Факультет компьютерных наук

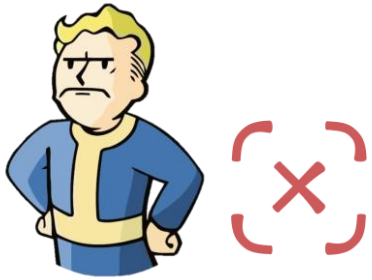
Машинное обучение
и высоконагруженные системы

Июнь, 2025

Система мультимодального анализа и поиска видеоконтента по пользовательским запросам

Владислав Панфиленко
студент МОВС, 2 курс

Преамбула



Если вас тоже это раздражает:

- Недавно смотрели смешное TikTok, Shorts, Reels, но сейчас не можете его найти;
- У вас есть идеальный короткий видеофрагмент для презентации, только не помните в каком из сотни видео он находится;
- В вашем архиве хранится множество коротких видео из отпуска, но не можете быстро найти те, где вы были на пляже.

То данная система решает подобные проблемы с помощью «умного поиска», который понимает и видит, что происходит в ваших видео!

Цель и задачи

Цель: разработка системы мультимодального анализа и поиска видео с использованием векторной базы данных Qdrant, способной предоставлять контекстно-релевантные результаты на запросы пользователей.

Задачи:

- Найти датасет для экспериментов;
- Протестировать различные модели/подходы создания/поиска по эмбеддингам;
- Разработать отдельные модули системы;
- Объединить компоненты в единую систему.

