Разработка сервиса автоматизированного мониторинга рекламы в печатных изданиях

Описание проблемы

Для мониторинга рекламодателей, оценки активности конкурентов и анализа рекламного рынка заказчик использует данные внешнего аналитического сервиса. При этом приобретаемый сервис охватывает значительно больше изданий, чем требуется для решения задач заказчика, что приводит к избыточным расходам на получение аналитики.

Исходные материалы — электронные версии и сканы печатных изданий — содержат необходимые сведения, однако их преобразование в структурированную базу требует распознавания содержания страниц, выявления рекламных размещений, определения их размеров и классификации рекламируемых объектов.

Задачу усложняют разнообразие оформления журналов, наличие рекламных статей и немаркированной рекламы, сканы разворотов, использование русского и казахского языков, а также различные варианты написания названий рекламодателей и брендов.

Необходим собственный сервис, который будет автоматически обрабатывать выбранный перечень изданий и формировать нормализованные данные для последующего анализа с минимальным участием человека.

Ожидаемый эффект

Реализация проекта позволит:

- Организовать регулярный мониторинг рекламы в необходимом заказчику перечне печатных изданий.
- Оптимизировать расходы на получение рекламной аналитики и снизить зависимость от внешних поставщиков данных.
- Создать единую базу рекламных размещений с историей, подтверждающими изображениями и связью с исходными файлами.
- Обеспечить анализ активности рекламодателей, брендов и товарных категорий в разрезе изданий и периодов.
- Автоматизировать не менее 95% обработки, ограничив ручную верификацию долей не более 5%.
- Обнаруживать не менее 98% рекламных размещений, соответствующих условиям мониторинга.
- Получать результаты обработки отдельного номера не позднее одного часа после его загрузки.

Цель и описание задачи проекта

Разработать и внедрить сервис для регулярного автоматизированного анализа рекламы в печатных изданиях на основе технологий искусственного интеллекта, распознавания текста и анализа изображений.

Сервис должен включать средства загрузки и обработки файлов, базу данных, интерфейс аналитика, раздел верификации и инструменты ведения справочников.

1. Загрузка и подготовка материалов

Сервис должен принимать качественные PDF и сканы разрешением от 300 dpi на русском и казахском языках. Каждый загружаемый файл соответствует одному номеру издания.

Планируемый охват — до 50 изданий с периодичностью выхода до одного раза в неделю и объёмом от 50 до 300 страниц в номере.

Необходимо обеспечить:

- Определение общего количества журнальных полос, включая обложку.
- Распознавание и разделение сканов разворотов, содержащих две журнальные страницы на одной странице PDF.
- Сохранение соответствия между страницами исходного файла и журнальными полосами.
- Определение номера, даты выпуска и принадлежности файла к изданию.
- Отображение статуса обработки и возникших ошибок.

2. Выявление и классификация рекламы

Сервис должен выявлять рекламные макеты, рекламные статьи, партнёрские материалы, рекламу самого издания и немаркированные материалы рекламного характера.

При анализе необходимо учитывать:

- Обозначения «реклама», «на правах рекламы» и аналогичные формулировки на поддерживаемых языках.
- Указания в выходных данных о рекламном характере отдельных рубрик.
- Содержание, визуальное оформление и другие признаки рекламного размещения.

Для каждого результата должны сохраняться тип материала и основание его отнесения к рекламе. Неоднозначные случаи направляются в раздел верификации.

3. Определение характеристик размещений

Для каждого размещения необходимо определять и сохранять:

- Издание, номер и дату выпуска.

- Рекламодателя — юридическое лицо, если его возможно достоверно установить.
- Рекламируемый бренд или несколько брендов.
- Объект рекламирования и его категорию.
- Тип материала и наличие рекламной маркировки.
- Размер размещения в долях полосы.
- Номер страницы PDF и напечатанный номер журнальной страницы, если он присутствует.
- Ссылку на исходный файл и изображение размещения.
- Статус автоматической обработки и верификации.

