

**КОММЕРЧЕСКОЕ ПРЕДЛОЖЕНИЕ**

**на разработку**

**Информационной системы интеллектуального поиска в корпусе отраслевых текстов**

**ОБЗОР ПРОЕКТА**

Проект включает разработку **Информационной системы интеллектуального поиска в корпусе отраслевых текстов (документов)** (ИС), которая позволяет генерировать текстовые ответы на запросы из корпуса внутренних отраслевых документов. ИС будет работать в закрытом контуре предприятия, обеспечивая безопасность и конфиденциальность данных.

Возможности системы включают:

1. Прием внутренних текстовых документов в формате Word или PDF.
2. Индексацию документов для эффективного поиска.
3. Использование модели обработки естественного языка для генерации текстовых ответов на основе проиндексированных документов.
4. Поддержку документов и запросов только на русском языке.

Проект реализуется в три этапа:

1. **Проектирование системы**
2. **Разработка системы**
3. **Внедрение и обучение**

**ОБЪЕМ УСЛУГ**

**Этап 1: Проектирование системы (2 недели)**

* Разработка архитектуры и прототипов.
* Создание макетов веб-интерфейса, включая модуль «Вопрос-Ответ».
* Выбор подходящих инструментов и фреймворков для реализации.

**Этап 2: Разработка системы (8 недель)**

* Разработка ИС с следующими функциями:
  + Веб-интерфейс для ввода текстовых запросов и отображения ответов.
  + Сервис индексации и переиндексации документов.
  + Интеграция OCR с использованием Tesseract для обработки сканированных PDF-файлов.
  + Развертывание большой языковой модели (LLM) в закрытом контуре предприятия.
* Проведение предварительных испытаний системы.

**Этап 3: Внедрение и обучение (6 недель)**

* Создание руководств для пользователей и администраторов.
* Проведение обучающих сессий для конечных пользователей и администраторов.
* Развертывание ИС на инфраструктуре заказчика.
* Проведение приемочных испытаний и устранение замечаний.

**ТЕХНИЧЕСКОЕ ПРЕДЛОЖЕНИЕ**

**Функциональные возможности системы**

1. **Веб-интерфейс**:
   * Удобный модуль «Вопрос-Ответ» для ввода запросов на естественном языке (до 300 символов).
   * Результаты поиска на основе проиндексированных документов.
   * Сообщения об ошибках для неподдерживаемых запросов или отсутствия ответов.
2. **Обработка документов**:
   * Поддержка форматов Word и PDF.
   * Использование OCR для сканированных PDF-файлов с применением **Tesseract**. Tesseract — это мощный инструмент для распознавания текста на изображениях, который преобразует текстовые данные в машинно-читаемый формат. Это особенно важно для сканированных документов, которые не содержат встроенного текста.
3. **Индексация и поиск**:
   * Предварительная индексация загруженных документов для ускорения ответов.
   * Переиндексация при добавлении новых документов.
   * Использование базы данных с полнотекстовым и векторным поиском. **Полнотекстовый поиск** позволяет находить ключевые слова и фразы в документах, а **векторный поиск** — искать семантически похожие результаты, даже если слова запроса и текста отличаются.
4. **Генерация ответов**:
   * Ответы на естественном языке на основе фрагментов документов.
   * Применение подхода **RAG (Retrieval-Augmented Generation)**. Этот подход сочетает поиск информации с помощью базы данных и генерацию текстов с использованием языковой модели. Это позволяет создавать осмысленные и точные ответы на основе актуальных данных.
   * Использование, вероятнее всего, моделей семейства LLaMa, которые обеспечивают широкий контекст и высокое качество генерации текстов на русском языке. Эти модели адаптированы для работы с корпоративными данными, что гарантирует безопасность и точность.
5. **Языковая поддержка**:
   * Только русский язык для документов и запросов.
6. **Внедрение**:
   * Работа в закрытой сети предприятия.
   * Развертывание на инфраструктуре заказчика.
7. **Сервис управления документами и их переиндексации**:
   * Управление документами, включая их добавление, удаление и обновление.
   * Автоматическая переиндексация для обеспечения актуальности данных. Система автоматически обновляет индексы при внесении изменений, что гарантирует корректность поиска.

**РЕЗУЛЬТАТЫ ПРОЕКТА**

* Полностью функциональная Информационная система интеллектуального поиска.
* Руководства для пользователей и администраторов.
* Обучающие сессии для команды заказчика.
* Развертывание на инфраструктуре заказчика.
* Приемочные испытания и передача документации.

**ГРАФИК РЕАЛИЗАЦИИ**

| Этап | Срок |
| --- | --- |
| Этап 1: Проектирование системы | 2 недели |
| Этап 2: Разработка системы | 8 недель |
| Этап 3: Внедрение и обучение | 6 недель |
| **Итого** | **16 недель** |

**РАСЧЕТ СТОИМОСТИ**

**Этап 1: Проектирование системы**

Команда:

* Тимлид
* Руководитель проекта
* Системный администратор
* UX/UI дизайнер

Стоимость этапа: 1 379 840,00 рублей

**Этап 2: Разработка системы**

Команда:

* Тимлид
* Разработчики
* Инженеры по тестированию
* Руководитель проекта
* Системный администратор

Стоимость этапа: 10 348 800,00 рублей

**Этап 3: Внедрение и обучение**

Команда:

* Разработчики
* Инженеры по тестированию
* Руководитель проекта
* Системный администратор

Стоимость этапа: 3 363 360,00 рублей

**ОБЩАЯ СТОИМОСТЬ ПРОЕКТА: 15 092 000,00 рублей**

* Окончательная стоимость может измениться в зависимости от изменений объема работ.
* Все цены указаны без учета применимых налогов и сборов.