119270, Москва, Лужнецкая наб., д. 6,
стр.1, офис 214, ООО «ЭР СИ О»
Тел./факс: (495) 287-98-87
E-mail: info@rco.ru
http://www.rco.ru

Руководство администратора

RCO Deduplicator – программа для выявления дубликатов документов

Версия 1.1 (Microsoft Windows)

Москва, 2020

В содержание данного документа могут быть внесены изменения без предварительного уведомления. Названия организаций, имена и даты, используемые в качестве примеров, являются вымышленными, если не оговорено обратное.

Содержание

Общие сведения

Назначение

Библиотека предназначена для выделения из текстовых документов признаков, позволяющих выявлять перепечатки (дубли) новостных сообщений путём сопоставления наборов признаков. Под перепечатками понимается небольшое изменение текста или оформления сообщения без изменения фактологической или смысловой составляющих.

Выявление дублей загружаемого документа среди имеющихся в базе данных (БД) необходимо для очистки результатов поиска от лишней информации и, следовательно, упрощения аналитической работы с базой.

Процедура избавления от дубликатов двухэтапная. Первый этап – выявление важных для обнаружения дубликатов характеристик поступившего в систему документа. Второй – поиск дубликатов.

Использование библиотеки **RCO Deduplicator** возможно лишь при наличии работающей версии программы **RCO Fact Extractor**.

Требования к квалификации персонала, необходимом для поддержки ПО

Администратор ПО должен иметь следующие навыки:

- Администрирования информационных систем;
- Технического обслуживания средств вычислительной техники, на которых устанавливается ПО;
- Работы с операционной системой Microsoft Windows.

Системные требования

Требования к характеристикам вычислительной техники зависят от количества и объёма обрабатываемых документов.

Минимальные требования: 1 ядро процессора частотой 2 ГГц, 4 Гб оперативной памяти (из которых 300 Мб непосредственно на запуск одного экземпляра), 300 Мб на жёстком диске.

Рекомендуемые требования, исходя из потока 50 тыс. документов в час общим объёмом 100 Мб: 2 ядра процессора частотой 3 ГГц (2 обработчика - по 1 ядру на каждый), 5 Гб оперативной памяти (по 500 Мб на каждый обработчик), 1 Гб на жёстком диске. При увеличении объёма потока документов необходимо пропорционально увеличить производительность ядра процессора или количество обработчиков, скорректировав соответствующим образом требования к памяти и количеству ядер процессора.

Подготовка к работе

Состав дистрибутива

Дистрибутив содержит также бинарные файлы и лингвистические ресурсы библиотеки RCO Fact Extractor SDK.

- DeDuplicator.dll и DeDuplicatorx64.dll - 32-х и 64-х разрядные сборки библиотеки под Windows
- RCO_Deduplicator.pdf - руководство разработчика
- папка include - заголовочные файлы API библиотеки (файлы декларации интерфейса библиотеки)
- папка include_fx - заголовочные файлы библиотеки RCO Fact Extractor (файлы декларации интерфейса библиотеки RCO Fact Extractor)
- папка lib - входные файлы для компоновщика C++ при линковке библиотеки
- папка lib_fx - входные файлы для компоновщика C++ при линковке библиотеки RCO Fact Extractor
- папка src - файл DeDuplicatorDllTest.cpp с примером кода C++ работы с библиотекой
- папка test_stand - сборка исполняемых файлов, лингвистических ресурсов и тестовых файлов для тестирования библиотеки, содержит:
- DeDuplicator.dll и DeDuplicatorx64.dll - 32-х и 64-х разрядные сборки библиотеки под Windows
- DeDuplicatorTest.exe и DeDuplicatorTestx64.exe - 32-х и 64-х разрядная исполняемая программа, собранная по коду DeDuplicatorDllTest.cpp - обрабатывает файлы из заданного каталога и печатает результат в файл doc_features.xls и в папку DuplicatesResult
- RCOFXRu.dll и RCOFXRux64.dll - 32-х и 64-х разрядные сборки библиотеки RCO Fact Extractor под Windows
- test.bat - Примеры вызова программы DeDuplicatorTest.exe
- dedup.ini - параметры извлечения характеристик и сравнения документов на дубликаты для программы DeDuplicatorTest.exe или DeDuplicatorTestx64.exe
- папка dic и файл fx.ini - лингвистические ресурсы и настройки конфигурации библиотеки RCO Fact Extractor
- fx_log.txt - файл журнала работы библиотеки RCO Fact Extractor (имя файла задаётся в fx.ini)
- benchmark_time.txt - отчёт о времени работы программы (формируется после завершения работы с библиотекой)
- папка Text - папка с документами, которые подаются на обработку в программы DeDuplicatorTest.exe и DeDuplicatorTestx64.exe
- папка DuplicatesResult и файл doc_features.xls - результаты работы программ DeDuplicatorTest.exe и DeDuplicatorTestx64.exe

Установка ПО

- Скопировать дистрибутив на локальный диск компьютера.
- Установить драйвер защиты Sentinel HASP с сайта производителя (https://thales-sentinel.ru/helpdesk/download-space/) или из папки Sentinel_LDK_Run-time_setup, сопровождающей дистрибутив.
- Вставить ключ защиты в usb-порт компьютера и убедиться, что он виден на странице http://localhost:1947/_int_/devices.html

Проверка работоспособности

Исполнить test.bat в папке test_stand. После этого запускается тестовое приложение DeDuplicatorTest.exe, которое считывает html и txt файлы из папки Text, передаёт их на анализ в библиотеку. Результаты анализа записываются в файл doc_features.xls, группы дублей входных документов записываются в файл index_duplicates.htm в папке DuplicatesResult.

Дополнительные настройки

Путь к обрабатываемым файлам и к файлу с результатами анализа можно менять, редактируя файл test.bat, параметры работы тестового приложения DeDuplicatorTest.exe задаются в файле dedup.ini, исходный код тестового приложения можно найти в дистрибутиве – файл DeDuplicatorTest.cpp в папке src дистрибутива.

Удаление ПО

Для удаления программы достаточно удалить содержимое корневой папки, созданной в процессе установки ПО.

Регламентное обслуживание

Устранение неисправностей в ходе эксплуатации ПО

Отчёт об ошибках, возникающих при работе программы, печатается в файл fx.log (путь к файлу задаётся в файле fx.ini). Отчёт содержит наименование модуля, дату/время, текстовое описание ошибки и имя документа, при обработке которого возникла ошибка. Ошибки, связанные с лингвистическими ресурсами или входными документами, устранимы на месте. По остальным ошибкам следует обращаться в техническую поддержку разработчика ПО.

При выставлении параметра <level> в файле fx.ini в значение verbose или info в файле fx.log дополнительно будут печататься сообщения о начале и окончании обработки входных файлов. В файле benchmark_time.txt печатается отчёт о процессорном времени, затраченном на обработку входных файлов.

Обновление ПО

Для обновления нужно остановить работу программы, перезаписать обновляемые файлы и снова запустить программу.

Совершенствование ПО

Производитель, ООО «ЭР СИ О», периодические выпускает новые версии ПО, содержащие новые функциональные возможности.

Для получения всех обновленных версий ПО, по мере их выхода, необходимо наличие действующего договора на техническую поддержку ПО с ООО «ЭР СИ О».