Размер рассчитывается относительно рабочей области страницы. Для рекламной статьи учитывается вся площадь материала, включая текст и иллюстрации.

Учитываются форматы: одна полоса, ½, ⅓, ¼, ⅛, 1/16 полосы, а также многополосные размещения. Нестандартные размеры приводятся к ближайшему стандартному формату. Размещения менее 1/16 полосы исключаются из мониторинга.

Одна запись в БД соответствует одному рекламному размещению. Для материалов на нескольких страницах сохраняются перечень страниц и суммарный объём. Отдельные рекламные блоки на одной странице учитываются отдельными записями.

4. Нормализация данных и развитие справочников

Необходимо обеспечить нормализацию рекламодателей, изданий, номеров выпусков, категорий и брендов.

Справочники должны формироваться и дополняться в процессе эксплуатации. При обработке очередного номера система должна предлагать в специальном разделе:

- Добавление новых рекламодателей, брендов, изданий и категорий.
- Объединение вариантов написания существующих наименований.
- Изменение связей между рекламодателями, брендами и категориями.
- Исправление ранее выполненной классификации.

Для уточнения данных допускается использование открытых интернет-источников. Недостоверные или отсутствующие сведения не должны подменяться предположениями: такие записи отмечаются как требующие проверки.

5. Интерфейс аналитика

Интерфейс должен поддерживать работу до 10 пользователей и обеспечивать:

- Загрузку файлов и контроль обработки.
- Поиск и фильтрацию по периоду, изданию, рекламодателю, бренду, категории и формату.
- Просмотр найденного размещения в контексте исходной страницы.
- Получение сводных показателей по количеству и объёму размещений.
- Проверку спорных результатов и исправление данных.
- Работу с предложениями по изменению справочников.

- Выгрузку результатов в Excel.

Необходимо предусмотреть роли администратора и аналитика, а также историю ручных изменений.

6. Хранение и эксплуатация

Результаты должны храниться в структурированной БД с возможностью прямого доступа для аналитических запросов. Исходные файлы и изображения размещений сохраняются в организованной структуре на диске с привязкой к записям БД.

Развёртывание — на внутреннем сервере или в облаке с использованием Docker. Допускается применение внешних ИИ-сервисов.

Необходимо обеспечить резервное копирование БД и возможность восстановления данных.

Время от загрузки номера до получения результата, включая ожидание в очереди, — не более 60 минут. При поступлении нескольких файлов должна поддерживаться параллельная обработка. Конфигурация вычислительных ресурсов и нагрузочный сценарий фиксируются при внедрении.

7. Проверка качества и приёмка

Приёмку проводит аналитик заказчика на согласованной контрольной выборке. Материалы для разработки и проверки предоставляются по запросу.

Контрольная выборка должна охватывать оба языка, электронные PDF, сканы, развороты и различные виды рекламы. Правила разметки, сопоставления результатов и расчёта показателей согласуются до итоговых испытаний.

Основные критерии:

- **Полнота обнаружения — не менее 98%:** доля найденных рекламных размещений от всех размещений в эталонной разметке, подлежащих мониторингу.
- **Автоматизация — не менее 95%:** доля размещений, для которых система корректно формирует итоговую запись без ручной корректировки.
- **Ручная верификация — не более 5% размещений.**
- **Срок обработки номера — не более 60 минут** в согласованном нагрузочном сценарии.

Дополнительно проверяются ложные обнаружения, правильность размеров, нормализация данных, работа интерфейса, экспорт, прямой доступ к БД и восстановление из резервной копии.

8. Срок и результаты проекта

Срок реализации — **до трёх месяцев с начала проекта**.

По итогам должны быть предоставлены:

- Работоспособный сервис с БД и интерфейсом аналитика.
- Комплект для развёртывания с использованием Docker.
- Инструкция по установке, настройке, резервному копированию и восстановлению.
- Документация пользователя и описание структуры БД.
- Отчёт об испытаниях с результатами проверки согласованных показателей.