ООО «ЭЛЕТЕК»

Руководство по установке и эксплуатации программного обеспечения «ЛАН.Студия управления знаниями»

2024

Аннотация

Данный документ содержит информацию о назначении, а также описание функций и принципов эксплуатации программного обеспечения «ЛАН.Студия управления знаниями», предназначенного для создания и редактирования специализированных классификаторов и словарей, в дальнейшем используемых в процессе автоматической обработки и структурирования отобранной информации по различным тематикам, а также выделения фактографических данных.

Содержание

1 Установка, обновление и удаление

1.1 Установка

Для установки программного обеспечения «ЛАН.Студия управления знаниями» требуется запустить установочный файл. Цифры в названии установочного файла означают номер версии и могут отличатся от примера, приведенного в данном руководстве.

В результате запуска установочного файла откроется мастер установки программного компонента.

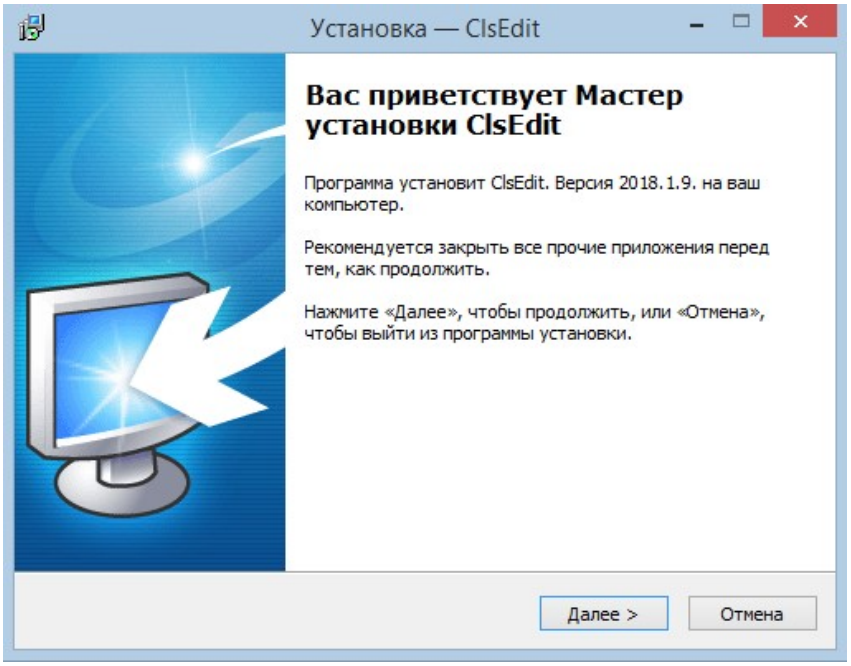


Рисунок 1 – Оно приветствия мастера установки

Следуя инструкциям данного мастера, необходимо выбрать директорию установки программы (по умолчанию «C:\Program Files\Lan\ClsEdit»), и нажать на расположенную в нижней части окна мастера кнопку «Далее».

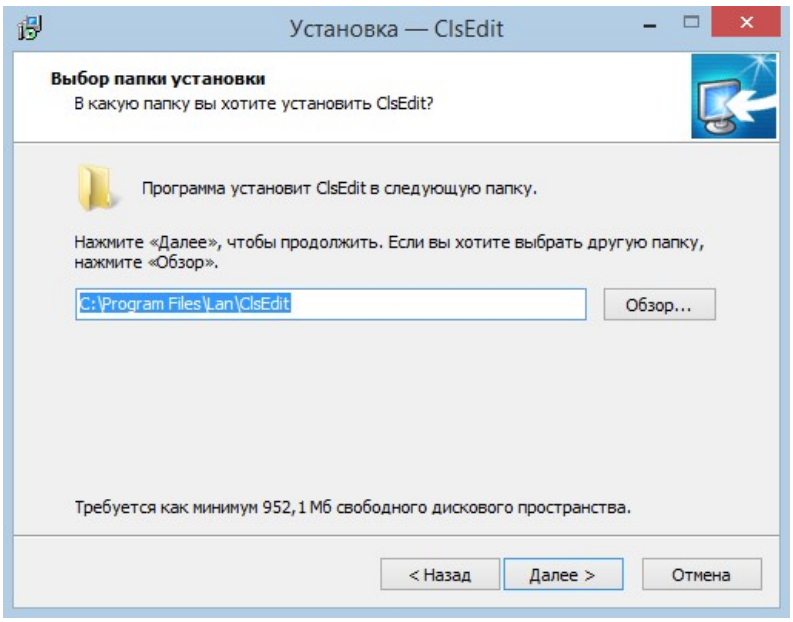


Рисунок 2 – Окно выбора директории установки программы

Следующее окно мастера установки позволяет выбрать тип установки. В том случае, если программное средство никогда не было установлено ранее, необходимо выбрать пункт «Полная установка». Соответственно, если мастер установки был запущен с целью обновления версии программного средства до более новой, необходимо осуществить выбор пункта «Обновление установленной программы».

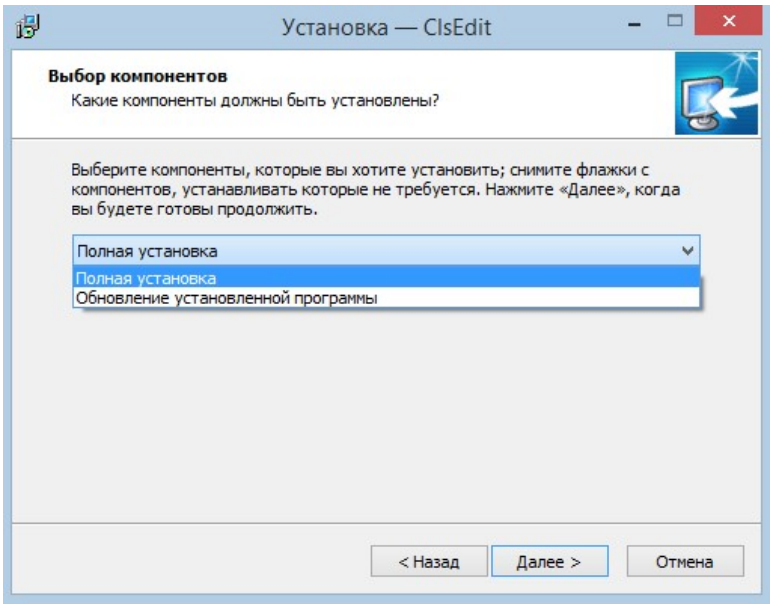


Рисунок 3 – Окно выбора типа установки

После указания всех параметров, которые необходимы мастеру, следует

запустить процесс установки с помощью кнопки «Установить»

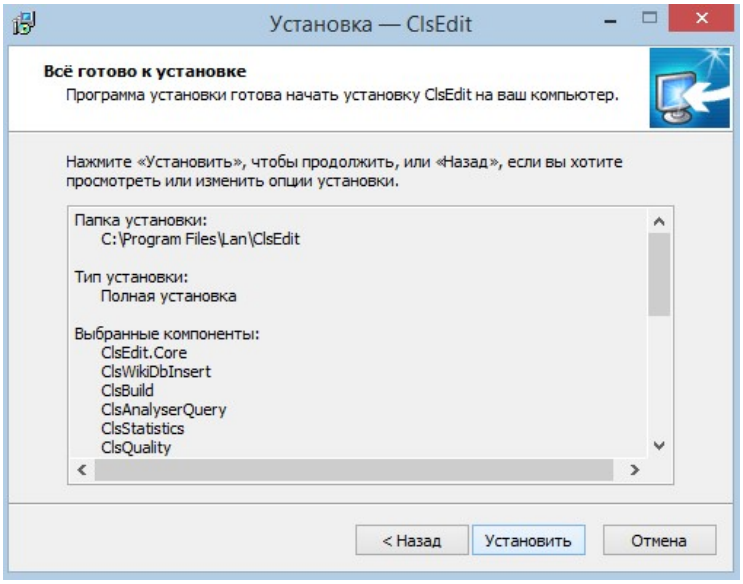


Рисунок 4 – Окно запуска процесса установки программы

В результате выполнения данного действия появится окно, в котором отображается процесс установки. В случае успешной инсталляции программного средства появится окно, в котором необходимо нажать на кнопку «Завершить» для закрытия мастера установки.

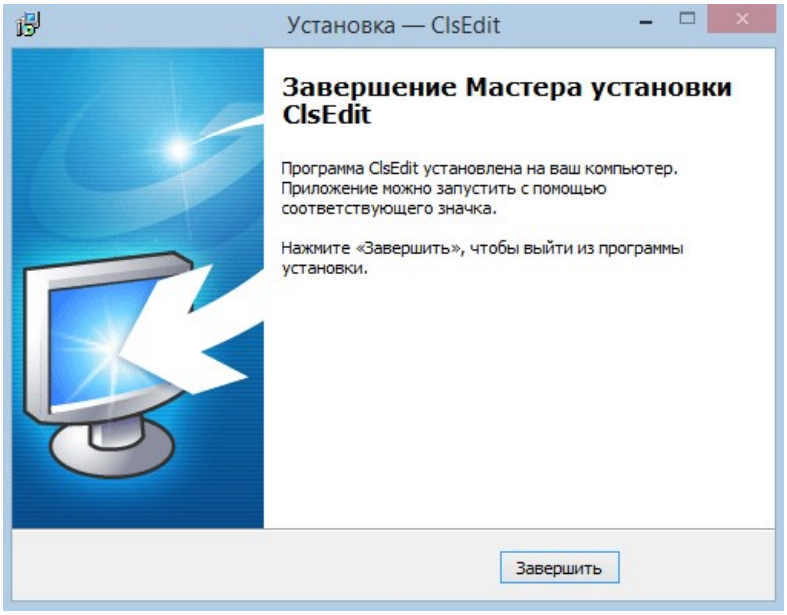


Рисунок 5 – Окно завершения установки программы

Результатом установки программного средства будет его отображение в панели меню «Пуск».

1.2 Обновление

Обновление программного обеспечения «ЛАН.Студия управления знаниями» производится путем запуска установочного файла. Перед началом обновления необходимо закрыть программу, если она была запущена.

Двойное нажатие левой клавиши «мыши» на установочный файл вызывает мастер установки программного средства.

Следуя инструкциям данного мастера, необходимо выбрать директорию установки программы (по умолчанию «C:\Program Files\Lan\ClsEdit»), и нажать на расположенную в нижней части окна мастера кнопку «Далее».

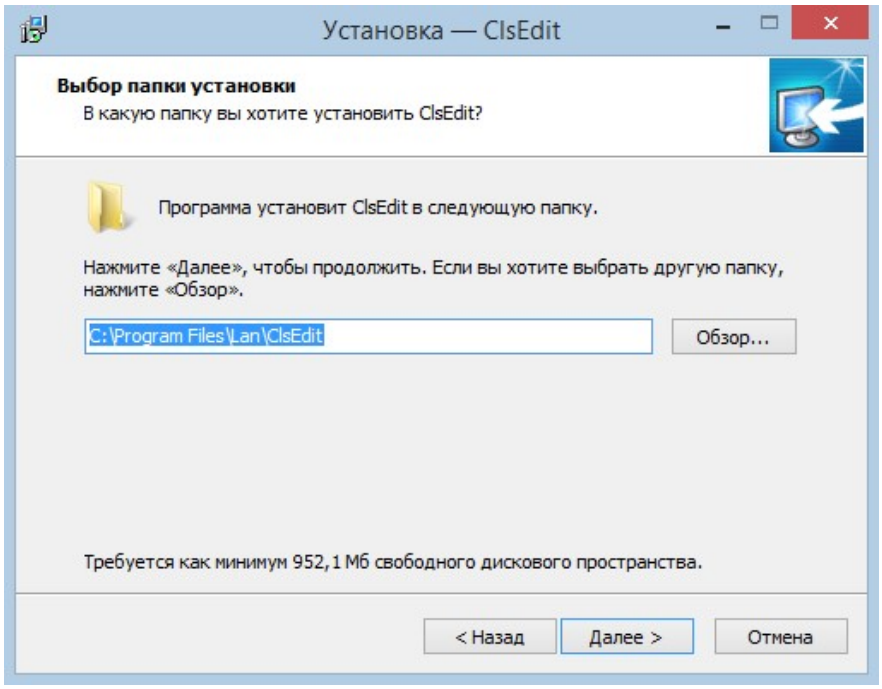


Рисунок 6 – Окно выбора директории установки программы

Следующее окно мастера установки позволяет выбрать тип установки. В том случае, если программное средство никогда не было установлено ранее, необходимо выбрать пункт «Полная установка». Соответственно, если мастер установки был запущен с целью обновления версии программного средства до более новой, необходимо осуществить выбор пункта «Обновление установленной программы».

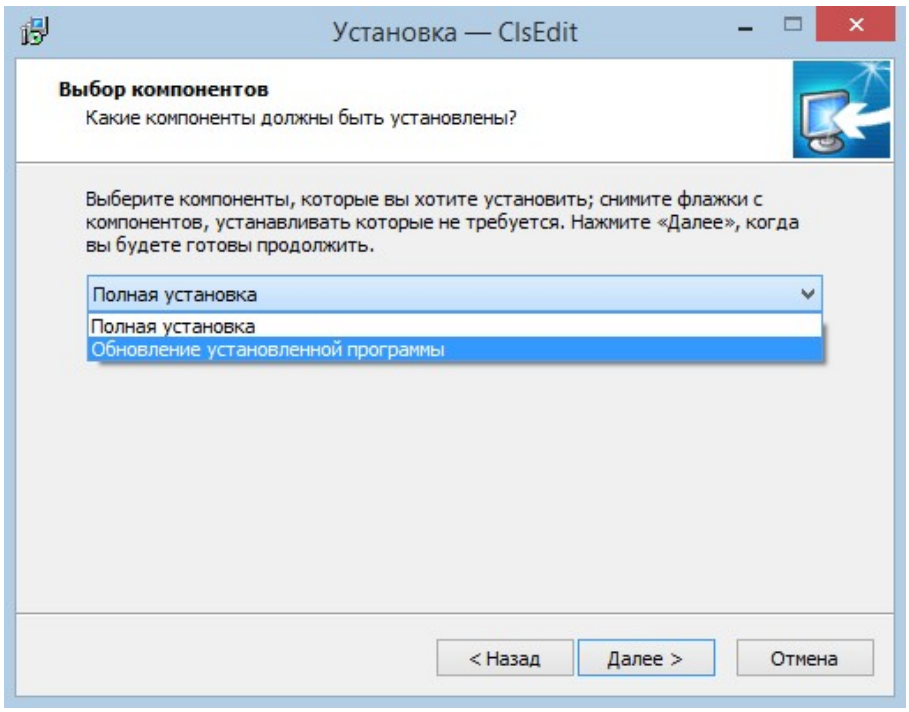


Рисунок 7 – Окно выбора типа установки

После указания всех параметров, которые необходимы мастеру, следует запустить процесс установки с помощью кнопки «Установить».

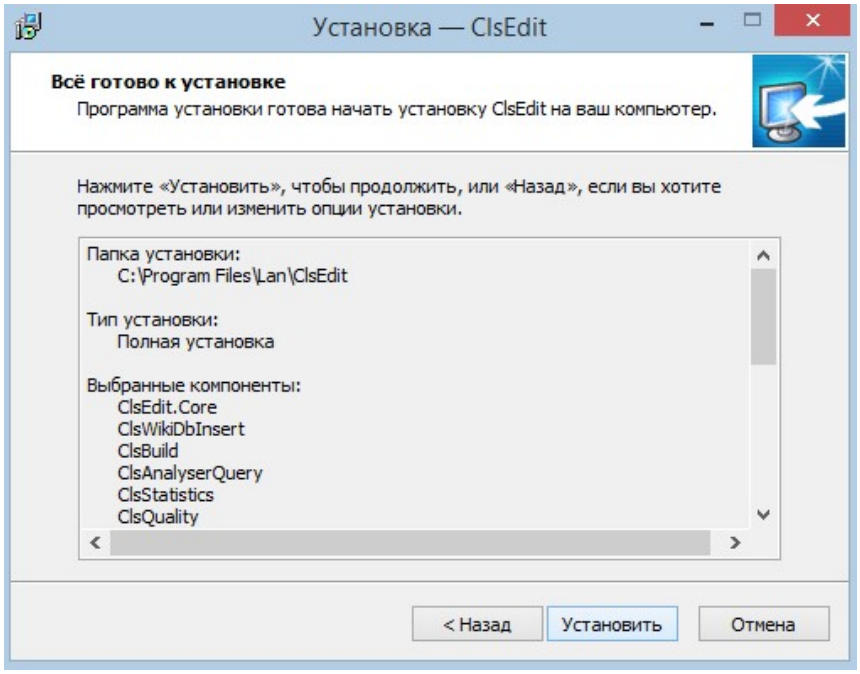


Рисунок 8 – Окно запуска процесса установки программы

В результате выполнения данного действия появится окно, в котором отображается процесс установки. В случае успешной инсталляции программного средства появится окно, в котором необходимо нажать на кнопку «Завершить» для

закрытия мастера установки.

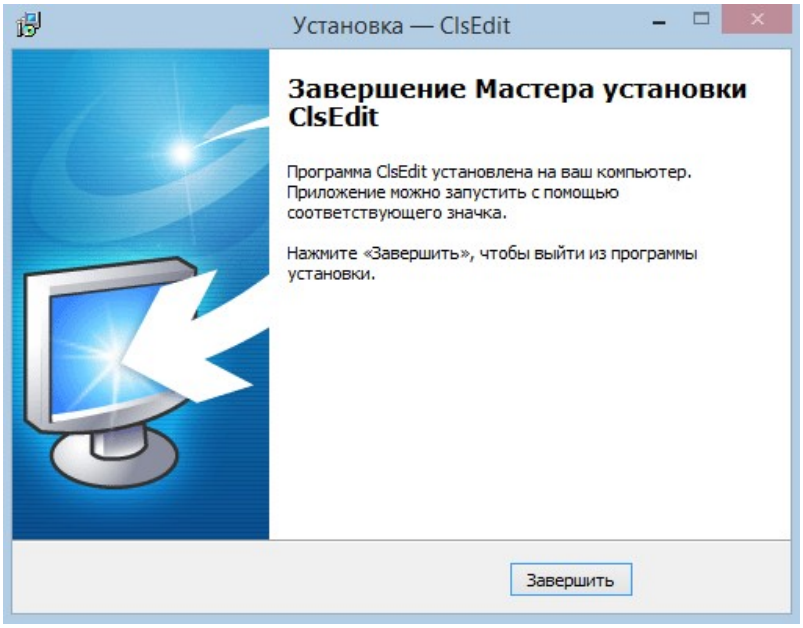


Рисунок 9 – Окно завершения установки программы

Результатом установки программного средства будет его отображение в панели меню «Пуск».

1.3 Удаление

Для деинсталляции программного обеспечения необходимо выполнить следующие действия:

- перейти в панель «Пуск», найти в списке директорию, в которую установлена программа (по умолчанию «ClsEdit») и запустить мастер деинсталляции («Удалить ClsEdit») либо запустить мастер удаления программного компонента через компонент «Установка и удаление программ» панели управления Windows;

- подтвердить деинсталляцию в открывшемся диалоговом окне, нажав на кнопку «Да»;

- дождаться завершения деинсталляции, процесс которой будет отображен в окне мастера;

- при необходимости удалить вручную файлы, которые не смог удалить мастер деинсталляции, если мастер деинсталляции сообщит об их наличии.

Результатом деинсталляции будет полное удаление исполняемых файлов

программы с устройства. Данные, созданные пользователем в результате работы с программой (проекты классификаторов, пользовательские настройки и пр.), удалению не подлежат.

1.4 Очистка остаточных данных

Очистка пользовательских данных программы производится путем удаления содержимого директории, которая расположена по адресу «C:\Users\User\AppData\Roaming\Cls».

Также необходимо проверить существование директории установки (по умолчанию «C:\Program Files\Lan\ClsEdit») и удалить, если она присутствует.

2 Эксплуатация

2.1 Общий графический интерфейс

Главное окно программы, представленное на рисунке 10, включает в себя следующие разделы: меню для выбора подсистемы (1 на рисунке Рисунок 10), панель инструментов (2 на рисунке 10), а также панели «Решение» (3 на рисунке 10) и «Проекты» (4 на рисунке 10).

По умолчанию при открытии программы отображается подсистема «Классификаторы». На панели «Решение» отображается структура открытого проекта классификатора и его свойства (возможно открытие нескольких проектов одновременно), на панели «Проекты» отображается список проектов, которые располагаются в каталоге «ClassifierProjects» (для подсистемы «Классификаторы»), а также действия для создания нового или открытия существующего проекта из другой директории.

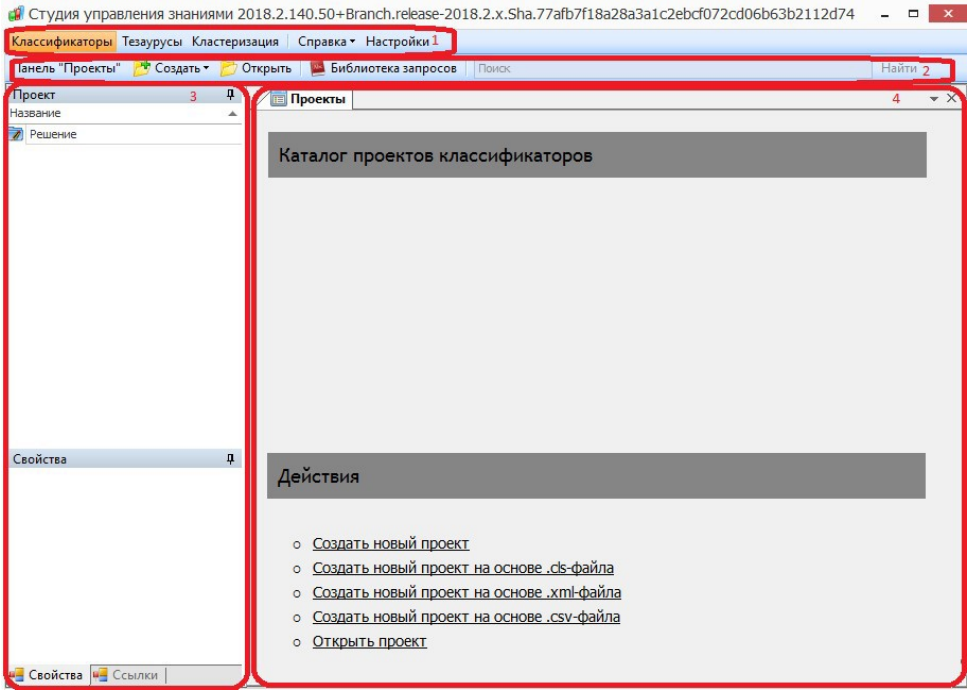


Рисунок 10 – Главное окно программы

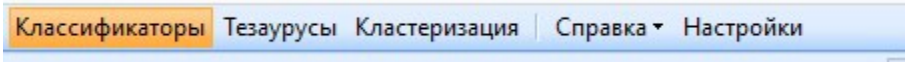


Рисунок 11 – Основное меню для выбора подсистемы

- подсистема «Классификаторы» предназначена для создания,

тестирования и отладки классификаторов для классификации текстовых документов;

- подсистема «Тезаурусы» предназначена для создания и редактирования тезаурусов необходимых для расширения правил классификации «текстовых» классификаторов;

- подсистема «Кластеризация» предназначена для настройки и тестирования конфигурационного файла для автоматической кластеризации (сюжетирования) текстовых документов;

- меню «Справка» содержит информацию о версионности программы и внутренних библиотек «О программе», а также текущее руководство пользователя «Помощь F1» в pdf-формате;

- кнопка «Настройки» открывает окно «Настройки приложения» (см. рисунок 12):

- «Общие» (1 на рисунке 12) – в данной вкладке можно изменить рабочие пути каталогов для сохранения и открытия проектов («Каталог проектов») и классификаторов («Каталог опубликования CLS»), прописать путь к каталогу с оценочной подборкой документов «Каталог с оценочными документами» для оценки качества правил классификации, а также указать интервал автоматического бекапирования (создания резервной копии) запросов.

- «Поведение» (2 на рисунке 12) – вкладка, используемая для выбора операций, которые будут выполняться после двойного нажатия правой клавишей «мыши» по рубрике: открытие поля для редактирования запросов («Редактировать запрос») или открытие поля для работы с документами рубрики («Показать документы»). Также на вкладке имеется доступ к включению/отключению функции проверки времени работы регулярных выражений, соответствующих запросам, во время компиляции. В случае включения функции запросы классификатора будут преобразованы в регулярные выражения (если это возможно), и во время компиляции будет выполнен подсчет времени работы регулярных выражений. Информация о времени работы регулярных выражений будет отражена в отчете о результатах обучения классификатора (появляется после

выполнения команды «Компиляция») в пунктах «Время классификации подборки (рег.)», «Скорость классификации подборки (рег.)», «Время обработки эталонов (рег.)». Также доступна функция включения/отключения режима проверки орфографии;

- «Источник (каталог)» (3 на рисунке 12) – в данной вкладке производятся создание и настройка пути к каталогу, в котором располагается тестовая подборка документов;
- «Источник (БД)» (4 на рисунке 12) – вкладка для настройки базы данных классификатора (данные для настройки предоставляются Администратором);
- «Тезаурус» (5 на рисунке 12) – вкладка для настройки базы тезаурусов (словарей). Данные для настройки предоставляются Администратором. При нажатии на кнопку «Применить» заданные настройки вступают в силу. При нажатии на кнопку «Обновить» происходит обновление базы тезаурусов. Данная функция используется для актуализации базы в случае, если со времени подключения к программе в базу были внесены изменения. С помощью кнопки «Создать» можно создать новую базу данных с заданными настройками. Кнопка «Удалить» выполняет удаление базы данных с сервера (после удаления данные не восстанавливаются);
- «Шаблоны запросов» (6 на рисунке 12) – в данной вкладке прописывается запрос «по умолчанию», который автоматически будет подставляться в каждую пустую рубрику, кроме рубрики «Другие», при компиляции проекта;
- «Кластеризация» (7 на рисунке 12) – вкладка для настройки пути к каталогу с подборками для кластерного анализа.

