**Эссе о проекте: от идеи до реализации**

Наш проект представляет собой комплексное решение, интегрирующее современные технологии для обработки и поиска данных. Его цель — создать систему, которая преобразует PDF-документы и изображения в векторное представление, используя модель ColPali, а затем предоставляет возможность искать по этим данным с помощью методов семантического поиска и RAG (Retrieval-Augmented Generation). Архитектура проекта включает фронтенд на Vue.js, бэкенд на FastAPI, сервис работы с языковыми моделями (LLM), а также preprocessing service для подготовки данных.

**Идея и цель проекта**

Основной задачей было разработать инструмент, который позволит организациям эффективно управлять документами и находить нужную информацию даже в сложных контекстах. Например, пользователь может искать документы, содержащие упоминания определённых концепций, даже если те выражены синонимами или описаны в косвенной форме. Кроме того, система должна справляться как с текстовыми документами, так и с изображениями.

**Процесс разработки**

**1. Архитектурное проектирование**

Для достижения целей мы выбрали модульный подход:

* **Фронтенд на Vue.js** отвечает за пользовательский интерфейс. Здесь реализована возможность отправлять запросы, просматривать результаты поиска и работать с файлами.
* **Бэкенд на FastAPI** обрабатывает запросы от фронтенда и взаимодействует с другими сервисами.
* **Preprocessing Service** выполняет ключевую роль в обработке данных, преобразуя документы и изображения в векторное представление.
* **LLM-сервис (Llama 3.1)** позволяет выполнять семантический поиск и уточнять результаты запросов.

На этапе проектирования мы уделили особое внимание интеграции всех компонентов, чтобы минимизировать точки отказа и обеспечить масштабируемость системы.

**2. Реализация ключевых модулей**

* **Preprocessing Service**: Основным вызовом здесь была работа с изображениями и большими PDF-документами. Мы интегрировали модель ColPali для получения векторных представлений с размерностью 128, оптимизировав процесс батчевой обработки для повышения производительности.
* **Семантический поиск**: Для обеспечения релевантности результатов мы подключили Llama 3.1, которая использует векторный индекс, созданный через RAG. Это позволило объединить классический поиск по ключевым словам с семантическим анализом текста.
* **Фронтенд**: Vue.js был выбран за его гибкость и возможность создавать быстрые, отзывчивые интерфейсы. Мы разработали простой и понятный UI, обеспечивающий удобство работы с системой.

**Сложности и их преодоление**

Разработка проекта не обошлась без вызовов, которые потребовали от нас глубокой проработки решений:

1. **Оптимизация производительности**: Одной из главных проблем стало время обработки больших объемов данных, особенно PDF с сотнями страниц. Использование ColPali в связке с оптимизированными библиотеками, такими как PyTorch и NumPy, PyPDF2, позволило существенно сократить время на обработку.
2. **Интеграция LLM и RAG**: Совмещение семантического поиска с извлечением информации из индекса потребовало тонкой настройки. Мы экспериментировали с различными параметрами модели Llama 3.1 и векторами, чтобы добиться наилучшей релевантности.
3. **Работа с ограничениями инфраструктуры**: На этапе тестирования мы столкнулись с нехваткой ресурсов для работы всех сервисов. В качестве решения была внедрена контейнеризация через Docker, что упростило развертывание и управление инфраструктурой.
4. **Обработка изображений**: Преобразование графических данных в векторное представление потребовало подключения дополнительных библиотек для предварительной обработки (ColPali, byaldi) и решения вопросов с качеством.

**Результаты и перспективы**

На текущий момент проект позволяет эффективно находить информацию в документах по смыслу, обеспечивая высокую точность поиска. Мы видим потенциал для дальнейшего развития:

* Добавление поддержки мультиязычных моделей для работы с текстами на разных языках.
* Интеграция новых модулей для более сложной аналитики.
* Масштабирование системы для обработки данных в реальном времени.

Проект дал нам бесценный опыт работы с современными технологиями, показал, насколько важна слаженная командная работа и тщательное планирование. Мы гордимся достигнутыми результатами и уверены, что это только начало большого пути.