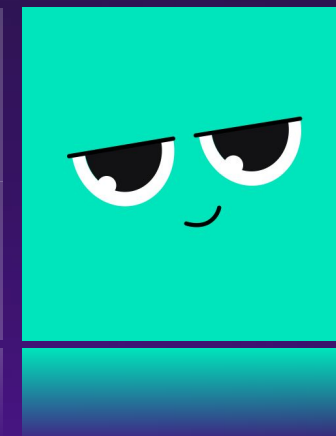




Yappu

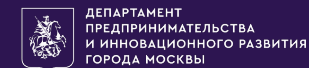
Команда **GO_HACK**

Задача №15 - Сервис текстового поиска по медиаконтенту





Команда GO_HACK



**Александр
Валясин**

- MLOps
- GlowByte (Sber)
- @alexander_zxcc



**Кирилл
Богатырёв**

- Backend developer
- ex-Yandex
- @fizzzzgen



**Богдан
Онищенко**

- Data Scientist
- Sber
- @yourbg000



**Никита
Молчанов**

- Data Scientist
- Sber
- @lusm554



**Денис
Самаркин**

- Data Engineer, Data Scientist
- Sber
- @DenisSamarkin





Уникальность решения:



ПРОЕКТ
МЭРА
МОСКВЫ



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ

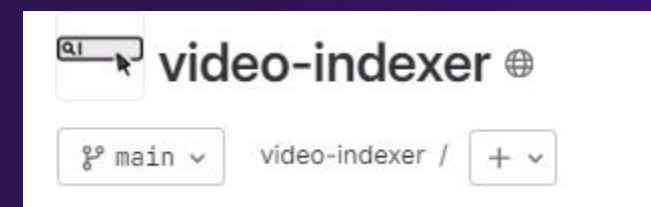


АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ

Решение **GO_HACK**:

Комбинирует визуальный и AI-текстовый поиск:

- 1 Визуальный поиск по самому часто встречающейся сцене на видео.
- 2 Текстовый — ищет по ключевым словам в описании, транскрибации аудио, по надписям/субтитрам и символам в видео.
- 3 Это позволяет быстро находить наиболее релевантные видео по запросу.



Подробнее: gitlab.com/fizzzzgen/video-indexer



ПРОЕКТ
МЭРА
МОСКВЫ

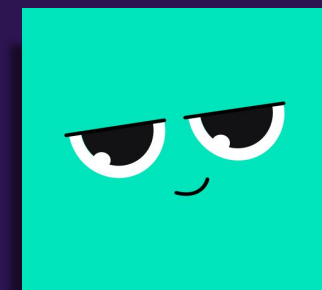


ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ

GO_DEMO?



0.2511475086212158s. 0.07930827140808105s. on vector search. 0.1477830410003662s. on embedding.

Text & Audio



Distance:
0.23094600439071655

Мне нравится спорт, спорт, мотивация, тренировка и мотивация.



Distance:
0.3248252868652344

Вот уже пять дней о спорте, растяжке, спорте и мотивации.



Distance:

Visual



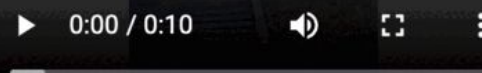
Distance:
0.6183792948722839

В этом блоге мы собрали лучшие новости о спорте.



Distance:
0.61963951587677

В этом блоге мы собрали тысячи цитат, которые хотели бы услышать сами.



Distance:



Скорость работы:




ПРОЕКТ
МЭРА
МОСКВЫ



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ

 Flower Workers Tasks Broker Documentation

Show tasks

Name	UUID	State	args
app.video_2_text	9f7e9a35-b7d8-4813-89d4-4b0be7ebb4a1	SUCCESS	a/0f/48/8

Showing 1 to 1 of 1 tasks

api/index

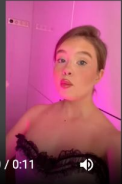
avg: 35 sec

py & GO_HACK team

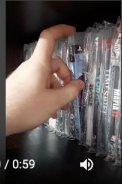
моргенштерн

Поиск


0.23342156410217285s. 0.06386804580688477s. on vector search. 0.14606046676635742s. on embedding.




Distance:
0.4290253520011902
В этом блоге мы собрали лучшие новости о свадьбе Моргенштерна и Дилара.



Distance:
0.515743613243103
Олег Здравое и его коллега Персик - настоящий фанат зомби.



Distance:
0.5565165281295776
В этом блоге мы собрали лучшие фильмы и сериалы, которые вы сможете посмотреть на канале Би-Си-си.