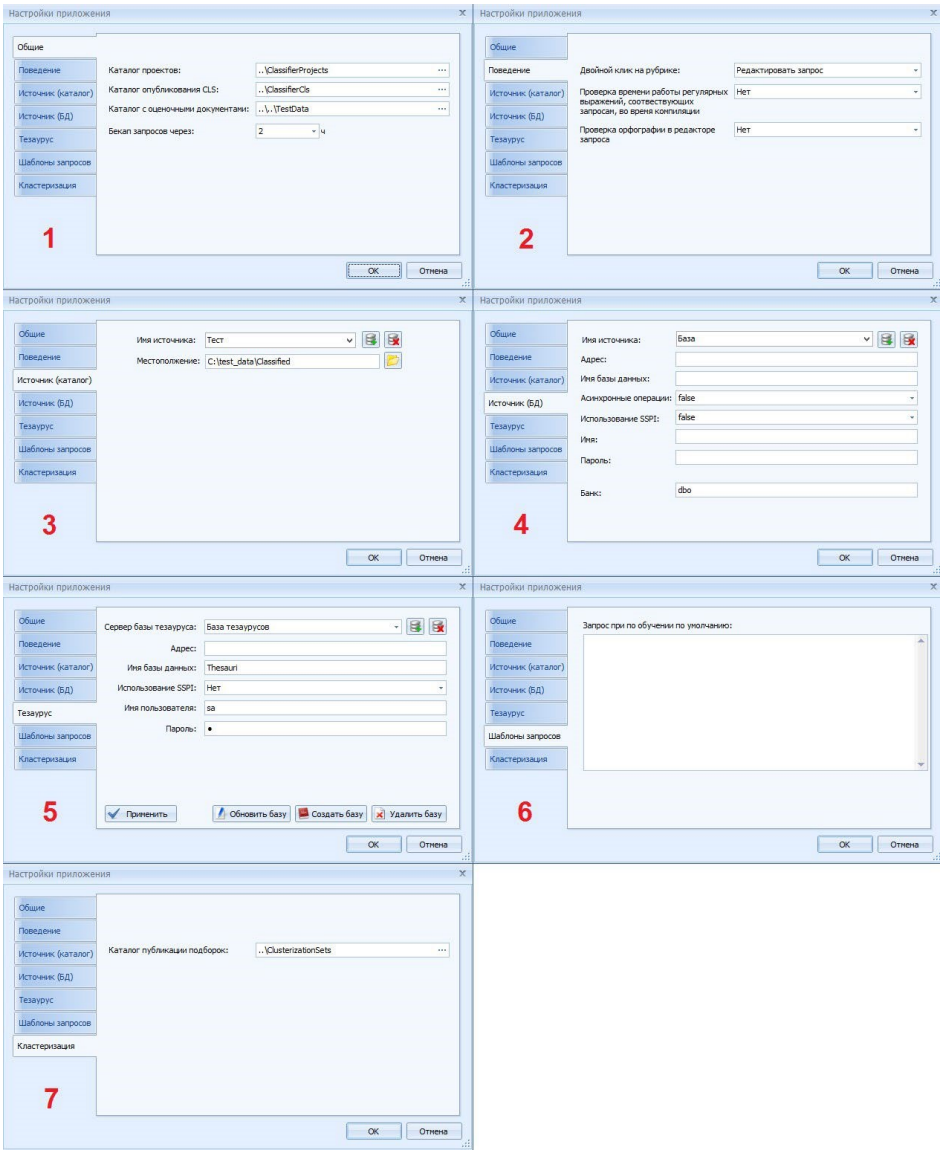


Рисунок 12 – Окно «Настройки приложения»

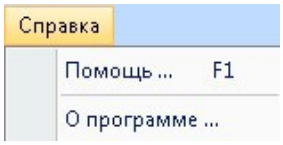


Рисунок 13 – Меню «Справка»

2.2 Настройки

Для указания параметров приложения необходимо произвести его запуск и в основном меню выбрать пункт «Настройки» (см. рисунок 14).

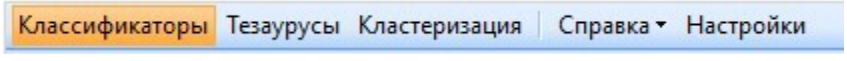


Рисунок 14 – Открытие настроек приложения

В результате описанных выше действий откроется окно настроек приложения, показанное на рисунке 15. По умолчанию открыта вкладка «Общих настроек». Из данного окна также можно переключится на вкладки «Поведение», «Источник (каталог)», «Источник (БД)», «Тезаурус», «Шаблоны запросов» и «Кластеризация».

2.2.1 Общие настройки приложения

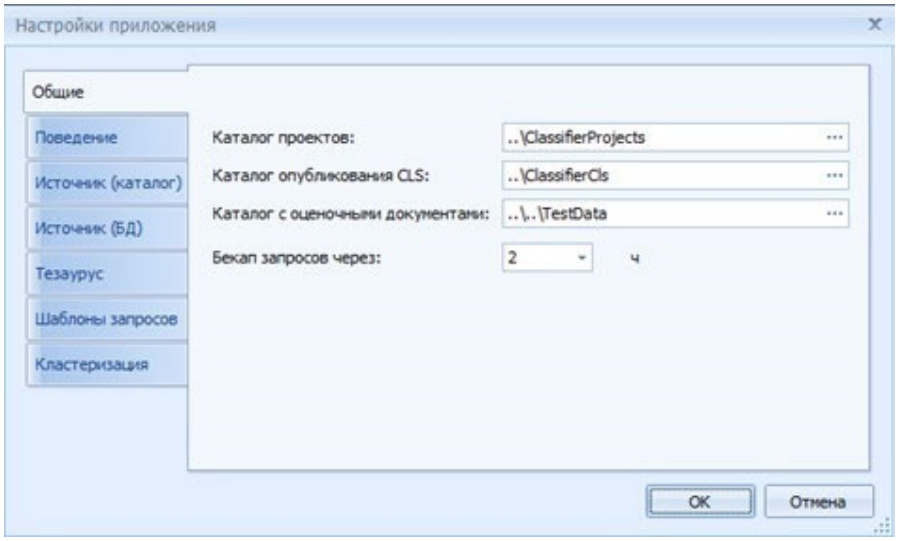


Рисунок 15 – Окно настроек приложения – Общие настройки

Вкладка «Общие» позволяет:

- изменить рабочие пути каталогов для сохранения и открытия проектов («Каталог проектов») и классификаторов («Каталог опубликования CLS»);
- прописать путь к каталогу с оценочной подборкой документов «Каталог с оценочными документами» для оценки качества правил классификации;
- указать интервал автоматического бекапирования (создания резервной копии) запросов.

После того, как настройки будут изменены, необходимо нажать на кнопку «ОК».

2.2.2 Настройки поведения

Вкладка «Поведение» представлена на рисунке 16 и используется для:

- выбора операций, которые будут выполняться после двойного нажатия

правой клавишей «мыши» по рубрике: открытие поля для редактирования запросов («Редактировать запрос») или открытие поля для работы с документами рубрики («Показать документы»);

- включения/отключения функции проверки времени работы регулярных выражений, соответствующих запросам, во время компиляции. В случае включения функции запросы классификатора будут преобразованы в регулярные выражения (если это возможно), и во время компиляции будет выполнен подсчет времени работы регулярных выражений. Информация о времени работы регулярных выражений будет отражена в отчете о результатах обучения классификатора (появляется после выполнения команды «Компиляция») в пунктах «Время классификации подборки (рег.)», «Скорость классификации подборки (рег.)», «Время обработки эталонов (рег.)»;
- включения/отключения режима проверки орфографии.

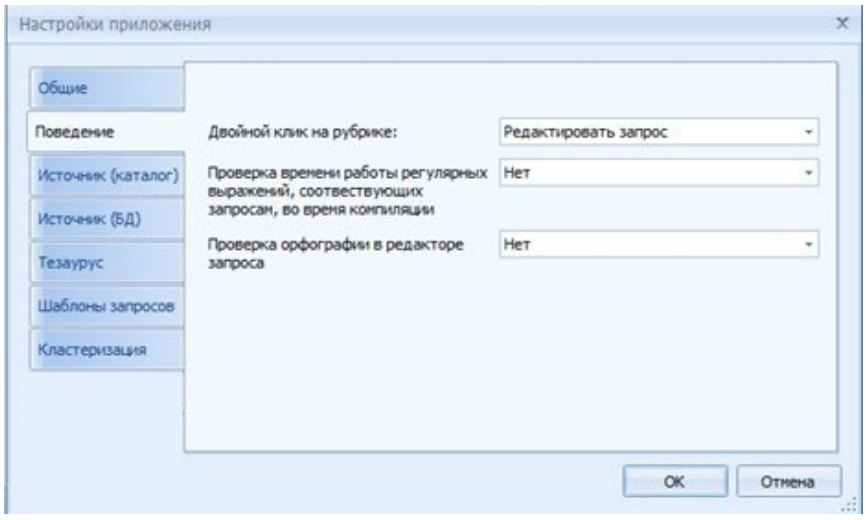


Рисунок 16 – Окно настроек приложения «Поведение»

2.2.3 Настройки пути к каталогу

Во вкладке «Источник (каталог)», представленной на рисунке 17, производятся создание и настройка пути к каталогу, в котором располагается тестовая подборка документов.

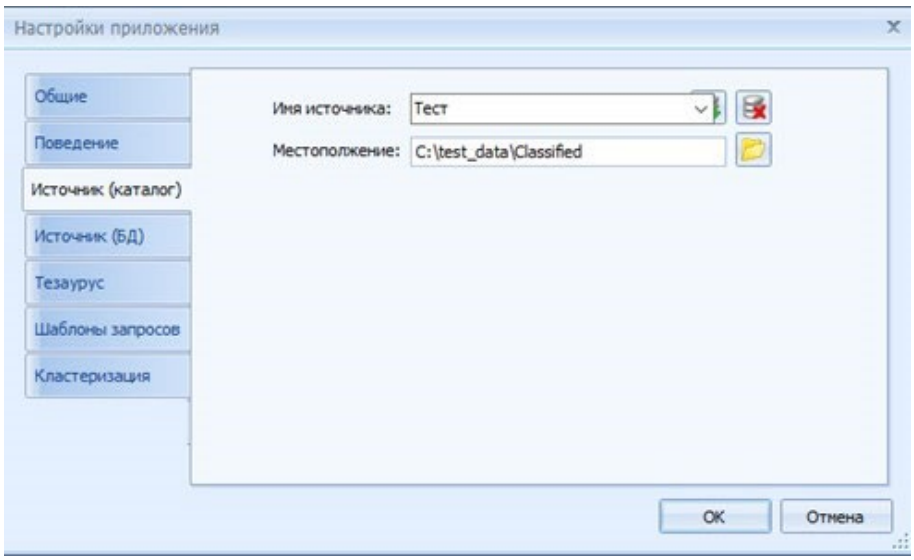


Рисунок 17 – Окно настроек приложения – Источник (каталог)

2.2.4 Настройки пути к Базе данных

Во вкладке «Источник (БД)», представленной на рисунке 18, происходит настройка Базы данных классификатора. Данные для настройки предоставляются Администратором.

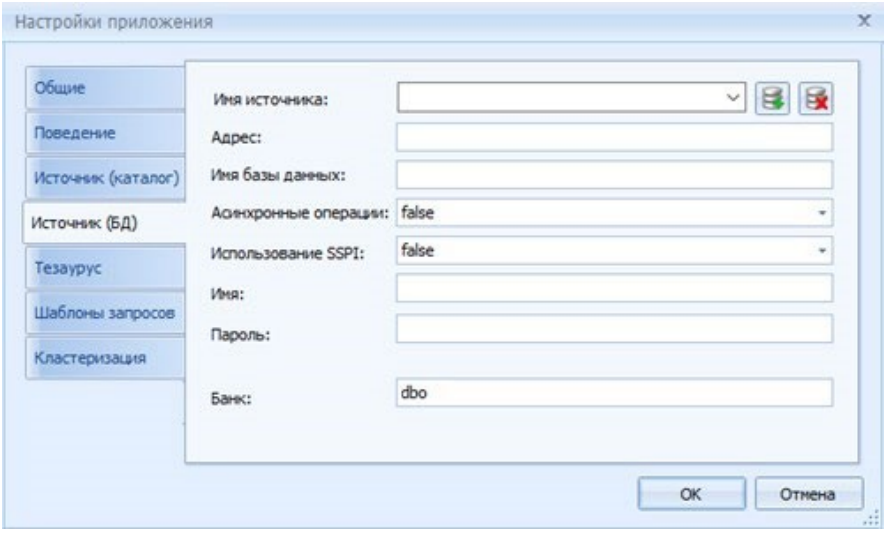


Рисунок 18 – Окно настроек приложения «Источник (БД)»

2.2.5 Настройки Базы тезаурусов

«Тезаурус» – вкладка для настройки Базы тезаурусов (словарей), представленная на рисунке 19. Данные для настройки предоставляются Администратором.

При нажатии на кнопку «Применить» запускаются заданные настройки. При нажатии на кнопку «Обновить» происходит обновление Базы тезаурусов. Данная функция используется для актуализации базы в случае, если со времени подключения к программе в базу были внесены изменения.

С помощью кнопки «Создать» можно создать новую Базу тезаурусов с заданными настройками. Кнопка «Удалить» выполняет удаление Базы тезаурусов с сервера.

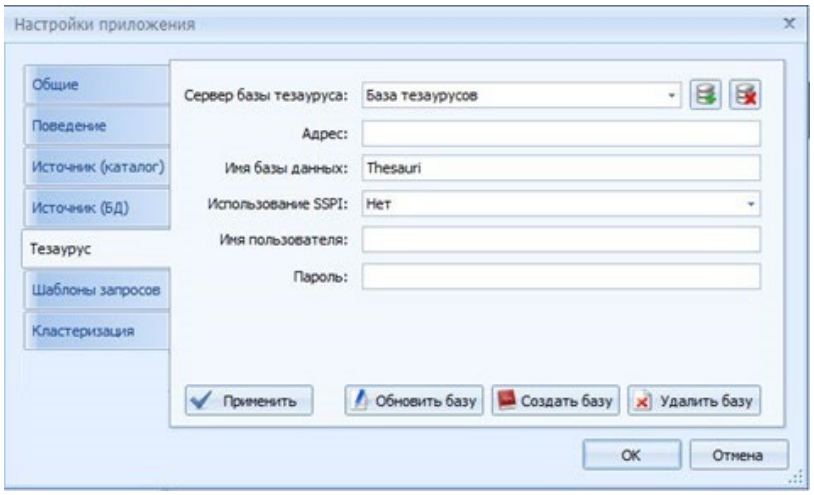


Рисунок 19 – Окно настроек приложения «Тезаурус»

2.3 Резервирование

Основные настройки программного средства хранятся в директории, которая расположена по адресу «C:\Users\User\AppData\Roaming\Cls». После переустановки программы настройки автоматически восстанавливаются.

2.4 Шаблоны запросов

Во вкладке «Шаблоны запросов», представленной на рисунке 20, прописывается запрос «по умолчанию», который автоматически будет подставляться в каждую пустую рубрику, кроме рубрики «Другие», при компиляции проекта.

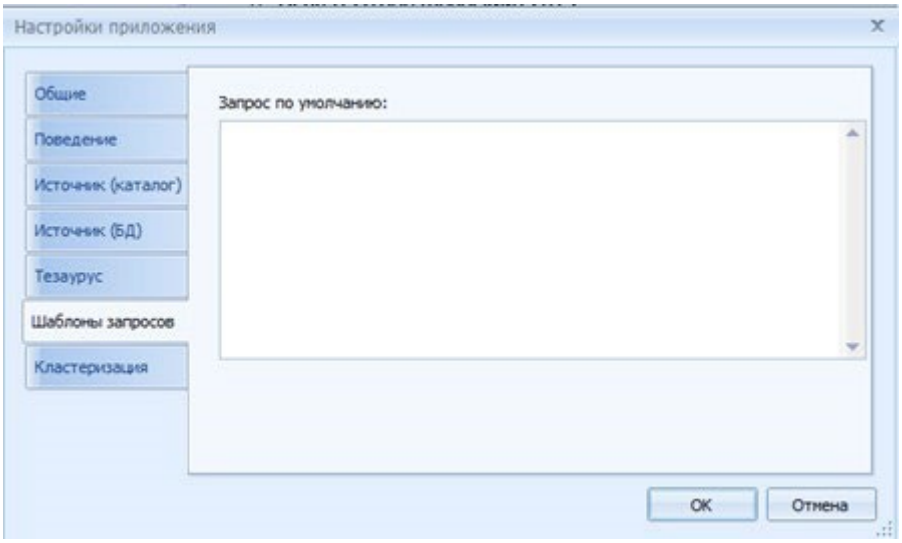


Рисунок 20 – Окно настроек приложения «Шаблоны запросов»

2.5 Кластеризация

Во вкладке «Кластеризация», представленной на Рисунок 21, прописывается путь к каталогам с подборками документов на которых будет производиться обучение для формирования правильных кластеров.

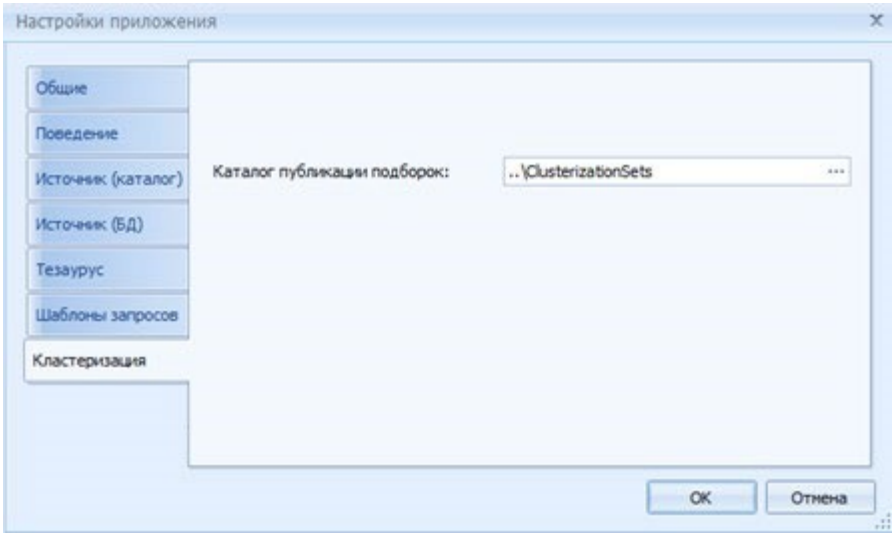


Рисунок 21 – Окно настроек приложения «Кластеризация»

2.6 Подсистема «Классификаторы»

Подсистема «Классификаторы» предназначена для создания, тестирования и отладки классификаторов для классификации текстовых документов.

2.6.1 Графический интерфейс

На рисунке 10 представлен графический интерфейс подсистемы «Классификаторы» при первичной установке программы.

Панель инструментов (2 на Рисунок 10) содержит:

- кнопку «Панель “Проектов”» для открытия панели со списком проектов классификаторов;
- меню «Создать», в котором перечислены варианты создания проекта (все функции из списка продублирована на панели «Проекты» | «Действия):
  - «Создать Ctrl+N» – создание нового проекта классификатора;
  - «Создать из .cls» – создание нового проекта на основе .clsфайла;
  - «Создать из .xml» – создание нового проекта на основе .xmlфайла, который содержит в себе структуру классификатора;
  - «Создать из .csv» - создание нового проекта на основе .csvфайла – excel - таблицы;
- кнопку «Открыть» для выбора проекта классификатора из другой директории;
- кнопку «Библиотека запросов»;
- поисковую строку для контекстного поиска.

После открытия проекта классификатора панель инструментов дополняется следующими инструментами (Рисунок 22):

- кнопка «Сохранить» для сохранения внесенных изменений проекта;
- кнопка «Компилировать» для запуска процесса компиляции текущего проекта (компиляция правил классификации, проверка на наличие синтаксических ошибок в правилах, а также оценка их качества);
- кнопка «Обучить» для запуска процесса обучения классификатора на основе документов классификатора и правил, заданных в конфигурационном файле;
- кнопка «Опубликовать» для сохранения проекта в формате cls в заданной директории;

- кнопка «Перевод» – перевод классификатора с одного языка на другой;
- кнопка «Оценка» – оценка качества правил классификации по документам из внешнего источника;
- кнопка «Удалить с диска» – полное удаление проекта классификатора из директории.

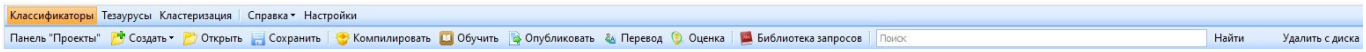


Рисунок 22 – Панель «Классификаторы»

2.6.2 Основные функции

Проект классификатора включает в себя структуру рубрик, правила классификации, а также эталонные примеры документов.

Процесс создания проекта классификатора состоит из следующих этапов:

- создание или открытие проекта классификатора;
- структурирование классификатора;
- написание и редактирование правил классификации;
- работа с эталонными документами классификатора;
- анализ и отладка правил классификации;
- компиляция классификатора и формирование отчета о результатах обучения;
- выгрузка классификатора.

Для работы с проектом классификатора предусмотрен следующий набор функций:

- создание нового проекта классификатора;
- создание проекта классификатора из cls-файла;
- создание проекта классификатора из xml-файла;
- создать проект классификатора из csv-файла;
- открытие проекта классификатора;
- сохранение проекта;
- закрытие проекта;
- удаление проекта.

Для структурирования классификатора предусмотрено:

- создание новой рубрики/подрубрики;
- копирование/вставка существующей рубрики в пределах одного классификатора;
- изменение названия рубрики или других ее свойств;
- удаление рубрики.

Для написания и редактирования правил классификации предусмотрен следующий ряд функций:

- создание правил классификации с использование языка SCATQL (глобального запроса или запроса отдельной рубрики);
- создание правила классификации на основе переменных из тезауруса;
- расширение правила классификации с помощью тезауруса;
- использование инструмента «Статистика» для формирования и расширения правила классификации для рубрики;
- автоматическое обучение классификатора;
- проверка синтаксических ошибок в запросе;
- использование внешних файлах в рамках запроса;
- поиск понятий;
- перевод правил классификации на другие языки.

Для работы с эталонными документами классификатора предусмотрен следующий ряд функций:

- прикрепление эталонных документов к рубрике (каталогом или пофайлово);
- прикрепление текста из буфера обмена;
- прикрепление сгруппированных эталонов;
- тестирование внешних источников (тестирование с помощью БД или внешнего каталога, «Оценка»);
- просмотр эталонных документов;
- просмотр фрагментов в тексте (разметка по запросу, пользовательская разметка);

- полнотекстовый поиск по документам классификатора;
- ручное перераспределение документов по рубрикам;
- выгрузка документов классификатора;
- удаление документов;
- просмотр неиспользуемых документов классификатора;
- удаление неиспользуемых документов классификатора;
- выгрузка неиспользуемых документов классификатора;
- задание рубрики для неиспользованного документа классификатора.

Для анализа и отладки правил классификации предусмотрен следующий ряд функций:

- «Анализ»;
- «Оценка»;
- «Тестирование»;
- «Отладка».

Для компиляции классификатора и формирования отчета о результатах обучения предусмотрена функция «Компилировать», а для выгрузки классификатора (сохранение в cls формате) функция «Опубликовать».

2.6.2.1 Создание, открытие и удаление проекта классификатора С проектом классификатора можно осуществить ряд действий:

- создать - новый или на основе уже существующего классификатора из нескольких форматов;
- - открыть из выбранной директории;
- - произвести полное удаление проекта и всех его файлов с диска.

2.6.2.2 Создание нового проекта классификатора

Для создания нового проекта классификатора необходимо выбрать команду «Создать Ctrl+N» (см. Рисунок 22) на панели инструментов подсистемы «Классификаторы», либо на панели «Проекты | Действия» «Создать новый проект».

После чего откроется диалоговое окно «Создать новый проект классификатора», представленное на рисунке 23, в котором необходимо задать «Название классификатора» и «Комментарий» к классификатору, а также определить путь к каталогу проекта. Путь проекта можно прописать самостоятельно (вручную) или оставить по умолчанию (основывается на названии создаваемого классификатора). Рекомендуется оставлять путь по умолчанию, т.к. в этом случае созданный классификатор будет включен в каталог проектов, показываемый при открытии приложения.

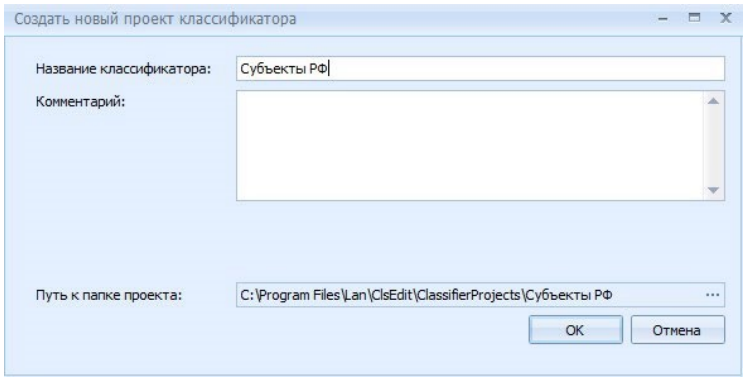


Рисунок 23 – Создание нового проекта классификатора

После нажатия на кнопку «ОК» будет создан проект классификатора, содержащий по умолчанию единственную рубрику «Другие», как показано на рисунке 24. В данную рубрику будут помещаться документы, не относящиеся ни к одной из тематических рубрик. Она работает по принципу «Корзины» и содержит нерелевантные документы.

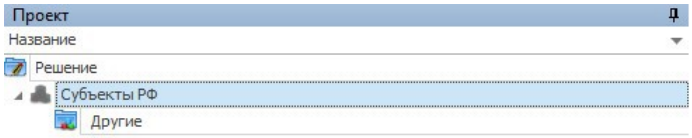


Рисунок 24 – Новый проект классификатора

Для создания проекта классификатора из cls-файла необходимо в меню «Создать» выбрать команду «Создать из .cls», либо на панели «Проекты | Действия» «Создать новый проект на основе .cls-файла», как показано на рисунке 25.

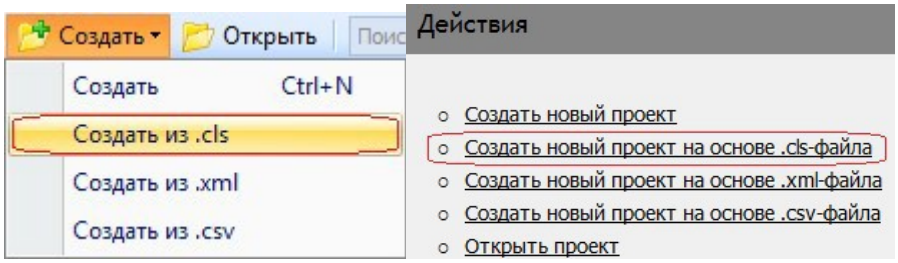


Рисунок 25 – Инструмент для создания нового проекта из cls-файла

В результате выполнения данного действия откроется окно для выбора clsфайла, представленное на рисунке 26.

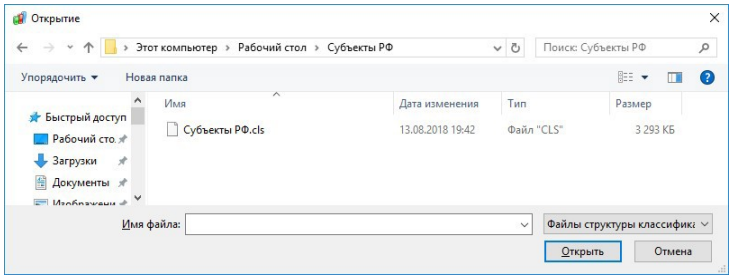


Рисунок 26 – Диалоговое окно для выбора cls-файла

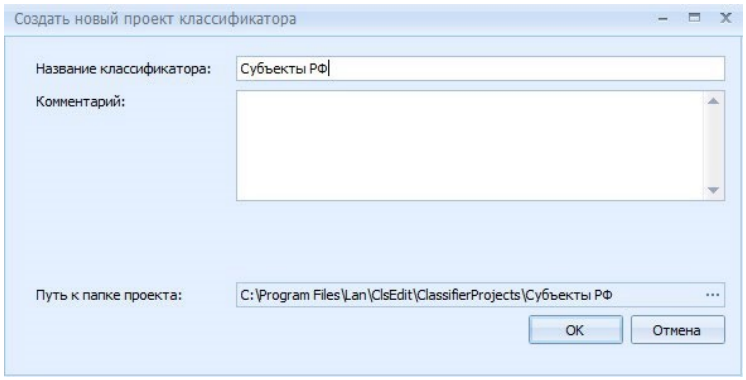


Рисунок 27 – Окно «Создать новый проект классификатора»

2.6.2.3 Создание проекта классификатора из xml-файла

Для создания проекта классификатора из xml-файла необходимо в меню «Создать» выбрать команду «Создать из .xml», либо на панели «Проекты | Действия» «Создать новый проект на основе .xml-файла», как показано на рисунке 28.

