**.

Столбец Библиотека

В этом столбце выводится библиотека, в которой хранится термин. Чтобы изменить библиотеку термина, его можно перетащить в другой тип на панели дерева библиотек.

Чтобы добавить в словарь типов один термин:

1. На панели дерева библиотек выберите словарь типов, куда вы хотите добавить термин.
2. В списке терминов на центральной панели введите термин в первой доступной пустой ячейке и задайте для него все нужные вам опции.

Чтобы добавить в словарь типов несколько терминов:

1. На панели дерева библиотек выберите словарь типов, куда вы хотите добавить термины.
2. В меню выберите **Инструменты > Новые термины**. Откроется диалоговое окно Добавить новые термины.
3. Введите термины, которые вы хотите добавить в выбранный словарь типов, введя эти термины или скопировав и вставив их набор. Если вводится несколько терминов, нужно либо ввести их через разделитель, определяемый в диалоговом окне Опции, либо добавить каждый термин в новой строке. Дополнительную информацию смотрите в разделе “Настройка опций” на стр. 82.

4. Нажмите кнопку **OK**, чтобы добавить термины в словарь. В качестве опции сопоставления будет автоматически задана опция по умолчанию для этой библиотеки типов. Диалоговое окно закроется, и новые термины появятся в словаре.

## Принудительное назначение типов терминам

Если вы хотите. чтобы термину был назначен конкретный тип, его можно добавить в соответствующий словарь типов. Однако при наличии нескольких типов с одним и тем же именем механизм извлечения должен знать, какой тип следует использовать. Поэтому будет предложено выбрать, какой тип использовать. Это называется **принудительным назначением** типа термину. Наибольший прок от этой опции - при определении назначения типа из скомпилированного (внутреннего, нередактируемого) словаря. В целом, мы рекомендуем вообще избегать повторений терминов.

Принудительное назначение типов не *удалит* другие вхождения этого термина; вместо этого они будут игнорироваться механизмом поиска. Позднее можно будет изменить вхождение, которое должно использоваться при принудительном или непринудительном назначении типа термину. Может также потребоваться принудительно назначить термин в словарь типов при добавлении общедоступной библиотеки или ее обновлении.

Посмотреть, каким терминам был принудительно назначен тип, а какие будут игнорироваться, можно в столбце Принудительно (втором столбце на панели терминов). Если выводится значок канцелярской кнопки, это означает, что данному вхождению термина был принудительно назначен тип. Если выводится чёрный значок X, это означает, что во время извлечения это вхождение термина будет игнорироваться, поскольку ему был принудительно назначен тип в другом месте. Кроме того, после принудительного назначения типа термину он будет выводиться цветом для типа, который ему был принудительно назначен. Это означает, что если термину, которому назначен и `Тип 1`, и `Тип 2`, вы принудительно назначите `Тип 1`, он будет всё время выводиться в окне шрифтом, цвет которого определен для `Типа 1`.

Дважды щёлкнув по значку, можно изменить состояние. Если термин встречается в другом месте, откроется диалоговое окно, в котором можно выбрать, какое вхождение следует использовать.

## Переименование типов

Отредактировав свойства типов, можно переименовать словарь типов или изменить другие параметры словаря.

**Важно!** Мы рекомендуем вам не использовать в именах типов пробелы, особенно если несколько имен типов начинаются с одного и того же слова. Мы также рекомендуем не переименовывать типы в библиотеках Core и Opinions и не изменять их атрибуты сопоставления по умолчанию.

Чтобы переименовать тип:

1. На панели дерева библиотек выберите словарь типов, который вы хотите переименовать.
2. Щелкните правой кнопкой мыши и в контекстном меню выберите **Свойства**. Откроется диалоговое окно Свойства типа.
3. Введите новое имя для словаря типов в текстовом поле Имя.
4. Нажмите кнопку **OK** для принятия нового имени. Новое имя словаря типов появится на панели дерева библиотек.

## Перемещение типов

Словарь типов можно перетащить в другое положение в библиотеке или другую библиотеку в дереве.

Чтобы изменить порядок вывода типа в библиотеке:

1. На панели дерева библиотек выберите словарь типов, который вы хотите переместить.

2. В меню выберите **Изменить > Переместить вверх**, чтобы переместить словарь типов на панели дерева библиотек на одну позицию вверх, или **Изменить > Переместить вниз**, чтобы переместить его на одну позицию вниз.

Чтобы переместить тип в другую библиотеку:

1. На панели дерева библиотек выберите словарь типов, который вы хотите переместить.
2. Щелкните правой кнопкой мыши и в контекстном меню выберите **Свойства**. Откроется диалоговое окно Свойства типа. (Тип можно также перетащить в другую библиотеку).
3. В окне списка "Добавить в" выберите библиотеку, куда вы хотите переместить словарь типов.
4. Нажмите кнопку **OK**. Диалоговое окно закроется, и тип появится в выбранной вами библиотеке.

## Отключение и удаление типов

Если вы хотите временно удалить словарь типов, его можно отключить, отключив переключатель слева от имени этого словаря на панели дерева библиотек. Тем самым вы укажете, что хотите сохранить словарь в библиотеке, но хотите, чтобы его содержимое игнорировалось при проверке конфликтов и во время процесса извлечения.

Словари типов можно также удалить из библиотеки навсегда.

Чтобы отключить словарь типов:

1. На панели дерева библиотек выберите словарь типов, который вы хотите отключить.
2. Нажмите клавишу пробела. Переключатель слева от имени типа будет отключен.

Чтобы удалить словарь типов:

1. На панели дерева библиотек выберите словарь типов, который вы хотите удалить.
2. В меню выберите **Изменить > Удалить**, чтобы удалить словарь типов.

---

## Словари подстановок/синонимов

**Словарь подстановок** - это собрание терминов, помогающее сгруппировать схожие термины под одним целевым термином. Элементы управления словарями синонимов находятся на нижней панели вкладки Ресурсы библиотек. Вы можете вызвать это представление, выбрав **Вид > Редактор ресурсов** в меню, если у вас запущен интерактивный сеанс инструментальной среды. Другой вариант - редактировать словари для определенного шаблона в Редактор шаблонов.

В этом словаре можно определить две формы подстановок: **синонимы** и **необязательные элементы**. Щелкая на этой панели по вкладкам, можно переключаться между ними.

После выполнения процесса извлечения для текстовых данных можно найти несколько понятий, являющихся синонимами или флективными формами других понятий. Определив необязательные элементы и синонимы, с помощью механизма извлечения их можно отобразить на единственный целевой термин.

Подстановка с применением синонимов и необязательных элементов сокращает число понятий на панели Результаты извлечения благодаря их объединению в более значимые представительные понятия с подсчетом документов.

*Примечание:* для ресурсов на японском языке необязательные элементы не применяются и недоступны. Кроме того, синонимы для текстовых данных на японском языке обрабатываются немного по-другому.

Синонимы

Синонимы связывают несколько слов, у которых одинаковые значения. Кроме того, с помощью синонимов можно сгруппировать термины с их сокращениями или сгруппировать слова, в которых часто допускают ошибки, с их правильным написанием. Эти синонимы можно определить на вкладке Синонимы.

Определение синонимов состоит из двух частей. Первая часть - это **целевой** термин, то есть термин, вокруг которого вы хотите группировать все термины синонимов. Если только этот целевой термин не используется в качестве синонима другого целевого термина и если только он не исключен, скорее всего этот термин станет понятием, выводящимся на панели Результаты извлечения. Вторая часть - это список синонимов, которые будут сгруппированы под этим целевым термином.

Например, если вы хотите, чтобы термин `automobile` был заменен термином `vehicle`, то `automobile` будет считаться синонимом, а `vehicle` - целевым термином.

В столбец **Синоним** можно вводить любые слова, но если слово не будет найдено вовремя извлечения, а у термина будет опция сопоставления со значением `Весь`, никакая постановка не выполняется. Но чтобы сгруппировать вокруг целевого термина синонимы, извлекать это термин не нужно.

Необязательные элементы

Необязательные элементы определяют необязательные слова в составном термине, которые во время извлечения могут быть проигнорированы, чтобы похожие термины оставались вместе, даже если в тексте они немного отличаются. Необязательные элементы - это отдельные слова, удаление которых из составного термина может привести к образованию соответствия с другим термином. Эти отдельные слова могут появляться в любом месте составного термина (в начале, середине или конце). Необязательные элементы можно определить на вкладке Необязательные.

Например, чтобы сгруппировать друг с другом термины `ibm` и `ibm corp`, нужно объявить, что термин `corp` будет рассматриваться в этом случае как необязательный элемент. В другом примере, если необязательным элементом назначить термин `access`, а во время извлечения будут найдены и `internet access speed`, и `internet speed` они будут сгруппированы друг с другом под термином, встречающимся наиболее часто.

*Примечание:* Для текстовых ресурсов на японском языке вкладка Необязательные элементы отсутствует, поскольку необязательные элементы не применяются.

## Определение синонимов

На вкладке Синонимы можно ввести определение синонимов в пустой строке в начале таблицы. Начните с определения термина назначения и его синонимов. Можно также выбрать библиотеку, в которой вы хотите хранить это определение. При извлечении все вхождения синонимов будут сгруппированы в окончательном извлечении под этим термином назначения. Дополнительную информацию смотрите в разделе “Добавление терминов” на стр. 192.

Например, если ваши текстовые данные содержат много информации, связанной с телекоммуникациями, могут использоваться следующие термины: `сотовый телефон`, `беспроводной телефон` и `мобильный телефон`. В этом примере может потребоваться определить слова `сотовый` и `мобильный` как синонимы слова `беспроводной`. Если определить эти синонимы, всякое извлеченное появление терминов `сотовый телефон` и `мобильный телефон` будет рассматриваться как появление термина `беспроводной телефон`, и все эти термины будут совместно представлены в списке терминов.

При построении словарей типов можно ввести термин, а затем рассмотреть возможность добавления трех-четырех синонимов для этого термина. В этом случае в словарь подстановки можно ввести все термины, а затем выбранный целевой термин, после чего перетащить синонимы.

*Примечание*: в текстах на японском языке синонимы обрабатываются несколько иначе.

Подстановка синонимов применяется также в изменяемых формах синонима (например, во множественном числе). В зависимости от контекста может понадобиться ввести ограничения на способ подстановки терминов. Определенные символы можно использовать для установления пределов того, насколько далеко должна зайти обработка синонимов:

- **Восклицательный знак (!).** Если восклицательный знак расположен непосредственно перед синонимом (`!синоним`), это означает, что никакие изменяемые формы синонима не будут подставляться термином назначения. Однако восклицательный знак непосредственно перед термином назначения (`!термин-назначения`) означает, что не допускается дальнейшая подстановка для любой части составного термина назначения или его вариантов.
- **Звездочка (*).** Звездочка непосредственно после синонима (`синоним*`) означает, что вы хотите заменить это слово на целевой термин. Например, если вы определили `manage*` как синоним, а `management` - как целевой термин, термин `associate managers` будет замещен термином `associate management`. Можно добавить также пробел между словом и звездочкой (`синоним *`), например, `internet *`. Если вы определили целевой термин как `internet`, а синонимы как `internet * *` и `web *`, термины `internet access card` и `web portal` будут заменяться на `internet`. В этом словаре нельзя начинать слово или строку с символа подстановки звездочка.
- **Каре (^).** Символ каре и пробел перед синонимом (`^ synonym`) означают, что группировка синонимов применима только в случае, когда термин начинается с синонима. Например, если определить `^ wage` как синоним, а `income` - как целевой термин, и извлекаются оба термина, они будут сгруппированы вместе вокруг термина `income`. Однако если извлекаются термины `minimum wage` и `income`, они не будут группироваться вместе, так как `minimum wage` не начинается со слова `wage`. Пробел должен размещаться между этим знаком и синонимом.
- **Обозначение доллара ($).** Пробел и обозначение доллара после синонима (`synonym $`) означают, что группировка синонимов применима только в том случае, когда термин оканчивается на этот синоним. Например, если определить `cash $` как синоним, а `money` - как целевой термин, и извлекаются оба термина, они будут сгруппированы вместе вокруг термина `money`. Однако если извлекаются термины `cash cow` и `money`, они не будут группироваться вместе, так как `cash cow` не оканчивается на слово `cash`. Пробел должен размещаться между этим знаком и синонимом.
- **Каре (^) и обозначение доллара ($).** Если знаки каре и доллара используются совместно (`^ синоним $`), термин совпадает с синонимом только при точном соответствии. Это означает, что никакие слова не могут появиться до или после синонима в извлеченном термине, чтобы произошла группировка синонимов. Например, вы можете определить `^ van $` как синоним, а `truck` - как целевой термин, так что только `van` будет с группирован с `truck`, а `marie van guerin` останется без изменений. Кроме этого, при всяком определении синонима со знаками каре и доллара в случае появления данного слова с любом месте исходного текста синоним будет извлекаться автоматически.

*Примечание*: в текстах на японском языке эти специальные символы и знаки подстановки не поддерживаются.

Чтобы добавить запись синонимов:

1. При выведенной панели подстановки щелкните по вкладке **Синонимы** в нижнем левом углу.
2. В пустой строке наверху в таблице введите целевой термин в столбце Назначение. Вводимый целевой термин выводится цветным шрифтом. Этот цвет представляет тип, в котором представлен или куда вставляется термин. Если термин выводится черным шрифтом, это означает, что его нет ни в каких словарях типов.
3. Щелкните в ячейке справа от термина назначения (второй ячейке) и введите набор синонимов. Разделяйте каждую из записей глобальным разделителем, задаваемым в диалоговом окне Опции. Дополнительную информацию смотрите в разделе “Настройка опций” на стр. 82. Вводимые термины выводятся цветным шрифтом. Этот цвет представляет тип, в котором появляется термин. Если термин выводится черным шрифтом, это означает, что его нет ни в каких словарях типов.
4. Щелкнув в последней ячейке, выберите библиотеку, где вы хотите хранить это определение синонимов.

*Примечание*: Эти инструкции показывают, как вносить изменения в представление Редактор ресурсов или в Редактор шаблонов. Имейте в виду, что вы можете выполнять такую тонкую настройку непосредственно из

панели Результаты извлечения, панели Данные, панели Категории или диалогового окна Определения кластеров в прочих представлениях. Дополнительную информацию смотрите в разделе “Уточнение результатов извлечения” на стр. 95.

## Определение необязательных элементов

На вкладке Необязательные элементы можно определить необязательные элементы для любой библиотеки. Эти записи группируются совместно для каждой библиотеки. Как только библиотека добавляется на панель дерева библиотек, на вкладке Необязательные элементы добавляется пустая строка для необязательных элементов.

Все записи автоматически преобразуются в слова в нижнем регистре. Механизм извлечения будет искать совпадения записей со словами в тексте и в верхнем, и в нижнем регистре.

*Примечание:* для ресурсов на японском языке необязательные элементы не применяются и недоступны.

*Примечание*: термины разделяются с использованием разделителя, определенного в диалоговом окне Опции. Дополнительную информацию смотрите в разделе “Настройка опций” на стр. 82. Если вводимый вами необязательный элемент включает в себя такой же разделитель в качестве части термина, перед ним должен стоять знак обратной дробной черты.

Чтобы добавить запись:

1. При выведенной панели подстановки щелкните по вкладке Необязательные элементы в нижнем левом углу редактора.
2. Щелкните по ячейке в столбце Необязательные элементы для библиотеки, в которую вы хотите добавить эту запись.
3. Введите необязательный элемент. Разделяйте каждую из записей глобальным разделителем, задаваемым в диалоговом окне Опции. Дополнительную информацию смотрите в разделе “Настройка опций” на стр. 82.

## Отключение и удаление подстановок

Можно временно удалить запись, отключив ее в вашем словаре. При отключении записи она будет игнорироваться при извлечении.

В вашем словаре подстановок можно удалить также любые устаревшие записи.