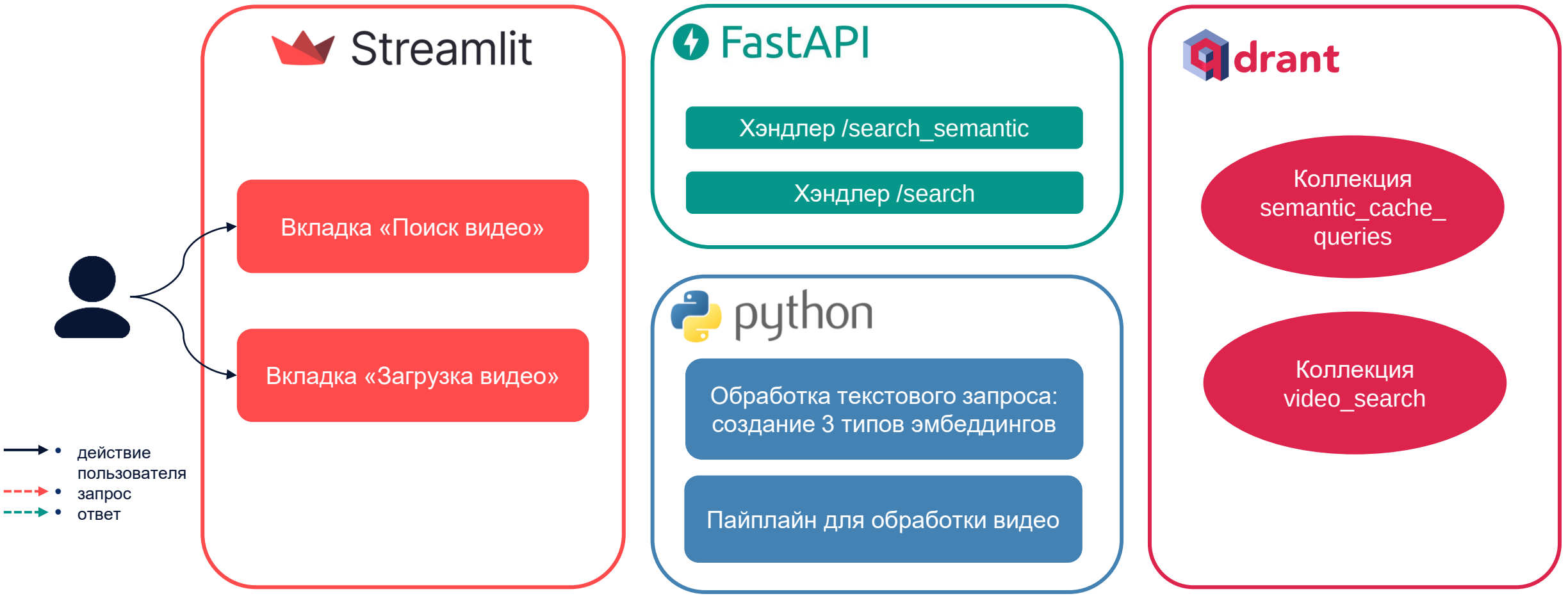
Ключевые особенности системы

- **Мультимодальный подход**: анализ видео и аудио содержимого
- **Векторная база** данных для быстрого поиска
- **Удобный веб-интерфейс** для поиска и загрузки новых видео
- Полностью контейнеризированное решение

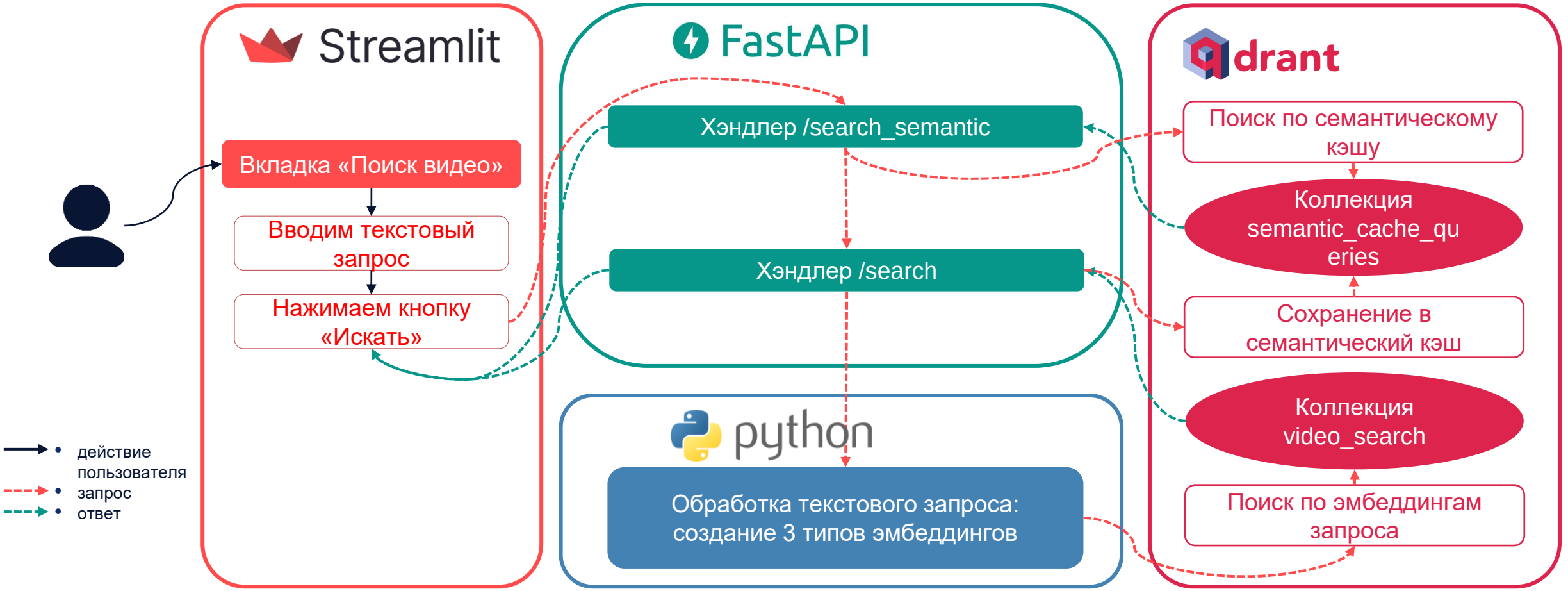
Преимущества мультимодального подхода

- Повышенная точность поиска: анализ как визуальной, так и текстовой информации
- Понимание контекста: системы способна понимать смысл запроса, а не только ключевые слова
- Гибкость запросов: возможность искать видео даже при отсутствии точных текстовых совпадений
- Улучшенное ранжирование: более релевантные результаты благодаря комбинированию разных типов эмбеддингов