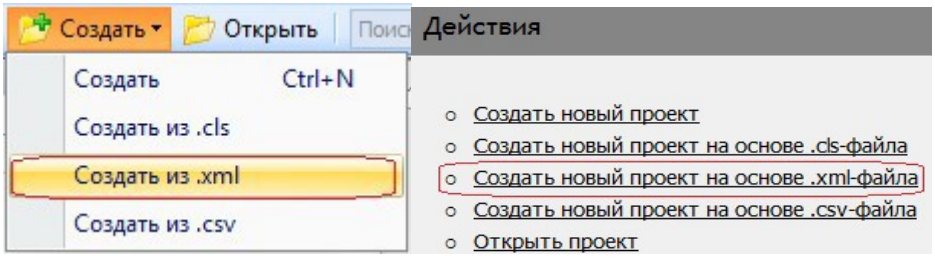


Рисунок 28 – Инструмент для создания классификатора из xml-файла

В результате выбора данного пункта откроется окно выбора файла и окно «Создать новый проект классификатора», представленные на рисунке 29.

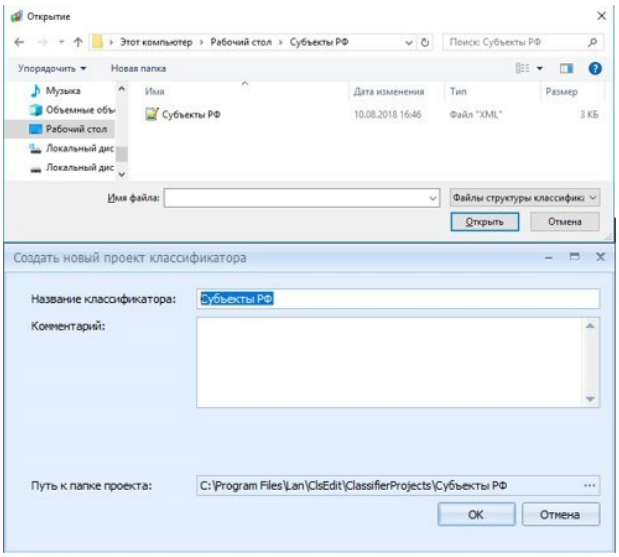


Рисунок 29 – Создание проекта классификатора из xml-файла

В результате будет создан проект классификатора, содержащий структуру рубрик из xml-файла, представленный на рисунке 30.

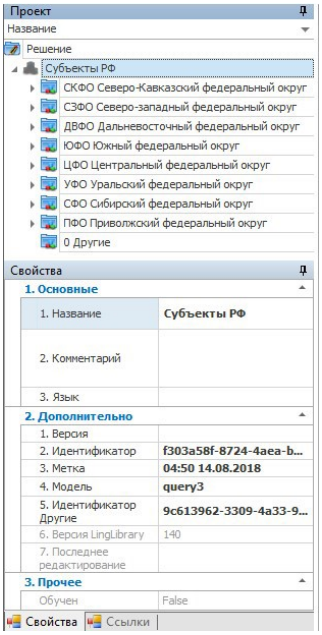


Рисунок 30 – Проект из xml-файла

2.6.2.4 Создание проекта классификатора из csv-файла

Для создания проекта классификатора из csv-файла необходимо в меню «Создать» выбрать команду «Создать из .csv», либо на панели «Проекты | Действия» «Создать новый проект на основе .csv-файла».

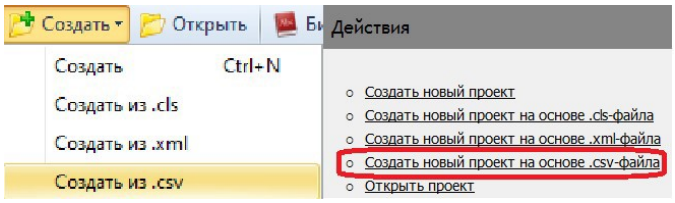


Рисунок 31 – Инструмент для создания нового проекта из csv-файла

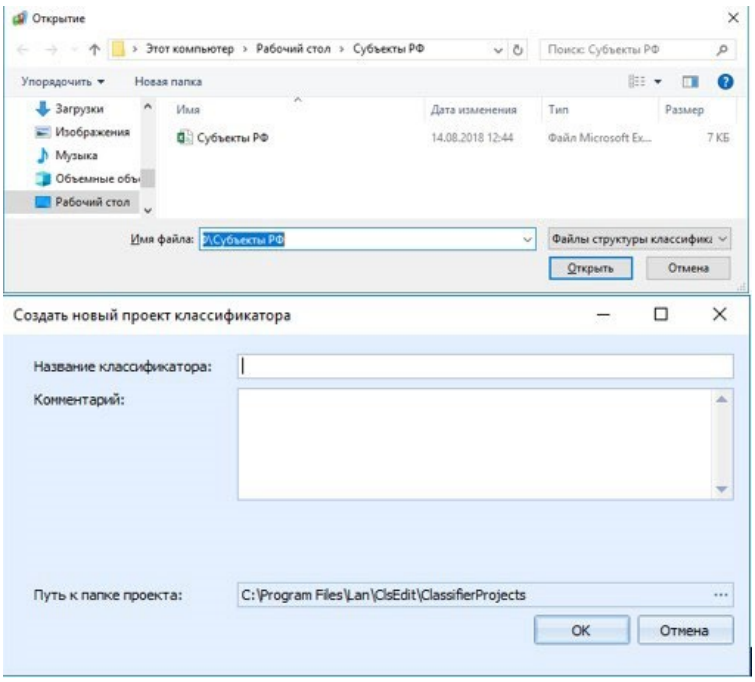


Рисунок 32 – Создание проекта классификатора из csv-файла

В окне «Создать новый проект классификатора» задаются название, комментарий и путь, куда будет сохраняться проект классификатора. По умолчанию в качестве директории для сохранения проекта указывается «ClassifierProjects».

В результате будет создан проект классификатора, содержащий структуру рубрик из csv-файла, представленный на рисунке 33.

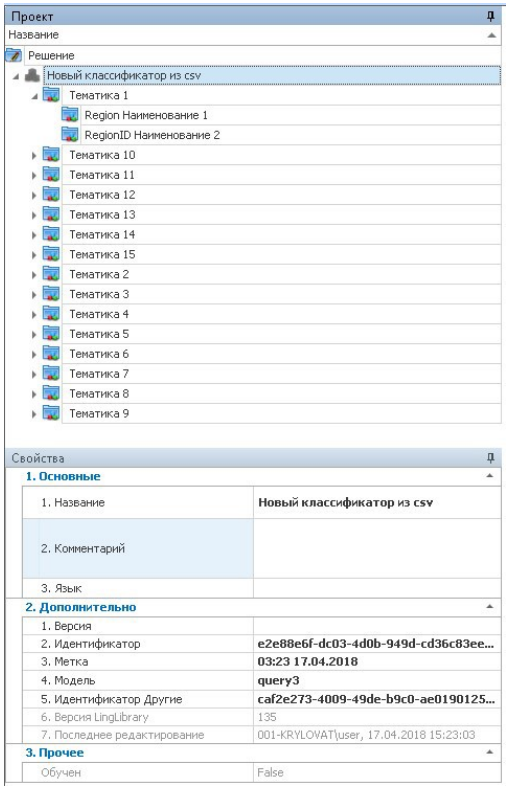


Рисунок 33 – Проект из csv-файла

A	B	C	D
Тематика	Наименование	Поле	Рубрика
Тематика 1	Наименование 1	Region	
Тематика 1	Наименование 2	RegionID	
Тематика 2	Наименование 3	District	
Тематика 2	Наименование 4	DistrictID	
Тематика 2	Наименование 5	DateEvent	
Тематика 2	Наименование 6	TimeEvent	
Тематика 2	Наименование 7	FromDoc	
Тематика 2	Наименование 8	RefDoc	
Тематика 2	Наименование 9	DateDoc	
Тематика 2	Наименование 10	Built	
Тематика 2	Наименование 11	Changed	
Тематика 2	Наименование 12	Okrug	
Тематика 2	Наименование 13	About	
Тематика 2	Наименование 14	WhoBuilt	
Тематика 3	Наименование 15	WhoChanged	
Тематика 3	Наименование 16	StatTA	

Рисунок 34 – Содержимое csv файла

Для открытия уже существующего проекта необходимо на панели инструментов выбрать команду «Открыть», либо на панели «Проекты | Действия» выбрать «Открыть проект».

Для закрытия проекта в контекстном меню, выбранного классификатора необходимо выбрать функцию «Убрать классификатор», представленной на рисунке 35 в контекстном меню классификатора.

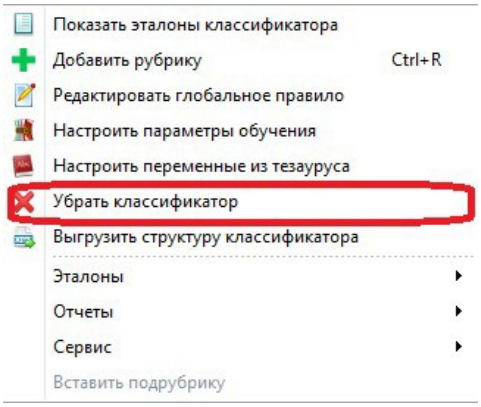


Рисунок 35 – Контекстное меню классификатора | «Убрать классификатор»

Рисунок 36 – Панель инструментов | «Удалить с диска»

Структура классификатора определяется иерархическим деревом содержащихся в нем рубрик.

2.6.2.5 Создание новой рубрики или подрубрики

Для создания новой рубрики/подрубрики необходимо в контекстном меню выбрать функцию «Добавить рубрику/подрубрику», как показано на рисунке 37, либо воспользоваться горячим сочетанием клавиш «Ctrl + R».

Важно! При использовании горячих клавиш фокус должен быть на том узле (название классификатора или рубрике), в котором будет создаваться дочерняя рубрика.

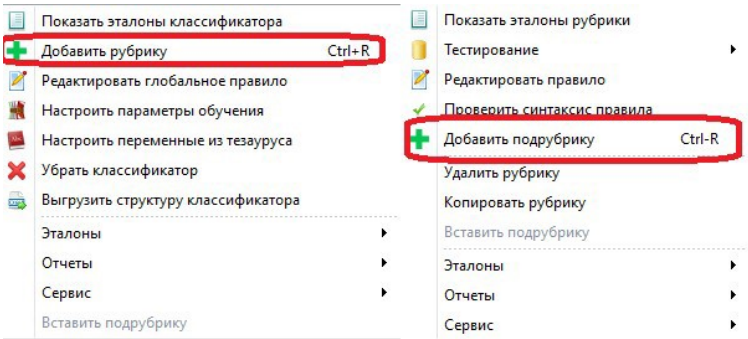


Рисунок 37 – Создание новой рубрики

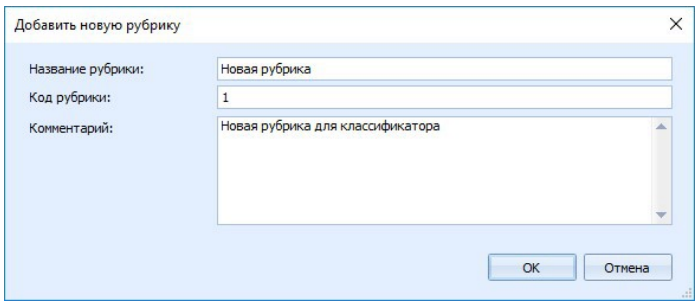


Рисунок 38 – Диалоговое окно для создания новой рубрики

2.6.2.6 Удаление рубрики

Для удаления рубрики необходимо в контекстном меню рубрики выбрать пункт «Удалить рубрику» (рисунок 39). После чего подтвердить свои действия в открывшемся диалоговом окне, как показано на рисунке 40.

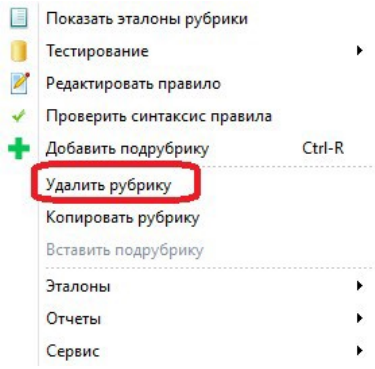


Рисунок 39 – Удаление рубрики

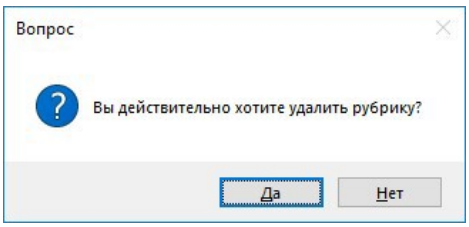


Рисунок 40 – Диалоговое окно «Вопрос»

Эталоны – это массив «хороших» документов, которые наиболее полно раскрывают тематику классификатора и были подобраны экспертом вручную. Эталонная (экспертная) подборка необходима для написания и отладки правил классификации. Чем полнее и качественнее подборка, тем выше показатели качества классификатора.

2.6.2.7 Прикрепление эталонных документов к рубрике

Для добавления эталонных документов в рубрику необходимо в контекстном меню рубрики выбрать команду «Эталоны | Прикрепить эталоны к рубрике», как показано на рисунке 41, либо воспользоваться горячим сочетанием клавиш «Ctrl + D»

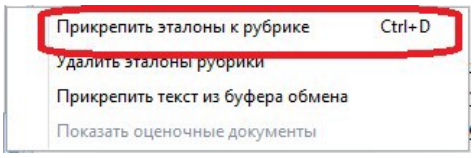


Рисунок 41 – Функция «Прикрепить эталоны к рубрике Ctrl-D»

После чего откроется диалоговое окно «Прикрепить эталоны», представленное на рисунке 42.

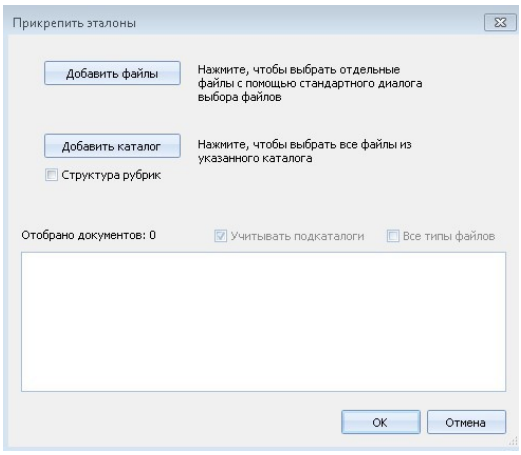


Рисунок 42 – Диалоговое окно «Прикрепить эталоны»

«Добавить файлы» – добавление отдельного(ых) файла(ов) в форматах: TXT, DOC, DOCX, UIML, XML и др.

«Добавить каталог» – добавление всех файлов из указанного каталога.

Также доступна метка «Структура рубрик», в сочетании с которой функция «Добавить каталог» создает структуру подрубрик из подкаталогов выбранной директории с добавлением в качестве эталонов документов, содержащихся в подкаталогах.

Для добавления отдельного файла или нескольких файлов необходимо в диалоговом окне «Прикрепить эталоны» нажать на кнопку «Добавить файлы»,

после чего откроется окно, в котором можно выбрать необходимый(ые) файл(ы). После того, как файлы были выбраны, снова отобразится диалоговое окно «Прикрепить эталоны», в котором будет указано, сколько документов отобрано и каких форматов. После нажатия на кнопку «ОК» все выбранные документы будут прикреплены к рубрике. Процесс добавления отдельных документов к рубрике представлен на рисунке 43.

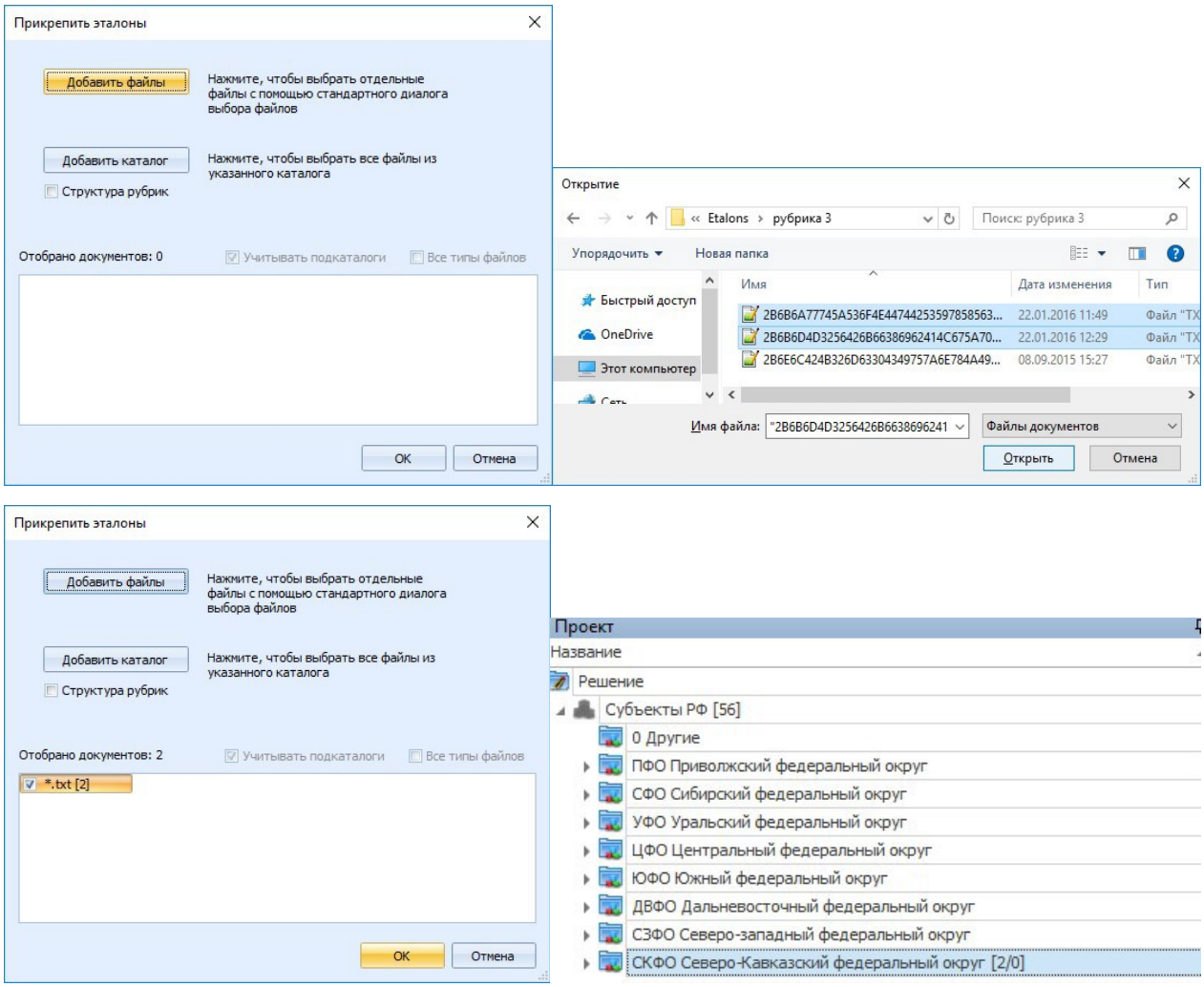


Рисунок 43 – Процесс добавления отдельных документов к рубрике

Для добавления содержимого всего каталога в диалоговом окне «Прикрепить эталоны» необходимо нажать на кнопку «Добавить каталог», после чего откроется окно «Обзор папок» для выбора необходимого каталога. После того, как каталог был выбран, также, как и в случае добавления отдельных документов, откроется окно «Прикрепить эталоны», в котором будут отображены количество и форматы прикрепляемых файлов из выбранного каталога. После нажатия на

кнопку «ОК» все документы из выбранного каталога будут прикреплены к рубрике.

Метки «Учитывать подкаталоги» и «Все типы файлов» указывают все ли подкаталоги и форматы будут учитываться при загрузке эталонов в рубрику. Процесс добавления каталога с документами к рубрике представлен на рисунке 44.

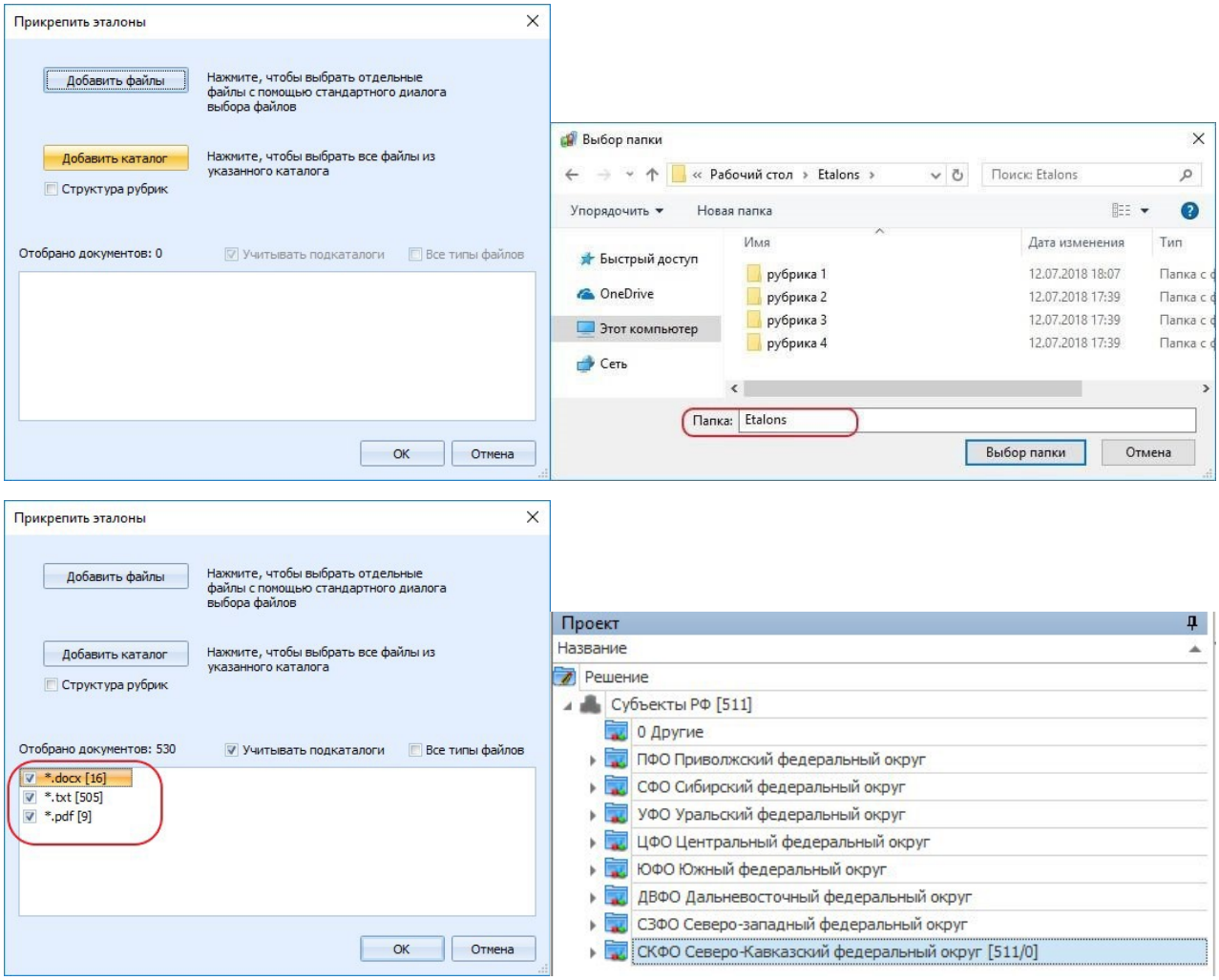


Рисунок 44 – Процесс добавления каталога с документами к рубрике

После добавления документов они будут скопированы в папку «Etalons» и добавлены в рубрику в качестве эталонных.

В случае, если добавляемые документы уже были ранее прикреплены к рубрике, будет выведено сообщение о том, что данные документы уже добавлены в качестве эталонных, как показано на рисунке 45.

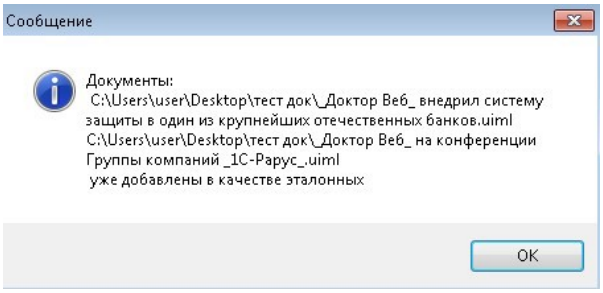


Рисунок 45 – Добавление эталонов

Количество прикрепленных документов показывается после названия рубрики в квадратных скобках «[a/b]», где a – эталоны, прикрепленные вручную; b – эталоны, прикрепленные автоматически при компиляции или обучении классификатора.

Общее количество документов в проекте показывается после названия в корне дерева проекта.

При написании классификатора может возникнуть потребность в создании иерархии подрубрик конкретной рубрики и заполнении их эталонами. Для ускорения работы над данным процессом в программном комплексе предусмотрена функция «Прикрепить сгруппированные эталоны».

В диалоговом окне «Прикрепить эталоны» необходимо пометить чекбокс «Структура рубрик», после чего нажать на кнопку «Добавить каталог» и в открывшемся окне «Обзор папок» выбрать необходимый каталог, который содержит иерархическую структуру папок, каждая из которых в свою очередь содержит набор документов. После этого во вновь открывшемся окне «Добавить каталог» следует установить метку «Учитывать подкаталоги», если она отсутствует по умолчанию.

В результате в выбранной рубрике будет создана иерархическая структура подрубрик, аналогичная структуре выбранного каталога, а к каждой подрубрике будут прикреплены документы в качестве эталонов из соответствующего подкаталога.

Процесс добавления структуры подкаталогов в качестве подрубрик представлен на рисунке 46.

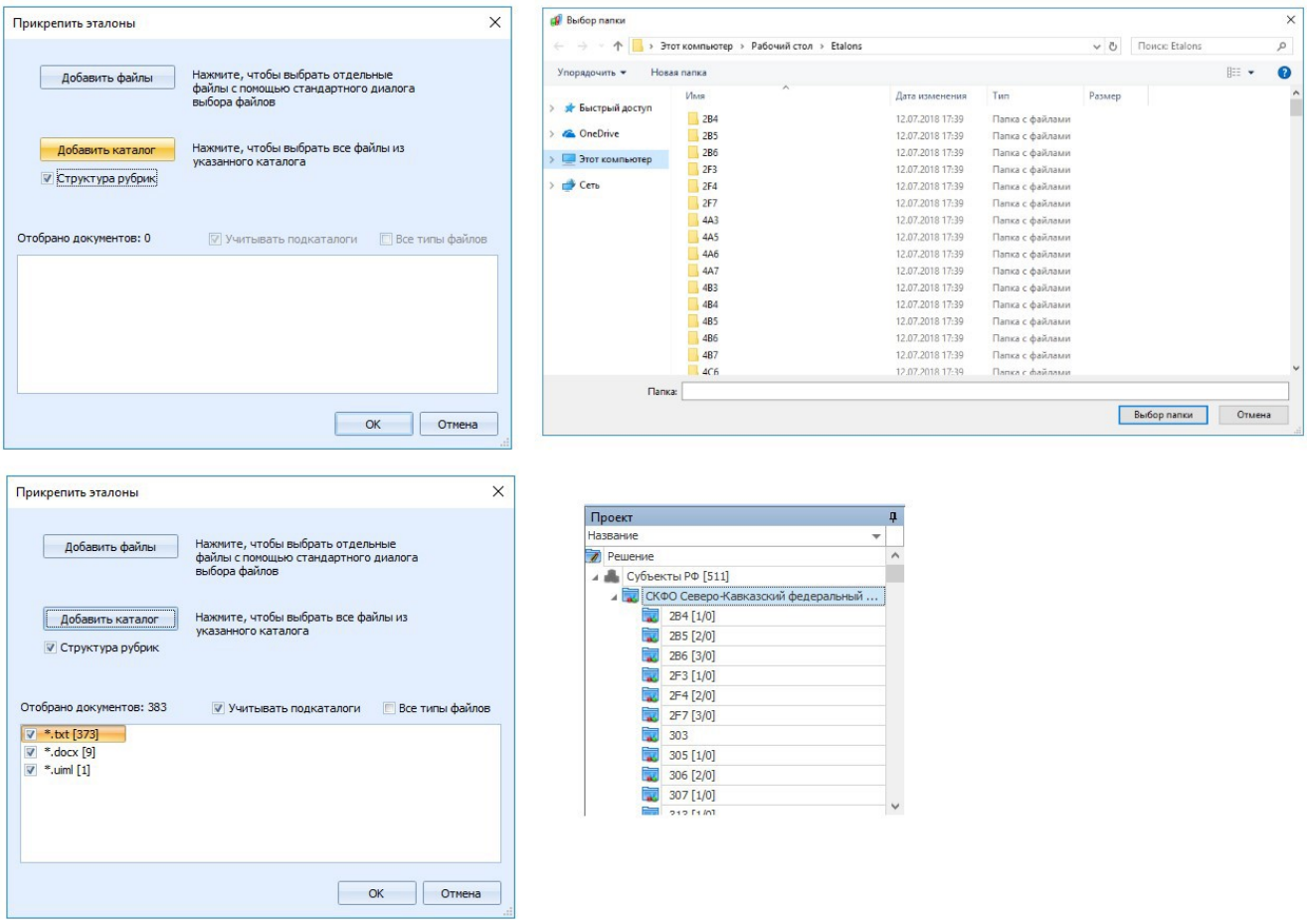


Рисунок 46 – Процесс добавления структуры подкаталогов в качестве подрубрик

В качестве эталонного документа также можно добавить текст из буфера обмена. Для этого необходимо скопировать интересующий текст и в контекстном меню рубрики выбрать соответствующий инструмент «Эталоны | Прикрепить текст из буфера обмена», после чего отобразится окно «Вопрос» с запросом на подтверждение добавления содержимого буфера к рубрике.