Чтобы отключить запись:

1. В вашем словаре выберите запись, которую нужно отключить.
2. Нажмите клавишу пробела. Переключатель слева от записи будет выключен.

*Примечание*: для отключения можно также выключить переключатель слева от записи.

Чтобы удалить запись синонима:

1. В вашем словаре выберите запись, которую нужно удалить.
2. Выберите в меню опцию **Изменить > Удалить** или нажмите клавишу **Delete** на клавиатуре. Этой записи больше не будет в словаре.

Чтобы удалить запись необязательного элемента:

1. В вашем словаре дважды щелкните по записи, которую нужно удалить.
2. Удалите этот термин вручную.
3. Нажмите клавишу Enter, чтобы применить изменение.

## Словари исключения

**Словарь исключения** - это список слов, словосочетаний или частичных строк. Любые термины, совпадающие с записью в словаре исключения или содержащие запись из этого словаря, будут игнорироваться или исключаться из процесса извлечения. Словари исключения управляются на правой панели редактора. Обычно добавляемые в этот список термины - это заполняющие слова или словосочетания, которые используются для связности текста, но в действительности не добавляют в него никакой важной информации, а только мешают восприятию результатов извлечения. Добавление этих терминов в словарь исключения гарантирует, что они никогда не попадут в число извлеченных.

Управление словарями исключения происходит на верхней правой панели вкладки Ресурсы библиотек в редакторе. Вы можете вызвать это представление, выбрав **Вид > Редактор ресурсов** в меню, если у вас запущен интерактивный сеанс инструментальной среды. Другой вариант - редактировать словари для определенного шаблона в Редактор шаблонов.

В словарь исключения можно ввести слово, словосочетание или часть строки в пустую строку в верхней части таблицы. Строки символов можно добавить в ваш словарь исключения как одно или несколько слов или даже как часть слов с использованием звездочки как символа подстановки. Объявленные в словаре исключений записи будут использоваться для блокировки понятий при извлечении. Если запись объявлена еще где-то в интерфейсе, например, в словаре типов, она показывается перечеркнутой в других словарях, что указывает на ее текущее исключение. Эта строка не должна появляться в текстовых данных или объявляться как часть любого словаря типов, который будет применяться.

*Примечание*: если добавить в словарь исключения понятие, которое представляет собой также целевой объект в записи синонимов, этот целевой объект и все его синонимы также будут исключены. Дополнительную информацию смотрите в разделе “Определение синонимов” на стр. 197.

Использование символов подстановки (*)

Для текстов на всех языках, кроме японского, можно использовать символ подстановки (звездочку *) для обозначения, что вы хотите рассматривать запись исключения как частичную строку. Все термины, обнаруженные механизмом извлечения и содержащие слово, начинающееся и оканчивающееся строкой, которая есть в словаре исключения, будут исключены из окончательных результатов извлечения. Однако есть два случая, когда использование символа подстановки не допускается:

- Символ тире (-) после символа подстановки в виде звездочки, *-
- Апостроф (') с предшествующим символом подстановки в виде звездочки, например, *'s

*Таблица 39. Примеры записей исключения.*

| Запись | Пример | Результаты |
|---|---|---|
| word | *next* | Никакие понятия (или их термины) не будут извлекаться, если они содержат слово *next*. |
| phrase | *for example* | Никакие понятия (или их термины) не будут извлекаться, если они содержат словосочетание *for example*. |
| partial | *copyright** | Будет исключены все понятия (или их термины), совпадающие со словом *copyright* или содержащие это слово или его вариации, такие как `copyrighted`, `copyrighting`, `copyrights` или `copyright 2010`. |
| partial | **ware* | Будут исключены все понятия (или их термины), совпадающие со словом *ware* или содержащие это слово или его вариации, такие как `freeware`, `shareware`, `software`, `hardware`, `beware` или `silverware`. |

Чтобы добавить записи:

1. В пустой строке в начале таблицы введите термин. Вводимый термин выделяется цветом. Этот цвет представляет тип, в котором появляется термин. Если термин выводится черным шрифтом, это означает, что его нет ни в каких словарях типов.

Чтобы отключить записи

Можно временно удалить запись, отключив ее в вашем словаре исключения. При отключении записи она будет игнорироваться при извлечении.

1. В вашем словаре исключений выберите запись, которую нужно отключить.
2. Нажмите пробел. Переключатель слева от записи выключится.

*Примечание*: для отключения можно также выключить переключатель слева от записи.

Чтобы удалить записи:

Можно удалить любые ненужные записи в вашем словаре исключения.

1. В вашем словаре исключения выберите запись, которую нужно удалить.
2. В меню выберите **Изменить > Удалить**. Этой записи больше не будет в словаре.

# Глава 18. О расширенных ресурсах

В дополнение к словарям типов, исключений и подстановок, можно работать также с разнообразными параметрами расширенных ресурсов, такими как параметры нечеткой группировки, или с определениями нелингвистических типов. С этими ресурсами можно работать на вкладке Расширенные ресурсы в представлении Редактор шаблонов или Редактор ресурсов.

**Важно!** Эта вкладка недоступна для ресурсов, настроенных под японский текст.

При переходе на вкладку Расширенные ресурсы можно изменить следующую информацию:

- **Язык назначения для ресурсов.** Используется для выбора языка, на котором будут создаваться и настраиваться ресурсы. Дополнительную информацию смотрите в разделе “Язык назначения для ресурсов” на стр. 205.
- **Нечеткая группировка (исключения).** Используется для исключения пар слов из алгоритма нечеткой группировки (исправление ошибок произношения). Дополнительную информацию смотрите в разделе “Нечеткая группировка” на стр. 206.
- **Нелингвистические объекты.** Используется для включения и отключения нелингвистических объектов для извлечения, а также регулярных выражений и правил нормализации, которые будут применяться при их извлечении. Дополнительную информацию смотрите в разделе “Нелингвистические объекты” на стр. 206.
- **Языковая обработка.** Используется для объявления специальных способов структурирования предложений (паттерны извлечения и принудительные определения) и использования аббревиатур для выбранного языка. Дополнительную информацию смотрите в разделе “Языковая обработка” на стр. 211.
- **Идентификатор языка.** Используется для конфигурирования идентификатора языка, вызываемого, когда для языка задано значение **Все**. Дополнительную информацию смотрите в разделе “Идентификатор языка” на стр. 212.

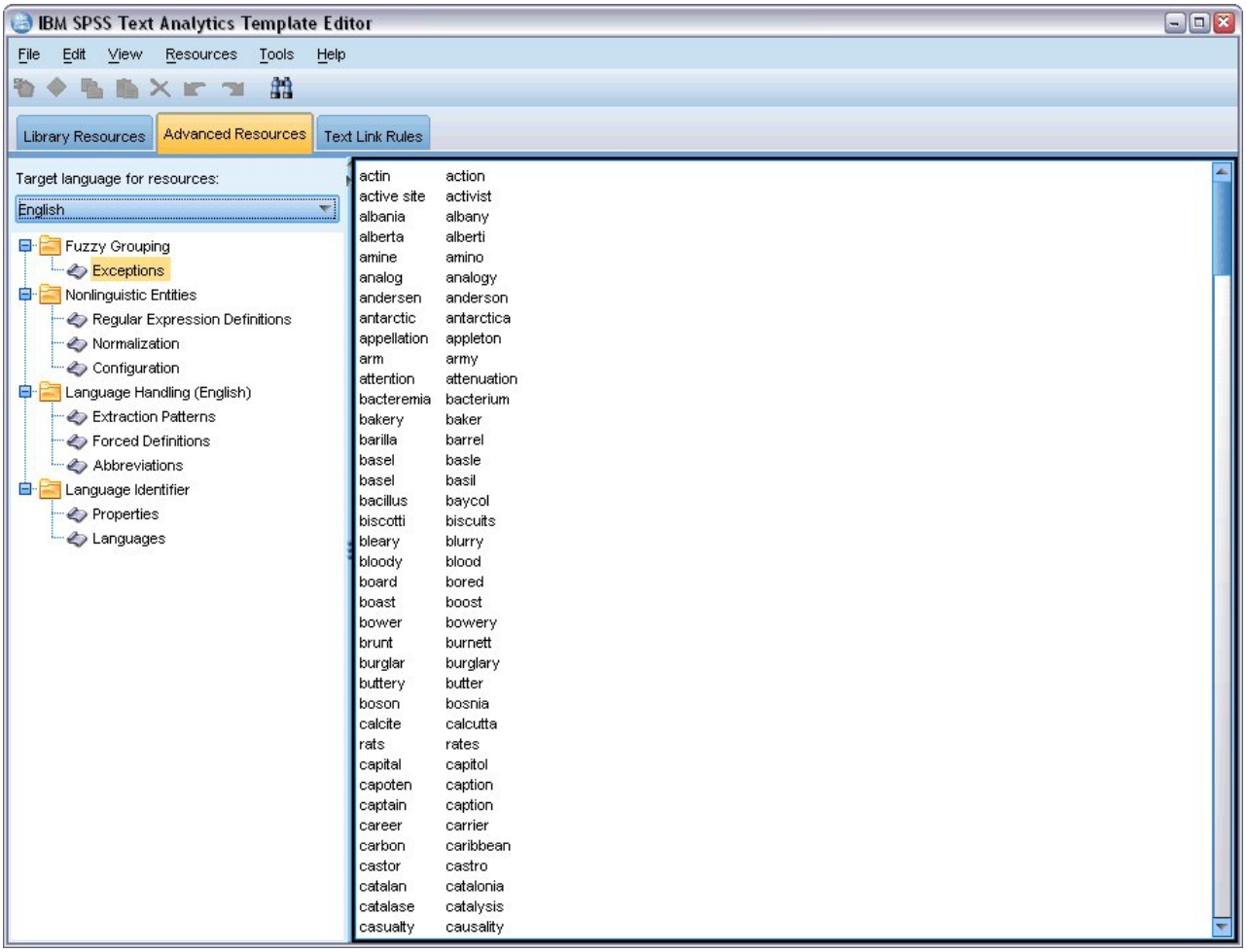



*Рисунок 41. Редактор шаблонов исследования текста - вкладка Расширенные ресурсы*

*Примечание*: Чтобы быстро найти информацию или внести общие изменения во всем разделе, можно использовать панель инструментов Найти/заменить. Дополнительную информацию смотрите в разделе “Замена” на стр. 205.

**Изменить расширенные ресурсы**

1. Найдите и выберите раздел ресурсов, который вы хотите изменить. Этот контент появится на правой панели.
2. Используйте меню или кнопки панели инструментов, чтобы при необходимости контент вырезать, скопировать или вставить из буфера.
3. Измените выбранные файлы, используя правила форматирования из этого раздела. Ваши изменения будут сразу сохранены. Используйте стрелки undo (отменить) или redo (повторить) на панели инструментов, чтобы вернуться к предыдущим изменениям.

## Поиск

В некоторых случаях нужно быстро найти информацию в конкретном разделе. Например, при анализе текстовых связей иногда используются сотни определений макросов и паттернов.При помощи функции поиска можно быстро найти конкретное правило. Для поиска информации в некотором разделе можно использовать панель инструментов Поиск.

Чтобы использовать функцию поиска

1. Найдите и выберите раздел ресурсов, в котором хотите вести поиск. Этот контент появится на правой панели редактора.
2. В меню выберите **Изменить > Найти**. Панель инструментов Поиск выводится в верхнем правом углу диалогового окна Изменить расширенные ресурсы.
3. Введите строку искомого слова в текстовом окне. Кнопками на панели поиска можно задать учет регистра, частичное соответствие и направление поиска.
4. Нажмите кнопку **Найти** для запуска поиска. Если соответствие найдено, текст выделяется в окне.
5. Снова нажмите кнопку **Найти**, чтобы найти следующее вхождение.

**Примечание**: При работе на вкладке Правила текстовых связей опция Поиск доступна только в представлении исходного кода.

## Замена

В некоторых случаях нужны изменения в обширном списке дополнительных ресурсов. Функция замены полезна для однотипных изменений содержимого.

Чтобы использовать функцию замены

1. Найдите и выберите раздел ресурсов, в котором хотите провести поиск и замену. Этот контент появится на правой панели редактора.
2. В меню выберите **Изменить > Заменить**. Откроется диалоговое окно Замена.
3. В текстовом окне **Найти** введите строку искомого слова.
4. В текстовом окне **Заменить на** введите строку, на которую нужно заменять найденное.
5. Включите переключатель **Соответствие только слова целиком**, если нужно найти или заменить только слова целиком.
6. Включите переключатель **Учитывать регистр**, если нужно найти или заменить только слова с таким же регистром букв.
7. Нажмите кнопку **Найти далее**, чтобы найти соответствие. Если соответствие найдено, текст выделяется в окне. Если это соответствие заменять не нужно, продолжайте нажимать кнопку **Найти далее**, пока не найдете то соответствие, которое нужно заменить.
8. Нажмите кнопку **Заменить**, чтобы заменить выбранное соответствие.
9. Нажмите кнопку **Заменить**, чтобы заменить все соответствия в разделе. Откроется сообщение с числом выполненных замен.
10. Завершив замены, нажмите кнопку **Закрыть**. Диалоговое окно закроется.

*Примечание*: Если вы сделали ошибочную замену, ее можно отменить, закрыв диалоговое окно и выбрав **Правка > Отмена** в меню. Это действие нужно выполнить по одному разу для каждой отменяемой замены.

## Язык назначения для ресурсов

Ресурсы создаются для конкретного языка текстов. Язык, под который настроены ресурсы, определяется на вкладке Расширенные ресурсы. При необходимости можно переключиться на другой язык, выбрав его в окне со списком **Язык назначения для ресурсов**. Кроме того, выведенный здесь язык будет выводиться как язык для любых пакетов TAP (text analysis package), которые вы создадите для этих ресурсов.

**Важно!** Случаи, когда нужно изменить язык ресурсов, весьма редки. Такое изменение может вызвать ошибки, потому что ресурсы перестанут соответствовать языку извлечения. Иногда язык меняют, планируя использовать опцию языка `BCE` при извлечении, поскольку ожидается текст на нескольких языках. После изменения языка можно открыть, например, ресурсы обработки языка для паттернов извлечения, сокращений и принудительных определений для нужного вам вторичного языка. Однако имейте в виду, что перед публикацией или сохранением изменений в ресурсах или запуском другой извлечения нужно вернуться к основному языку извлечения.

## Нечеткая группировка

В узле Исследование текста и в Параметрах извлечения, если включен переключатель **Допускать орфографические ошибки при минимуме символов корня**, доступен алгоритм нечеткой группировки.

Нечеткая группировка помогает сгруппировать слова, которые часто пишутся с ошибками, а также вариативные написания слов; для этого перед сравнением извлеченных слов из них временно удаляются все гласные (кроме первой) и двойные или тройные согласные. Во время извлечения функция нечеткой группировки применяется к извлеченным терминам, и результаты сравниваются в поисках соответствий. Если соответствие найдено, исходные термины группируются в итоговом списке извлечения. Главным представителем группы служит термин, встречающийся в данных чаще всего.

*Примечание*: Если два сравниваемых термина - разных типов, отличных от типа `<Неизвестный>`, к такой паре метод нечеткой группировки не применяется. Другими словами, метод применяется только к терминам одного типа или типа `<Неизвестный>`.

Если вы активировали эту функцию и обнаружили два ошибочно сгруппированных слова, близких по написанию, их можно исключить из нечеткой группировки. Для этого можно ввести ошибочно составленные пары в разделе Исключения на вкладке Расширенные ресурсы. Дополнительную информацию смотрите в разделе Глава 18, “О расширенных ресурсах”, на стр. 203.