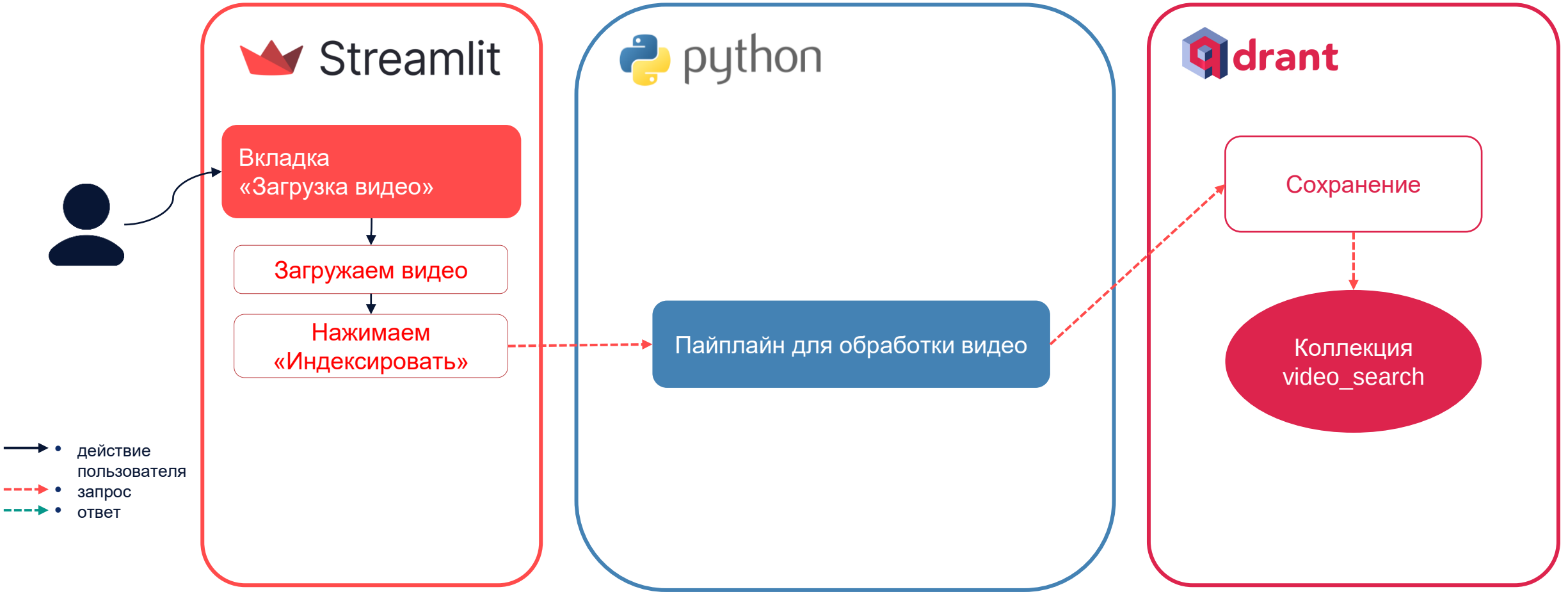
Архитектура системы



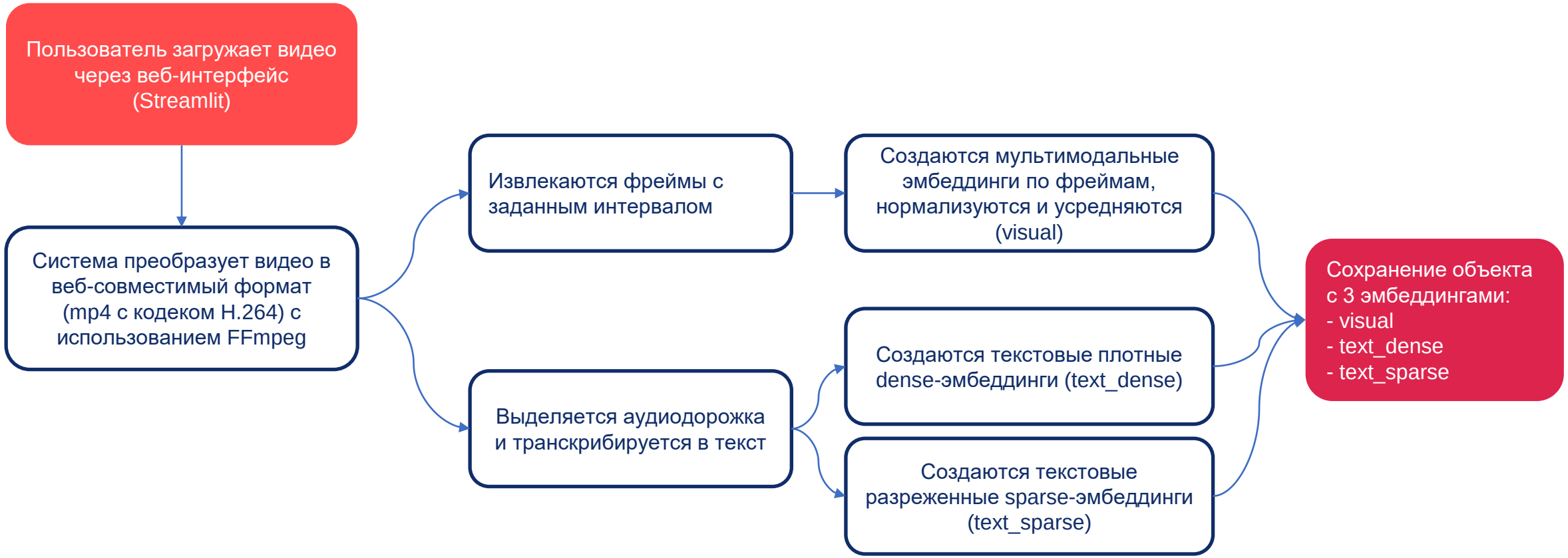
Архитектура системы – клиентский путь поиска



Архитектура системы – клиентский путь загрузки



Пайплайн обработки видео при работе с векторной БД



Формирование датасета

1. Для экспериментов был использован датасет bytedance/Shot2Story в котором уже были размеченные видеоролики.
2. Из них были отобраны случайным образом **172 видеоролика** продолжительностью **до 30 секунд** из различных категорий.
3. Из отобранных случайным образом было выбрано **100 видео,** к которым **вручную** были сделаны **запросы**, таким образом чтобы каждый запрос был **наиболее релевантным к конкретному видео**.

В дальнейшем на основе отобранных видео и написанных вручную запросов проводил эксперименты разных моделей/подходов и оценивал их качество для релевантности выдачи в поиске.

Проведение экспериментов: 1-2 итерации

Для оценки качества моделей на разных этапах экспериментов использовались метрики:

- recall@k – доля **релевантных** видео, которые попали в топ-k результатов поиска
- MRR@k - средняя позиция **первого** релевантного видео в результатах поиска по топ-k.