После подтверждения выбранный текст будет сохранен в виде файла формата .txt и прикреплен в качестве эталона к рубрике.

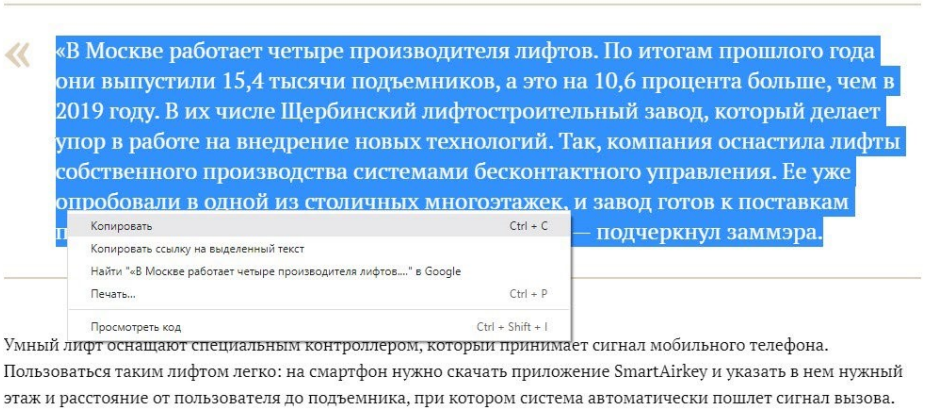


Рисунок 47 – Прикрепление текса из буфера обмена

Функция «Прикрепить сгруппированные эталоны» вызывается из контекстного меню классификатора и позволяет сократить время создания структуры классификатора и заполнения ее эталонами.

После запуска процесса прикрепления открывается окно «Обзор папок», в котором нужно выбрать каталог с внутренней структурой папок, которую необходимо добавить в качестве структуры рубрик. В результате папки из выбранного каталога будут добавлены в классификатор в виде рубрик, а документы, содержащиеся в них, в виде эталонов.

Процесс прикрепления сгруппированных эталонов представлен на рисунке 48.

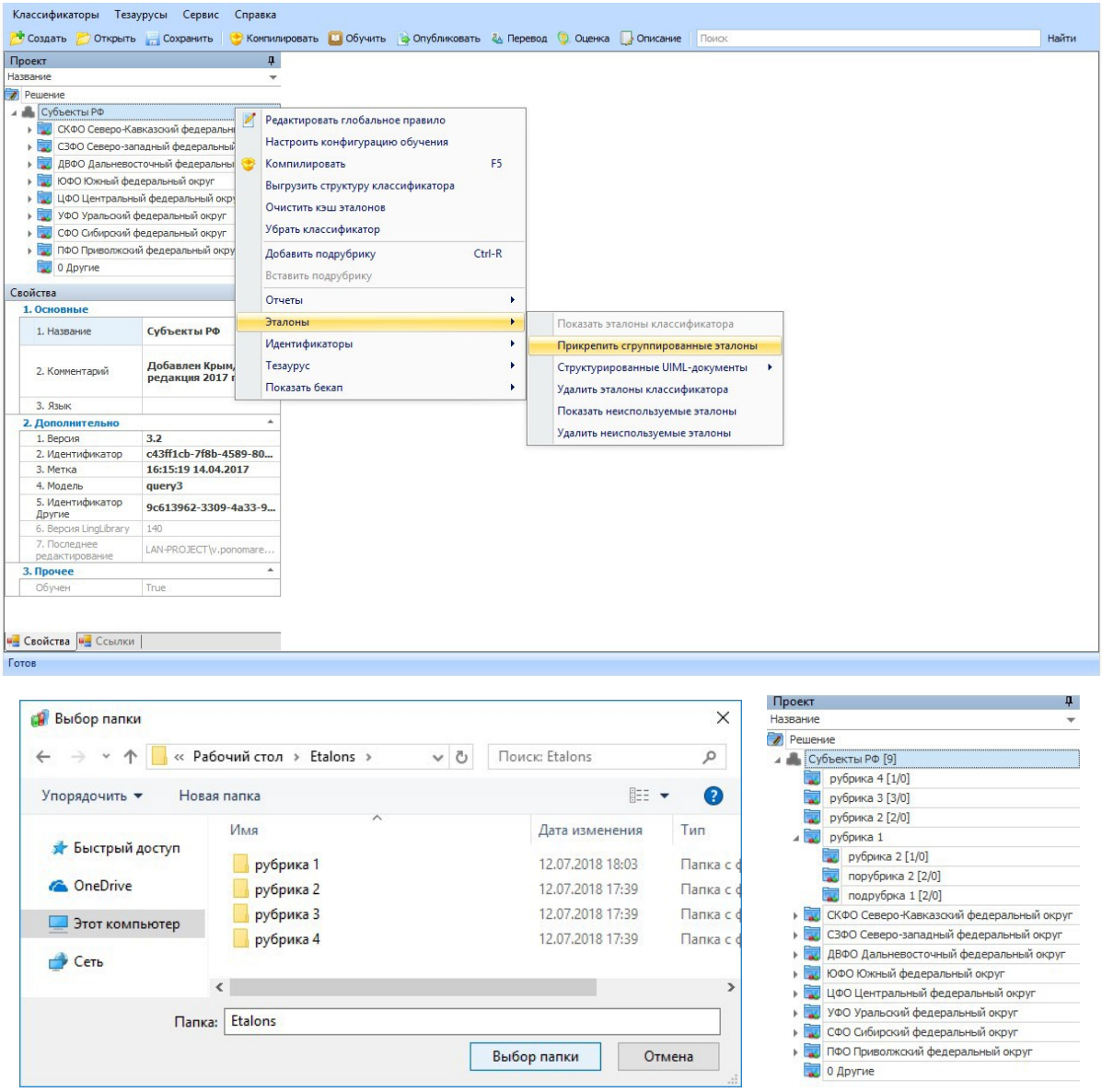


Рисунок 48 – Процесс прикрепления сгруппированных эталонов

2.6.2.8 Тестирование внешних источников

Функция «Тестировать» вызывается из контекстного меню рубрики и позволяет протестировать правила классификации отдельной рубрики на документах, не прикрепляя их в качестве эталонных или тестовых документов к классификатору, что позволяет облегчить «вес» проекта и скорость его обработки.

Для тестирования правил классификации отдельной рубрики по документам из внешней директории необходимо произвести ряд настроек. В меню «Сервис | Настройка» выбрать вкладку «Источник (каталог)», в которой необходимо создать «новый источник», нажав на кнопку  . После этого в поле «Имя источника»

необходимо ввести название тестовой подборки, а в поле «Местоположение» ввести вручную или выбрать путь к тестовой подборке, нажав на кнопку  . На рисунке 49 данные указаны по умолчанию.

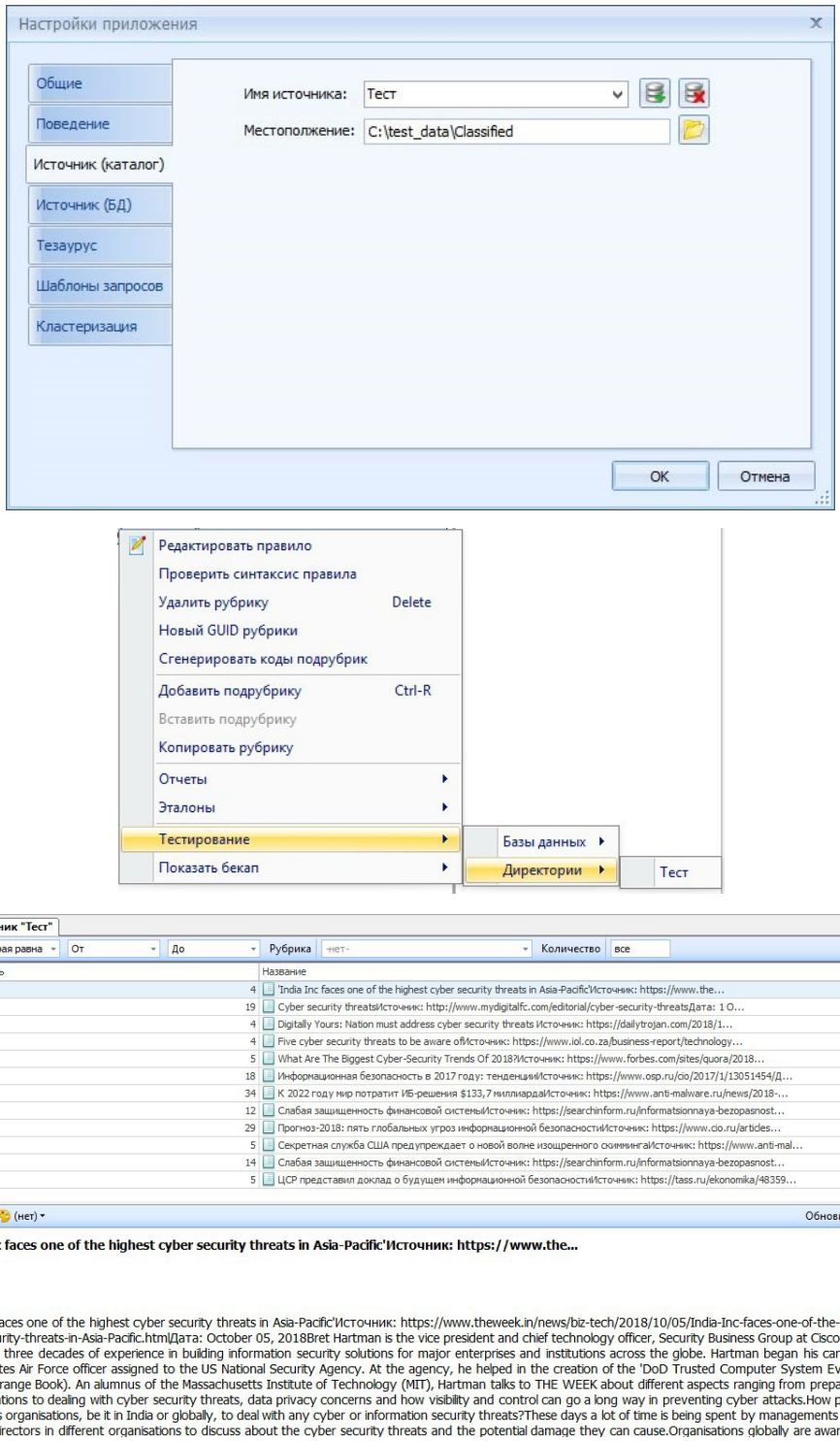


Рисунок 49 – Процесс тестирования внешнего источника

2.6.2.9 Просмотр эталонных документов

Для совершения различных манипуляций с документами всего классификатора или отдельной рубрики необходимо открыть область работы с документами, для этого в контекстном меню классификатора или рубрики необходимо выбрать «Эталоны | Показать эталоны классификатора» или «Эталоны | Показать эталоны рубрики» соответственно, как показано на рисунке 50.

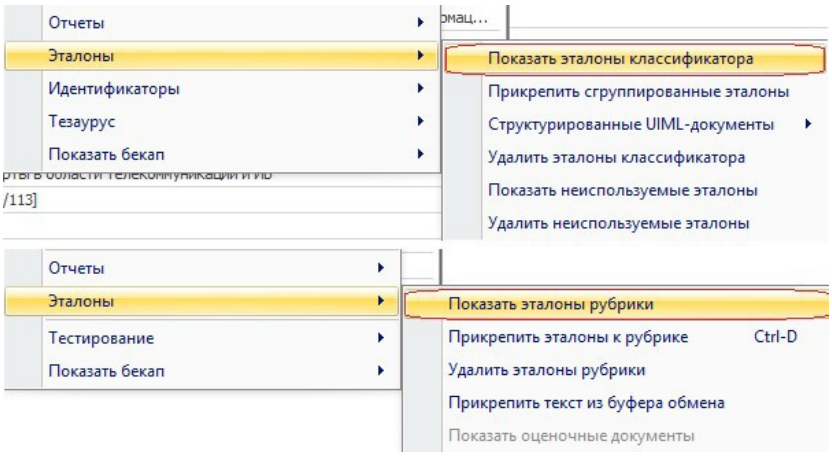


Рисунок 50 – Инструменты для показа документов всего классификатора или отдельной рубрики

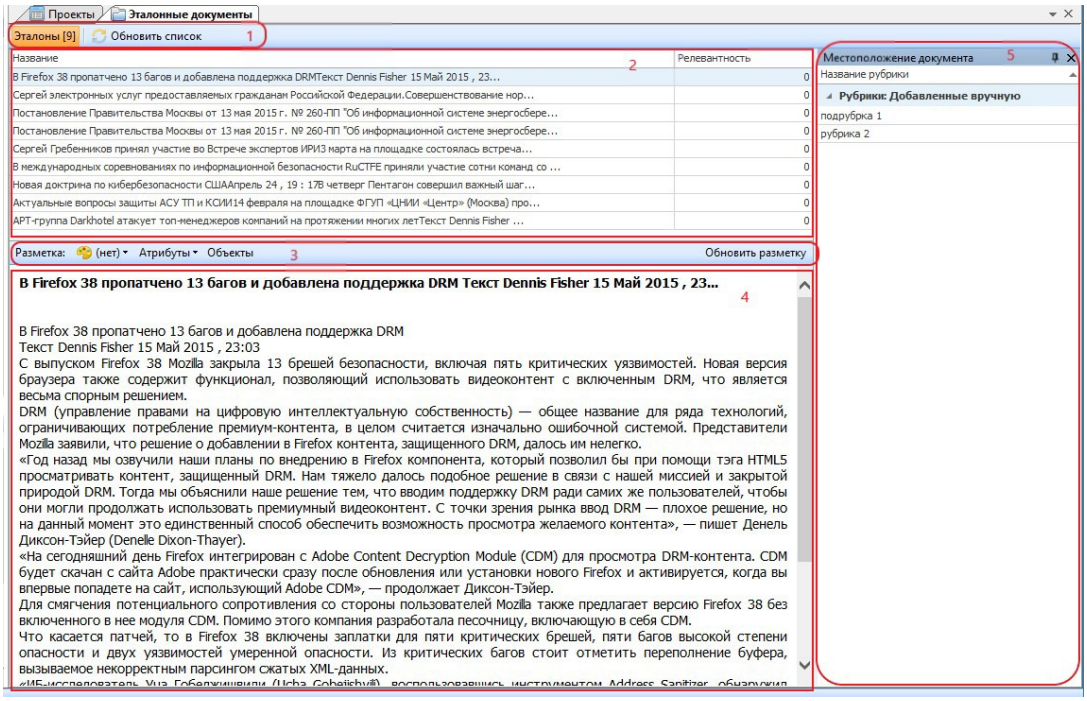


Рисунок 51 – Окно работы с документами для всего классификатора

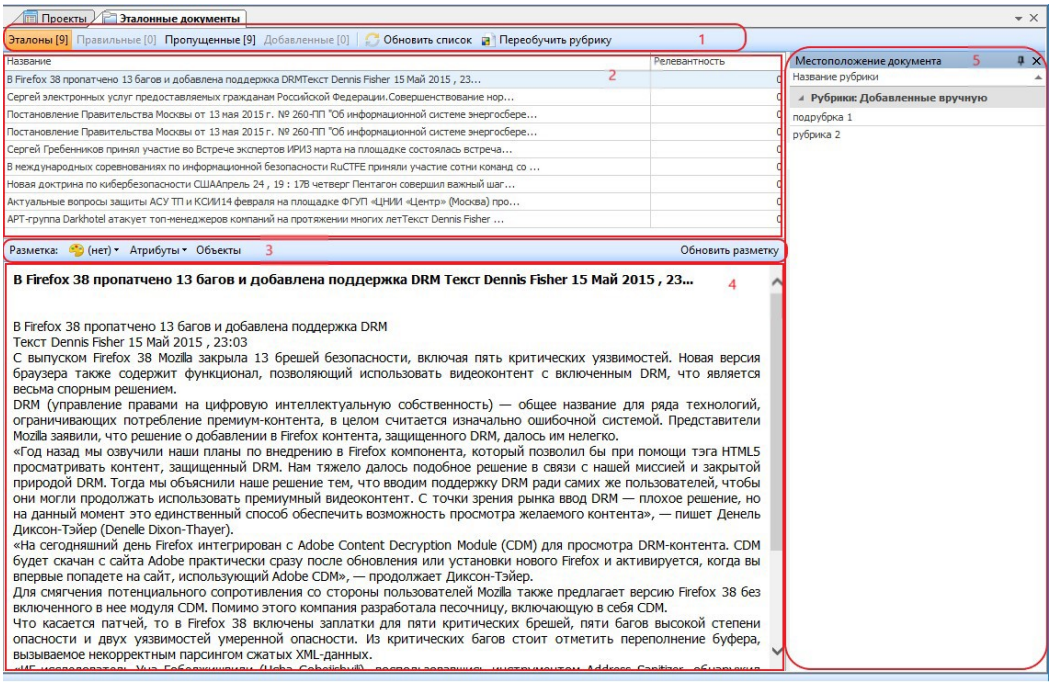


Рисунок 52 – Окно работы с документами для отдельной рубрики

Панель инструментов для работы с документами включает в себя:

- «Эталоны» – отображается список эталонных документов, которые были добавлены пользователем вручную, по-другому - экспертная подборка;

- «Правильные» – отображается список документов, который соответствует пересечению документов, добавленных пользователем вручную (эталоны) и отобранных при автоматической классификации по запросу (добавленные);

- «Пропущенные» – отображается список документов, отобранных пользователем (экспертом), но не отобранных при автоматической классификации документов (не удовлетворяют запросу рубрики);

- «Добавленные» – отображается список документов, отобранных при автоматической классификации документов;

- «Обновить список» – данный инструмент необходим для обновления списка документов при перераспределении документов по рубрикам или очередном добавлении (удалении) документов;

- «Переобучить рубрику» – обновление разметки документов и результатов автоматической классификации конкретной рубрики при изменении правил без компиляции всего проекта.

Панель инструментов для работы с фрагментами документа включает в себя:

- «Атрибуты» – открывается список атрибутов документа (пользовательские и по запросу), если они есть. Атрибут – это фрагмент текста, выделенный по запросу с заданным типом (названием атрибута). Атрибут и его тип задается с помощью оператора «#attribute», например, «#attribute PLACE #sourcestr(Россия)», – в данном случае в качестве типа атрибута выступает «PLACE», а его содержимого – «Россия», оператор «#sourcestr» указывает на то, что «Россия» в атрибут будет записываться в том же виде, что и в тексте. Пользовательские атрибуты – атрибуты, добавленные пользователем в качестве эталонных. По запросу – атрибуты, добавленные автоматически в результате обработки запроса классификатора;
- «Объекты» – открывается список объектов документа, если они есть в тексте. Объект – это фрагмент или несколько логически связанных элементов текста, которые выделяются по запросу с заданным типом (названием объекта).

Объекты могут иметь связи, образовывать иерархическую структуру.

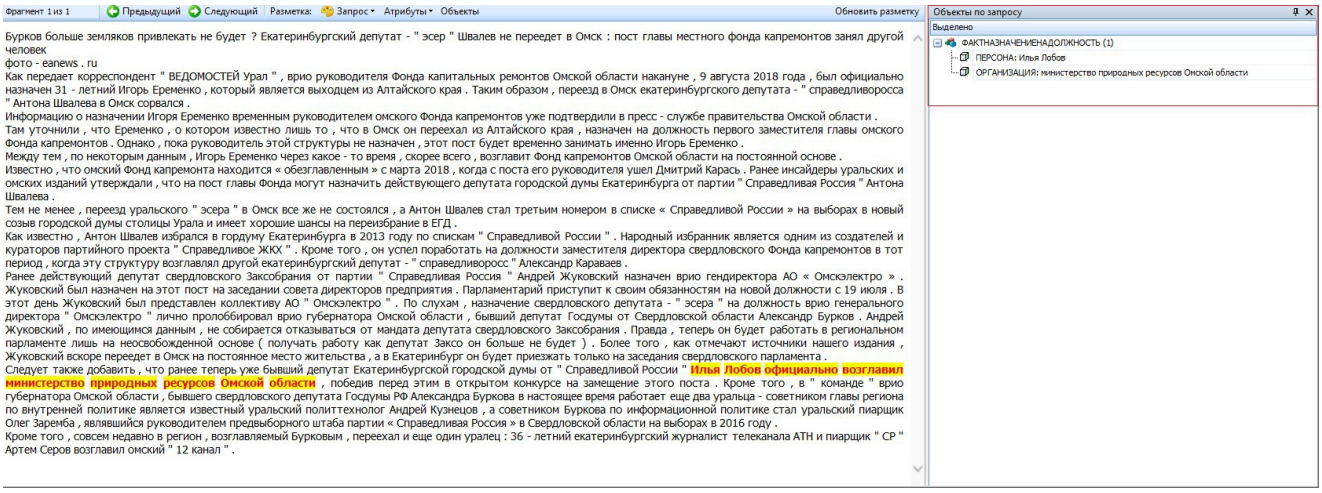


Рисунок 53 – Окно отображения объектов

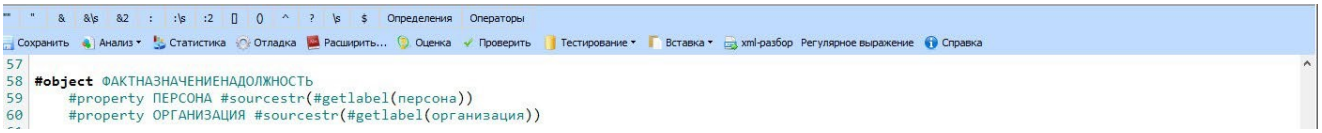


Рисунок 54 – Пример задания объекта со структурированными элементами.

Пример размеченного документа представлен на рисунке 55.

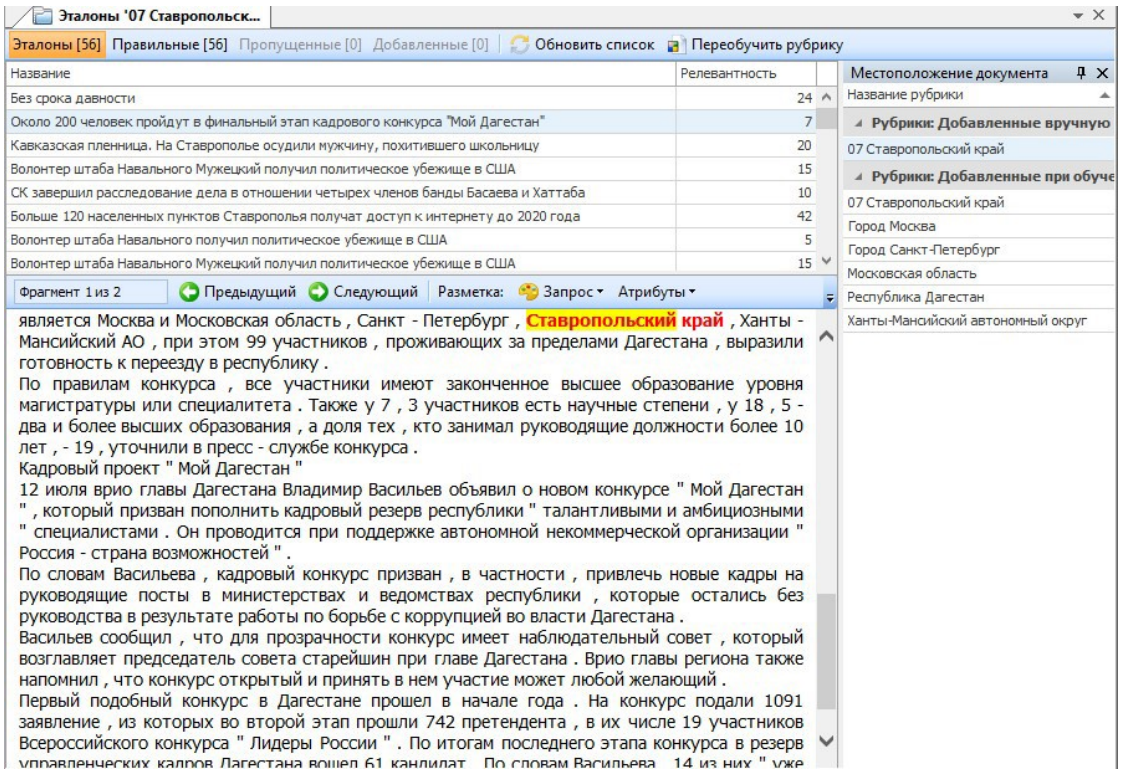


Рисунок 55 – Пример размеченного документа

Вкладка «Местоположение документа» отображает, к какой(им) рубрике(ам) был отнесен документ при различной классификации документов – ручной или автоматической («Рубрики: Добавленные вручную» и «Рубрики: Добавленные при обучении» соответственно).

2.6.2.10 Просмотр фрагментов в тексте (разметка по запросу, пользовательская разметка)

Для просмотра фрагментов текста с учетом различной разметки (по запросу или пользовательской) необходимо открыть панель работы с документами, после чего на панели инструментов для работы с фрагментами документа выбрать инструмент «Разметка» и соответствующий тип разметки, как показано на рисунке 56.

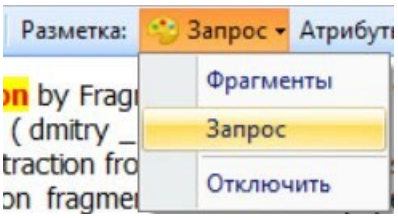


Рисунок 56 – Инструмент «Разметка»

- «Фрагменты» – разметка, которую назначает сам пользователь;

- «Запрос» – разметка, соответствующая запросу рубрики и отображающая, по какому признаку документ был отобран в данную рубрику;

- «Отключить» – отключает разметку.

При включении разметки на панели инструментов появляется дополнительный функционал для просмотра всех фрагментов выбранного документа, представленный на рисунке 57 и позволяющий переключаться с фрагмента на фрагмент с помощью кнопок «Предыдущий» и «Следующий», а также отображающий общее количество фрагментов в выбранном документе и номер текущего фрагмента.