В следующем примере показано, как работает нечеткая группировка. Если нечеткая группировка активирована, следующие слова будут считаться одинаковыми и сопоставляться, как показано:

```
color  -> colr                mountain -> montn
colour -> colr                montana  -> montn

modeling  -> modlng           furniture -> furntr
modelling -> modlng           furnature -> furntr
```

В приведенном выше примере есть смысл исключить группировку `mountain` и `montana`. Поэтому можно их можно ввести в разделе исключений в следующей форме:

```
mountain      montana
```

**Важно!** В некоторых случаях исключения из нечеткой группировки не предотвращают объединения двух слов в пару из-за действия определенных правил синонимов. В этом случае есть смысл попробовать ввести синонимы с символом подстановки восклицательный знак (`!`), чтобы не дать этим словам стать синонимам в выходной информации. Дополнительную информацию смотрите в разделе “Определение синонимов” на стр. 197.

Правила форматирования для исключений из нечеткой группировки

- Определяйте только одну пару исключений на строке.
- Используйте простые или составные слова.
- Используйте в словах только символы нижнего регистра. Слова в верхнем регистре будут проигнорированы.
- Используйте символ `TAB` как разделитель слов в паре.

## Нелингвистические объекты

При работе с некоторыми видами данных бывает нужно извлечь даты, номера социальной страховки, проценты и другие нелингвистические объекты. Эти объекты явным образом декларируют в файле конфигурации, в котором их можно включать или выключать. Дополнительную информацию смотрите в разделе “Конфигурация” на стр. 210. Чтобы оптимизировать информацию, выводимую механизмом извлечения, входную информацию от нелингвистических процессов нормализуют путем группировки сходных объектов согласно предварительно определенным форматам. Дополнительную информацию смотрите в разделе “Нормализация” на стр. 209.

*Примечание*: Извлечение нелингвистических объектов можно включить и выключить в параметрах извлечения.

Доступные нелингвистические объекты

Нелингвистические объекты, для которых возможно извлечение, приведены в следующей таблице. Имя типа указано в скобках.

*Таблица 40. Нелингвистические объекты, которые можно извлекать*

| | |
|---|---|
| Адреса | (<Address>) |
| Аминокислоты | (<Aminoacid>) |
| Валюты | (<Currency>) |
| Даты | (<Date>) |
| Задержка | (<Delay>) |
| Цифры | (<Digit>) |
| Адреса электронной почты | (<email>) |
| Адреса HTTP/URL | (<url>) |
| IP-адрес | (<IP> |
| Организации | (<Organization>) |
| Проценты | (<Percent>) |
| Продукты | (<Product>) |
| Белки | (<Gene>) |
| Номера телефонов | (<PhoneNumber>) |
| Время | (<Time>) |
| Номера социального страхования США | (<SocialSecurityNumber>) |
| Веса и меры | (<Weights-Measures>) |

Очистка текста для обработки

Для извлечения нелингвистических объектов входной текст необходимо сначала очистить. На этом шаге вносятся следующие временные изменения, которые служат для идентификации и извлечения нелингвистических элементов:

- Все последовательности из нескольких пробелов заменяются одиночными пробелами.
- Табуляции заменяются пробелом.
- Одиночный символ или одиночная последовательность символов, обозначающие конец строки, заменяются пробелом, а несколько последовательностей, обозначающих конец строки, помечаются как конец абзаца. Конец строки может обозначаться как возврат каретки (carriage return, CR), как перевод строки (line feed, LF) или как оба символа вместе.
- Теги HTML и XML временно удаляются и игнорируются.

## Определения регулярных выражений

При извлечении нелингвистических объектов иногда нужно отредактировать или добавить определения регулярных выражений, использующиеся для идентификации регулярных выражений. Это делается в разделе **Определения регулярных выражений** на вкладке Расширенные ресурсы. Дополнительную информацию смотрите в разделе Глава 18, “О расширенных ресурсах”, на стр. 203.

Файл разбит на разделы. Первый раздел называется [macros] (макросы). Помимо этого раздела, могут существовать дополнительные разделы, по одному для каждого нелингвистического объекта. В этот файл

можно добавлять разделы. В пределах каждого раздела правила нумеруются (*regexp1*, *regexp2* и так далее). Эти правила должны нумероваться последовательно, от 1 до *n*. Любой пропуск в нумерации приведет к остановке обработки этого файла.

В некоторых случаях объект зависит от языка. Объект считается зависящим от языка, если его параметр языка в файле конфигурации принимает значение, отличное от 0. Дополнительную информацию смотрите в разделе “Конфигурация” на стр. 210. Когда объект зависит от языка, в имени раздела должен быть префикс языка, например, `[english/PhoneNumber]`. Этот раздел будет содержать правила, применяемые только к английским номерам телефонов, когда для объекта `PhoneNumber` задан язык 2.

**Важно!** Если после изменений в этом файле или любом другом, внесенным в редакторе, механизм извлечения перестал работать, как нужно, используйте опцию **Сброс до исходного содержимого** на панели инструментов, чтобы восстановить содержимое файла, которое было в момент поставки. Для редактирования этого файла нужны навыки работы с регулярными выражениями. Если вам нужна дополнительная помощь в этой области, обратитесь в IBM Corp..

Специальные символы . [] {} () \ * + ? | ^ $

Все символы соответствуют сами себе, кроме следующих специальных символов, которые служат для специальных целей в выражениях: `.[{()\*+?|^$` Чтобы использовать эти символы в определении как представляющие самих себя, перед ними нужно добавлять обратную дробную черту (\).

Например, если вы пытаетесь извлечь веб-адреса, для объекта важны стоп-символы, который необходимо предварять обратной дробной чертой:

```
www\.[a-z]+\.[a-z]+
```

Операции повтора и квантификаторы ? + * {}

Для большей гибкости определений доступны несколько символов подстановки, обычных в регулярных выражениях. Это `* ? +`

- *Звездочка* * означает *ноль или более* вхождений предшествующей строки. Например, `ab*c` соответствует "*ac*", "*abc*", "*abbbc*" и так далее.
- *Плюс* + означает *одно или несколько* вхождений предшествующей строки. Например, `ab+c` соответствует "*abc*", "*abbc*", "*abbbc*", но не "*ac*".
- *Знак вопроса ?* означает *ноль или одно* вхождение предшествующей строки. Например, `пиксель?` соответствует и "*пиксель*", и "*пиксел*".
- *Квадратные скобки задают повторение {} раз*, с указанием границ повторяемого. Например,

  `[0-9]{n}` соответствует *n* цифрам. Например, `[0-9]{4}` соответствует “*1998*”, но не “*33*” или “*19983*”.

  `[0-9]{n,}` соответствует *n или более* цифрам. Например, `[0-9]{3,}` соответствует “*199*” или “*1998*”, но не “*19*”.

  `[0-9]{n,m}` соответствует строке, содержащей *от n до m* цифр включительно. Например, `[0-9]{3,5}` соответствует “*199*”, “*1998*” или “*19983*”, но не “*19*” и не “*199835*”.

Необязательные пробелы и дефисы

В некоторых случаях нужно включить в определение необязательный пробел. Например, если нужно извлечь денежные единицы "*тайские баты*", "*тайский бат*", "*таиландские баты*", "*таиландский бат*", "*баты*" или "*бат*", нужно учесть, что часть вариантов представляют собой два слова через пробел. В этом случае нужное определение имеет вид `(тайские |тайский |таиландские |таиландский )?баты?`. Поскольку после слов *тайские*, *тайский* и так далее перед словом *баты*/*бат* требуется пробел, нужно задать необязательный пробел в последовательности вариантов `(тайские |тайский |таиландские |таиландский )`. Если пробел поместить за пределами последовательности вариантов, например, задать `(тайские|тайский|таиландские|таиландский)? баты?`, не будет найдено соответствие с “*баты*” и “*бат*”, поскольку пробел окажется обязательным.

Если вы ищете ряд объектов, включая дефис (-), его нужно задавать в списке последним. Например, если вы ищете запятую (,) или дефис (-), то задайте `[,-]`, а не `[-,]`.

Порядок строк в списках и макросах

Длинные последовательности следует определять до коротких; иначе возможное соответствие длинной последовательности останется непрочитанным, потому что сначала будет обнаружено соответствие короткой последовательности. Например, если вы ищите строки "*расчет*" и "*счет*", строку "*расчет*" нужно задать до строки "*счет*". Например, можно задать `(расчет|счет)`, но не `(счет|расчет)`. То же применимо к макросам, поскольку они представляют собой списки строк.

Порядок правил в разделе определений

Определяйте только одно правило на строке. В пределах каждого раздела правила нумеруются (*regexp1*, *regexp2* и так далее). Эти правила надо нумеровать последовательно, от 1 до *n*. Любой пропуск в нумерации приведет к остановке обработки этого файла. Чтобы отключить запись, добавьте символ # в начало строки, в которой определяется регулярное выражение. Чтобы включить запись, удалите символ # в начале этой строки.

Для правильно обработки в каждом разделе более конкретные правила следует определять до более общих. Например, если вы ищите дату в формате "*месяц год*" и в формате "*месяц*", то правило "*месяц год*" должно быть определено до правила "*месяц*". Определять нужно так:

```
#@# January 1932
regexp1=$(MONTH),? [0-9]{4}

#@# January
regexp2=$(MONTH)
```

а не так:

```
#@# January
regexp1=$(MONTH)

#@# January 1932
regexp2=$(MONTH),? [0-9]{4}
```

Использование макросов в правилах

Если конкретная последовательность используется в различных правилах, можно использовать макрос. Тогда, если понадобится изменить определение этой последовательности, достаточно будет одного изменения на все правила, в которых используется изменяемая последовательность. Допустим, например, что у вас есть следующий макрос:

```
MONTH=((январь|февраль|март|апрель|июнь|июль|август|сентябрь|октябрь|
ноябрь|декабрь)|(янв|фев|мар|апр|май|июн|июл|авг|сен|окт|ноя|дек)(\.)?)
```

При использовании макроса вокруг имени нужно добавлять символы `$()`, например: `regexp1=$(MONTH)`

Все макросы должны быть определены в разделе `[macros]`.

## Нормализация

При извлечения нелингвистических объектов обнаруженные объекты нормализуются, объединяясь в группу похожих объектов согласно предварительно определенным форматам. Например, символы денежных единиц и эквивалентные им слова обрабатываются как одинаковые объекты. Объекты нормализации хранятся в разделе **Нормализация** на вкладке Расширенные ресурсы. Дополнительную информацию смотрите в разделе Глава 18, "О расширенных ресурсах", на стр. 203. Файл разбит на разделы.

**Важно!** Этот файл предназначен только для опытных пользователей. Скорее всего, вам не понадобится вносить изменения в этот файл. Если вам нужна дополнительная помощь в этой области, обратитесь в IBM Corp..

Правила форматирования для раздела Нормализация

- Добавляйте только одну запись нормализации на строку.
- Строго соблюдайте разделы в этом файле. Новые разделы добавлять нельзя.
- Чтобы отключить запись, добавьте символ # в начало строки. Чтобы включить запись, удалите символ # в начале этой строки.

Английские даты в нормализации

По умолчанию даты в английском шаблоне распознаются в формате даты в американском стиле, то есть в порядке месяц, день, год. Если нужно заменить этот формат на день, месяц, год, отключите строку "format:US" (добавив символ # в начало строки) и включите "format:UK" (удалив символ # из этой строки).

## Конфигурация

Включать и выключать типы нелингвистических объектов, которые нужно извлекать, можно в файле конфигурации нелингвистических объектов. Отключая ненужные объекты, можно уменьшить время, необходимое для обработки. Это делается в разделе **Конфигурация** на вкладке Расширенные ресурсы. Дополнительную информацию смотрите в разделе Глава 18, “О расширенных ресурсах”, на стр. 203. Если извлечение нелингвистических объектов включено, в процессе извлечения механизм извлечения читает файл конфигурации, чтобы узнать, какие типы нелингвистических объектов нужно извлекать.

Синтаксис этого файла:

*#имя*<TAB>*Язык*<TAB>*Код*

*Таблица 41. Синтаксис для файла конфигурации.*

| Метка столбца | Описание |
|---|---|
| #имя | Это название используется для обозначения нелингвистических объектов еще в двух файлах, необходимых для извлечения нелингвистических объектов. Имена здесь задаются с учетом регистра. |
| Язык | Язык документов . Лучше выбрать конкретный язык, но есть также опция **Любой**. Возможные опции: 0 = Любой используемый, если регулярное выражение не зависит от языка и может использоваться в различных шаблонах с разными языками, например, адреса IP, URL и электронной почты; 1 = французский; 2 = английский; 4 = немецкий; 5 = испанский; 6 = голландский; 8 = португальский; 10 = итальянский. |
| Код | Код части речи. Для большинства объектов, за несколькими исключениями, допустимо значение “s”. Возможные значения: s = стоп-слово; a = прилагательное; n = существительное. Если разрешено, сначала извлекаются нелингвистические объекты, и применяются паттерны извлечения, чтобы идентифицировать роль объекта в более широком контексте. Например, для процентов задается значение “a”. Пусть 30% извлекается как нелингвистический объект. Этот элемент будет идентифицирован как прилагательное. Тогда, если текст содержит словосочетание "30% повышение оклада", нелингвистический объект “30%” соответствует паттерну частей речи “ann” (прилагательное существительное существительное). |

Порядок в определении объектов

Порядок, в котором объекты декларируются в этом файле, существен и влияет на извлечение. Декларации применяются в том порядке, в котором перечислены. Изменение порядка изменит результаты. Более конкретные нелингвистические объекты следует определять до более общих.

Например, нелингвистический объект “Aminoacid” (аминокислота) определяется так:

```
regexp1=($(AA)-?$(NUM))
```

где $(AA) соответствует “(ala|arg|asn|asp|cys|gln|glu|gly|his|ile|leu|lys|met|phe|pro|ser)”, набор трехбуквенных аббревиатур для конкретных аминокислот.

С другой стороны, более общий нелингвистический объект "Gene" (ген) определяется так:

```
regexp1=p[0-9]{2,3}
regexp2=[a-z]{2,4}-?[0-9]{1,3}-?[r]
regexp3=[a-z]{2,4}-?[0-9]{1,3}-?p?
```

Если "Gene" определить до "Aminoacid" в разделе Конфигурация, соответствий объекту "Aminoacid" вообще не будет, поскольку сначала будет найдено соответствие regexp3 из объекта "Gene".

Правила форматирования для раздела Конфигурация

- Используйте символ TAB для отделения каждой записи в столбце.
- Не удаляйте строки.
- Соблюдайте синтаксис, показанный в предыдущей таблице.
- Чтобы отключить запись, добавьте символ # в начало строки. Чтобы включить объект, удалите символ # в начале этой строки.

---

## Языковая обработка

У каждого живого языка есть свои особые способы для выражения идей, структурирования предложений и использования сокращений. В разделе Языковая обработка можно отредактировать паттерны извлечения, задать принудительные определения для этих паттернов и объявить сокращения для языка, выбранного в выпадающем списке Язык.

- Паттерны извлечения
- Принудительные определения
- Сокращения

## Паттерны извлечения

При извлечении информации из документов механизм извлечения применяет набор паттернов извлечения частей речи к "стеку" слов в тексте, чтобы идентифицировать термины (слова и словосочетания) - кандидаты на извлечение. Вы можете добавлять и редактировать паттерны извлечения.

Части речи включают в себя грамматические элементы, такие как существительные, прилагательные, причастия, детерминативы, предлоги, координаторы, имена, инициалы и частицы. Паттерн извлечения частей речи представляет собой последовательность таких элементов. Для удобства задания паттерна в продуктах исследования текстов IBM Corp. каждая часть речи представлена одним символом. Например, прилагательное (adjective) представлено строчной буквой *a*. Чтобы облегчить понимание использования кодов набор поддерживаемых кодов выводится по умолчанию в верхней части каждого раздела паттернов извлечения наряду с набором паттернов и примерами по каждому паттерну.

Правила форматирования для паттернов извлечения

- Один паттерн на строке.
- Чтобы выключить паттерн, используйте # в начале строки.