Из подходов ранжирования результатов поиска использовались:

- **RRF** (Reciprocal Rank Fusion)
- **DBSF** (Distributed Biased Score Fusion)

1. Мультимодальные эмбеддинги (фреймы)

Модель	Recall@3	MRR@3	Время обработки на одно видео, сек.
openai_clip_vit_base_patch32	0,7500	0,6633	0,0108
laion_CLIP_ViT_B_32_laion2B_s34B_b79K	0,7500	0,6725	0,0083
google_siglip_base_patch16_224	0,5067	0,4517	0,0076
microsoft_git_base	0,3850	0,3408	0,0075
facebook_flava_full	0,3160	0,2780	0,0102
openai_clip_vit_large_patch14	0,3933	0,3561	0,0088

2. Транскрибация звука в текст

Модель	Среднее BLEU	Среднее WER	Средняя косинусная близость	Время обработки на одно видео, сек.
faster_whisper_medium	0,7678	0,2969	0,8945	12,46
openai_whisper_large_v3	0,8081	0,4556	0,9052	21,09
faster_whisper_base	0,7094	0,5949	0,8581	1,77
facebook_wav2vec2_base_960h	0,4837	0,6925	0,7242	0,29

Проведение экспериментов: 3-4 итерации

3. Текстовые dense-эмбеддинги

Модель	Recall@3	MRR@3
sentence-transformers/all-MiniLM-L6-v2	0,7500	0,7167
mixedbread-ai/mxbai-embed-large-v1	0,7750	0,7317
nomic-ai/nomic-embed-text-v1.5	0,7700	0,7333
Snowflake/snowflake-arctic-embed-l	0,7150	0,6696
BAAI/bge-large-en-v1.5	0,7240	0,6810
thenlper/gte-large	0,7400	0,6983

