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 Categorization Engine – библиотека категоризации текстов

Версия 2.0

(Microsoft Windows, Unix)

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

Обзор

Назначение

Библиотека категоризации текстов позволяет решать следующие задачи:

- На основании лексических профилей эффективно определять принадлежность текста к заданному множеству категорий;
- Для каждого термина из лексических профилей, обнаруженного в тексте, получить количество его вхождений в текст, а также позиции терминов в тексте.

Основными областями применения библиотеки являются:

- Тематическая категоризация текстов в электронных библиотеках, информационно-поисковых и информационно-аналитических системах;
- Тематический таргетинг в баннерных сетях;
- Мониторинг ключевых слов и словосочетаний в системах мониторинга и сбора информации.

Требования к квалификации

Администратор ПО должен иметь следующие навыки:

- администрирования информационных систем;
- технического обслуживания средств вычислительной техники, на которых устанавливается ПО;
- работы с операционной системой 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 200 Мб непосредственно на запуск одного экземпляра), 100 Мб на жёстком диске.

Рекомендуемые требования, исходя из потока 100 тыс. документов в час общим объёмом 500 Мб: 1 ядро процессора частотой 3 ГГц, 4 Гб оперативной памяти, 500 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.

Оценки даны для лексических профилей из дистрибутива. При увеличении количества профилей, их объёма или сложности входящих в них терминов требования к ресурсам могут повыситься.

Подготовка к работе

Состав дистрибутива

В состав дистрибутива входят следующие файлы и папки:

- gpcat.dll и gpcatx64.dll - 32-х и 64-х разрядные сборки библиотеки под Windows
- gpcat.lib и gpcatx64.lib - 32-х и 64-х разрядные файлы для компоновщика C++ при линковке библиотеки
- RCO_GP_Cat_API.pdf - документация (наиболее актуальную документацию можно найти на странице http://www.rco.ru/?page_id=4848)
- папка include - заголовочные файлы API библиотеки
- папка test_stand - пример работы с библиотекой
- папка TestGPCAT - исходный код программы TestGPCAT, иллюстрирующей вызовы функций API библиотеки

Содержание папки test_stand:

- TestGPCAT.exe и TestGPCATx64.exe - 32-х и 64-х разрядная исполняемая программа TestGPCAT, загружает профили из заданного каталога, обрабатывает файлы из заданного каталога и печатает результат в формате XML в файл resultGPCAT.xml
- test.bat и testx64.bat - примеры вызова программы TestGPCAT.exe и TestGPCATx64.exe
- gpcat.dll и gpcatx64.dll - 32-х и 64-х разрядные сборки библиотеки под Windows
- resultGPCAT.xml - файл с результатом работы программ TestGPCAT.exe и TestGPCATx64.exe
- папка dic - файлы с морфлогией (русской и английской)
- папка samples - папка с документами, которые подаются на обработку в программы TestGPCAT.exe и TestGPCATx64.exe

Содержание папки TestGPCAT:

- TestGPCAT.cpp, stdafx.h и stdafx.cpp - исходные коды на C++ базовой версии программы TestGPCAT, иллюстрирующей вызовы функций API библиотеки
- папка FXWrapper - исходные коды расширения программы TestGPCAT, которое позволяет использовать в качестве классификационных признаков сущности, выделенные из текста библиотекой RCO Fact Extractor
- файл FXWrapper/ExtractParams.txt - параметры генерации имени сущности библиотекой RCO Fact Extractor, используемые при её подключении

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 run.bat или runx64.bat в папке test_stand. После этого запускается тестовое приложение TestGPCAT.exe или TestGPCATx64.exe, которое загружает лексические профили из папки Profiles, считывает html и txt файлы из папки samples, передаёт их на анализ в библиотеку. Результаты анализа записываются в формате XML в файл resultGPCAT.xml.

Дополнительные настройки

Путь к лексическим профилям, обрабатываемым файлам и к файлу с результатами анализа можно менять, редактируя файл run.bat и runx64.bat. Дополнительно можно менять параметры распечатки результатов в resultGPCAT.xml, включать в процесс анализа библиотеку RCO Fact Extractor и прочее. Полный список возможностей можно посмотреть в файле TestGPCAT.cpp в папке TestGPCAT.

Удаление ПО

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

Регламентное обслуживание

Устранение неисправностей в ходе эксплуатации

Ошибки, возникающие в процессе работы программы, печатаются в консоль. При необходимости можно изменить способ журналирования ошибок, модифицировав код тестового приложения TestGPCAT.exe. Выдаётся код ошибки и её текстовое описание. Ошибки, связанные с лексическими профилями или входными документами, устранимы на месте. По остальным ошибкам следует обращаться в техническую поддержку разработчика ПО.

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».