Порядок, в котором перечислены паттерны извлечения, важен, поскольку данная последовательность слов прочитывается механизмом извлечения только один раз, и для нее задаются первые паттерны извлечения, для которых механизм нашел соответствие.

## Принудительные определения

При извлечении информации из документов механизм извлечения просматривает текст и идентифицирует часть речи для каждого встреченного слова. В некоторых случаях слово может соответствовать нескольким возможным ролям в зависимости от контекста. Если нужно принудительно задать слову конкретную роль как части речи или вовсе исключить слово из обработки, это можно сделать в разделе **Принудительное определение** вкладки Расширенные ресурсы. Дополнительную информацию смотрите в разделе Глава 18, "О расширенных ресурсах", на стр. 203.

Чтобы принудительно задать для слова некоторую роль как части речи, нужно добавить в этот раздел строку, используя такой синтаксис:

*термин*`:`*код*

*Таблица 42. Описание синтаксиса.*

| Запись | Описание |
|---|---|
| `term` | Имя термина. |
| `code` | Однобуквенный код, представляющий роль слова как части речи. Можно перечислить до шести кодов различных частей речи для одного отдельного термина. Кроме того, можно отменить извлечение слова в составные слова / словосочетания, задав строчную букву `s`, например `additional:s`. |

Правила форматирования для принудительных определений

- Одна строка на слово.
- Термины не могут содержать двоеточие.
- Используйте в качестве кода части речи строчную букву `s`, чтобы слово не извлекалось вовсе.
- Используйте до шести кодов части речи в одной строке. Поддерживаемые коды частей речи выводятся в разделе Паттерны извлечения. Дополнительную информацию смотрите в разделе "Паттерны извлечения" на стр. 211.
- Используйте звездочку (*) как символ подстановки в конце строки для поиска частичных соответствий. Например, если ввести `add*:s`, то такие слова, как `add`, `additional`, `additionally`, `addendum`, и `additive` не будут извлекаться как термин или часть составного слова-термина. Несмотря на это, слово все же будет извлечено, если соответствие слову явным образом объявлено как термин в скомпилированном словаре или в принудительных определениях. Например, если ввести и `add*:s`, и `addendum:n`, то при нахождении в тексте слова `addendum` оно будет извлечено.

## Сокращения

Когда механизм извлечения обрабатывает текст, точка обычно считается концом предложения. Это верно в типичном случае, но не тогда, когда текст содержит аббревиатуры с точками.

Если при извлечении терминов из текста обнаруживается неправильно обработанное сокращение, объявите это сокращение явным образом в этом разделе.

*Примечание*: Если аббревиатура уже есть в списке синонимов или определена как термин в словаре типов, добавлять здесь значение аббревиатуры не требуется.

Правила формата для сокращений

- Определяйте только одну аббревиатуру на строке.

---

## Идентификатор языка

Хотя для анализа текстовых данных лучше, если выбран конкретный язык, иногда текст написан на нескольких языках или на неизвестном языке, и на этот случай предусмотрена опция **Все**. Языковая опция **Все** использует механизм автоматического распознавания языка, который называется Идентификатор языка. Идентификатор языка просматривает документы , чтобы идентифицировать текст на поддерживаемом

языке, и при извлечении автоматически применяет для каждого файла оптимальный внутренний словарь. Опция **Все** управляется параметрами в разделах свойств.

## Свойства

Идентификатор языка конфигурируется с использованием параметров в этом разделе. В следующей таблице описаны параметры, которые можно задавать в разделе **Идентификатор языка - Свойства** на вкладке Расширенные ресурсы. Дополнительную информацию смотрите в разделе Глава 18, “О расширенных ресурсах”, на стр. 203.

*Таблица 43. Описания параметров*

| Параметр | Описание |
|---|---|
| NUM_CHARS | Задает число символов, которое должен прочитать механизм извлечения при определении языка, на котором написан текст. Чем меньше это число, тем быстрее будет идентифицирован язык. Чем больше это число, тем точнее будет идентифицирован язык. Если задать 0, будет прочитан весь текст документы . |
| USE_FIRST_SUPPORTED _LANGUAGE | Задает, должен ли механизм извлечения использовать первый поддерживаемый язык, найденный идентификатором языка. Если задать 1, используется первый поддерживаемый язык. Если задать 0, используется базовы язык. |
| FALLBACK_LANGUAGE | Задает так называемый базовый язык, то есть язык, используемый в том случае, когда идентификатор вернул неподдерживаемый язык. Возможные значения - english (английский), french (французский), german (немецкий), spanish (испанский), dutch (голландский), italian (итальянский) и ignore (игнорировать). Если задать значение ignore, документ на неподдерживаемом языке будет проигнорирован. |

## Языки

Идентификатор языка поддерживает многие языки. Список языков можно отредактировать в разделе **Идентификатор языка - Языки** на вкладке Расширенные ресурсы.

Есть смысл устранить из списка ненужные языки, поскольку чем больше языков, тем выше вероятность ложно-положительных результатов и тем ниже производительность. Однако добавить новые языки в этот файл нельзя. Есть смысл поместить наиболее вероятные языки в начало списка, чтобы помочь идентификатору языка быстрее находить соответствие для ваших документов.

# Глава 19. О правилах текстовых связей

Анализ текстовых связей (text link analysis, TLA) - это метод сопоставления паттернов для извлечения взаимосвязей, которые ищутся в тексте по некоторому набору правил. Если при извлечении включен TLA, текстовые данные сравниваются с паттернами по этим правилам. При обнаружении соответствия паттерн TLA извлекается и предъявляется. Правила определяются на вкладке Правила текстовых связей.

Например, иногда недостаточно извлекать понятия, представляющие простые идеи о некоторой организации, тогда как при помощи TLA удается узнать о связях между организациями или о людях, ассоциированных с данной организацией. Кроме того, при помощи TLA можно извлекать мнения о тех или иных темах, например, как относятся люди к данному товару и каковы их впечатления от него.

Чтобы воспользоваться преимуществами TLA, нужны ресурсы, содержащие правила текстовых связей (правила TLA). Когда вы выбираете шаблон, вы можете видеть, у каких шаблонов есть правила TLA, по наличию значка в столбце TLA.

Паттерны анализа текстовых связей ищутся в текстовых данных во время извлечения на фазе сопоставления паттернов. На этой фазе правила сравниваются с текстовыми данными и, когда найдено соответствие, информация извлекается как паттерн. Иногда нужно получить дополнительные сведения от анализа текстовых связей или изменить способ сопоставления. В таких случаях можно уточнить правила согласно конкретным потребностям. Это делается на вкладке Правила текстовых связей.

*Примечание*: Поддержка переменных была прекращена в версии 13. Вместо этого используйте макросы. Дополнительную информацию смотрите в разделе “Работа с макросами” на стр. 220.

## Где работать с правилами текстовых связей

Изменять и создавать правила можно непосредственно на вкладке Правила текстовых связей в представлении Редактор шаблонов или Редактор ресурсов. Запуск имитации на этой вкладке поможет увидеть, как правила могут искать соответствия в тексте. При имитации извлечение запускается только для выборки данных имитации, и правила текстовых связей применяются для определения, существуют ли какие-то совпадающие паттерны. Любые правила, дающие результат для этого текста, показываются затем на панели имитации. На основании полученных совпадений можно выбрать изменения правил и макросов, чтобы совпадающим был другой текст.

В отличии от других расширенных ресурсов, правила TLA зависят от библиотеки, и поэтому одновременно можно использовать только правила TLA из одной библиотеки. Из Редактор шаблонов или Редактор ресурсов перейдите на вкладку **Правила текстовых связей**. На этой вкладке можно выбрать библиотеку в шаблоне, содержащую правила TLA, которые нужно использовать или редактировать. По этой причине настоятельно рекомендуется, если нет веских оснований поступить иначе, хранить все правила в одной библиотеке.

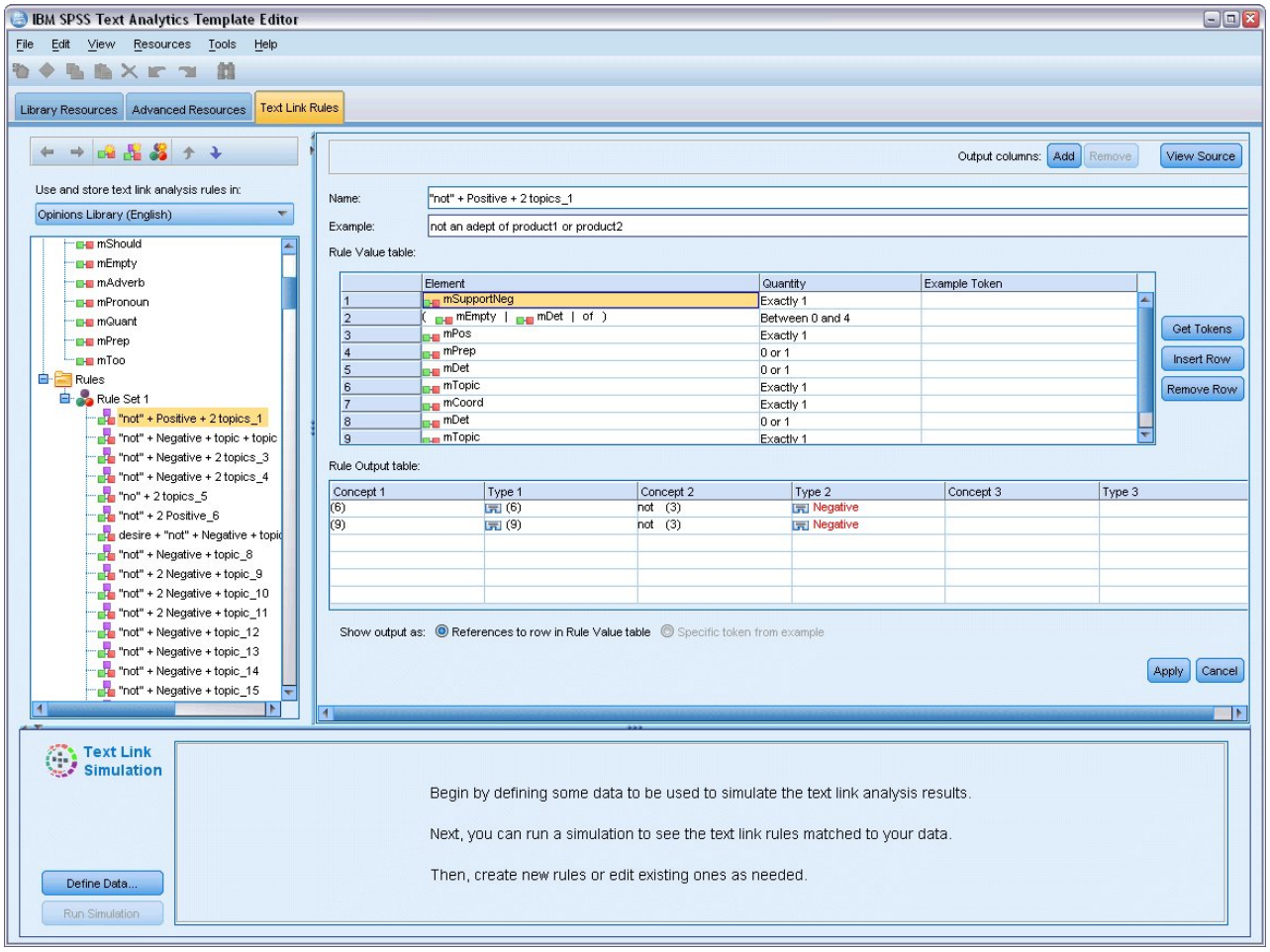



*Рисунок 42. Вкладка Правила текстовых связей*

**Важно!** Эта вкладка недоступна для ресурсов на японском языке.

## С чего начинать работу

Есть несколько способов начала работы в редакторе на вкладке Правила текстовых связей:

- Начните с имитации результатов, используя какой-то образец текста, и измените или создайте правила соответствия на основании того, как текущий набор правил извлекает паттерны из данных имитации.
- Создайте новое правило с нуля или измените существующее правило.
- Работайте непосредственно в представлении исходного кода.

## Когда изменять или создавать правила

Хотя предоставляемые с каждым шаблоном правила анализа текстовых связей часто оказываются адекватными для извлечения многих простых и сложных взаимосвязей из вашего текста, бывают ситуации, когда могут потребоваться некоторые изменения этих правил или создание ваших собственных правил. Например:

- Захватить смысл или отношение, не извлекаемые существующими правилами, создав новое правило или макрос.
- Изменить поведение по умолчанию для типа, который вы добавили к ресурсам. Обычно для этого требуется изменение макроса, такого как `mTopic` или `mNonLingEntities`. Дополнительную информацию смотрите в разделе “Специальные макросы: mTopic, mNonLingEntities, SEP” на стр. 222.

- Добавить новые типы к существующим макросам и правилам анализа текстовых связей. Например, если вы оцениваете тип <Организация> как слишком широкий, можно создать новые типы для организаций в нескольких разных отраслях, такие как <Фармацевтическая компания>, <Автопроизводитель>, <Финансовая компания> и так далее. В этом случае необходимо изменить правила анализа текстовых связей и/или создать макрос, чтобы учитывать эти новые типы и соответствующим образом их обрабатывать.
- Добавить новые типы к существующему правилу анализа текстовых связей. Допустим, например, что у вас есть правило, захватывающее следующий текст - john doe called jane doe, но вы хотите, чтобы это правило, захватывающее сведения о телефонных разговорах, распространялось и на обмен сообщениями электронной почты. Вы могли бы добавить в правило тип нелингвистического объекта для электронных адресов, чтоб захватывался такой текст как johndoe@ibm.com emailed janedoe@ibm.com.
- Немного изменить существующее правило вместо создания нового. Допустим, например, что у вас есть правило, соответствующее следующему тексту xyz is very good, но вы хотели бы захватывать также такой текст как xyz is very, very good.

## Имитация результатов анализа текстовых связей

Использование части текста и запуск имитации часто может помочь в определении новых правил текстовых связей и в понимании сопоставления определенных предложений при анализе текстовых связей. При имитации извлечение запускается только для выборки данных имитации с использованием текущего набора лингвистических ресурсов и текущих параметров извлечения. Цель этой процедуры - получение результатов имитации и их использование для усовершенствования существующих и создания новых правил или для лучшего понимания хода поиска совпадений. Для каждого элемента текста (в зависимости от контекста - для предложения, слова или условия) выход имитации показывает собрание маркеров и все правила TLA, приведшие к обнаружению паттерна в тексте. **Маркер** определяется как любое слово или сочетание слов, определенное в процессе извлечения.

В отличии от других расширенных ресурсов, правила TLA зависят от библиотеки, и поэтому одновременно можно использовать только правила TLA из одной библиотеки. Из Редактор шаблонов или Редактор ресурсов перейдите на вкладку **Правила текстовых связей**. На этой вкладке можно выбрать библиотеку в шаблоне, содержащую правила TLA, которые нужно использовать или редактировать. По этой причине настоятельно рекомендуется, если нет веских оснований поступить иначе, хранить все правила в одной библиотеке.

**Важно!** При использовании файла данных для минимизации времени обработки настоятельно рекомендуется предварительно убедиться, что содержащийся в нем текст короткий. Цель имитации - увидеть, как интерпретируется часть текста, и понять, как правила сопоставляются с этим текстом. Эта информация поможет вам писать и изменять правила. Используйте узел анализа текстовых связей или запустите поток в интерактивном сеансе с включенным извлечением TLA, чтобы получить результаты для более полного набора данных. Эта имитация предназначена только для целей проверки и разработки правил.

## Определение данных для имитации

Запуск имитации с использованием выборки данных поможет увидеть, как правила могут сопоставляться с текстом. Первый шаг - это определение данных.