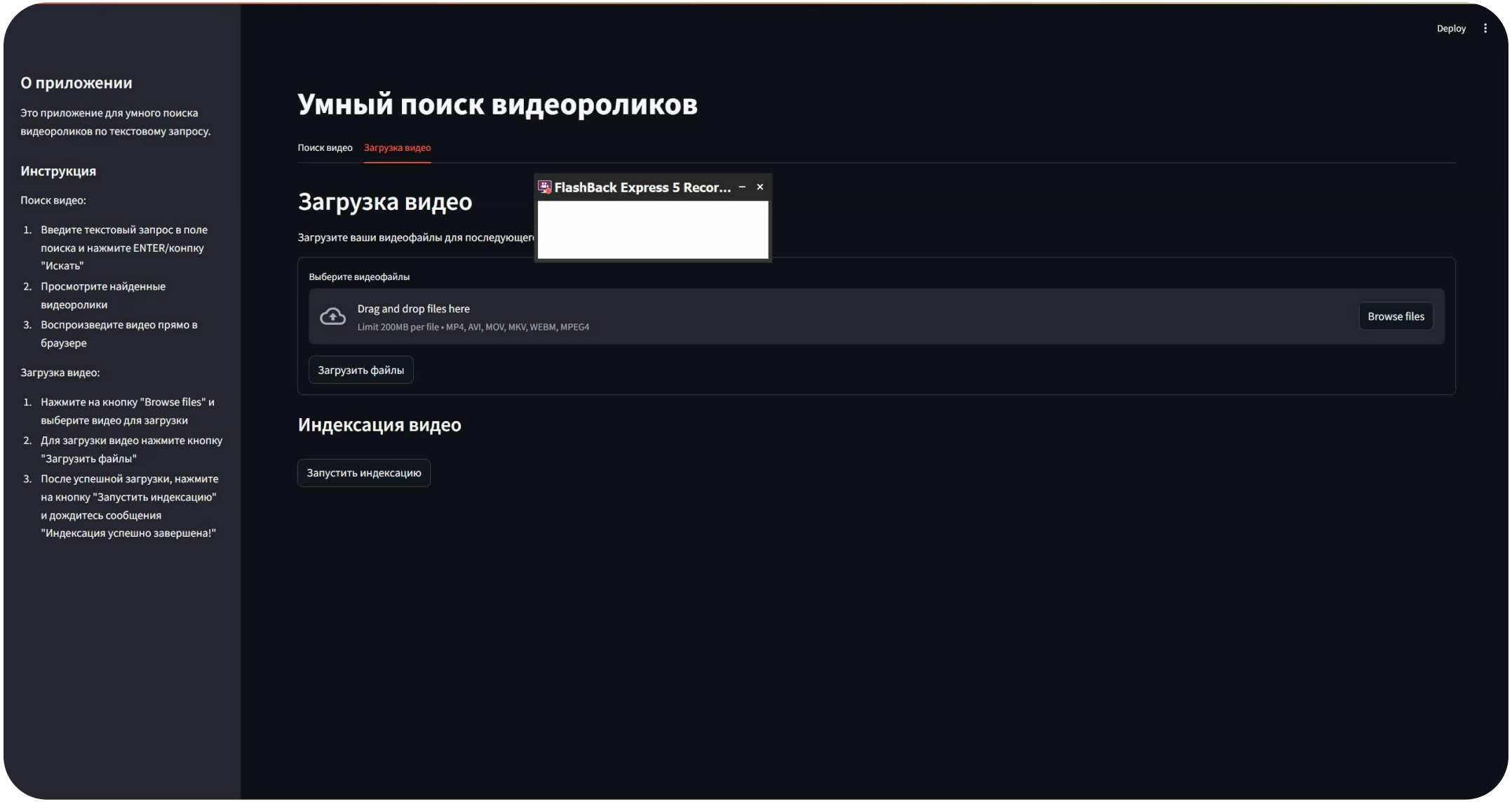
4. Текстовые sparse-эмбеддинги

Модели/подходы к ранжированию	Recall@3	MRR@3
dense (mixedbread-ai/mxbai-embed-large-v1)	0,8000	0,7467
dense (mixedbread-ai/mxbai-embed-large-v1) + fastembed sparse (Qdrant/bm25) RRF	0,7850	0,7358
dense (mixedbread-ai/mxbai-embed-large-v1) + fastembed sparse (Qdrant/bm25) DBSF	0,7833	0,7389
dense (mixedbread-ai/mxbai-embed-large-v1) + splade sparse (naver/splade-cocondenser-ensembledistil) RRF	0,7825	0,7308
dense (mixedbread-ai/mxbai-embed-large-v1) + splade sparse (naver/splade-cocondenser-ensembledistil) DBSF	0,7800	0,7250

Проведение экспериментов: 5 итерация

5. Финальная итерация

Финальная конфигурация	Recall@3	Recall@5	Recall@10	MRR@3	MRR@5	MRR@10
visual multimodal + dense RRF	0.83	0.83	0.85	0.7717	0.7517	0.7748
visual multimodal + dense DBSF	0.83	0.83	0.85	0.7658	0.7567	0.7723
visual multimodal + dense + sparse RRF	0.83	0.83	0.85	0.7778	0.7717	0.7845
visual multimodal + dense + sparse DBSF	0.83	0.83	0.85	0.7837	0.7792	0.7906


Deploy
О приложении
Это приложение для умного поиска видеороликов по текстовому запросу.
Инструкция
Поиск видео:
1. Введите текстовый запрос в поле поиска и нажмите ENTER/конпку "Искать"
2. Просмотрите найденные видеоролики
3. Воспроизведите видео прямо в браузере
Загрузка видео:
1. Нажмите на кнопку "Browse files" и выберите видео для загрузки
2. Для загрузки видео нажмите кнопку "Загрузить файлы"
3. После успешной загрузки, нажмите на кнопку "Запустить индексацию" и дождитесь сообщения "Индексация успешно завершена!"
Умный поиск видеороликов
Поиск видео
Загрузка видео
Загрузка видео
FlashBack Express 5 Recor...
Загрузите ваши видеофайлы для последующег
Выберите видеофайлы
Drag and drop files here
Limit 200MB per file • MP4, AVI, MOV, MKV, WEBM, MPEG4
Browse files
Загрузить файлы
Индексация видео
Запустить индексацию

Спасибо за внимание!