Distance:
0.5816481113433838
У Львов - одна из самых популярных соцсетей.

api/search

avg: 0.3 sec

32 CPU + SSD
NO GPU



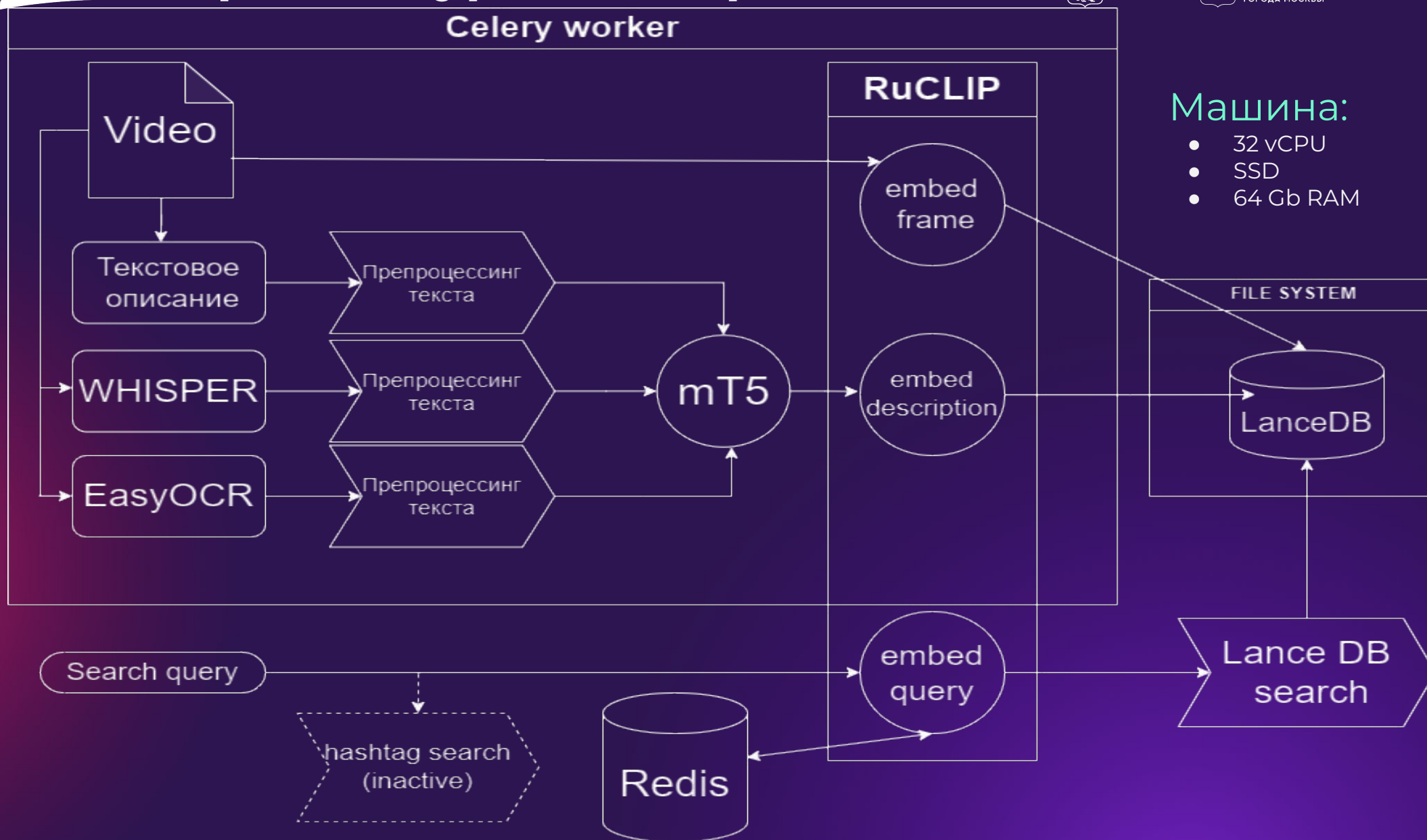
Архитектура веб-сервиса



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ





ML-инструменты:



ПРОЕКТ
МЭРА
МОСКВЫ



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ

01 **mT5_multilingual_XLSum**

Многоязычная модель для автоматического суммирования текстов. Суммируем текст на EN + RU.

02 **Ruclip-vit-base-patch16-384**

Используем для генерации эмбеддингов из текстов и изображений с одинаковым размером. Из-за широкого применения модели, одновременно не перегружаем архитектуру и получаем когерентные эмбеддинги.

03 **EasyOCR – cyrillic_g2**

Модель OCR для распознавания кириллического текста из изображений. После оптимизации показала лучшие результаты.

EasyOCR – CRAFT

Модель для детекции текстов в изображениях, отдельных символов и текстовых областей.

04 **Whisper**

Модель для автоматического распознавания речи в текст. Одна модель для распознавания, перевода, определения языка. Также работаем в основном на RU+EN.