### Задание свойств данных

1. Щелкните по ссылке **Определить данные** на панели имитации в нижней части вкладки **Правила текстовых связей**. Как вариант, если данные были определены ранее, выберите пункт меню **Инструменты > Запустить имитацию**. Откроется мастер по данным имитации.
2. Укажите тип данных, выбрав одну из следующих возможностей:
   - **Непосредственный ввод или вставка текста** Предоставляется текстовое поле, в которое можно вставить текст из буфера обмена или ввести нужный для обработки текст вручную. Можно вводить по одному предложению на строку или использовать знаки препинания (точки или запятые) для разбивки предложения. После ввода текста можно начать процесс имитации, нажав кнопку **Запустить имитацию**.

- **Указать источник данных файла** Эта опция означает, что вы хотите обработать содержащий текст файл. Нажмите кнопку **Далее** для перехода к шагу мастера, где можно определить файл для обработки. После выбора файла можно начать имитацию, нажав кнопку **Запустить имитацию**. Поддерживаются типы файлов `.txt` и `.text`. Во время имитации выбранный файл данных читается 'как есть'. Весь файл обрабатывается так же, как при наличии соединения узла Список файлов с узлом Исследование текста.

**Важное замечание:** При использовании файла данных для минимизации времени обработки настоятельно рекомендуется предварительно убедиться, что содержащийся в нем текст короткий. Цель имитации - увидеть, как интерпретируется часть текста, и понять, как правила сопоставляются с этим текстом. Эта информация поможет вам писать и изменять правила. Используйте узел анализа текстовых связей или запустите поток в интерактивном сеансе с включенным извлечением TLA, чтобы получить результаты для более полного набора данных. Эта имитация предназначена только для целей проверки и разработки правил.

3. Чтобы начать процесс имитации, нажмите кнопку **Запустить имитацию**. Появится диалоговое окно с ходом выполнения. При работе в интерактивном сеансе используемые для имитации параметры извлечения - это параметры, в настоящее время выбранные в этом интерактивном сеансе (смотрите пункт меню **Инструменты > Параметры извлечения** в представлении Понятия и категории). При работе в Редактор шаблонов используемые для имитации параметры извлечения - это параметры извлечения по умолчанию, совпадающие с параметрами, показанными на вкладке Эксперт узла Анализ текстовых связей. Дополнительную информацию смотрите в разделе “Как понять результаты имитации”.

## Как понять результаты имитации

Запуск имитации с использованием выборки данных и пересмотр результатов поможет увидеть, как правила могут сопоставляться с текстом. С учетом этого вы можете изменить набор правил, лучше отвечающих вашим данным. После завершения извлечения и процесса имитации вам будут представлены результаты имитации.

Для каждого “предложения”, идентифицированного при извлечении, вам будет представлено несколько частей информации, в том числе точное "предложение", разбиение по маркерам, обнаруженным в этом входящем текстовом предложении, и окончательные правила, соответствующие тексту этого предложения. Здесь "**предложением**" может считаться слово, предложение или условие в зависимости от того, как механизм извлечения разбивает текст на читаемые чанки.

**Маркер** определяется как любое слово или словосочетание, идентифицированное при извлечении. Например, в предложении *Мой дядя приехал в Нижний Новгород* при извлечении могут быть идентифицированы маркеры *мой*, *дядя*, *приехал*, *в* и *нижний новгород*. Кроме того, *дядя* может быть извлечен как понятие, которому будет задан тип `<Неизвестный>`, а *нижний новгород* может быть извлечен как понятие, которому задан тип `<Положение>`. Все понятия - маркеры, но все маркеры - понятия. Маркеры могут быть также макросами, литеральными строками и промежутками между словами. Понятиями могут быть только те слова и словосочетания, для которых задан тип.

При нахождении в интерактивном сеансе или в редакторе ресурсов вы работаете на уровне понятий. Правила TLA более детализированы, и отдельные маркеры в предложении можно использовать в определении правила, даже если они никогда не извлекаются и не вводятся. Возможность использования маркеров, которые не представляют собой понятия, делает правила еще более гибкими при захвате сложных взаимосвязей в вашем тексте.

Если в данных для имитации несколько предложений, перемещаться вперед и назад по результатам можно, нажимая кнопки **Следующее** и **Предыдущее**.

В тех случаях, когда предложение не соответствует ни одному правилу TLA в выбранной библиотеке (смотрите имя библиотеки над деревом на этой вкладке), результаты считаются неподходящими, и включаются кнопки **Следующее несогласование** и **Предыдущее несогласование**, показывающие, что есть текст, для которого согласования с правилом не найдено, и позволяющие быстро перейти к этим примерам.

После создания новых или изменения существующих правил, а также после изменения ресурсов или параметров извлечения, может потребоваться повторный запуск имитации. Для перезапуска имитации нажмите кнопку **Запустить имитацию** на панели имитации, и будут снова использованы те же входные данные.

Следующие поля и таблицы выводятся в результатах имитации:

**Входной текст**. Фактическое 'предложение', идентифицированное в процессе извлечения среди данных имитации, определенных вами в мастере. Здесь предложением может считаться слово, предложение или условие в зависимости от того, как механизм извлечения разбивает текст на читаемые чанки.

**Представление системы**. Собрание маркеров, идентифицированных процессом извлечения.

- **Маркер входного текста.** Каждый маркер, найденный во входном тексте. Маркеры были определены ранее в этой теме.
- **Определенный тип.** Если маркер был идентифицирован как понятие с назначенным типом, в этом столбце будет показано соответствующее имя типа (такое как `<Неизвестно>`, `<Человек>`, `<Положение>`).
- **Совпадающий макрос.** Если маркер соответствует существующему макросу, в этом столбце выводится соответствующее имя макроса.

**Соответствующие правила во входном тексте.** В этой таблице показываются все правила TLA, для которых были найдены соответствия во входном тексте. Для каждого подошедшего правила будет показано имя этого правила в столбце **Выход правила** и его связанные выходные значения (пары Понятие + тип). Можно дважды щелкнуть по имени соответствующего правила, чтобы открыть его на панели редактора над панелью имитации.

Кнопка **Генерировать правило**. Если нажать эту кнопку на панели имитации, откроется новое правило на панели редактора правил над панелью имитации. Входной текст будет принят как пример для редактора. Аналогично любой маркер с назначенным типом или с найденным при имитации соответствием макросу будет автоматически вставлен в столбец Элементы в таблице **Значения правил**. Если маркеру был назначен тип *и* он был сопоставлен с макросом, значением макроса будет использоваться в правиле для его упрощения. Например, предложению "*I like pizza*" при имитации мог быть назначен тип `<Неизвестный>` и найдено соответствие с макросом `mTopic`, если использовались ресурсы на базовом английском языке. В этом случае `mTopic` будет использоваться как элемент в сгенерированном правиле. Дополнительную информацию смотрите в разделе "Работа с правилами TLA" на стр. 223.

## Навигация по дереву правил и макросов

Если при извлечении анализируются текстовые связи, используются правила TLA, хранящиеся в той библиотеке, которая выбрана на вкладке **Правила текстовых связей**.

В отличии от других расширенных ресурсов, правила TLA зависят от библиотеки, и поэтому одновременно можно использовать только правила TLA из одной библиотеки. Из Редактор шаблонов или Редактор ресурсов перейдите на вкладку **Правила текстовых связей**. На этой вкладке можно выбрать библиотеку в шаблоне, содержащую правила TLA, которые нужно использовать или редактировать. Поэтому настоятельно рекомендуется, если нет веских причин поступить иначе, хранить все правила в одной библиотеке.

Чтобы указать библиотеку, с которой нужно работать на вкладке Правила текстовых связей, можно выбрать нужную библиотеку на этой вкладке в выпадающем списке **Использовать и сохранять правила TLA в:**. Если при извлечении анализируются текстовые связи, используются правила TLA, хранящиеся в той библиотеке, которая выбрана на вкладке **Правила текстовых связей**. Поэтому, если вы определили правила TLA в нескольких библиотеках, в анализе связей будет использоваться только первая библиотека, в которой найдены правила TLA. По этой причине настоятельно рекомендуется, если нет веских оснований поступить иначе, хранить все правила в одной библиотеке.

Когда вы выбираете в дереве макрос или правило, его содержимое выводится на панели редактора справа. Если щелкнуть правой кнопкой по элементу в дереве, откроется контекстное меню со списком других допустимых задач, таких как:

- Создать новый макрос в дереве и открыть его в редакторе справа.
- Создать новое правило в дереве и открыть его в редакторе справа.
- Создать новый набор правил в дереве.
- Вырезать, скопировать и вставить элементы для удобства редактирования.
- Удалить макросы, правила и наборы правил, чтобы удалить их из ресурсов.
- Отключить макросы, правила и наборы правил, чтобы указать, что при обработке их следует игнорировать.
- Переместить правила вверх и вниз, чтобы изменить порядок обработки.

Предупреждения в дереве

Предупреждения выводятся в дереве при помощи желтого треугольника и сигнализируют о возможной проблеме. Остановите указатель мыши над проблемным макросом или правилом, чтобы вывести всплывающее объяснение. В большинстве случаев это выглядит примерно так: **Предупреждение: не задан ни один пример; введите пример**. Это значит, что нужно ввести пример.

Поскольку без примера, или если пример не соответствует правилу, недоступна функция Получить маркеры, рекомендуется ввести хотя бы один пример для каждого правила.

Если правило выделено желтым, это значит, что редактору TLA неизвестен тот или иной тип или макрос. Сообщение будет выглядеть примерно так: **Предупреждение: неизвестный тип или макрос**. Оно сигнализирует о том, что некоторый элемент будет определен как `$something` в представлении источника, например, если элемент `$myType` не окажется ни унаследованным типом в вашей библиотеке, ни макросом.

Чтобы обновить средство проверки синтаксиса, достаточно переключиться в другое правило или макрос; перекомпилировать что-либо не требуется. Таким образом, если, например, для правила `A` выведено предупреждение об отсутствии примера, нужно добавить пример, щелкнуть по правилу выше или ниже данного правила, потом вернуться к правилу `A` и проверить, стало ли оно правильным.

## Работа с макросами

При помощи макросов удается упростить вид правил TLA благодаря группировке типов, других макросов и литеральных строк (слов) с использованием операции `ИЛИ` (`|`). Преимущество использования макросов не только в том, что макросы можно использовать повторно в нескольких правилах TLA, упрощая эти правила, но и в возможности вносить изменения в один макрос вместо изменений по всем правилам TLA. Большинство поставляемых правил TLA содержат предварительно определенные макросы. Макросы выводятся на верхнему уровне дерева на панели в левой части вкладки Правила текстовых связей.

Следующие поля и таблицы выводятся в результатах имитации:

**Имя**. Уникальное имя, идентифицирующее этот макрос. Для удобства чтения правил рекомендуется начинать имена макросов с префикса в виде строчной буквы m. При вставке ссылок на макросы в правило вручную (путем редактированием строки или в представлении источника) нужно добавлять префикс в виде символа $, чтобы процесс извлечения нашел такое специальное имя. Если перетащить имя макроса или выбрать макрос в контекстном меню, продукт автоматически распознает элемент как макрос, а символ $ добавлен не будет.

Таблица **Значения макроса**.

- Несколько строк, представляющих все возможные значения, представляемые этим макросом. Эти значения задаются с учетом регистра.

- Такое значение может содержать, по отдельности или в сочетании, такие элементы, как типы, литеральные строки, промежутки между словами и макросы. Дополнительную информацию смотрите в разделе “Поддерживаемые элементы для правил и макросов” на стр. 230.
- Чтобы ввести значение для элемента в макросе, щелкните дважды по нужной строке. Откроется окно редактируемого текста, в котором можно ввести ссылку на тип, ссылку на макрос, литеральную строку или промежуток между словами. Другой вариант - щелкнуть правой кнопкой мыши по ячейке, чтобы вывести контекстное меню со списком обычных макросов, имен типов и имен нелингвистических типов. В ссылке на тип или макрос перед их именем нужно добавлять символ ‘$’, например, ввести `$mTopic` для макроса `mTopic`. В случае сочетания аргументов нужно группировать их, заключая в скобки `( )`, и разделять символом `|`, означающим логическую операцию `ИЛИ`.
- Можно добавлять и удалять строки в таблице значений макроса, нажимая кнопки справа.
- Вводите каждый элемент в отдельной строке таблицы. Например, если нужно создать макрос, представляющий 3 литеральные строки `был ИЛИ есть ИЛИ будет`, каждую литеральную строку нужно ввести в отдельной строке представления, так что таблица макроса будет содержать три строки.

## Создание и редактирование макросов

Вы можете создавать новые макросы или изменять существующие. Следуйте рекомендациям и описаниям для редактора макросов. Дополнительную информацию смотрите в разделе “Работа с макросами” на стр. 220.

Создание новых макросов

1. В меню выберите **Инструменты > Создать макрос**. Другой вариант - щелкните по значку Создать макрос на панели инструментов дерева для открытия нового макроса в редакторе.
2. Введите уникальное имя и определите элементы значений макроса.
3. Закончив, щелкните по **Применить**, чтобы проверить на наличие ошибок.

Правка макросов

1. Щелкните по имени макроса в дереве. Макрос откроется в панели редактора справа.
2. Выполните свои изменения.
3. Закончив, щелкните по **Применить**, чтобы проверить на наличие ошибок.

## Отключение и удаление макросов

Отключение макросов

Если нужно игнорировать макрос при обработке, его можно отключить. При этом могут появиться сообщения о предупреждениях или ошибках в правилах, которые по-прежнему ссылаются на отключенный макрос. С осторожностью отключайте или удаляйте макросы.

1. Щелкните по имени макроса в дереве. Макрос откроется в панели редактора справа.
2. Щелкните правой кнопкой по имени.
3. В контекстном меню выберите **Отключить**. Значок макроса станет затененным, а сам макрос будет недоступен для изменений.

Удаление макроса

Если макрос больше не нужен, его можно удалить. При этом могут возникнуть ошибки в правилах, которые по-прежнему ссылаются на этот макрос. С осторожностью отключайте или удаляйте макросы.

1. Щелкните по имени макроса в дереве. Макрос откроется в панели редактора справа.
2. Щелкните правой кнопкой по имени.
3. В контекстном меню выберите **Удалить**. Макрос исчезнет из списка.

## Проверка ошибок, сохранения и отмены

Применение изменений макросов

Если щелкнуть мышью вне редактора макросов или нажать кнопку **Применить**, макрос будет автоматически проверен на наличие ошибок. Если ошибка обнаружится, нужно исправить ее, прежде чем переходить к другой части прикладной программы.

Однако при обнаружении менее серьезных ошибок выводится только предупреждение. Например, сообщение с предупреждением появится, если в вашем макросе есть неполные определения типов или других макросов или определения без ссылок. После нажатия кнопки **Применить**, если какие-то предупреждения не исправлены, появится значок предупреждения слева от имени макроса на левой панели дерева Правила и макросы.

Применение макроса не означает, что он будет навсегда сохранен. При применении инициируется процесс проверки ошибок и предупреждений.

Сохранение ресурсов в сеансе интерактивной инструментально среды

1. Чтобы сохранить изменения, внесенные в ваши ресурсы в сеансе интерактивной инструментальной среды, и обеспечить возможность их использования при следующем запуске потока, необходимо:
   - Изменить ваш узел моделирования для обеспечения возможности получения тех же ресурсов при следующем выполнении потока. Дополнительную информацию смотрите в разделе “Обновление узлов моделирования и сохранение” на стр. 84. После этого сохраните ваш поток. Чтобы сохранить поток, сохранение нужно выполнить в главном окне IBM SPSS Modeler после обновления узла моделирования.
2. Чтобы сохранить изменения, внесенные в ваши ресурсы в сеансе интерактивной инструментальной среды, и обеспечить возможность их использования в других потоках, можно сделать следующее:

- Изменить использованный шаблон или создать новый. Дополнительную информацию смотрите в разделе “Создание и изменение шаблонов” на стр. 167. При этом изменения для текущего узла сохранены не будут (смотрите предыдущий шаг)
- Другой вариант - изменить используемый TAP. Дополнительную информацию смотрите в разделе “Изменение пакетов анализа текста” на стр. 142.