Рисунок 57 – Инструменты для просмотра фрагментов текста

Для осуществления полнотекстового поиска по документам классификатора необходимо в поле «Поиск» (на панели инструментов) ввести интересующий запрос, далее нажать на кнопку «Найти» или «Enter», после чего откроется вкладка «панель работы с документами», где в списке документов будут только те документы, которые удовлетворяют запросу, как показано на рисунке 58.

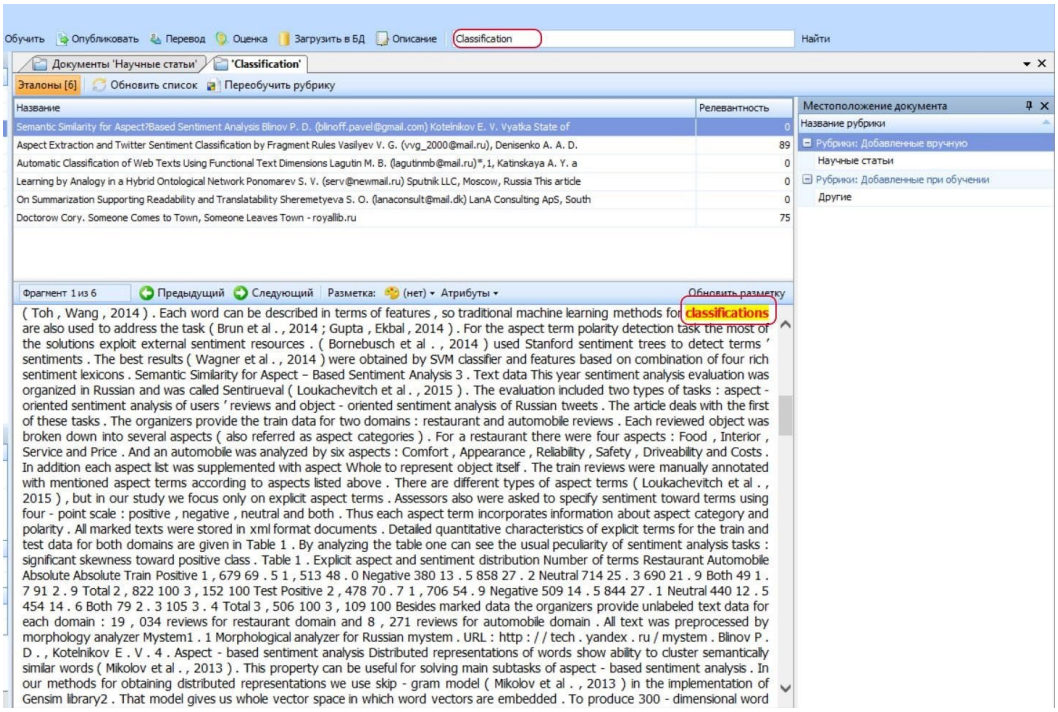


Рисунок 58 – Результат полнотекстового поиска по документам

Чтобы изменить набор рубрик для выделенного документа, необходимо вызвать контекстное меню рубрик и выбрать пункт «Задать рубрики…», после чего откроется окно «Местоположение документа», представленное на рисунке 59. В данном окне следует установить метки перед названиями рубрик, к которым относится документ.

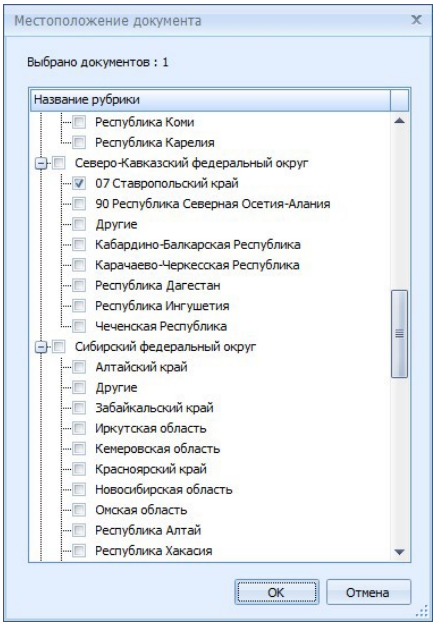


Рисунок 59 – Окно «Местоположение документа»

При работе с документами может возникнуть потребность в их сохранении для дальнейшего использования. Для этого в поле отображения документов необходимо выделить документ правой клавишей «мыши» и в открывшемся контекстном меню выбрать один из следующих пунктов:

- «Выгрузить документы…» - выгружает документ в том же самом формате, в котором он изначально был загружен в классификатор;
- «Выгрузить UIML-документы…» - выгружает документ в uiml-формате;
- «Выгрузить UIML-документы с атрибутами…» - выгружает документ в uimlформате с сохранением в нем атрибутов;
- «Выгрузить UIML-документы со структурами по запросу…» - выгружает документ в uiml-формате, записывая в получившийся файл информацию о структуре классификатора, рубриках.

После выбора одного из этих пунктов откроется окно «Обзор папок» для выбора каталога, в который будет сохранен необходимый документ. По

завершению процесса выгрузки документов отобразится окно «Информация» с подтверждением сохранения документов на диск.

Процесс выгрузки документов представлен на рисунке 60.

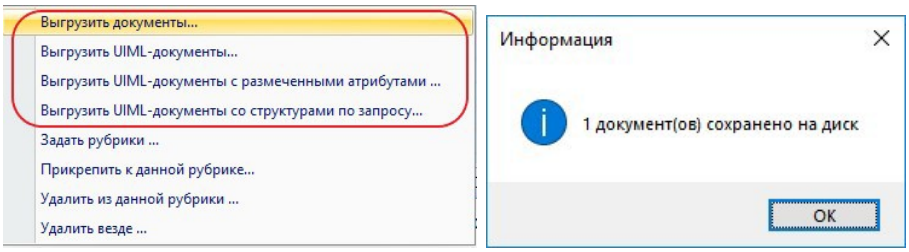


Рисунок 60 – Процесс выгрузки документов

2.6.2.11 Удаление документов

Для удаления всех документов классификатора или рубрики необходимо в контекстном меню выбрать «Эталоны | Удалить эталоны классификатора» или «Эталоны | Удалить эталоны рубрики», как показано на рисунке 61.

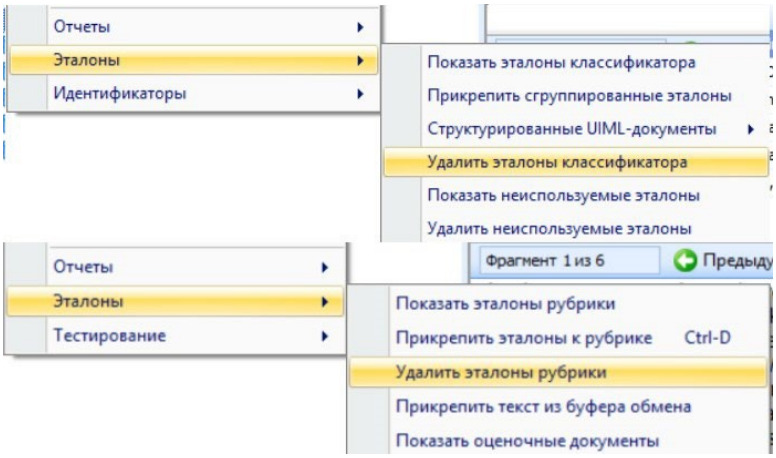


Рисунок 61- Удаление эталонов рубрик

Для удаления конкретного документа необходимо открыть панель работы с документами, далее правой клавишей «мыши» открыть контекстное меню документа и выбрать пункт «Удалить из данной рубрики» или «Удалить везде», выделенные на рисунке 62. В результате выбора одного из пунктов данный документ будет удален либо в заданной рубрике, либо во всем классификаторе.

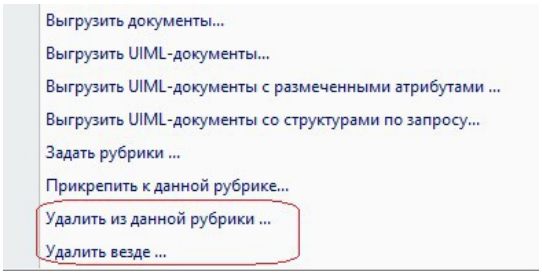


Рисунок 62 - Удаление из рубрик

2.6.2.12 Просмотр неиспользованных документов классификатора

Кроме основных документов классификатора (эталонов), в нем могут находиться и другие документы (неиспользованные).

Неиспользованные документы – это документы, которые были добавлены в папку «Etalons», но по какой-либо причине не были использованы (классифицированы).

Для их просмотра необходимо открыть контекстное меню классификатора и выбрать «Эталоны | Показать неиспользуемые эталоны», для удаления – «Эталоны | Удалить неиспользуемые эталоны». Оба пункта контекстного меню выделены на рисунке 63.

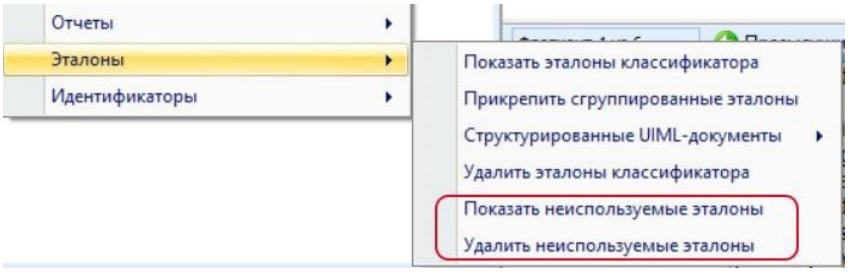


Рисунок 63 – Неиспользуемые эталоны

В результате выбора пункта «Показать неиспользуемые эталоны» должна открыться вкладка просмотра неиспользованных документов «Неиспользуемые документы», представленная на рисунке 64.

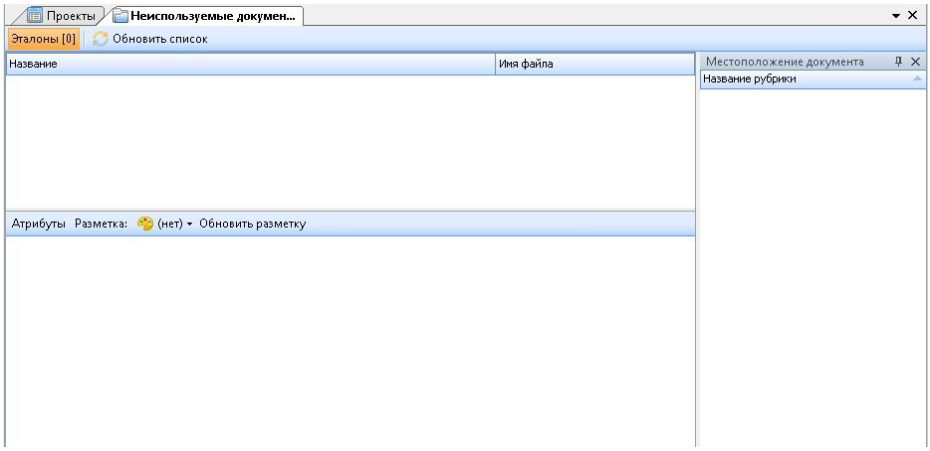


Рисунок 64 - Неиспользуемые документы

С неиспользованными документами можно совершать те же манипуляции, что и с эталонными документами, т.е. просматривать, удалять, выгружать и задавать рубрику.

2.6.2.13 Написание и редактирование запросов

Запрос – это определенное правило, по которому будет производиться классификация документов.

На данный момент существуют два варианта написания запросов – глобальный (правила, написанные в глобальном запросе, могут использоваться во всех рубриках классификатора) и локальный (правила, написанные для каждой рубрики отдельно), поэтому в программе предусмотрены:

- область для редактирования текстов глобального запроса;
- область для редактирования текстов запроса отдельной рубрики.

2.6.2.14 Создание и редактирование глобального запроса

Для написания или редактирования правил глобального запроса необходимо открыть область для редактирования глобального запроса, представленную на рисунке 65.

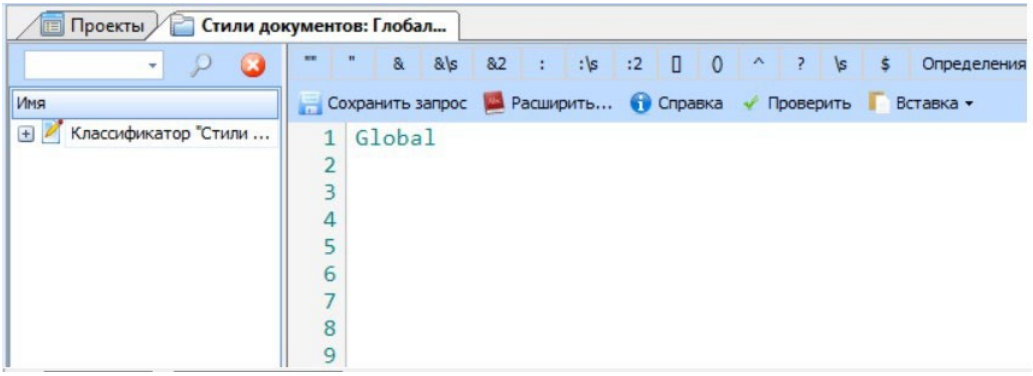


Рисунок 65 - Редактирование глобального запроса

Открыть данную область можно двумя способами:

- выделить название классификатора двойным нажатием левой клавишей «мыши»;
- выбрать функцию «Редактировать глобальное правило…», в контекстном меню классификатора, как показано на рисунке 66.

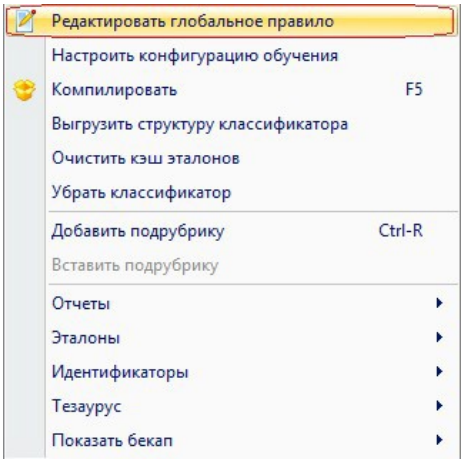


Рисунок 66 - Редактирование глобального правила

При написании глобального запроса в первую строку в обязательном порядке необходимо ввести слово «Global», в противном случае классификатор не будет корректно скомпилирован. Начиная с версии 2.3, слово «Global» формируется автоматически, в более ранних – его необходимо вводить вручную.

2.6.2.15 Создание и редактирование запросов отдельной рубрики

Для написания или редактирования правил отдельной рубрики необходимо открыть область для редактирования текста отдельной рубрики, представленную

на рисунке 67.

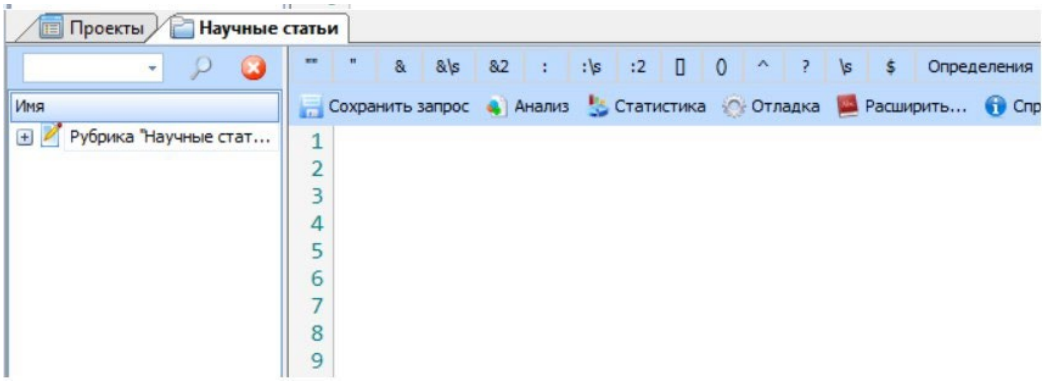


Рисунок 67 - Редактирование текста отдельной рубрики

Открыть данную область можно двумя способами:

- выделить название рубрики двойным нажатием левой клавишей «мыши»;
- в контекстном меню рубрики выбрать функцию «Редактировать правило», как показано на рисунке 68.

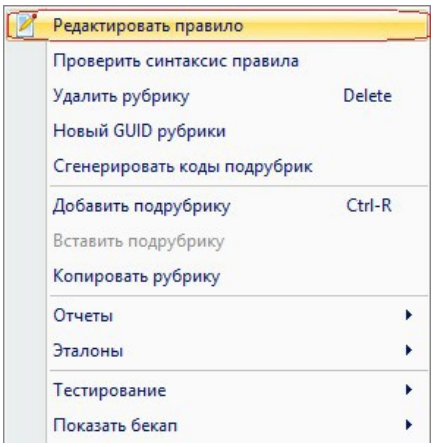


Рисунок 68 - Редактирование правил

2.6.2.16 Использование панели инструментов

Панель инструментов представлена на рисунке 69 и предназначена для упрощения работы по редактированию или созданию правил классификации. Она состоит из:

- набора зарезервированных элементов синтаксиса языка SCATQL;
- кнопки «Сохранить запрос» для сохранения изменений в запросе;
- инструментов для анализа и отладки правил – «Анализ», «Статистика», «Отладка» и «Оценка»;

- кнопки для дополнения правил классификации «Расширить»;

- кнопки для проверки на наличие синтаксических ошибок в запросе «Проверить»;

- кнопка «Тестирование» для доступа к тестированию классификаторов по базам данных и директориям;

- кнопки «Вставка» для упрощения вставки списков в текущий запрос;

- кнопка «xml-разбор» – преобразование запроса в xml-файл, содержащий дерево разбора согласно грамматике;

- кнопка «Регулярные выражения» – запрос преобразуется в регулярные выражения (если синтаксис запроса позволяет совершить эту операцию), на выходе получается текстовый файл с регулярными выражениями;

- раздела «Справка», который содержит краткую справку по языку запросов SCATQL в формате pdf.

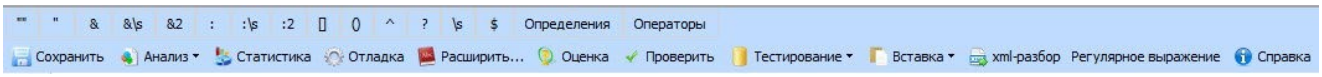


Рисунок 69 - Панель инструментов

При наведении на любой из элементов синтаксиса языка SCATQL отобразится его краткое описание в виде всплывающего сообщения, как показано на рисунке 70. Одиночное нажатие левой клавиши «мыши» по элементу добавляет его в заданное место (местоположение курсора), как показано на рисунке 71.

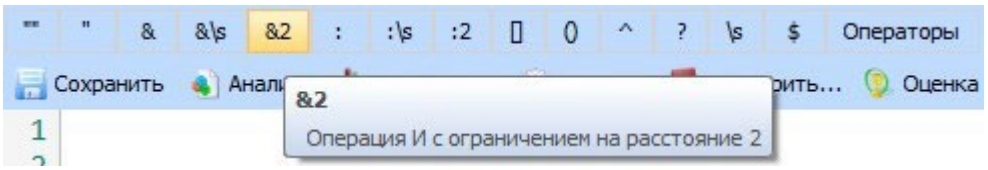


Рисунок 70

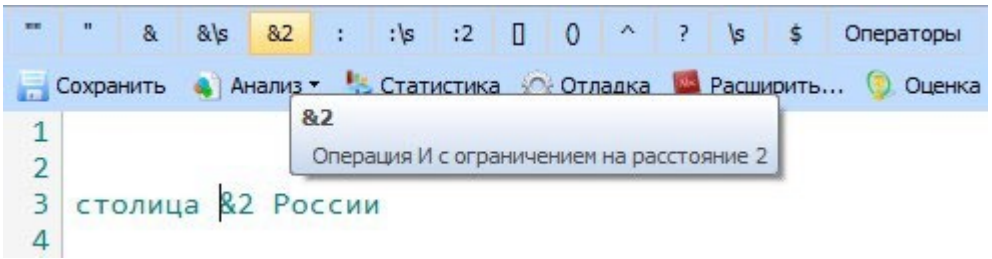


Рисунок 71

Меню «Операторы» содержит в себе список операторов языка SCATQL по категориям.

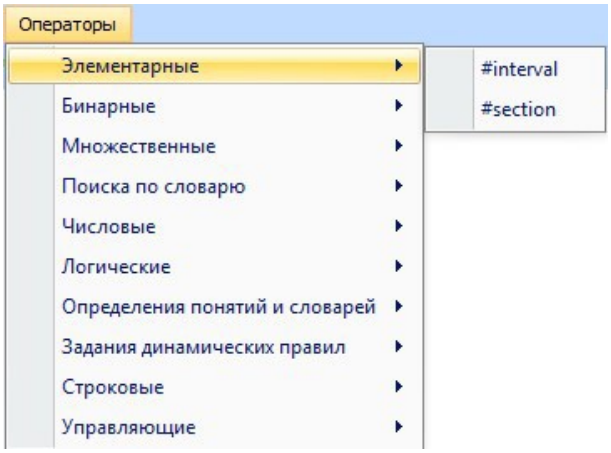


Рисунок 72 -Категории языка SCATQL

2.6.2.17 Проверка синтаксических ошибок в запросе

Для проверки на наличие ошибок в конкретной рубрике необходимо в контекстном меню рубрики выбрать функцию «Проверить синтаксис правила» или на панели инструментов нажать на кнопку «Проверить» (кнопка «Проверить» работает и в глобальном запросе). Оба инструмента выделены на рисунке 73.

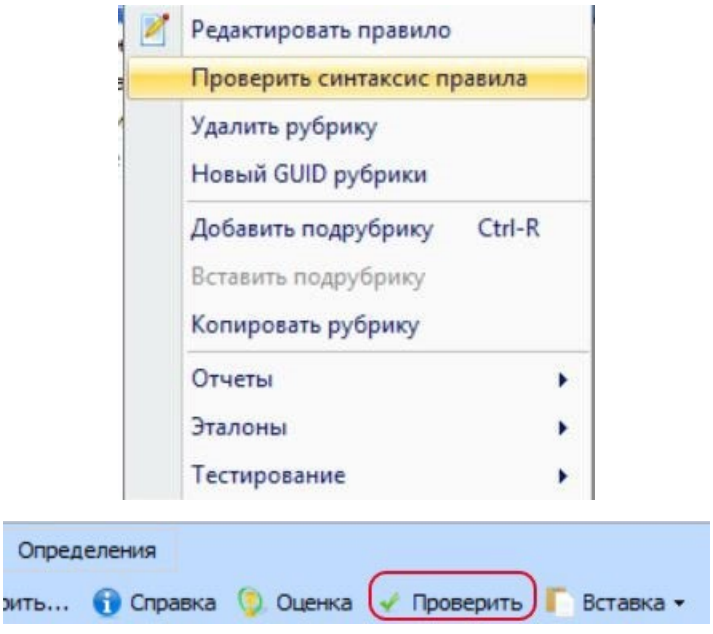


Рисунок 73 – Функция для проверки на наличие ошибок в конкретной рубрике

При наличии ошибок отобразится окно «Список ошибок» с их описанием и местоположением, как показано на рисунке 74.

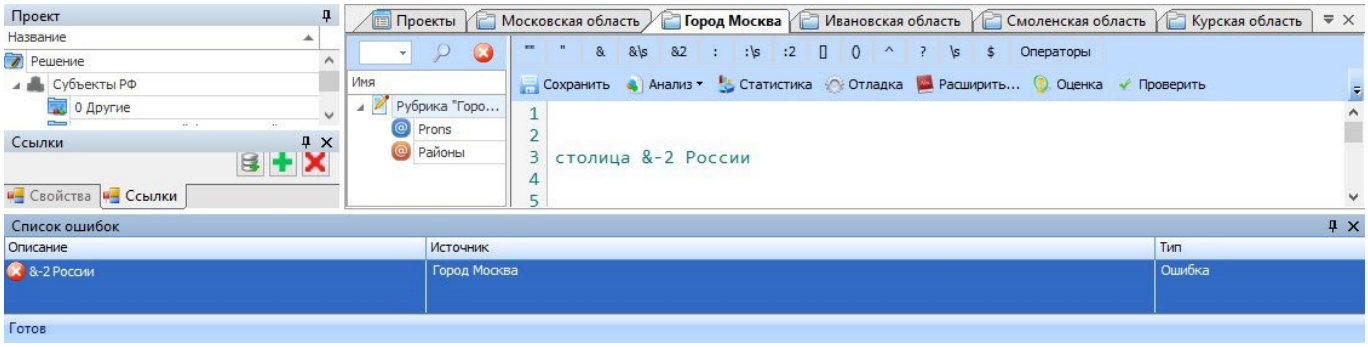


Рисунок 74 – Отображение ошибок

2.6.2.18 Создание и редактирование пространства имен

Пространство имен (переменных) представляет собой систему взаимосвязанных понятий, необходимых для создания правил классификации.

Операция определения понятия (#define), пример применения которой представлен на рисунке 75, позволяет задать определение (понятие), которое при компиляции (обучении) классификатора будет подставляться вместо ссылки на него.

Использование данного оператора позволяет организовать формирование правил классификации путем первоначального создания системы понятий, а затем задания различных логических и контекстуальных условий на встречаемость их в тексте.

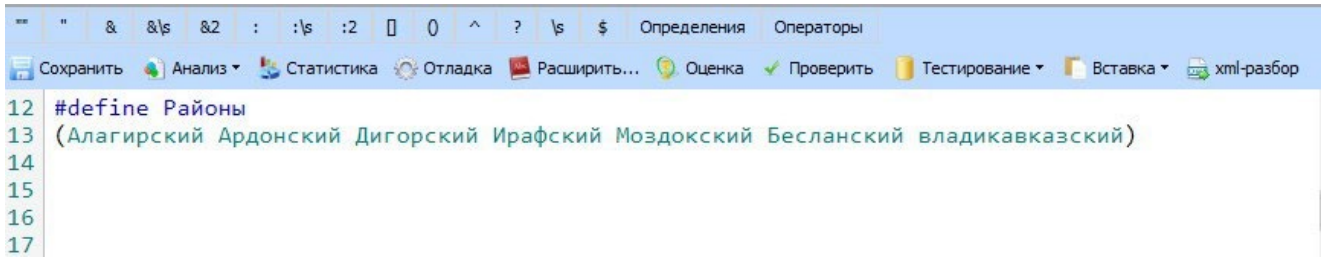


Рисунок 75 – Пример задания понятия «#define»

Операция задания набора фрагментов (#set), пример применения которой представлен на рисунке 76, позволяет при задании нескольких ссылок на одно определение при компиляции правила многократно повторить его содержимое. Это может привести к существенному возрастанию размера правил и, соответственно, замедлению скорости их выполнения при классификации. Операция задания

набора фрагментов в отдельной переменной, при обращении к которой вычисления уже не будут делаться повторно.

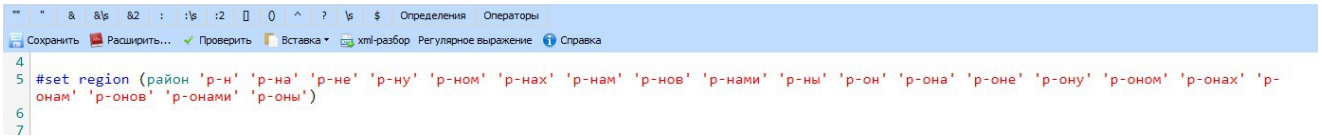


Рисунок 76 – Пример задания набора фрагментов (переменной) «#set»

Необходимо отметить, что в виде набора фрагментов можно задать только полностью определенные понятия, абстрактные понятия задавать стоит лишь при помощи #define, т.к. множество фрагментов, которые им соответствуют, зависят от контекста их использования.