Сохранение ресурсов в Редактор шаблонов

1. Сначала опубликуйте библиотеку. Дополнительную информацию смотрите в разделе “Публикация библиотек” на стр. 185.
2. Затем сохраните шаблон с помощью опции меню **Файл > Сохранить шаблон ресурсов**.

Отмена изменений макросов

1. Если нужно отбросить изменения, нажмите кнопку **Отмена**.

## Специальные макросы: mTopic, mNonLingEntities, SEP

Шаблон Opinions (и аналогичные шаблоны), а также шаблон Basic Resources поставляются с двумя специальными макросами с именами `mTopic` и `mNonLingEntities`.

mTopic

По умолчанию макрос `mTopic` группирует все типы, поставляемые в шаблоне, которые с высокой вероятностью относятся к мнению, например, следующие типы из библиотеки *Core*: `<Персональный>`, `<Организация>`, `<Положение>` и так далее, кроме типа, относящегося к типу мнений (например, `<Отрицательный>` или `<Положительный>`) или типа, определенного как нелингвистический объект в Расширенных ресурсах.

При создании нового типа в шаблоне Opinions (или аналогичном) продукт предполагает, что, если этот тип не задан в другом макросе или в разделе нелингвистических объектов на вкладке Расширенные ресурсы, он будет обрабатываться, как другие типы, определенные в макросе mTopic.

Пусть вы создали новые типы в ресурсах из некоторого шаблона Opinions: <Овощи> и <Фрукты>. Если не вносить изменений, новые типы обрабатываются как типы mTopic, так что можно автоматически открывать положительные, отрицательные, нейтральные и контекстные мнения о новых типах. Например, при извлечении из предложения "*Я обожаю брокколи, но ненавижу грейпфруты*" будут созданы 2 выходных паттерна:

```
брокколи <Овощи> + нравится <Положительный>

грейпфруты<Фрукты> + не_нравится <Отрицательный>
```

Но если нужно обработать эти типы не так, как остальные типы в mTopic, можно добавить имя типа в существующий макрос, например, mPos, группирующий типы положительных мнений, или создать новый макрос, на который затем можно ссылаться в одном или нескольких правилах.

**Важно!** Если создать новый тип, например, <Овощи>, он будет включен как тип в mTopic, но его имя не будет явным образом показано в определении макроса.

mNonLingEntities

Аналогичным образом, если добавить новые нелингвистические объекты в разделе **Нелингвистические объекты** на вкладке Расширенные ресурсы, они будут автоматически обрабатываться как mNonLingEntities, если не указать иное. Дополнительную информацию смотрите в разделе “Нелингвистические объекты” на стр. 206.

SEP

Кроме того, можно использовать предварительно определенный макрос SEP, который соответствует глобальному разделителю, определенному на компьютере; обычно это запятая (,).

## Работа с правилами TLA

Правило анализа текстовых связей (text link analysis, TLA) - это логический запрос для поиска соответствия некоторому предложению. Правила TLA содержат один или несколько из следующих аргументов: типы, макросы, литеральные строки или промежутки между словами. Для извлечения результатов TLA требуется хотя бы одно правило TLA.

На вкладке Правила текстовых связей в редакторе правил есть следующие области и поля:

Поле **Имя**. Уникальное имя для правила TLA.

Поле **Пример**. Можно включить в правило пример предложения или последовательности слов, которые будут захвачены этим правилом. Рекомендуется использовать примеры. В этом редакторе из такого примера текста можно сгенерировать маркеры и увидеть, как этот текст сопоставляется правилу и как он будет выведен. **Маркер** определяется как любое слово или словосочетание, идентифицированное при извлечении. Например, в предложении *Мой дядя приехал в Нижний Новгород* при извлечении могут быть идентифицированы маркеры *мой*, *дядя*, *приехал*, *в* и *нижний новгород*. Кроме того, *дядя* может быть извлечен как понятие, которому будет задан тип <Неизвестный>, а *нижний новгород* может быть извлечен как понятие, которому задан тип <Положение>. Все понятия - маркеры, но все маркеры - понятия. Маркеры могут быть также макросами, литеральными строками и промежутками между словами. Понятиями могут быть только те слова и словосочетания, для которых задан тип.

**Таблица значений правила**. Эта таблица содержит элементы правила, используемые для сопоставления правила предложению. Можно добавлять и удалять строки в таблице, нажимая на кнопки справа. Таблица состоит из 3 столбцов:

- Столбец **Элемент**. Введите по отдельности или в сочетании такие элементы, как типы, литеральные строки, промежутки между словами (<Any Token>) или макросы. Дополнительную информацию смотрите в разделе “Поддерживаемые элементы для правил и макросов” на стр. 230. Щелкните дважды по ячейке элемента, чтобы ввести информацию непосредственно. Другой вариант - щелкнуть правой кнопкой мыши по ячейке, чтобы вывести контекстное меню со списком обычных макросов, имен типов и имен нелингвистических типов. Не забывайте, что при вводе информации в ячейке с клавиатуры перед именем типа или макроса нужно добавлять символ ‘$’, например, ввести $mTopic для макроса mTopic. Порядок, в котором вы создаете строки элементов, имеет значение при поиске соответствий этому правилу в тексте. В случае сочетания аргументов нужно группировать их, заключая в скобки ( ), и разделять символом |, означающим логическую операцию ИЛИ. Следует иметь в виду, что значения регистрозависимы.
- Столбец **Количество**. Показывает минимально и максимально допустимое число вхождений элемента для установления соответствия. Например, если нужно определить промежуток между словами или последовательность слов между двумя другими элементами длиной от 0 до 3 слов, можно выбрать **От 0 до 3** в списке или ввести эти числа непосредственно в диалоговом окне. Значение по умолчанию ‘**Ровно 1**’. Иногда нужно сделать некоторый элемент необязательным. В этом случае задайте минимально допустимое количество 0 и максимально допустимое количество больше 0 (примеры: 0 или 1, от 0 до 2). Обратите внимание на то, что первый элемент в правиле не может быть необязательным, то есть не допускает количества 0.
- Столбец **Пример маркера**. При щелчке по **Получить маркеры**, программа разбивает текст, заданный как **Пример**, на маркеры, которыми заполняет этот столбец в соответствии с определенными вами элементами. Эти же маркеры можно посмотреть в таблице вывода.

**Таблица вывода правил** В каждой строке этой таблице определяется, в каком виде найденные паттерны TLA будут выводиться в результатах. Вывод правила может создавать паттерны, содержащие до шести пар столбцов Понятие/Тип, каждая из которых представляет собой *слот*. Например, паттерн типа <Положение> + <Положительный> - это паттерн с двумя слотами, то есть он состоит из двух пар столбцов Понятие/Тип.

**Примечание:** Термины в столбце **Элемент** в таблице **Значения правила** или любые столбцы **Понятие** в таблице **Вывод правила** не могут начинаться со следующих символов : `, #, %, ^, *, _, -, :, <, >, /, \ или ".

Поскольку язык предоставляет свободу выражать некоторые базовые идеи самыми разными путями, иногда нужно задать целый ряд правил, чтобы захватить одну базовую идею. Например, текст *"Париж - это город, который я обожаю"* и текст *"Мне очень, очень понравились Париж и Флоренция"* оба содержат ту базовую идею, что Париж понравился, но эта идея выражена по-разному, и для захвата обоих текстов потребуются два разных правила. В то же время удобно работать с результатами паттернов, где сходные идеи сгруппированы. Поэтому, используя два разных правила для захвата этих двух фраз, можно определить для обоих правил один и тот же вывод, например, паттерн типа <Положение> + <Положительный>, чтобы представить им оба текста. Таким образом, вывод не всегда повторяет структуру или порядок слов оригинального текста. Более того, такой паттерн типа может соответствовать и другим фразам и может генерировать паттерны понятия, такие как: париж + нравится и токио + нравится.

Чтобы быстрее и с меньшим числом ошибок определить вывод, можно использовать контекстное меню, в котором выбрать нужный элемент для вывода. Другой вариант - перетащить элементы из таблицы значений правила в вывод. Например, если у вас есть правило со ссылкой на макрос mTopic в строке 2 таблицы значений правила, и нужно включить это значение в вывод, можно просто перетащить элемент mTopic в первый столбец пары столбцов в таблице вывода правила. В результате будут автоматически заданы и Понятие, и Тип для выбранной пары. Или, если нужно начать вывод с типа, определенного третьим элементом (строка 3 таблицы значений правила), перетащите этот тип из таблицы значений правила в ячейку **Тип 1** в таблице вывода. В таблице появится ссылка на строку в скобках (3).

Другой вариант - ввести эти ссылки в таблицу вручную, щелкнув дважды по ячейке в каждом столбце **Понятие**, который нужно выводить, и введя символ $ и за ним номер строки. Например, $2 будет ссылаться

на элемент, определенный в строке 2 таблицы значений правила. Когда вы вводите информацию вручную, нужно также задавать столбец **Тип**. Введите символ # и за ним номер строки; например, #2 будет ссылаться на элемент, определенный в строке 2 таблицы значений правила.

Иногда нужно сочетать оба эти метода. Допустим, у вас был тип <Положительный> в строке 4 таблицы значений правила. Можно перетащить его в столбец Тип 2 и затем дважды щелкнуть по ячейке в столбце Понятие 2 и вручную ввести перед ним слово '*не*'. Тогда столбец вывода в таблице будет содержать не (4), а если вы работали в режиме редактирования или в режиме источника, то не $4. Затем можно щелкнуть правой кнопкой в столбце Тип 1 и выбрать, например, макрос с именем mTopic. В результате будет выводиться, например, паттерн понятия вида: автомобиль + плохой.

Большинство правил содержит только одну строку вывода, но иногда нужно выводить несколько. В таком случае определите по одному выводу для каждой строки в таблице вывода правила.

**Важное замечание:** Следует иметь в виду, что во время извлечения паттернов TLA выполняются и другие операции лингвистической обработки. Таким образом, если задан вывод t$3\t#3, это значит, что в итоговом паттерне будет выведен итоговый понятие для третьего элемента и итоговый тип для третьего элемента после применения всех лингвистических процедур (группировки синонимов и других видов группировки).

- **Показать вывод как**. По умолчанию опция **Ссылки на строку в таблице значений правила** включена, и вывод показан ссылками на номера строк, заданных на вкладке значений правила. Если вы ранее включили Получить маркеры, и в столбце Примеры маркеров в таблице значений правила есть маркеры, можно выбрать показ вывода этих конкретных маркеров, включив эту опцию.

**Примечание:** Если в таблице вывода показано недостаточно пар вывода понятие/тип, можно добавить еще одну пару, нажав кнопку Добавить на панели инструментов редактора. Если в настоящее время показано 3 пары, и вы нажали кнопку Добавить, в таблицу будут добавлены 2 столбца (Понятие 4 и Тип 4). Это значит, что вы увидите 4 пары в таблице вывода для всех правил. Кроме того, можно удалить ненужные пары, если их не использует никакое другое правило в наборе правил в этой библиотеке.

## Пример правила

Пусть ваши ресурсы содержат следующее правило анализа текстовых связей, и вы включили извлечение результатов TLA:

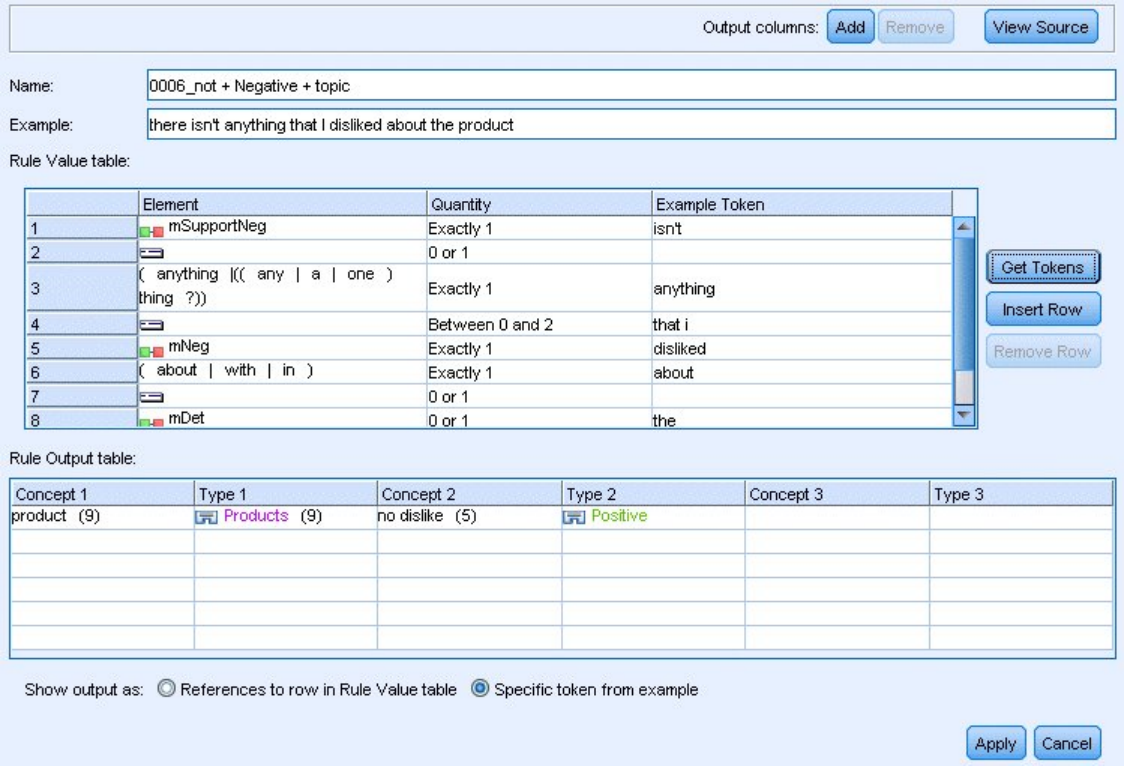



*Рисунок 43. Вкладка правил TLA: редактор правил*

При любого рода извлечения механизм извлечения прочитает каждое предложение и попытается найти соответствие следующей последовательности:

*Таблица 44. Пример последовательности извлечения*

| **Элемент (строка)** | **Описание аргументов** |
|---|---|
| 1 | Понятие из одного из типов, представленных макросом mPos, mNeg или с типом <Неопределенный>. |
| 2 | Понятие, введенный как один из типов, представленных макросом mTopic. |
| 3 | Одно из слов, представленных макросом mBe. |
| 4 | Необязательный элемент, 0 или 1 слово, который также называют промежуток между словами или <Any Token> |
| 5 | Понятие, введенный как один из типов, представленных макросом mTopic. |

Таблица вывода показывает, что от этого правила нужен только паттерн, где любой понятие или тип, соответствующий макросу mTopic, который был определен в строке 5 в таблице **Значения правила** + любой понятие или тип, соответствующий mPos, mNeg или <Неопределенный>, как было определено в строке 1 в таблице **Значения правила**. Например, это может быть сосиска + нравится или <Неизвестный> + <Положительный>.

## Создание и редактирование правил

Можно создавать новые правила или изменять существующие. Следуйте рекомендациям и описаниям для редактора правил. Дополнительную информацию смотрите в разделе “Работа с правилами TLA” на стр. 223.

Создание новых правил

1. В меню выберите **Инструменты > Создать правило**. Другой вариант - щелкните по значку Создать правило на панели инструментов дерева для открытия нового правила в редакторе.

2. Введите уникальное имя и определите элементы значений правила.
3. Закончив, щелкните по **Применить**, чтобы проверить на наличие ошибок.

Изменение правил

1. Щелкните по имени правила в дереве. Правило откроется в панели редактора справа.
2. Выполните свои изменения.
3. Закончив, щелкните по **Применить**, чтобы проверить на наличие ошибок.