Для обращения к ранее определенному пространству имен существует специальный оператор – ссылка (для #define обращение к переменой происходит при помощи @, а для #set – @@). Пример применения ссылок представлен на рисунках 77 и 780 .

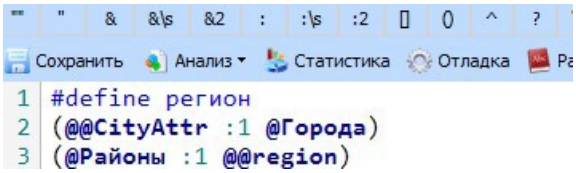


Рисунок 77 – Пример использования и вызова определения типа «define»

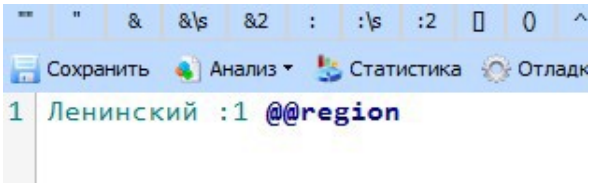


Рисунок 78 – Пример использования и вызова переменной типа «set»

2.6.2.19 Поиск понятий

Для осуществления поиска понятий (переменных) необходимо выполнить следующие действия:

- выделить двойным нажатием левой клавишей «мыши» название классификатора или рубрики (откроется поле для редактирования правил классификации);

- в поле поиска, находящемся над перечнем (списком) переменных, необходимо ввести интересующее понятие (название переменной);

- нажать на клавишу «Enter» или на кнопку с графическим обозначением поиска.

В результате должен отобразиться список переменных, в название которых входит искомое понятие, как показано на рисунке 79.

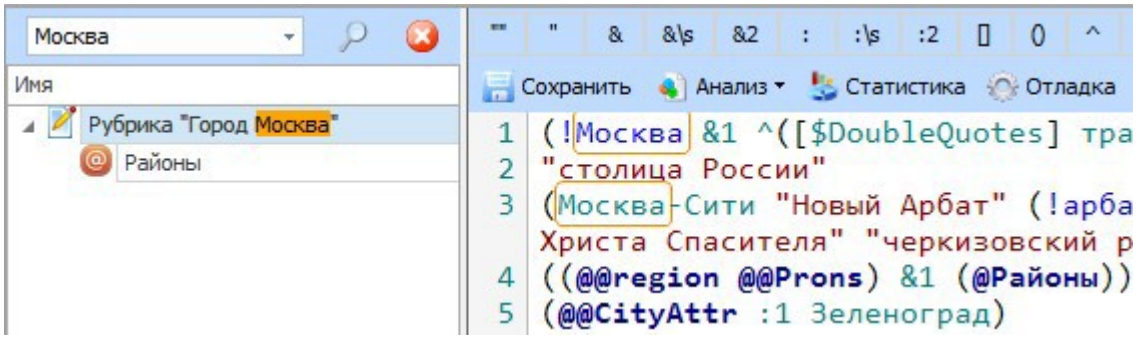


Рисунок 79 – Результат поиска понятий

2.6.2.20 Функция «Расширить»

Функция «Расширить» позволяет расширить содержимое переменной синонимами и словами, подходящими по смыслу к выделенному слову.

В области написания и редактирования запроса необходимо выделить слово, к которому будут подбираться синонимы, и слова (из тезаурусов), подходящие по смыслу, после чего нажать на кнопку «Расширить» на панели инструментов. В открывшемся окне «Найти по тезаурусу» пользователь может отметить подходящие ему слова и нажать на кнопку «Расширить», в результате выбранные слова добавятся в запрос. Процесс расширения переменной синонимами и словами, подходящими по смыслу, представлен на рисунке 80.

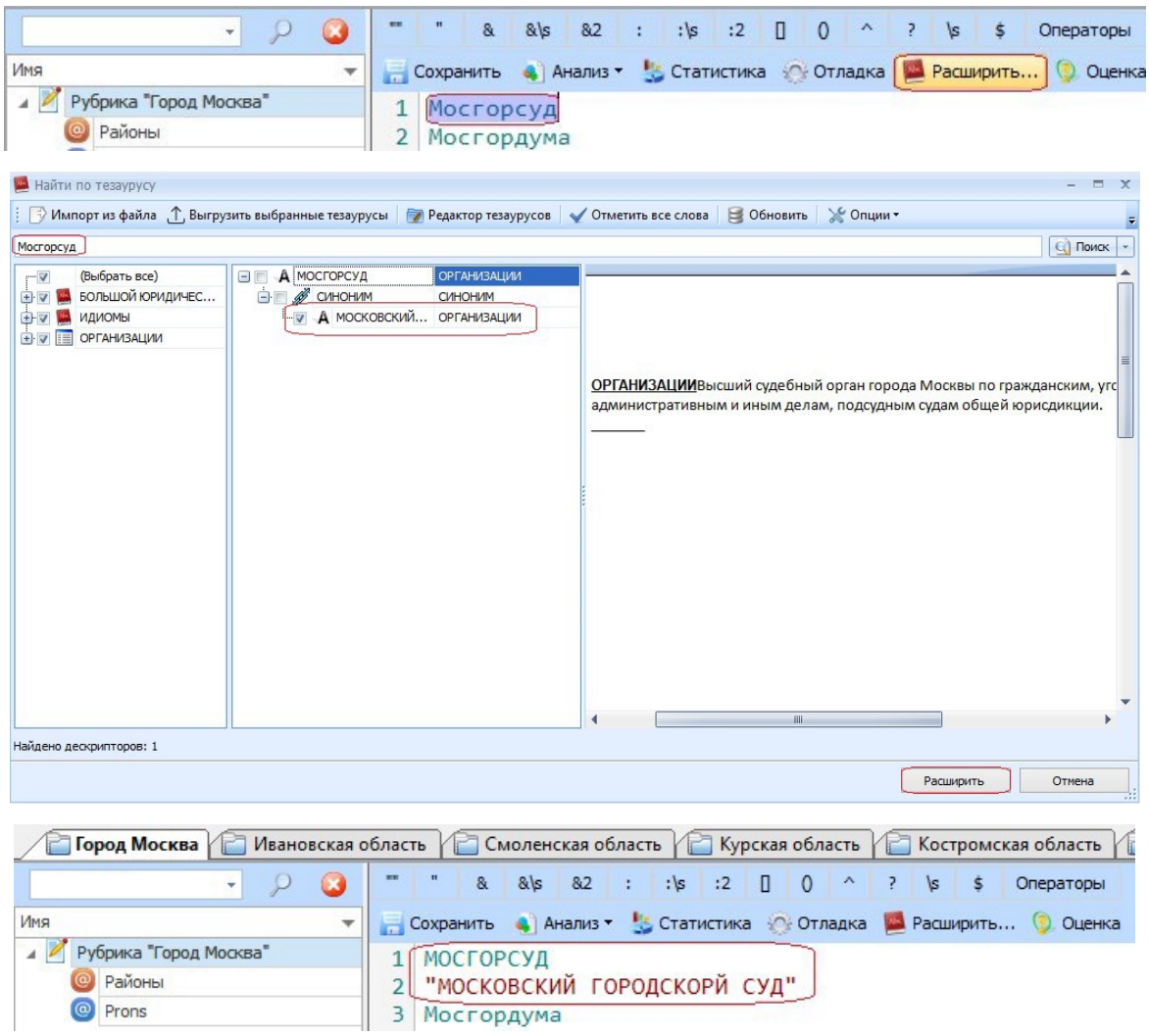


Рисунок 80 – Расширение переменной синонимами и словами подходящими по смыслу

2.6.2.21 Перевод правил классификации на другие языки

Для создания многоязычных классификаторов в программе «Студия управления знаниями» предусмотрена функция «Перевод», которая осуществляет перевод правил классификации и создает копию исходного классификатора на другом языке.

Для перевода классификатора на другой язык необходимо на панели инструментов выбрать инструмент «Перевод», выделенный на рисунке 81.

Рисунок 81 – Инструмент «Перевод»

После этого откроется окно «Создать новый проект классификатора», представленное на рисунке 82. В данном окне задаются название нового классификатора, комментарий и путь к нему, а также исходный язык и язык для перевода.

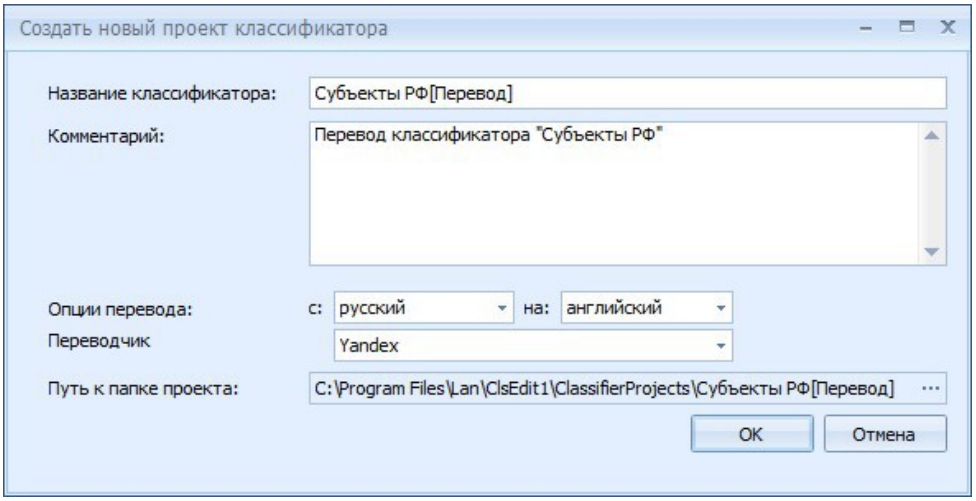


Рисунок 82 – Окно «Создать новый проект классификатора»

Список доступных языков представлен на рисунке 83.

русский	английский	азербайджанский	албанский	арабский	армянский
африкаанс	баскский	белорусский	бенгальский	бирманский	болгарский
боснийский	валлийский	венгерский	вьетнамский	галисийский	греческий
грузинский	гуджарати	датский	зулу	иврит	игбо
идиш	индонезийский	ирландский	исландский	испанский	итальянский
йоруба	казахский	каннада	каталанский	китайский	корейский
креольский	кхмерский	лаосский	латынь	латышский	литовский
македонский	малагасийский	малайский	малайялам	мальтийский	маори
маратхи	монгольский	немецкий	непали	нидерландский	норвежский
панджаби	персидский	польский	португальский	румынский	себуанский
сербский	сесото	сингальский	словацкий	словенский	сомали
суахили	суданский	тагальский	таджикский	тайский	тамильский
телугу	турецкий	узбекский	украинский	урду	финский
французский	хауса	хинди	хмонг	хорватский	чева
чешский	шведский	эсперанто	эстонский	яванский	японский

Рисунок 83 – Список языков

В результате будет создана копия исходного классификатора на выбранном языке для перевода.

2.6.2.22 Анализ и отладка правил классификации

Анализ и отладка правил классификации осуществляются при помощи следующих функций: «Анализ», «Статистика», «Оценка» и «Отладка».

Функция «Анализ» предназначена для выполнения анализа влияния

различных элементов выделенного выражения на итоговые показатели качества классификатора.

С использованием данной функции можно решать следующие задачи:

- определение подвыражений, которые отрицательно влияют на качество классификации и должны быть удалены из текста классификатора;
- определение подвыражений, которые не оказывают влияния на качество классификации и могут быть удалены для сокращения размера запроса;
- определение ограничений или логических условий, которые ухудшают качество классификации и должны быть ослаблены или удалены.

Функция «Статистика» предназначена для вывода статистики по встречаемости элементов текста (слов) в документах классификатора.

Функция «Оценка» предназначена для оценки качества написанных правил классификации.

Функция «Отладка» предназначена для отладки правил классификации отдельной рубрики.

2.6.2.22.1 Функция «Анализ»

Для начала анализа правил классификации необходимо открыть поле редактирования запроса конкретной рубрики и на панели инструментов выбрать «Анализ», как показано на рисунке 84.

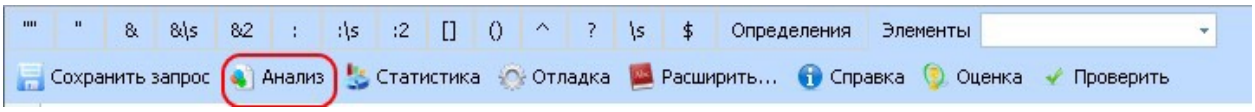


Рисунок 84 – Инструмент «Анализ»

После завершения процесса анализа запроса рубрики отобразится отчет, содержащий следующие данные exF, P, R, F, детальное описание которых представлено в таблице 1.

Таблица 1. Характеристики элементов выражения при выполнении анализа

Сокращение	Название	Комментарий
exF	Разница между значением F-меры для всего запроса с исключенным текущим выражением и значением F-меры всего запроса без исключения текущего выражения	Если exF принимает отрицательное значение, то это говорит о том, что исключение текущего выражения из запроса приводит к ухудшению качества классификации. Если exF принимает положительное значение, то при удалении текущего выражения качество классификации будет улучшено. Если exF принимает нулевое значение, то текущее выражение может быть удалено из запроса.
P	Точность классификации - это отношение числа документов, которые правильно попали в рубрику в результате автоматической классификации, к общему числу документов, которые попали в данную рубрику.	Чем больше значение точности, тем лучше, но при этом, если данному подвыражению соответствует мало документов, то оно может быть удалено, так как является слишком специфичным.
R	Полнота классификации - это отношение числа документов, которые	Чем больше значение полноты, тем лучше, но при этом, если у данного выражения маленькое значение

Сокращение	Название	Комментарий
	правильно попали в рубрику в результате автоматической классификации, к общему числу документов, которые были отнесены к рубрике экспертом (эталонам).	точности P, то оно должно уточняться с помощью дополнительных условий, которые накладывают ограничение на контекст его использования.
F	F-мера – это среднее гармоническое показателей точности и полноты.	Чем больше значение данного показателя, тем лучше. При анализе отчета на его значение необходимо обращать внимание в первую очередь.

Пример запроса рубрики и его анализа представлен на рисунке 85.

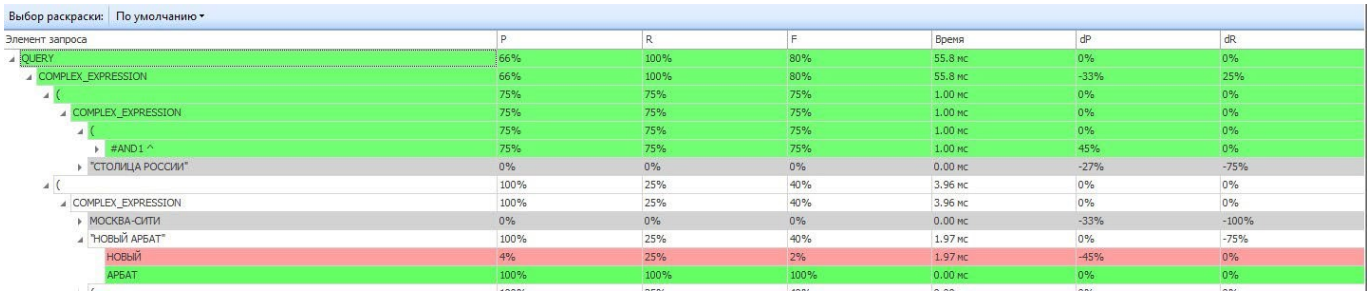


Рисунок 85 – Результат анализа правил классификации

Процесс анализа может быть достаточно длительным при большом числе документов или при сложной структуре выражения.

2.6.2.22.2 Функция «Статистика»

Функция «Статистика», представленная на рисунке 86, производит разбиение текста документов рубрики на отдельные элементы (слова) и отображает:

- «Точность» – отношение числа документов, которые правильно попали в

рубрику в результате автоматической классификации, к общему числу документов, которые попали в данную рубрику;

- «Полнота» – отношение числа документов, которые правильно попали в рубрику в результате автоматической классификации, к общему числу документов, которые были отнесены к рубрике экспертом (эталонам);

- «F-мера» – среднее гармоническое показателей точности и полноты;

- «Хи-квадрат», «прирост информации», «степень прироста информации» – это функции полезности термина относительно, обрабатываемого массива документов в зависимости от используемого метода анализа информации; - частота встречаемости слова в тексте.

После завершения статистического анализа документов, термины, которые наиболее актуальны для выбранной рубрики, можно перенести в запрос, нажав правой кнопкой «мыши» по термину и выбрав «Копировать в запрос».

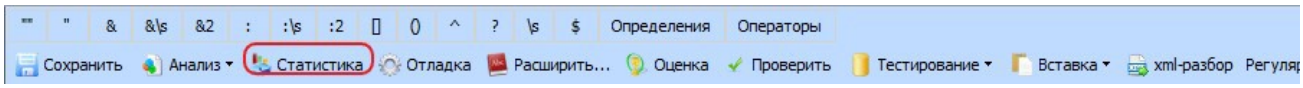


Рисунок 86 – Инструмент «Статистика»

После выбора элемента «Статистика» открывается окно статистики с названием текущей рубрики. В данном окне располагаются инструменты, представленные на рисунке 87:

- поле для ввода запроса, из которого нажатием на символ многоточия в правой части поля вызывается форма «Запрос»;

- функция «Выполнить» (в оригинальной или в нормальной форме слова, выделено на рисунке 87).

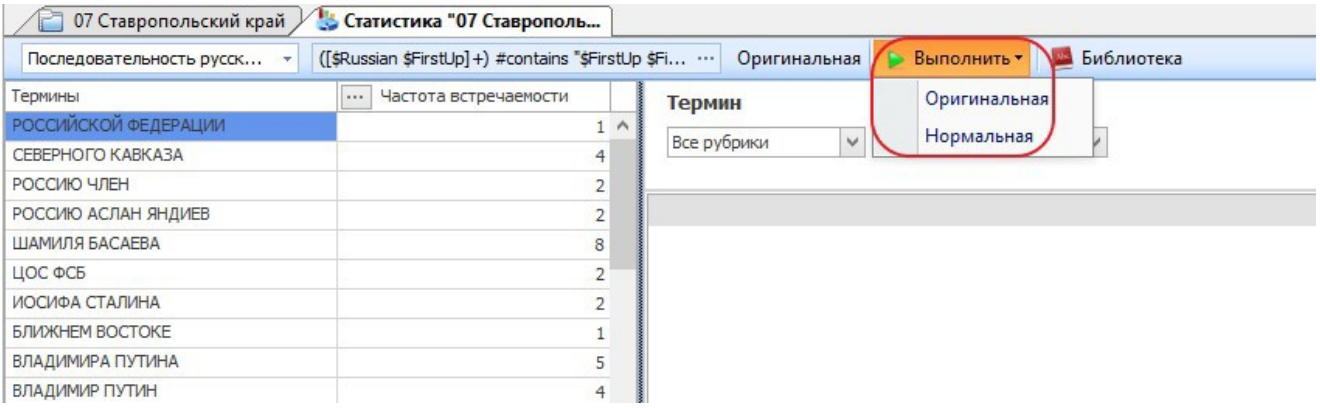


Рисунок 87 – Панель инструментов окна статистики

Форма «Запрос» позволяет отфильтровать статистику по необходимым параметрам. Например, при вводе в поле запроса «$Noun» и последующем выборе необходимой формы будут отображаться только существительные. Запрос можно как составить самостоятельно, так и выбрать из библиотеки запросов. Пример окна формы представлен на рисунке 88.

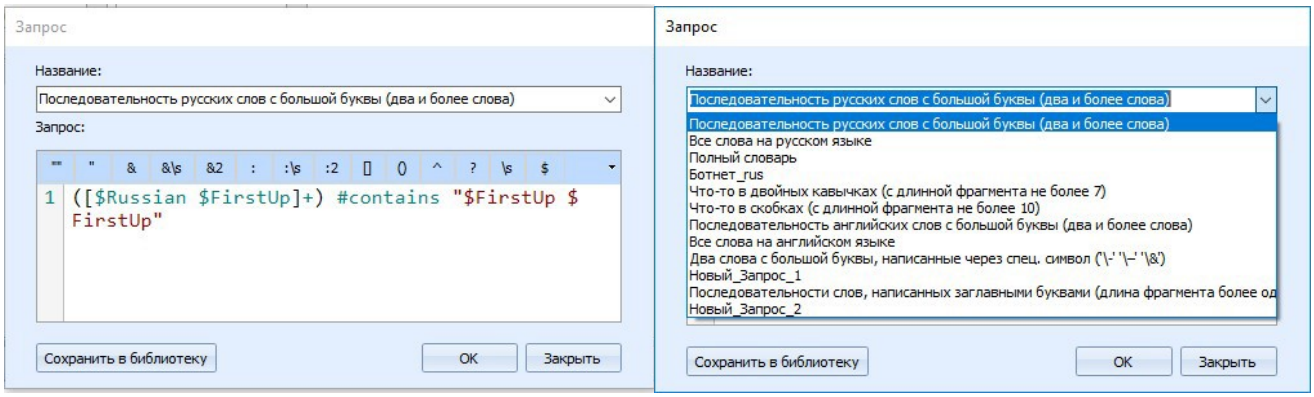


Рисунок 88 – Форма «Запрос»

Различие нормальной и оригинальной формы заключается в том, что в поле нормальной формы отображается инфинитив слов, например, существительные, встречающиеся в тексте, будут отображаться в нормальной форме (в ед.ч., Им.п.), а в поле оригинальной формы слова отображаются в том виде, в котором они встречались в тексте. Пример отображения статистики представлен на рисунке 89.

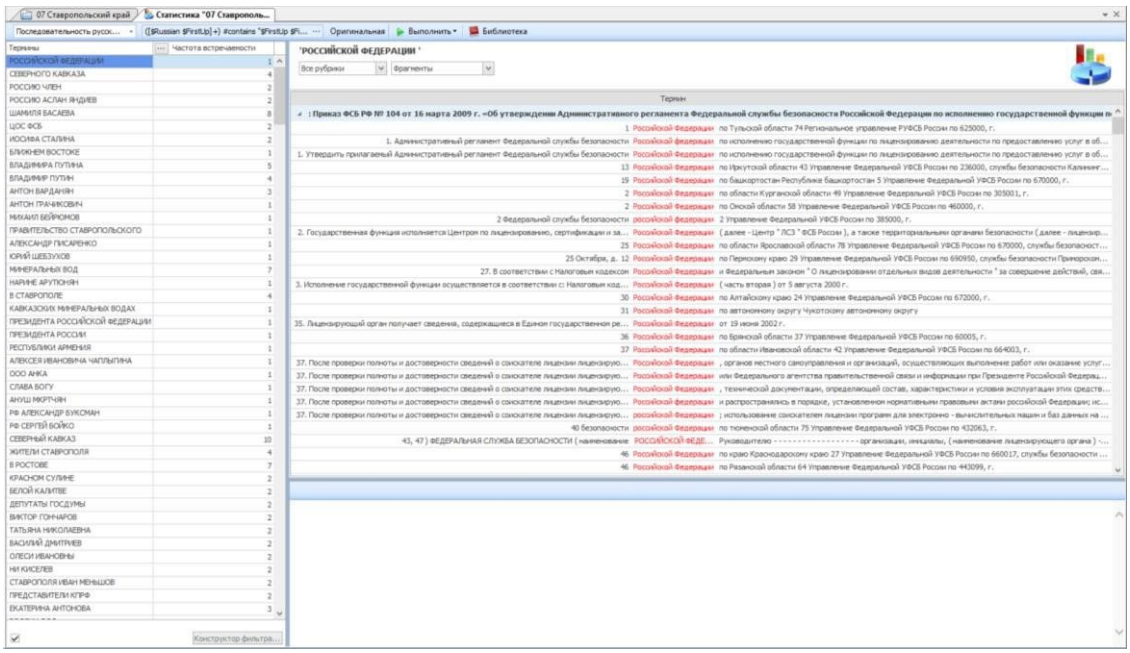


Рисунок 89 – Пример отображения статистики

2.6.2.22.3 Функция «Оценка»

Функция «Оценка» позволяет оценить качество написанных правил классификации для отдельной рубрики или классификатора в целом по документам классификатора или оценочному каталогу соответственно.

Для оценки качества правил отдельной рубрики необходимо открыть поле редактирования запроса рубрики, после чего нажать на кнопку «Оценка», как показано на рисунке 90.

Рисунок 90 – Инструмент «Оценка» для рубрики

В результате откроется окно с отчетом об оценке качества правил рубрики, представленное на рисунке 91.

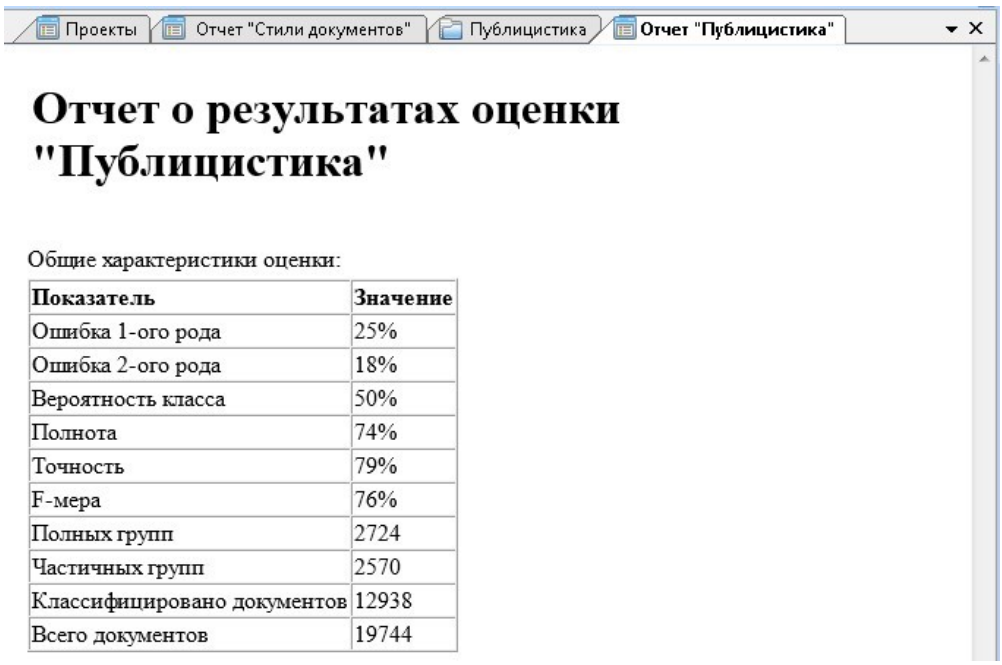


Рисунок 91 – Отчет о результатах оценки рубрики

Характеристики оценки имеют следующие значения:

- «Ошибка 1-ого и 2-ого рода» – показатели качества, выражаемые через отношение различных групп документов друг к другу;

- «Вероятность класса» - оценка того, на сколько правильно данный кластер (группа документов) отнесен к оцениваемому правилу (запросу);

- «Полнота» (R) – отношение числа документов, которые правильно попали в рубрику в результате автоматической классификации, к общему числу документов, которые были отнесены к рубрике экспертом (эталонам);

- «Точность» (P) – отношение числа документов, которые правильно попали в рубрику в результате автоматической классификации, к общему числу документов, которые попали в данную рубрику;

- «F-мера» – среднее гармоническое показателей точности и полноты;

- «Полных групп (кол-во)» – количество кластеров (групп) документов, в которых все документы удовлетворяют запросу;

- «Частичных групп (кол-во)» – количество кластеров (групп) документов, в которых только часть документов удовлетворяют запросу;

- «Классифицировано документов» – сколько всего было классифицировано документов;

- «Всего документов» – общее число документов.

Оценка правил классификации отдельной рубрики осуществляется по эталонным документам всего классификатора.

Для оценки качества правил классификации всего классификара в целом необходимо выбрать функцию «Оценка» на панели инструментов, как показано на рисунке 92.