## Отключение и удаление правил

Отключение правил

Если нужно игнорировать правило при обработке, его можно отключить. С осторожностью отключайте или удаляйте правила.

1. Щелкните по имени правила в дереве. Правило откроется в панели редактора справа.
2. Щелкните правой кнопкой по имени.
3. В контекстном меню выберите **Отключить**. Значок правила станет затененным, а само правило будет недоступно для изменений.

Удаление правил

Если правило больше не нужно, его можно удалить. С осторожностью отключайте или удаляйте правила.

1. Щелкните по имени правила в дереве. Правило откроется в панели редактора справа.
2. Щелкните правой кнопкой по имени.
3. В контекстном меню выберите **Удалить**. Правило исчезнет из списка.

## Проверка ошибок, сохранения и отмены

Применение изменений правил

Если щелкнуть мышью вне редактора правил или нажать кнопку **Применить**, правило будет автоматически проверено на наличие ошибок. Если ошибка обнаружится, нужно исправить ее, прежде чем переходить к другой части прикладной программы.

Однако при обнаружении менее серьезных ошибок выводится только предупреждение. Например, сообщение с предупреждением появится, если в вашем правиле есть неполные определения типов или макросов или определения без ссылок. После нажатия кнопки **Применить**, если какие-то предупреждения не исправлены, появится значок предупреждения слева от имени правила на дереве на левой панели.

Применение правила не означает, что оно будет навсегда сохранено. При применении инициируется процесс проверки ошибок и предупреждений.

Сохранение ресурсов в сеансе интерактивной инструментально среды

1. Чтобы сохранить изменения, внесенные в ваши ресурсы в сеансе интерактивной инструментальной среды, и обеспечить возможность их использования при следующем запуске потока, необходимо:
   - Изменить ваш узел моделирования для обеспечения возможности получения тех же ресурсов при следующем выполнении потока. Дополнительную информацию смотрите в разделе “Обновление узлов моделирования и сохранение” на стр. 84. После этого сохраните ваш поток. Чтобы сохранить поток, сохранение нужно выполнить в главном окне IBM SPSS Modeler после обновления узла моделирования.
2. Чтобы сохранить изменения, внесенные в ваши ресурсы в сеансе интерактивной инструментальной среды, и обеспечить возможность их использования в других потоках, можно сделать следующее:

- Изменить использованный шаблон или создать новый. Дополнительную информацию смотрите в разделе “Создание и изменение шаблонов” на стр. 167. При этом изменения для текущего узла сохранены не будут (смотрите предыдущий шаг)
- Другой вариант - изменить используемый TAP. Дополнительную информацию смотрите в разделе “Изменение пакетов анализа текста” на стр. 142.

Сохранение ресурсов в Редактор шаблонов

1. Сначала опубликуйте библиотеку. Дополнительную информацию смотрите в разделе “Публикация библиотек” на стр. 185.
2. Затем сохраните шаблон с помощью опции меню **Файл > Сохранить шаблон ресурсов**.

Отмена изменений правил

1. Если нужно отбросить изменения, нажмите кнопку **Отмена** на панели редактора.

## Порядок обработки для правил

Когда при извлечении производится анализ текстовых связей, "предложение" (условие, слово, словосочетание) будет по очереди сравниваться с каждым правилом, пока не будет найдено совпадение или не исчерпаются все правила. Расположение на дереве определяет порядок, в котором обрабатываются правила. Рекомендуемые приемы работы основаны на порядке обработки правил от наиболее конкретных до самых общих. Наиболее конкретные правила должны находиться в вершине дерева. Чтобы изменить расположение конкретного правила или набора правил, выберите опцию **Переместить вверх** или **Переместить вниз** в контекстном меню дерева Правила и макросы или воспользуйтесь стрелками вверх и вниз на панели инструментов.

Находясь в *представлении исходного кода*, невозможно изменить порядок правил, перемещая их в редакторе. Чем выше окажется правило в представлении исходного кода, тем раньше оно будет обработано. Чтобы исключить возможные проблемы при копировании и вставке, настоятельно рекомендуется переупорядочивать правила только с помощью дерева.

**Важно!** В предыдущих версиях IBM SPSS Modeler Text Analytics требовалось наличие уникального численного ID правила. Начиная с версии 18, можно указать только порядок обработки, перемещая правила вверх или вниз в дереве или располагая их нужным образом в представлении исходного кода.

Допустим, например, что в вашем тексте есть следующие предложения:

*I love anchovies*

*I love anchovies and green peppers*

Допустим кроме этого, что в существует два правила анализа текстовых связей со следующими значениями:

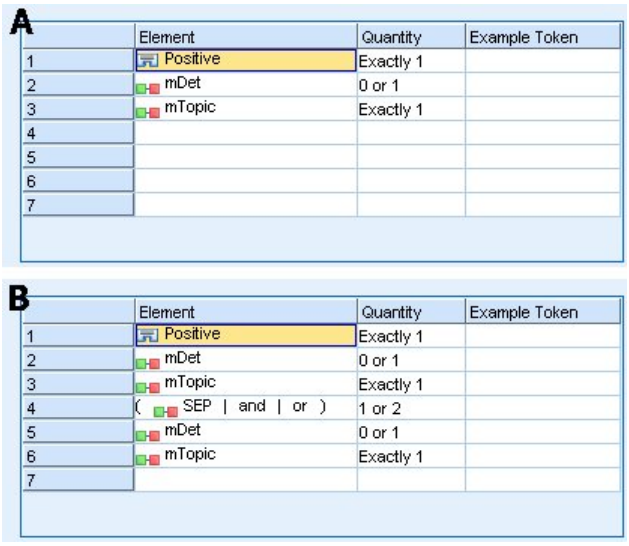



*Рисунок 44. Пример двух правил*

В представлении исходного кода значения правил могут выглядеть следующим образом:

*A:* `value = $Positive $mDet? $mTopic`

*B:* `value = $Positive $mDet? $mTopic ($SEP|and|or){1,2} $mDet? $mTopic`

Если правило **A** находится выше в дереве (ближе к вершине), чем правило **B**, правило **A** будет обрабатываться первым и предложение *I love anchovies and green peppers* будет первым соответствующим условию `$Positive $mDet? $mTopic`, и оно создаст на выходе неполный паттерн (`anchovies + like`), так как это предложение сравнивалось с правилом без поиска двух совпадений `$mTopic`.

Поэтому для захвата истинного значения текста более конкретное правило (в данное случае правило **B**) должно располагаться в дереве выше более общего правила (в нашем случае - правила **A**).

## Работа с наборами правил (несколько проходов)

Набор правил - полезный способ сгруппировать набор правил в дереве правил и макросов для обработки в несколько проходов. У набора правил нет собственного определения, не считая его имени, и он служит для организации правил в смысловые группы. В некоторых контекстах текст слишком богат и разнообразен, чтобы его можно было обработать за один проход. Например, при работе с разведданными служб безопасности текст может содержать связи между индивидуумами, не покрываемые методами контактов (*x звонил y*), через семейные связи (*x - свояк y*), через передачу денег (*x перевел 100 долларов y*) и так далее. В этом случае полезно создать специальные набора правил TLA, каждый из которых сфокусирован на своем типе взаимосвязи, например, один - на раскрытии контактов, другой - на раскрытии терминов семьи, и так далее.

Чтобы создать набор правил, выберите “Создать набор правил” в контекстном меню дерева правил и макросов или на панели инструментов. Затем можно непосредственно создать новые правила в дереве в подузле узла Набор правил или переместить существующие правила в Набор правил.

При извлечения с использованием ресурсов, в которых правила сгруппированы в наборы правил, механизм извлечения выполняет несколько проходов по тексту, и каждый раз ищет соответствия другому виду паттернов. Таким образом, некоторое "предложение" может быть найдено несколько раз, по правилу в разных наборах правил, в то время как без наборов правил оно может соответствовать только одному правилу.

*Примечание*: В один набор правил можно добавить до 512 правил.

Создание новых наборов правил

1. В меню выберите **Инструменты > Создать набор правил**. Другой вариант - щелкнуть по значку Новый набор правил на панели инструментов дерева. Набор правил появится в дереве правил.
2. Добавьте новые правила в этот набор правил или переместите в него существующие правила.

Отключение набора правил

1. Щелкните правой кнопкой по имени набора правил в дереве.
2. В контекстном меню выберите **Отключить**. Значок набора правил становится серым, и все содержащиеся в нем правила также отключаются и игнорируются во время обработки.

Удаление набора правил

1. Щелкните правой кнопкой по имени набора правил в дереве.
2. В контекстном меню выберите **Удалить**. Набор правил и все содержащиеся в нем правила удаляются из ресурсов.

## Поддерживаемые элементы для правил и макросов

Следующие аргументы принимаются как параметры в правилах анализа текстовых связей и макросах:

Макросы

Макрос можно использовать непосредственно в правиле TLA или в другом макросе. Если вы вводите имя макроса вручную или из представления источника (а не выбираете имя макроса в контекстном меню), не забудьте префикс - символ доллара ($); пример - `$mTopic`. Имя макроса зависит от регистра. При выборе макросов в контекстном меню доступны все макросы, определенные на текущей вкладке Правила текстовых связей.

Типы

Тип можно использовать непосредственно в правиле TLA или в макросе. Если вы вводите имя типа вручную или из представления источника (а не выбираете имя типа в контекстном меню), не забудьте префикс - символ доллара ($); пример - `$Person`. Имя типа зависит от регистра. Если вы используете контекстные меню, то можете выбрать любой тип из текущего используемого набора ресурсов.

Если сослаться на неизвестный тип, будет выведено предупреждение, а в дереве правил и макросов у такого правила появится предупреждающий значок, пока вы не исправите ошибку.

Литеральные строки

Чтобы включить информацию помимо извлеченного, можно определить литеральную строку, которую должен найти механизм извлечения. Для всех извлеченных слов и словосочетаний задан тот или иной тип; поэтому они не могут использоваться в литеральных строках. Если использовать слово, которое было извлечено, оно будет проигнорировано, даже если его тип - `<Неизвестный>`.

Литеральная строка может быть одним или несколькими словами. Следующие правила применяются при определении списка литеральных строк:

- Заключите список строк в скобки, например, (его). Если нужен выбор из литеральных строк, разделите их операцией ИЛИ, например, `(данный|этот|некоторый)` или `(его|ее|их)`.
- Используйте одиночные слова или словосочетания.
- Разделяйте слова в списке символом `|`, означающим логическую операцию `ИЛИ`.

- Если нужно соответствие формам единственного и множественного числа, введите обе формы. Автоматически разные формы слова не генерируются.
- Используйте только нижний регистр.
- Чтобы повторно использовать литеральные строки, определите их как макрос и затем используйте этот макрос в других макросах и в правилах TLA.
- Если искомая строка содержит точки или дефисы, их нужно включить. Например, чтобы найти в тексте соответствие `ч.д.а.`, введите как литеральную строку и буквы, и точки, `ч.д.а`.

Операция исключения

Восклицательный знак `!` служит операцией исключения, которая не допускает выражение отрицания занять конкретный слот. Операцию исключения можно добавить только вручную, редактируя ячейку (после двойного щелчка по ячейке в таблице значений правила или в таблице значений макроса) или в представлении источника. Например, если добавить `$mTopic @{0,2} !($Положительный) $Бюджет` в правило TLA, вы будете искать текст, который содержит (1) термин, для которого задан любой из типов в макросе `mTopic`, (2) промежуток между словами длиной от нуля до двух слов, (3) ни одного вхождения термина с типом `<Положительный>` и (4) термин с типом `<Бюджетный>`. Тогда возможен захват фразы "*автомобилям назначены неадекватные цены*", но будет проигнорирована фраза "*салон предлагает невероятные скидки*".

Для использования этой операции нужно ввести восклицательный знак и скобки вручную в ячейке элемента, дважды щелкнув по ячейке.

Промежутки между словами (<Any Token>)

Промежуток между словами, также называемый `<Any Token>`, задает диапазон для допустимого числа маркеров между двумя элементами. Промежутки между словами полезны при поиске соответствий с близкими словосочетаниями, различающимися дополнительными детерминативами, предложными группами, прилагательными и тому подобными словами.

*Таблица 45. Примеры элементов в таблице значений правила без промежутка между словами*

| # | Элемент |
|---|---|
| 1 |  `Нет данных` |
| 2 |  `mBeHave` |
| 3 |  `Положительные` |

*Примечание*: В представлении источника это значение определяется так: `$Неизвестный $mBeHave $Положительный`

Этому значению соответствуют такие предложения, как "*персонал отеля был вежлив*", где *персонал отеля* принадлежит типу `<Неизвестный>`, *был* входит в макрос `mBeHave`, а *вежлив* имеет тип `<Положительный>`. Но не будет соответствия предложению "*персонал отеля был очень вежлив*".

*Таблица 46. Пример элементов в таблице значений правила с промежутком между словами <Any Token>*

| # | Элемент |
|---|---|
| 1 |  `Нет данных` |

*Таблица 46. Пример элементов в таблице значений правила с промежутком между словами <Any Token> (продолжение)*

| | |
|---|---|
| 2 |  <br>mBeHave |
| 3 |  |
| 4 |  <br>Положительные |

*Примечание*: В представлении источника это значение определяется так: `$Неизвестный $mBeHave @{0,1} $Положительный`

Если добавить в значение правила промежуток между словами, соответствие будет найдено и с предложением "*персонал отеля был вежлив*", и с предложением "*персонал отеля был очень вежлив*".

В представлении источника и при редактировании строки синтаксис для промежутка между словами имеет формат `@{#,#}`, где `@` означает промежуток между словами, а `{#,#}` определяет минимально и максимально допустимое число слов между предшествующим и последующим элементами. Например, `@{1,3}` значит, что соответствие для двух заданных элементов будет признано, если между ними есть хотя бы одно и не больше трех слов. `@{0,3}` значит, что соответствие для двух заданных элементов будет признано, если число слов между ними 0, 1, 2 или 3, то есть не больше 3 слов.

# Просмотр данных и работа в режиме исходного кода

Для каждого правила и макроса редактор TLA генерирует базовый исходный код, используемый механизмом извлечения для поиска совпадений и создание выхода TLA. Если вы предпочитаете работать с самим исходным кодом, можно просматривать этот исходный код и непосредственно изменять его, нажав кнопку "Просмотр исходного кода" в верхней части редактора. Система перейдет в представление исходного кода, и будут выделено текущее выбранное правило или макрос. Однако для уменьшения вероятности ошибок мы рекомендуем использовать панели редактора.

После завершения просмотра или изменения исходного кода нажмите кнопку **Закрыть исходный код**. Если для правила был сгенерирован недопустимый синтаксис, его нужно будет исправить, прежде чем закрывать представление исходного кода.

**Важное замечание:** При изменениях в представлении исходного кода настоятельно рекомендуется изменять правила и макросы по одному. После изменения макроса проверьте результат, выполнив извлечение. Если вы удовлетворены результатами, рекомендуется сохранить этот шаблон, прежде чем выполнять другое изменение. Если вы не удовлетворены результатами или произошла ошибка, вернитесь к прежней версии сохраненных ресурсов.

## Макросы в представлении исходного кода

```
[macro]
name = имя_макроса
value = ([имя_типа|имя_макроса|литеральная_строка|промежуток_между_словами])
```

*Таблица 47. Записи макросов*

| | |
|---|---|
| `[macro]` | Каждый макрос должен начинаться со строки, обозначенной как [macro], что указывает на начало макроса. |
| `name` | Имя макроопределения. Каждое имя должно быть уникальным. |

*Таблица 47. Записи макросов (продолжение)*

| | |
|---|---|
| value | Комбинация одного или нескольких типов, литеральных строк, промежутков между словами или макросов. Дополнительную информацию смотрите в теме “Поддерживаемые элементы для правил и макросов” на стр. 230. При комбинировании аргументов необходимо использовать скобки ( ) собственно для групп аргументов и знаки \| для обозначения логического OR. |

В дополнение к рекомендациям и примерам синтаксиса, представленным в разделе о макросах, к представлению исходного кода относятся некоторые дополнительные рекомендации, не требующиеся при работе в представлении редактора. При работе с макросами в представлении исходного кода необходимо дополнительно учитывать следующее:

- Каждый макрос должен начинаться со строки, обозначенной как [macro], что указывает на начало макроса.
- Чтобы отключить какой-то элемент, разместите в начале каждой строки индикатор комментария (#).

**Пример**. В этом примере определяется макрос с названием mTopic. Значение для mTopic - это наличие термина, совпадающего с *одним* из следующих типов: <Продукт>, <Человек>, <Положение>, <Организация>, <Бюджет> или <Неизвестно>.

```
[macro]
name=mTopic
value=($Unknown|$Product|$Person|$Location|$Organization|$Budget|$Currency)
```

## Правила в представлении исходного кода

```
[pattern(ID)]
name = имя_паттерна
value = [$имя_типа|имя_макроса|промежутки_между_словами|литеральные_строки]
output = $digit[\t]#digit[\t]$digit[\t]#digit[\t]$digit[\t]#digit[\t]
```

*Таблица 48. Записи правил*

| | |
|---|---|
| [pattern (<ID>)] | Указывает на начало правила анализа текстовых связей и предоставляет уникальный численный ID, используемый для определения порядка обработки. |
| name | Предоставляет уникальное имя для этого правила анализа текстовых связей. |
| value | Предоставляет синтаксис и аргументы для поиска совпадений в тексте. Дополнительную информацию смотрите в разделе “Поддерживаемые элементы для правил и макросов” на стр. 230. |
| output | Выходной формат для итоговых найденных паттернов, обнаруженных в тексте. Выход не всегда отображает точное начальное положение элементов в исходном тексте. Кроме этого, для данного правила анализа текстовых связей может быть несколько выходных строк, если каждый вывод будет размещаться на отдельной строке.<br><br>Синтаксис вывода:<br>• Разделяйте вывод кодом табуляции \t, например, $1\t#1\t$3\t#3<br>• $ и число вызывает найденный термин, совпадающий с аргументом, определенным в параметре значения в этой позиции. Поэтому $1 означает термин, совпадающий с первым аргументом, определенным для значения.<br>• # и число вызывает имя типа элемента в этой позиции. Если элемент - это список литеральных строк, будет назначен тип <Неизвестно>.<br>• Значение Null\tNull не создаст никакого вывода. |

В дополнение к рекомендациям и примерам синтаксиса, представленным в разделе о правилах, к представлению исходного кода относятся некоторые дополнительные рекомендации, не требующиеся при работе в представлении редактора. При работе с правилами в режиме исходного кода необходимо дополнительно учитывать следующее:

- При всяком определении двух или более элементов они должны быть заключены в скобки независимо от того, обязательны они или нет (например, `($Negative|$Positive)` *или* `($mCoord|$SEP)?`). `$SEP` представляет запятую.
- Первый элемент в правиле анализа текстовых связей не может быть необязательным. Например, нельзя начать правило с `value = $mTopic?` или `value = @{0,1}`.
- Можно связать количество (или число экземпляров) с маркером. Это полезно для написания только одного правила, охватывающего все случаи, вместо написания по одному правилу для каждого случая. Например, можно использовать литеральную строку `($SEP|and)` при описании совпадения или с `,` (запятой), или с `and`. Если расширить этот пример, добавив количество, то есть превратить литеральную строку в `($SEP|and){1,2}`, будет получено совпадение с любым из следующих примеров: `","` `"and"` `", and"`.
- Не поддерживаются пробелы между символами `$` и `?` в значении `value` правила анализа текстовых связей.
- Пробелы не поддерживаются в выводе `output` правила анализа текстовых связей.
- Чтобы отключить какой-то элемент, разместите в начале каждой строки индикатор комментария (#).

**Пример**. Допустим, ваши ресурсы содержат следующее правило анализа текстовых связей TLA и включено извлечение результатов TLA:

```
## Jean Doe was the former HR director of IBM in France
[pattern(201)]
name= 1_201
value = $Person ($SEP|$mDet|$mSupport|as|then){1,2} @{0,1} $Function
 (of|with|for|in|to|at) @{0,1} $Organization @{0,2} $Location
output = $1\t#1\t$4\t#4\t$7\t#7\t$9\t#9
```

При любого рода извлечения механизм извлечения прочитает каждое предложение и попытается найти соответствие следующей последовательности:

*Таблица 49. Пример последовательности извлечения*

| Положение | Описание аргументов |
|---|---|
| 1 | Имя человека (`$Person`), |
| 2 | Один или два из следующих элементов: запятая (`$SEP`), разделитель (`$mDet`), вспомогательный глагол (`$mSupport`), строки “`then`” или “`as`”, |
| 3 | Число слов 0 или 1 (`@{0,1}`) |
| 4 | Функция (`$Function`) |
| 5 | Одна из следующих строк: “`of`”, “`with`”, “`for`”, “`in`”, “`to`” или “`at`”, |
| 6 | Число слов 0 или 1 (`@{0,1}`) |
| 7 | Название организации (`$Organization`) |
| 8 | Число слов 0, 1 или 2 (`@{0,2}`) |
| 9 | Имя положения (`$Location`) |

Этот образец правила анализа текстовых связей будет соответствовать предложениям или словосочетаниям, как следующие:

*Jean Doe, the HR director of IBM in France*

*Jean Doe was the former HR director of IBM in France*

*IBM appointed Jean Doe as the HR director of IBM in France*

Этот образец правила анализа текстовых связей создаст следующий вывод:

```
jean doe <Человек> hr director <Функция> ibm <Организация> france <Положение>
```

Здесь:

- jean doe - термин, соответствующий $1 (первый элемент в правиле анализа текстовых связей), а <Человек> - это тип для jean doe (#1),
- hr director - термин, соответствующий $4 (четвертый элемент в правиле анализа текстовых связей), а <Функция> - это тип для hr director (#4),
- ibm - термин, соответствующий $7 (седьмому элементу в правиле анализа текстовых связей), а <Организация> - это тип для ibm. (#7),
- france - термин, соответствующий $9 (девятому элементу в правиле анализа текстовых связей), а <Положение> - это тип для france (#9)

## Наборы правил в представлении исходного кода

```
[set(<ID>)]
```

Здесь [set (<ID>)] обозначает начало набора правил и предоставляет уникальный численный ID, используемый для определения порядка обработки наборов.

**Пример**. В следующем предложении содержится информация о людях, их функциях в компании и об операциях по слиянию/приобретению этой компании.

```
Org1 Inc has entered into a definitive merger agreement with Org2 Ltd, said John Doe, CEO of
Org2 Ltd.
```

Вы могли бы написать одно правило с несколькими выводами для обработки всех возможных выводов, таких как:

```
## Org1 Inc entered into a definitive merger agreement with Org2 Ltd, said
John Doe, CEO of Org2 Ltd.

[pattern(020)]
name=020
value = $Organization @{0,4} $ActionNouns @{0,6}  $mOrg @{1,2}
 $Person @{0,2} $Function @{0,1} $Organization
output = $1\t#1\t$3\t#3\t$5\t#5
output = $7\t#7\t$9\t#9\t$11\t#11
```

Этот код может создать следующие два выходные паттерна:

- org1 inc<Организация> + merges with <Активный глагол> + org2 ltd<Организация>
- john doe <Человек> + ceo <Функция> + org2 ltd<Организация>

**Важно!** Следует иметь в виду, что во время извлечения паттернов TLA выполняются и другие операции лингвистической обработки. В данной случае термин merger группируется со словосочетанием merges with в фазе группировки синонимов в процессе извлечения. А так как merges with принадлежит к типу <Активный глагол>, именно это имя типа появится в окончательном выводе паттерна TLA. Таким образом, если задан вывод t$3\t#3, это значит, что в итоговом паттерне будет выведен итоговый понятие для третьего элемента и итоговый тип для третьего элемента после применения всех лингвистических процедур (группировки синонимов и других видов группировки).

Вместо написания сложных правил, как было продемонстрировано в предыдущем примере, может быть проще управлять двумя правилами и работать с ними. Первое правило будет нацелено на обнаружение слияний/приобретений между компаниями:

```
[set(1)]
## Org1 Inc has entered into a definitive merger agreement with Org2 Ltd
[pattern(44)]
name=firm + action + firm_0044
value=$mOrg @{0,20} $ActionNouns @{0,6} $mOrg
output(1)=$1\t#1\t$3\t#3\t$5\t#5
```

Этот код может привести к выходному паттерну `org1 inc<Организация> + merges with <Активный глагол> + org2 ltd <Организация>`

Второе правило будет специализированным для сотрудников/их функций/компаний:

```
[set(2)]
## said John Doe, CEO of Org2 Ltd
[pattern(52)]
name=individual + role + firm_0007
value=$Person @{0,3} $mFunction (at|of)? ($mOrg|$Media|$Unknown)
output(1)=$1\t#1\t$3\tFunction\t$5\t#5
```

Итоговым паттерном может быть `john doe <Человек> + ceo <Функция> + org2 ltd <Организация>`

# Уведомления

Эта информация относится к продуктам и сервису, предлагаемым по всему миру.

IBM может не предоставлять в других странах продукты, услуги и аппаратные средства, описанные в данном документе. За информацией о продуктах и услугах, предоставляемых в вашей стране, обращайтесь к местному представителю IBM. Ссылки на продукты, программы или услуги IBM не означают и не предполагают, что можно использовать только указанные продукты, программы или услуги IBM. Разрешается использовать любые функционально эквивалентные продукты, программы или услуги, если при этом не нарушаются права IBM на интеллектуальную собственность. Однако ответственность за оценку и проверку работы любого продукта, программы или сервиса, не произведенного корпорацией IBM, лежит на пользователе.

IBM может располагать патентами или рассматриваемыми заявками на патенты, относящимися к предмету данного документа. Предъявление данного документа не предоставляет какую-либо лицензию на эти патенты. Вы можете послать письменный запрос о лицензии по адресу:

*IBM Director of Licensing*
*IBM Corporation*
*North Castle Drive, MD-NC119*
*Armonk, NY 10504-1785*
*US*

По поводу лицензий, связанных с использованием наборов двухбайтных символов (DBCS), обращайтесь в отдел интеллектуальной собственности IBM в вашей стране или направьте запрос в письменной форме по адресу:

*Intellectual Property Licensing*
*Legal and Intellectual Property Law*
*IBM Japan Ltd.*
*19-21, Nihonbashi-Hakozakicho, Chuo-ku*
*Tokyo 103-8510, Japan*

INTERNATIONAL BUSINESS MACHINES CORPORATION ПРЕДСТАВЛЯЕТ ДАННУЮ ПУБЛИКАЦИЮ "КАК ЕСТЬ", БЕЗ КАКИХ-ЛИБО ГАРАНТИЙ, КАК ЯВНЫХ, ТАК И ПОДРАЗУМЕВАЕМЫХ, ВКЛЮЧАЯ, НО НЕ ОГРАНИЧИВАЯСЬ ТАКОВЫМИ, ПРЕДПОЛАГАЕМЫЕ ГАРАНТИИ СОБЛЮДЕНИЯ ЧЬИХ-ЛИБО АВТОРСКИХ ПРАВ, ВОЗМОЖНОСТИ КОММЕРЧЕСКОГО ИСПОЛЬЗОВАНИЯ ИЛИ ПРИГОДНОСТИ ДЛЯ КАКИХ-ЛИБО ЦЕЛЕЙ И СООТВЕТСТВИЯ ОПРЕДЕЛЕННОЙ ЦЕЛИ. В некоторых странах для ряда сделок не допускается отказ от явных или предполагаемых гарантий; в таком случае данное положение к вам не относится.

Эта информация может содержать технические неточности и типографские ошибки. В представленную здесь информацию периодически вносятся изменения; эти изменения будут включаться в новые издания данной публикации. Фирма IBM может в любое время без уведомления вносить изменения и усовершенствования в продукты и программы, описанные в этой публикации.

Любые ссылки в этой публикации на сайты, не принадлежащие IBM, приведены только для удобства и никоим образом не означают их поддержки. Материалы на этих сайтах не входят в число материалов по данному продукту IBM, и весь риск пользования этими сайтами несете вы сами.

Любую предоставленную вами информацию IBM может использовать или распространять любым способом, какой сочтет нужным, не беря на себя никаких обязательств по отношению к вам.

Если обладателю лицензии на данную программу понадобятся сведения о возможности: (i) обмена данными между независимо разработанными программами и другими программами (включая данную) и (ii) совместного использования таких данных, он может обратиться по адресу:

*IBM Director of Licensing*
*IBM Corporation*
*North Castle Drive, MD-NC119*
*Armonk, NY 10504-1785*
*US*

Такая информация может быть доступна при соответствующих условиях и соглашениях, включая в некоторых случаях взимание платы.

Описанную в данном документе лицензионную программу и все прилагаемые к ней лицензированные материалы IBM предоставляет на основе положений Соглашения между IBM и Заказчиком, Международного Соглашения о Лицензиях на Программы IBM или любого эквивалентного соглашения между IBM и заказчиком.

Данные производительности и примеры клиентов представлены только для иллюстрации. Фактическая производительность зависит от конкретной конфигурации и условий работы.

Информация о продуктах других компаний (не IBM) получена от поставщиков этих продуктов, из их опубликованных объявлений или из иных общедоступных источников. IBM не производила тестирование этих продуктов и никак не может подтвердить информацию о их точности работы и совместимости, а также прочие заявления относительно продуктов других компаний (не IBM). Вопросы о возможностях продуктов других компаний (не IBM) следует направлять поставщикам этих продуктов.

Все утверждения о будущих планах и намерениях IBM могут быть изменены или отменены без уведомлений, и описывают исключительно цели фирмы.

Эти сведения содержат примеры данных и отчетов, используемых в повседневных деловых операциях. Чтобы проиллюстрировать их настолько полно, насколько это возможно, данные примеры включают имена индивидуумов, названия компаний, брендов и продуктов. Все эти имена являются вымышленными и любое их сходство с реальными именами и адресами предприятий является случайным.

## Товарные знаки

IBM, логотип IBM, и ibm.com являются товарными знаками или зарегистрированными товарными знаками компании International Business Machines Corp., зарегистрированными во многих странах мира. Прочие наименования продуктов и услуг могут быть товарными знаками, принадлежащими IBM или другим компаниям. Текущий список товарных знаков IBM смотрите на веб-сайте "Copyright and trademark information" (Информация об авторских правах и товарных знаках) по адресу www.ibm.com/legal/copytrade.shtml.

Intel, логотип Intel, Intel Inside, логотип Intel Inside, Intel Centrino, логотип Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium и Pentium являются товарными знаками или зарегистрированными товарными знаками компании Intel или ее дочерних компаний в Соединенных Штатах и других странах.

Linux является зарегистрированным товарным знаком Linus Torvalds в Соединенных Штатах и других странах.

Microsoft, Windows, Windows NT и логотип Windows являются товарными знаками корпорации Microsoft в Соединенных Штатах и других странах.

UNIX является зарегистрированным товарным знаком The Open Group в Соединенных Штатах и других странах.

Java и все основанные на Java товарные знаки и логотипы - товарные знаки или зарегистрированные товарные знаки Oracle и/или его филиалов.

Другие названия продуктов и услуг могут являться товарными знаками IBM или других компаний.

# Индекс

## Спец. символы



## F



## H



## I



## L



## M



## N



## T



## U



## А



## Б



## В



## Г



## Д

## З



## И



## К



## Л



## М



## Н

# О



# П



# Р

# C



# T

## У



## Ф



## Ц



## Ч



## Ш



## Э



## Я

Напечатано в Дании