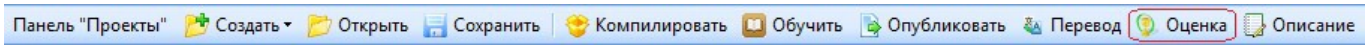


Рисунок 92 – Функция «Оценка» для классификатора

В результате откроется отчет об оценке правил классификации всего классификатора в целом, представленный на рисунке 93.

Общие характеристики оценки классификатора "Субъекты РФ"

Таблица 1. Общие качественные характеристики оценки классификатора на внешней подборке документов.

Название рубрики	Ошибка 1-ого рода	Ошибка 2-ого рода	Вероятность класса	F-мера	Полнота	Точность	Полных групп	Частичных групп	Классифицировано документов	Всего документов
0 Другие	0%	0%	0%	100%	100%	100%	0	0	0	0
КФО Крымский федеральный округ	0%	0%	0%	66%	99%	49%	4	10	25	198
35 Республика Крым	14%	0%	5%	85%	85%	86%	366	192	1452	2904
67 Севастополь	14%	0%	1%	80%	85%	75%	77	53	305	789
ПФО Приволжский федеральный округ	0%	0%	0%	85%	99%	74%	2	3	7	18
22 Нижегородская область	15%	0%	0%	80%	84%	77%	41	39	123	459
33 Кировская область	20%	0%	0%	80%	79%	83%	16	16	55	168
36 Самарская область	8%	0%	0%	80%	91%	72%	42	34	134	496
53 Оренбургская область	4%	0%	0%	64%	95%	48%	24	68	151	473
56 Пензенская область	4%	0%	0%	95%	95%	96%	14	3	37	41
57 Пермский край	11%	0%	0%	86%	88%	84%	23	14	59	106
63 Саратовская область	8%	0%	0%	87%	91%	83%	14	9	43	63
73 Ульяновская область	6%	0%	0%	91%	93%	89%	31	12	88	233
80 Республика Башкортостан	4%	0%	0%	87%	95%	80%	35	21	103	313
88 Республика Марий Эл	0%	0%	0%	88%	99%	79%	7	3	20	29
89 Республика Мордовия	7%	0%	0%	73%	92%	61%	11	13	41	110
92 Республика Татарстан (Татарстан)	13%	0%	0%	81%	86%	76%	48	47	171	597
94 Удмуртская Республика	2%	0%	0%	95%	97%	93%	13	4	40	106
97 Чувашская Республика - Чувашия	0%	0%	0%	98%	99%	96%	12	1	27	28
Другие	0%	0%	0%	100%	100%	100%	0	0	0	0
СФО Сибирский федеральный округ	0%	0%	0%	74%	99%	58%	1	4	7	38
01 Алтайский край	14%	0%	0%	84%	85%	83%	12	10	46	100
04 Красноярский край	6%	0%	0%	89%	93%	86%	26	16	76	136
25 Иркутская область	9%	0%	0%	86%	90%	83%	20	16	72	131
32 Кемеровская область	6%	0%	0%	89%	93%	86%	23	13	63	85
50 Новосибирская область	22%	0%	0%	73%	77%	68%	48	89	230	559
52 Омская область	19%	0%	0%	85%	80%	90%	19	11	47	98
69 Томская область	0%	0%	0%	90%	99%	82%	11	5	28	50
76 Забайкальский край	26%	0%	0%	83%	73%	96%	15	5	45	61
81 Республика Бурятия	0%	0%	0%	88%	99%	79%	10	5	32	48
84 Республика Алтай	15%	0%	0%	85%	84%	86%	14	9	38	101
93 Республика Тыва	4%	0%	0%	92%	95%	89%	9	5	33	50
95 Республика Хакасия	7%	0%	0%	93%	92%	95%	33	7	73	100

Рисунок 93 – Отчет об оценке правил классификации всего классификатора

Оценка правил классификации всего классификатора в целом осуществляется по документам из оценочного каталога, который настраивается в меню «Сервис | Настройки | Общие».

2.6.2.22.4 Функция «Отладка»

Для начала процесса отладки необходимо открыть поле редактирования запроса конкретной рубрики и на панели инструментов выбрать «Отладка», как показано на рисунке 94.

Рисунок 94 – Инструмент «Отладка»

В результате выполнения данного действия откроется окно отладки, представленное на рисунке 95.

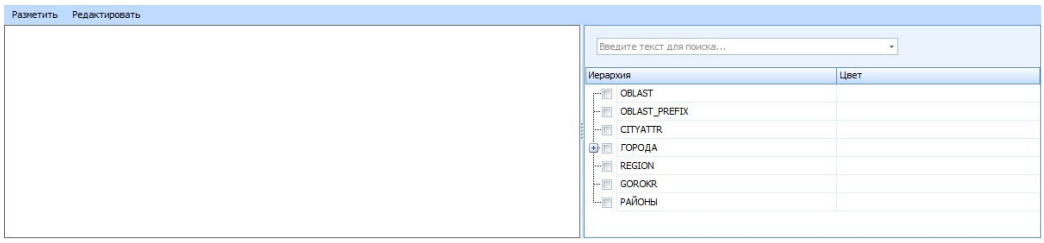


Рисунок 95 – Окно отладки

Окно отладки состоит из следующих элементов:

- поле ввода текста – в данное поле вводится текст, по которому будут производиться тестирование и отладка правил;
- поле иерархии переменных – в данном поле отображается иерархия переменных, содержащихся в правиле классификации, конкретной рубрике, в данном окне можно выбрать (отметить) переменные для отладки (каждая переменная имеет свой цвет);
- «Разметить» – выделяет в введенном тексте содержимое выбранных переменных;
- «Редактировать» – позволяет изменять ранее введенный текст.

Пример использования инструмента «Отладка» представлен на рисунке 96.

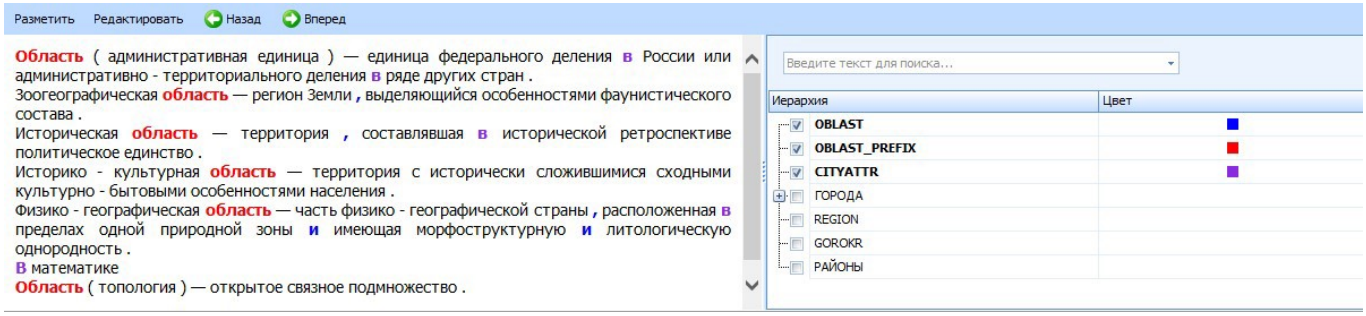


Рисунок 96 – Пример использования инструмента «Отладка»

Функция «Отладка» позволяет наглядно оценить качество написанных ранее правил классификации и словарей понятий и определить необходимость их редактирования.

2.6.2.23 Компиляция и обучение классификатора

Компиляция классификатора – это процесс, заключающийся в компиляции правил классификации, в проверке на наличие синтаксических ошибок в правилах, а также в оценке качества их работы.

Обучение классификатора – это процесс автоматической генерации правил классификации путем применения специальных элементов языка запросов, располагающихся на панели инструментов «Обучение».

2.6.2.23.1 Компиляция проекта

Запустить процесс компиляции проекта можно следующими способами:

- в контекстном меню классификатора выбрать «Компилировать F5»;
- в меню «Классификаторы» выбрать «Компилировать F5»;
- на панели инструментов выбрать «Компилировать»; - нажать «горячую» клавишу «F5».

После запуска процесса компиляции откроется окно «Компиляция проекта», представленное на рисунке 97 и отображающее детализированное описание хода выполнения процесса.

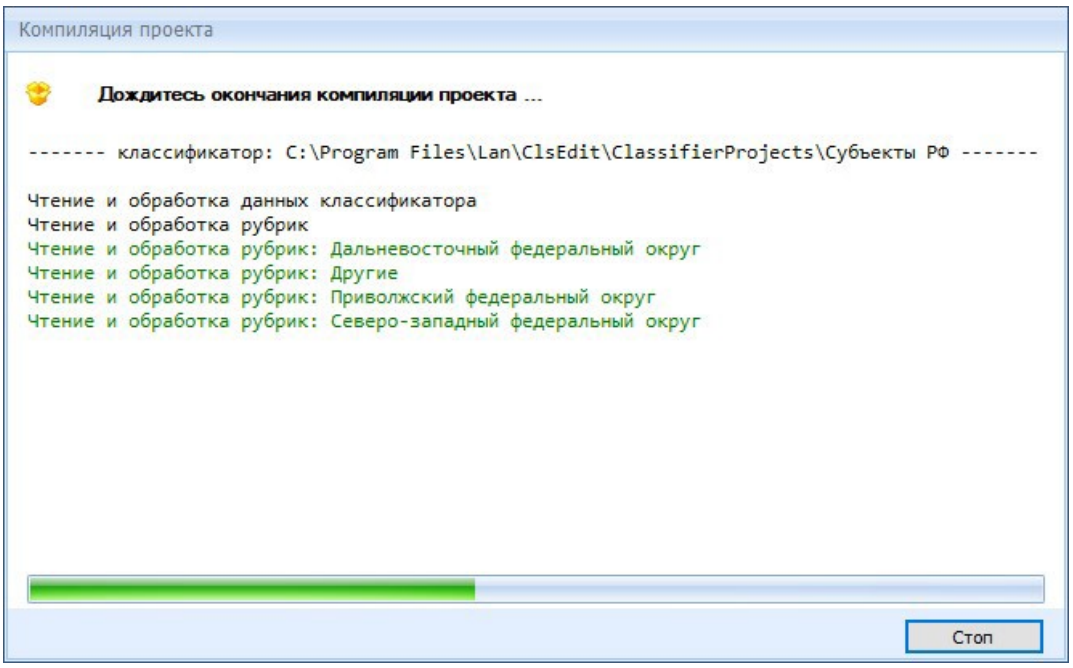
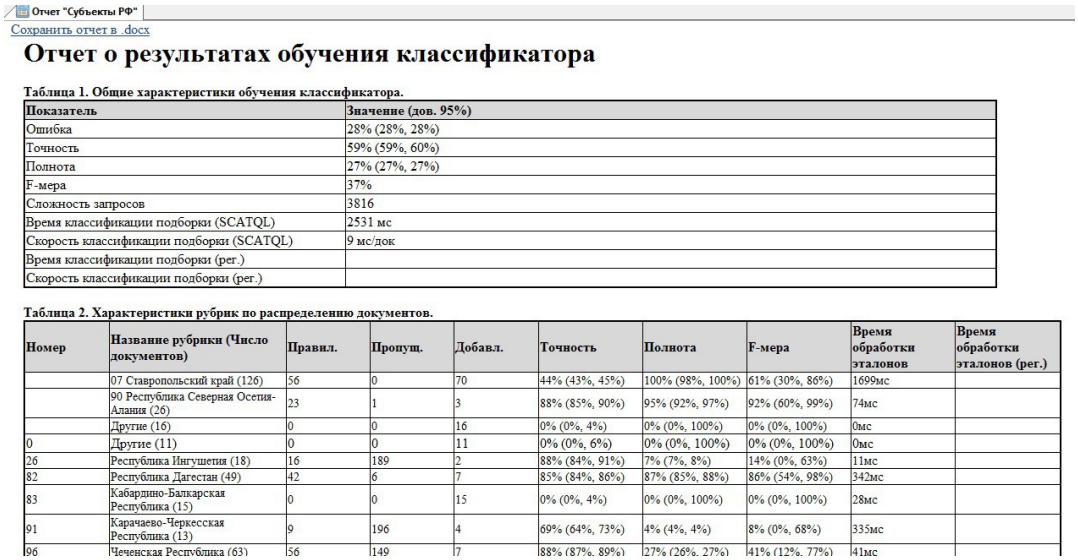


Рисунок 97 – Окно процесса компиляции

При успешном завершении компиляции откроется вкладка «Отчет о результатах обучения классификатора», представленный на рисунке 98 и состоящий из пунктов «Общие характеристики обучения классификатора» и «Основные характеристики рубрик».



Отчет "Субъекты РФ"

Сохранить отчет в .docx

Отчет о результатах обучения классификатора

Таблица 1. Общие характеристики обучения классификатора.

Показатель	Значение (дов. 95%)
Ошибка	28% (28%, 28%)
Точность	59% (59%, 60%)
Полнота	27% (27%, 27%)
F-мера	37%
Сложность запросов	3816
Время классификации подборки (SCATQL)	2531 мс
Скорость классификации подборки (SCATQL)	9 мс/док
Время классификации подборки (рег.)	
Скорость классификации подборки (рег.)	

Таблица 2. Характеристики рубрик по распределению документов.

Номер	Название рубрики (Число документов)	Правил.	Пропущ.	Добавл.	Точность	Полнота	F-мера	Время обработки эталонов	Время обработки эталонов (рег.)
	07 Ставропольский край (126)	56	0	70	44% (43%, 45%)	100% (98%, 100%)	61% (30%, 86%)	1699мс	
	90 Республика Северная Осетия-Алания (26)	23	1	3	88% (85%, 90%)	95% (92%, 97%)	92% (60%, 99%)	74мс	
	Другие (16)	0	0	16	0% (0%, 4%)	0% (0%, 100%)	0% (0%, 100%)	0мс	
0	Другие (11)	0	0	11	0% (0%, 6%)	0% (0%, 100%)	0% (0%, 100%)	0мс	
26	Республика Ингушетия (18)	16	189	2	88% (84%, 91%)	7% (7%, 8%)	14% (0%, 63%)	11мс	
82	Республика Дагестан (49)	42	6	7	85% (84%, 86%)	87% (85%, 88%)	86% (54%, 98%)	342мс	
83	Кабардино-Балкарская Республика (15)	0	0	15	0% (0%, 4%)	0% (0%, 100%)	0% (0%, 100%)	28мс	
91	Карачаево-Черкесская Республика (13)	9	196	4	69% (64%, 73%)	4% (4%, 4%)	8% (0%, 68%)	335мс	
96	Чеченская Республика (63)	56	149	7	88% (87%, 89%)	27% (26%, 27%)	41% (12%, 77%)	41мс	

Рисунок 98 – Отчет о результатах обучения классификатора

2.6.2.23.2 Обучение проекта

При обучении классификатора на основе примеров документов (эталонов) сами правила классификации формируются автоматически, однако существенное

влияние на итоговый результат оказывают следующие факторы:

- выбранный режим обучения классификатора;
- количество и качество подобранных эталонных документов;
- наличие дополнительных правил классификации в виде запросов на специальном языке.

Выбор режима обучения классификатора и настройка дополнительных параметров осуществляется с помощью конфигурационного файла представленный на рисунке 99.

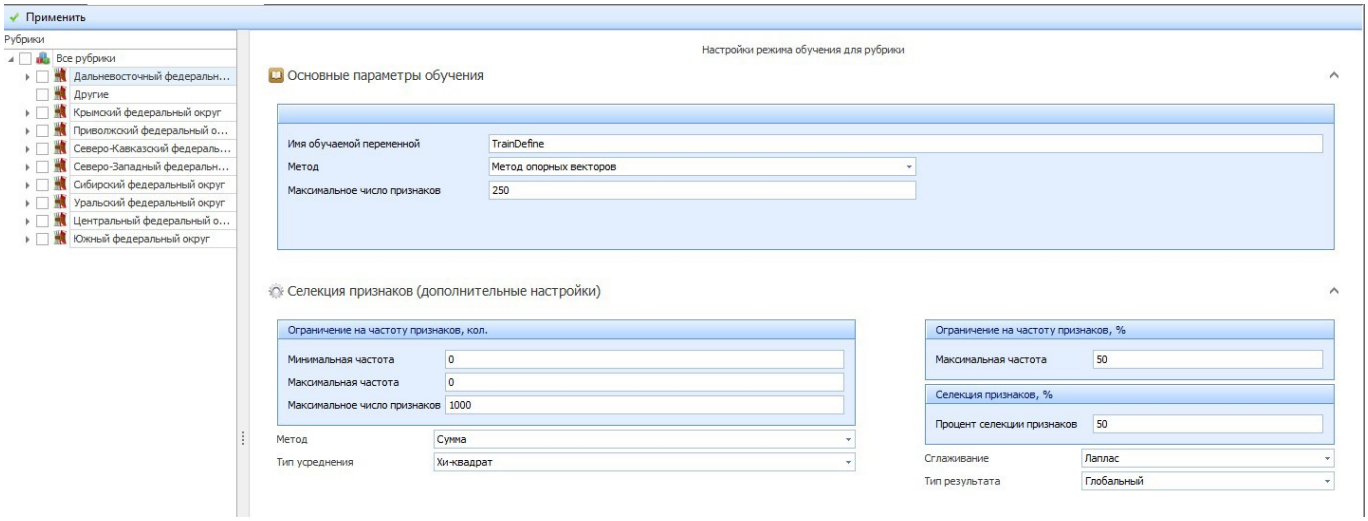


Рисунок 99 – Конфигурационный файл обучения рубрик классификатора

Для открытия окна настройки конфигурационного файла обучения рубрик классификатора необходимо кликнуть правой клавишей «мыши» по корню (названию) классификатора и в открывшемся контекстном меню выбрать пункт «Настроить конфигурацию обучения», после этого откроется окно «Настройки режима обучения для рубрики» (выбор метода обучения и настройка параметров осуществляется для каждой рубрики в отдельности).

Основные параметры обучения:

- «Имя обучаемой переменной» – название переменной типа «define», которая будет автоматически сформирована по завершению процесса обучения классификатора (значение по умолчанию «TrainDefine»)
- «Метод» – метод машинного обучения с помощью которого формируется правило. Интерфейс позволяет выбрать следующие методы:

- «Метод опорных векторов» – линейный метод машин опорных векторов (support vector machine, SVM). Реализует формирование списка условий включения документов в рубрику с локальными и глобальными весами на основе анализа всей эталонной коллекции. Сформированное правило представляется с помощью оператора #nscalar (см. справку по SCATQL соответствующий раздел), задающего линейную функцию. При обработке документа осуществляется подсчёт частот встречаемости перечисленных в операторе #nscalar условий, их значения подставляются в заданную оператором решающую функцию. Если полученное при подстановке значение больше 0, то документ относится к рубрике.
- «Список терминов» – метод решающего списка (decision list). Реализует формирование списка терминов, встречающихся в анализируемых текстах, по наличию которых можно судить об относимости текста к рубрике. Сформированное правило представляется с помощью оператора ИЛИ. Документ относится к рубрике, если он удовлетворяет хотя бы одному из выражений в списке:

а) «Максимальное число признаков» – (значение по умолчанию «250»).

б) Селекция признаков – дополнительные настройки (по умолчанию скрыты).

в) «Ограничение на частоту признаков, кол. | Минимальная частота» (значение по умолчанию «0», без ограничений).

г) «Ограничение на частоту признаков, кол. | Максимальная частота» (значение по умолчанию «0», без ограничений).

д) «Ограничение на частоту признаков, кол. | Максимальное число признаков» – (значение по умолчанию «1000»).

е) «Метод усреднения частот отдельных признаков» – Сумма (значение по умолчанию), взвешенная сумма и максимум.

ж) «Тип усреднения» – Хи-квадрат (значение по умолчанию), частота терминов, частота текстов, прирост информации, взаимная информация, кросс-энтропия, весомость признаков, коэффициент корреляции, упрощённый Хи-квадрат, коэффициент релевантности.

з) «Ограничение на частоту признаков, % | Максимальная частота» (значение по умолчанию «50»).

и) «Селекция признаков | Процент селекции признаков» (значение по умолчанию «50»).

к) «Сглаживание» – выбор способа сглаживания частоты признаков. Реализовано два режима «без сглаживания» и «Лапласа» (значение по умолчанию).

л) «Тип результата» – выбор способа итогового вычисления весов признаков, глобальный и локальный (по умолчанию выставлено «глобальное взвешивание»).

После завершения настройки параметров необходимо нажать на кнопку «Применить» для сохранения изменений, после на кнопку «Обучить» для автоматического формирования правил классификации.

2.6.2.23.3 Публикация классификатора

Для осуществления процесса публикации проекта в выходной каталог необходимо на панели инструментов или в контекстном меню «Классификаторы» выбрать пункт «Опубликовать», выделенный на рисунке 100.

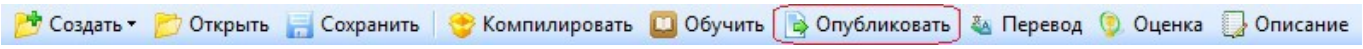


Рисунок 100 – Публикация классификатора

В результате выполнения данного действия откроется окно, представленное на рисунке 101, в котором следует задать «Имя файла» и путь для его сохранения (по умолчанию сохранение происходит в папку «ClassifierCls»).

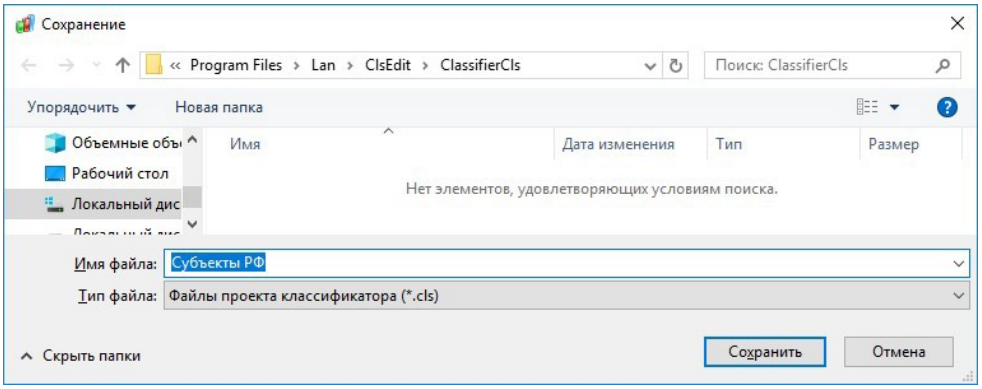


Рисунок 101 – Окно сохранения cls-файла 78

2.7 Подсистема «Тезаурусы»

2.7.1 Графический интерфейс

Компонент «Тезаурусы», представленный на рисунке 102, содержит в себе следующие инструменты:

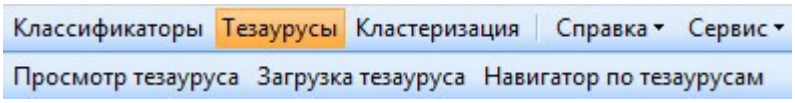


Рисунок 102 – Меню «Тезаурусы»

- «Просмотр тезауруса» – открытие окна для поиска и просмотра терминов по всем существующим в заданной базе тезаурусам;
- «Загрузка тезауруса» – загрузка тезаурусов из внешнего файла;
- «Навигатор по тезаурусам» – открытие окна для создания и редактирования тезаурусов и содержащихся в них терминов.

2.7.2 Основные функции

Функция «Тезаурус» предназначена для создания и редактирования пользовательских тезаурусов, которые необходимы для автоматического расширения правил классификации. Данная подсистема позволяет ускорить и тем самым упростить работу над созданием классификатора, представлена на рисунке 103.

Для начала работы с данной функцией сначала необходимо произвести настройку подключения к базе с хранящимся в ней тезаурусом.

Для этого необходимо зайти в «Сервис | Настройка», в открывшемся окне «Настройки приложения» перейти во вкладку «Тезаурус», во вкладке присутствуют следующие поля и кнопки для настройки:

- «Сервер базы тезауруса» – в данное поле вводится пользовательское имя используемой базы;
- «Адрес» – в данном поле указывается IP-адрес базы данных;
- «Имя базы данных» – по умолчанию все базы с тезаурусами имеют имя «Thesauri»;

- «Использование SSSI» – выбор типа подключения к базе;
- «Имя пользователя» и «Пароль» – имя и пароль для подключения к базе соответственно;
- «Применить» – сохранение внесенных изменений в поля настройки;
- «Обновить базу» – обновление текущей базы с тезаурусом до последней версии;
- «Создать базу» – создание новой базы с тезаурусом по заданному адресу;
- «Удалить базу» – удаление текущей базы с тезаурусом с сервера (данные после удаления не восстанавливаются).

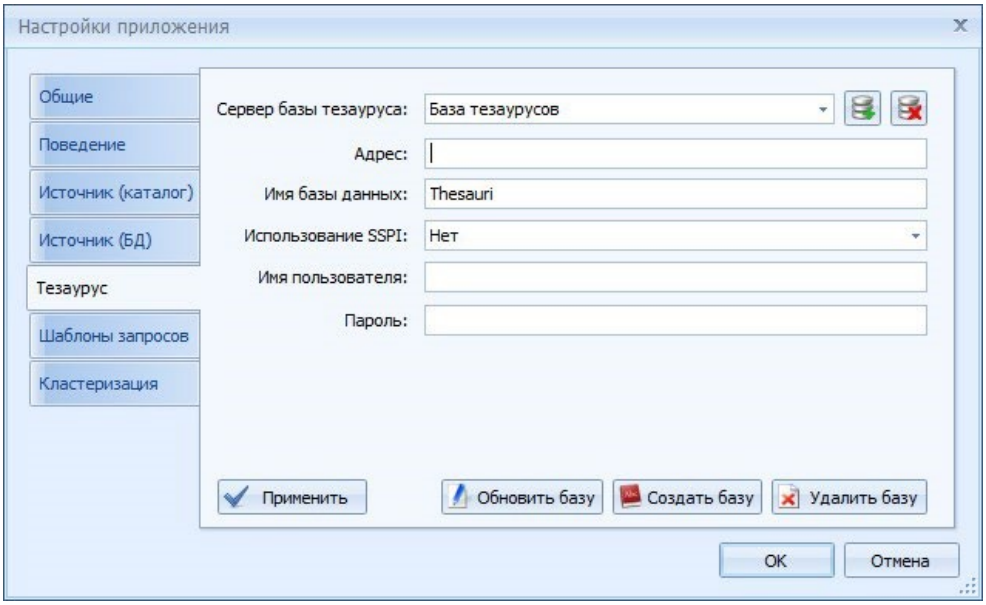


Рисунок 103 – Окно настройки базы тезаурусов

Для дальнейшей работы с тезаурусом необходимо воспользоваться меню «Тезаурусы», которое представлено на рисунке 104 и содержит в себе следующие функции:

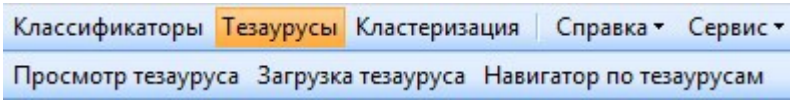


Рисунок 104 – Меню «Тезаурусы»

- «Просмотр тезауруса» – открытие окна для поиска и просмотра терминов по всем существующим в заданной базе тезаурусам. Окно представлено на рисунке 105;

- «Загрузка тезауруса» – загрузка тезауруса из внешнего файла;

- «Редактирование тезауруса» – открытие окна для создания и редактирования тезаурусов и содержащихся в них терминов.

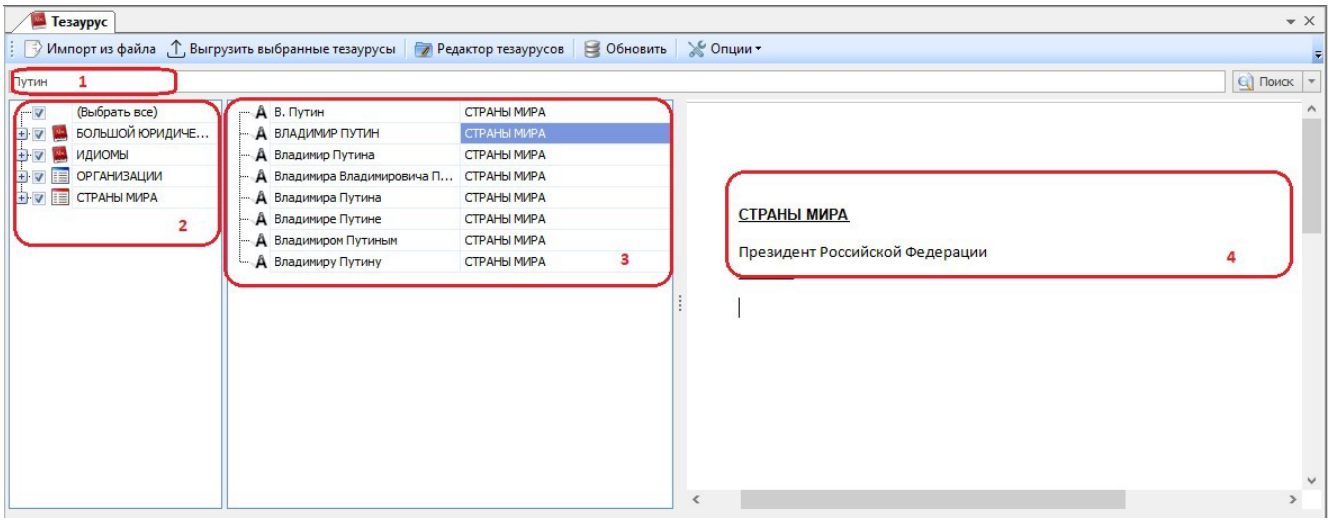


Рисунок 105 – Окно «Просмотр тезауруса»

Окно «Просмотр тезауруса» состоит из поля ввода искомого термина (1 на рисунке 105); списка тезаурусов, имеющихся в заданной базе (2 на рисунке 105); списка найденных по запросу терминов (3 на рисунке 105); окна описания найденного термина (4 на рисунке 105). Список имеющихся в заданной базе тезаурусов позволяет отмечать тезаурусы, по которым будет производиться поиск. В окне описания термина отображается описание термина, выбранного из списка найденных по запросу терминов.

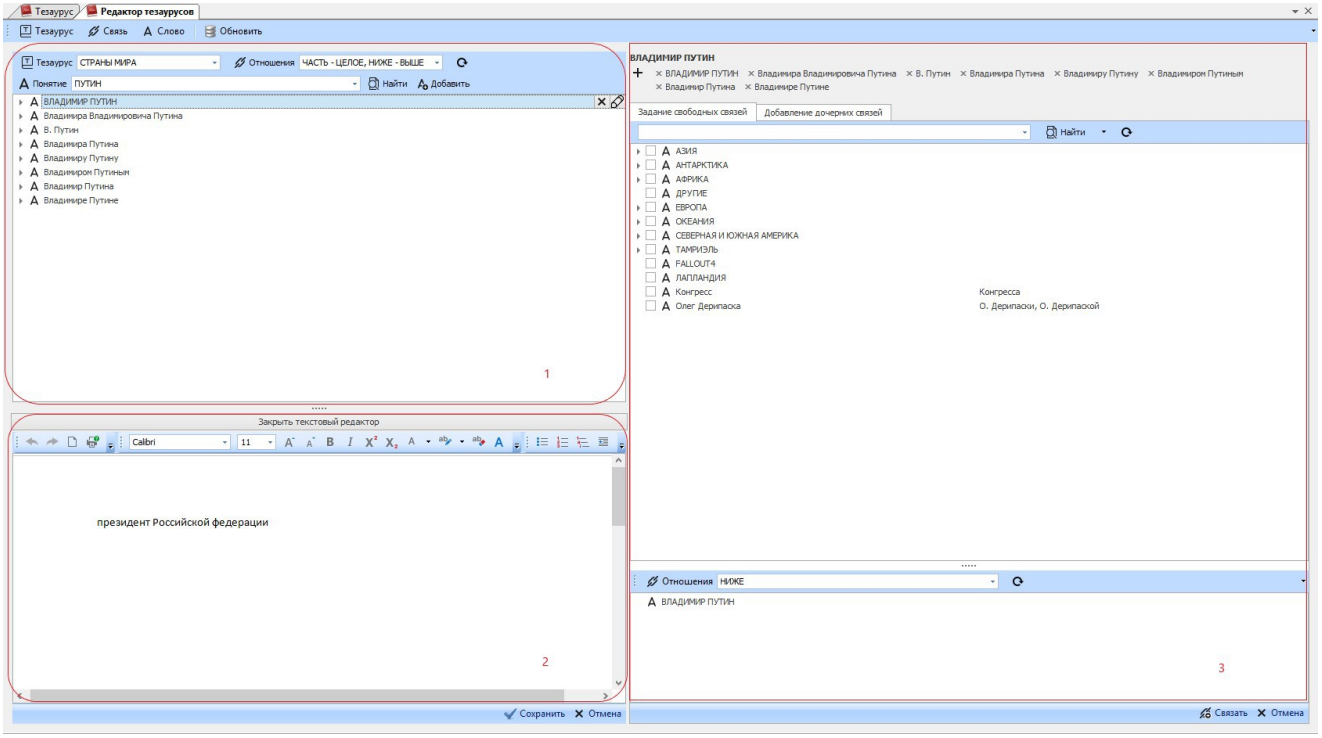


Рисунок 106 – Окно «Редактор тезаурусов»

Окно «Редактор тезаурусов» позволяет редактировать имеющиеся в базе тезаурусы и создавать новые тезаурусы. Окно «Редактор тезаурусов» состоит из окна со списком терминов выбранного тезауруса (1 на рисунке 106), встроенного текстового редактора, позволяющего редактировать выбранный термин (2 на рисунке 106) и окна задания связей между терминами тезауруса (3 на рисунке 106).

Меню «Тезаурус» редактора тезаурусов позволяет создавать новые тезаурусы, загружать тезаурусы из внешних источников, выгружать имеющиеся тезаурусы в файл, а также удалять их.

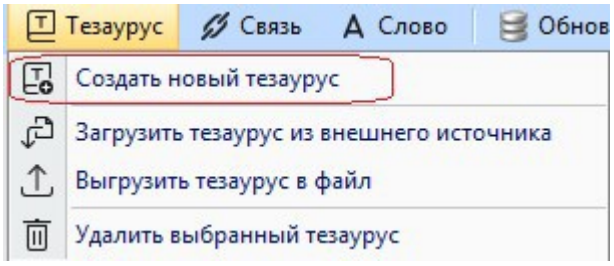


Рисунок 107 – Меню «Тезаурус» редактора тезаурусов

Для создания нового тезауруса необходимо выбрать пункт «Создать новый тезаурус» в меню «Тезаурус» редактора тезаурусов (выделено на рисунке 107). Данный пункт открывает доступ к окну создания нового тезауруса, где задаются

его параметры (рисунок 108).

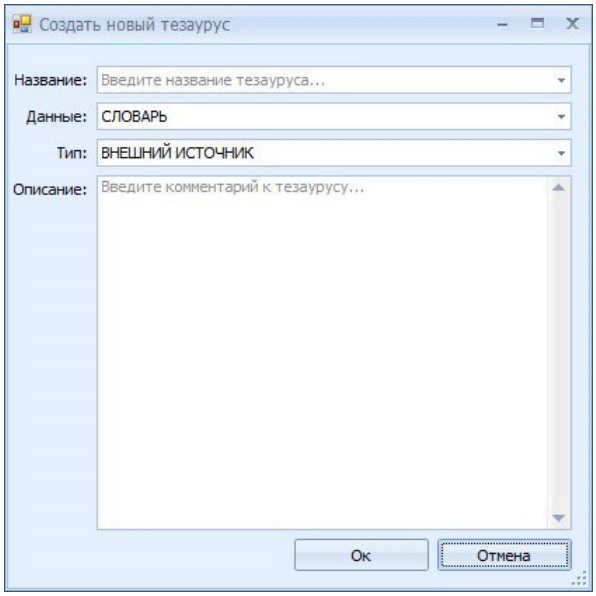


Рисунок 108 – Окно создания нового тезауруса.

Для редактирования терминов тезауруса используется окно редактирования терминов и встроенный текстовый редактор (рисунок 109). В окне редактирования можно искать имеющиеся в тезаурусе термины, добавлять новые термины и задавать отношения между терминами.

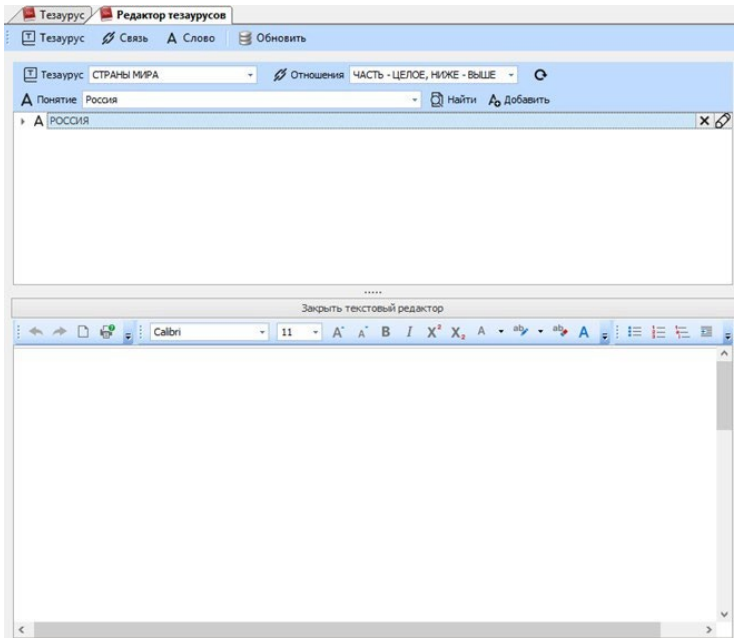


Рисунок 109 – Окно редактирования терминов и текстовый редактор

Для редактирования связей между терминами тезауруса используется окно редактирования связей (рисунок 110). Данное меню позволяет задавать свободные

связи и добавлять дочерние связи.

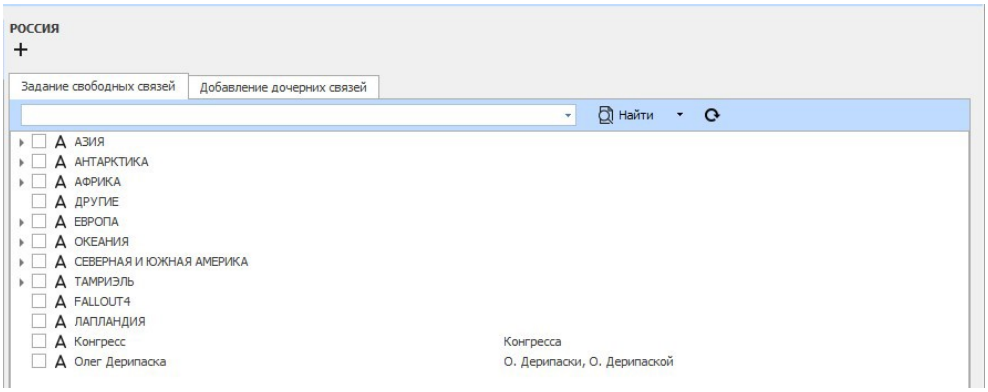


Рисунок 110 – Редактирование связей между терминами.

2.8 Подсистема «Кластеризация»

Подсистема «Кластеризация» необходима для создания и редактирования конфигурационных файлов, которые в дальнейшем будут использоваться для формирования кластеров документов.

Для формирования конфигурационного файла необходимо:

- подготовить эталонные классы (сюжеты) документов (кластеры, сформированные аналитиком), на которых будет производиться обучение (аналогично эталонным подборкам для классификаторов, но с более узким сюжетированием);

- подготовить набор тематических и атрибутивных классификаторов для выделения признаков из текста. По данным признакам будет производиться анализ похожести текстов.

- Произвести настройку параметров быстрого алгоритма, при необходимости скорректировать результаты вручную.

Для создания «Подборки» необходимо на панели инструментов, либо во вкладке «Подборки | Действия» выбрать «Создать новую подборку» или «Открыть подборку», в результате будет создана пустая подборка (только название подборки без классов) или открыта уже существующая подборка (с классами и уже настроенным конфигурационным файлом) соответственно. При создании или открытие подборки по умолчанию открывается вкладка для настройки конфигурационного файла, Рисунок 111.

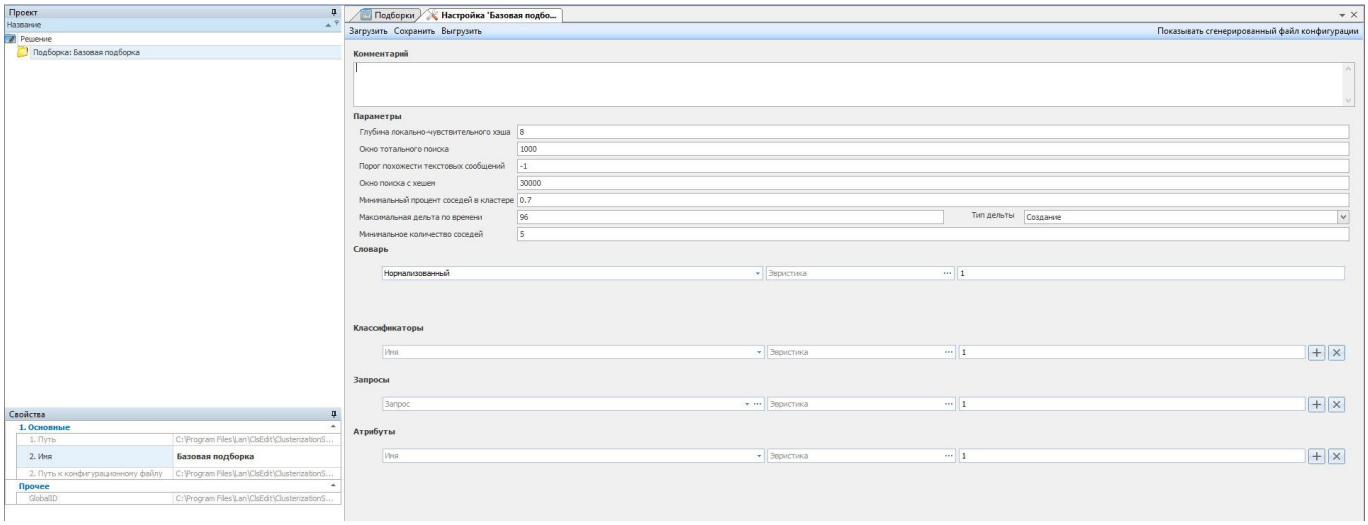


Рисунок 111 – Создание подборки

После создания подборки необходимо выбрать классификаторы, которые будут участвовать в извлечение признаков из текста. Для этого в контекстном меню подборки выбираем пункт «Задать классификаторы», после чего в открывшемся окне выбираем классификаторы для выделения фрагментов и атрибутов, Рисунок 112.

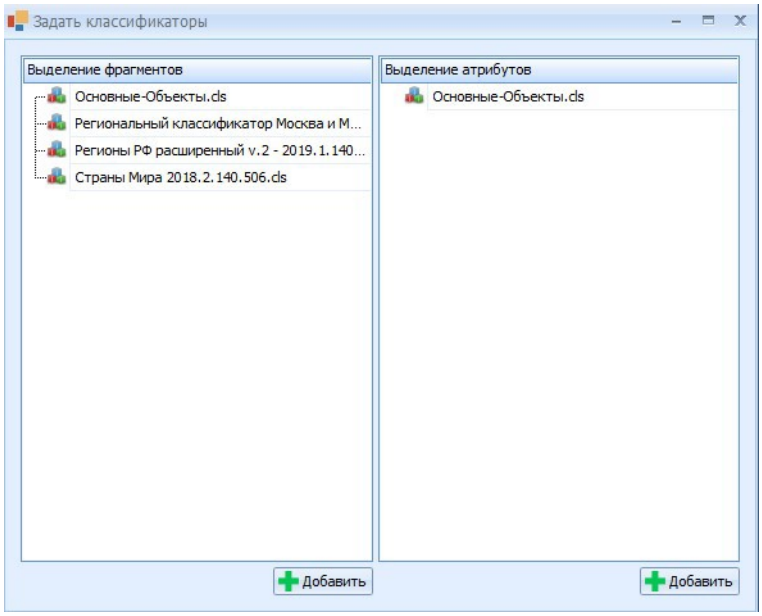


Рисунок 112 – Окно выбора классификаторов

После чего переходим к созданию эталонных классов (сюжетов) документов. Для этого в контекстном меню подборки выбираем «Создать класс», либо «Загрузить класс».

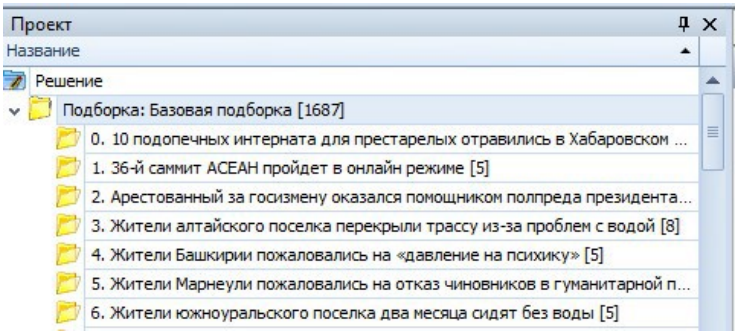


Рисунок 113 – Классы подборки «Базовая подборка»

Для настройки конфигурационного файла, представленного на рисунке 114 необходимо выбрать на панели инструментов «Настроить», либо в контекстном меню подборки (правой клавишей «мыши» по названию подборки | «Настроить») в результате откроется окно настройки конфигурационного файла.

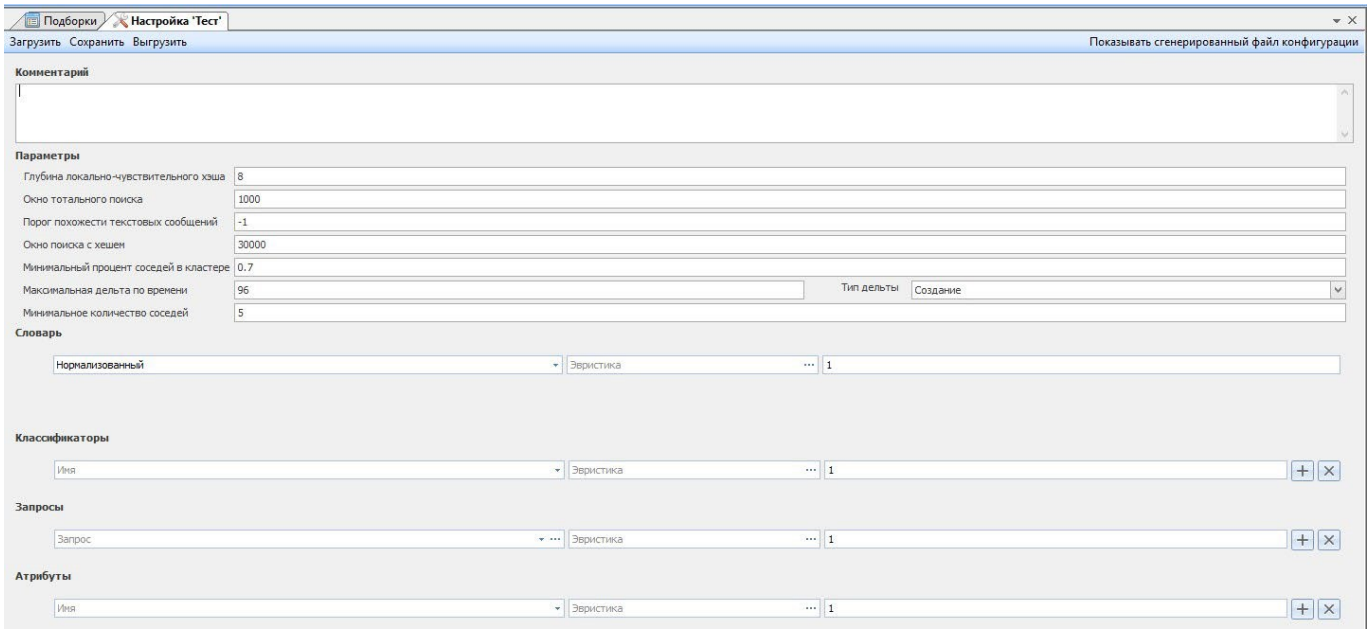


Рисунок 114 – Конфигурационный файл кластеризации без настроек

- «Комментарий» – в данное поле вводится важная информация по подборке и вводимым параметрам (необязательный параметр);

- «Параметры | Глубина локально-чувствительного хэша» – (значение по умолчанию «8»);

- «Параметры | Окно тотального поиска» – (значение по умолчанию «1000»);

- «Параметры | Порог похожести текстовых сообщений» – (значение по

умолчанию «-1»);

- «Параметры | Окно поиска с хешем» – (значение по умолчанию «30000»);
- «Параметры | Минимальный процент соседей в кластере» – (значение по умолчанию «0,7»);
- «Параметры | Максимальная дельта по времени» – (значение по умолчанию «96»);
- «Параметры | Тип дельты» – (значение по умолчанию «Создание»);
- «Параметры | Минимальное количество соседей» – (значение по умолчанию «5»);
- «Словарь» – (значение по умолчанию «Нормализованный»);
- «Классификаторы» – в данном поле осуществляется выбор классификаторов рубрики, которых будут выступать в качестве положительных признака во время кластерного анализа;
- «Запросы» – в данном поле осуществляется ввод поисковых запросов, которые будут выступать в качестве положительных признака во время кластерного анализа;
- «Атрибуты» – в данном поле осуществляется выбор атрибутов, которые будут выступать в качестве положительных признака во время кластерного анализа.

После завершения настройки параметров и выбора признаков необходимо сохранить изменения (кнопка «Сохранить»), после чего нажать на кнопку «Обучить» для автоматического вычисления веса каждого из признаков (по умолчанию все признаки имеют вес «1»).

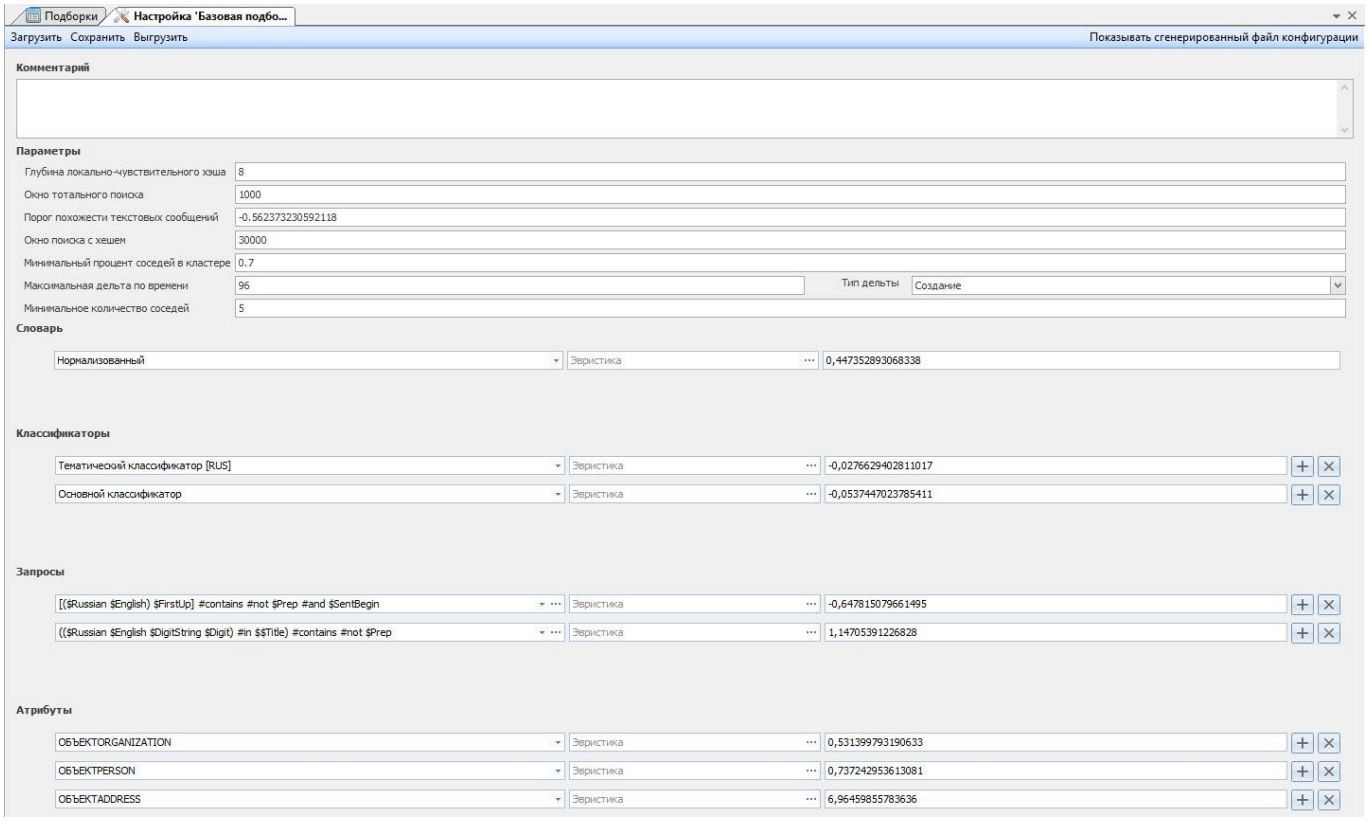


Рисунок 115 – Конфигурационный файл кластеризации после обучения

Далее для проверки качества произведенных настроек нажимаем на кнопку «Кластеризовать», в результате отобразится «Отчет о результатах кластеризации».

Отчет о результатах кластеризации

Время создания: 09:54 06.02.2019

Отчет получен при кластеризации со следующими параметрами:

Параметр	Значение
Глубина локально-чувствительного хэша	8
Окно тотального поиска	1000
Порог похожести текстовых сообщений	-0.56237323059211841
Окно поиска с хешем	30000
Минимальный процент соседей в кластере	0.7
Максимальная дельта между временем создания документов	96
Минимальное количество соседей	5

Словарь	Вес
Features	0.44735289306833803

Классификатор	Вес
Тематический классификатор [RUS]	-0.027662940281101694
Основной классификатор	-0.053744702378541107

Запрос	Вес
[($Russian $English) $FirstUp] #contains #not $Prep #and $SentBegin	-0.6478150796614951
(($Russian $English $DigitString $Digit) #in $$Title) #contains #not $Prep	1.1470539122682764

Атрибут	Вес
ОБЪЕКТORGANIZATION	0.53139979319063346
ОБЪЕКТPERSON	0.73724295361308123
ОБЪЕКТADDRESS	6.9645985578363625

Таблица 1. Общие характеристики результатов кластеризации.

Показатель	Значение
Точность	100%
Полнота	25%
F-мера	35%

33. Главные проблемы медиа в Казахстане.docx (5)	100%	50%	67%
44. Более 300 дагестанцев отбывают срок за экстремизм и терроризм.docx (1)	100%	10%	18%
46. ВСЕ КАК ОСЕНЬЮ 2017 ГОДА.docx (2)	100%	20%	33%
49. Правозащитники об угрозе активизации вооруженного подполья в Чечне.docx (1)	100%	10%	18%
52. Экстремистский журнал.docx (1)	100%	10%	18%

Рисунок 116 – Отчет о результатах кластеризации