ПРОЕКТ
МЭРА
МОСКВЫ



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



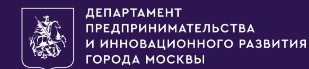
АГЕНТСТВО
ИННОВАЦИИ
МОСКВЫ



почему



Почему T5:



Model	XNLI	PAWS-X	WikiAnn-NER	XQuAD	MLQA	TyDiQA-GoldP
mBERT	65.4	81.9	62.2	64.5	61.4	59.7
XLM	69.1	80.9	61.2	59.8	48.5	43.6
InfoXLM	81.4	-	-	-	73.6	-
X-STILTs	80.4	87.7	64.7	77.2	72.3	76.0
XLM-R	79.2	86.4	65.4	76.6	71.6	65.1
VECO	79.9	88.7	65.7	77.3	71.7	67.6
RemBERT	80.8	87.5	70.1	79.6	73.1	77.0
mT5-Small	67.5	82.4	50.5	58.1	54.6	36.4
mT5-Base	75.4	86.4	55.7	67.0	64.6	59.1
mT5-Large	81.1	88.9	58.5	77.8	71.2	68.4
mT5-XL	82.9	89.6	65.5	79.5	73.5	77.8
mT5-XXL	<u>85.0</u>	<u>90.0</u>	<u>69.2</u>	<u>82.5</u>	<u>76.0</u>	<u>82.0</u>



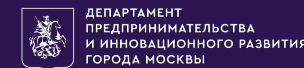
Почему RUCLIP:



Dataset	ruCLIP Base [vit-base-patch32-224]	ruCLIP Base [vit-base-patch16-224]	ruCLIP Large [vit-large-patch14-224]	ruCLIP Base [vit-base-patch32-384]	ruCLIP Large [vit-large-patch14-336]	ruCLIP Base [vit-base-patch16-384]	CLIP [vit-base-patch16-224] original
Food101	0.765	0.827	0.840	0.851	0.896	0.890	0.901
CIFAR10	0.917	0.922	0.927	0.934	0.943	0.942	0.953
CIFAR100	0.716	0.739	0.734	0.745	0.770	0.773	0.808
Birdsnap	0.347	0.503	0.567	0.434	0.609	0.612	0.664
SUN397	0.683	0.721	0.731	0.721	0.759	0.758	0.777
Stanford Cars	0.697	0.776	0.797	0.766	0.831	0.840	0.866
DTD	0.690	0.734	0.711	0.703	0.731	0.749	0.770
MNIST	0.963	0.974	0.949	0.965	0.949	0.971	0.989
STL10	0.957	0.962	0.973	0.968	0.981	0.974	0.982
PCam	0.827	0.823	0.791	0.835	0.807	0.846	0.830
CLEVR	0.356	0.360	0.358	0.308	0.318	0.378	0.604



Почему EasyOCR:



Библиотека / критерий	Работает из коробки*	Распознавание ru / en	Скорость на CPU	Предобработка изображений*	Лицензия
<u>EasyOCR</u>	Да	<u>Хорошее</u>	<u>Быстрая</u>	<u>Минимальная</u>	<u>Apache-2.0</u>
Tesseract OCR	Нет	Хорошее	Быстрая	Тщательная настройка	Apache-2.0
PaddleOCR	Нет	Умеренное. Нацелена на восточноазиатские языки	Умеренная	Умеренная	Apache-2.0
Keras-OCR	Да	Хорошее	Требуется настройки	Умеренная	MIT

*Работает из коробки - легко устанавливается, настраивается, доступная документация, имеет Python API.

*Предобработка изображений - требуется минимальная предобработка изображений для распознавания разных шрифтов, цветов на разных фонах.



Почему EasyOCR: заключение



ПРОЕКТ
МЭРА
МОСКВЫ



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ

EasyOCR:

1. Работает из коробки, минимальная настройка
2. Качественно распознает и обнаруживает символы на ru, en
3. Удобно настраивается в соотношении качество - скорость
4. В разы ускоряется на GPU
5. Доступная лицензия, open source, популярный и поддерживаемый проект

Почему не другие:

1. PaddleOCR - необходима настройка, нацелен преимущественно на восточноазиатские языки
2. Tesseract OCR - требует тщательной настройки, подходит для более глубокого контроля над процессом распознавания
3. Keras-OCR - отличный выбор для работы с Keras и TensorFlow, подходит больше для дообучения моделей
4. Облачные сервисы Google Cloud Vision OCR, AWS Textract, OCR.Space - сторонние API, закрытый код, платная подписка, ПО недружественных стран



Почему Whisper:



ПРОЕКТ
МЭРА
МОСКВЫ



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ

Библиотека / критерий	Легка в использовании	Распознавание ru / en	Скорость на CPU	Предобработка видеофайлов*	Распознает музыку
<u>Whisper</u>	<u>Да</u>	<u>Отличное</u>	<u>Быстрая</u>	<u>Не требуется</u>	<u>Да</u>
speechbrain	Нет	Хорошее	Быстрая	Требуется WAV	Нет
Google Speech-to-Text	Нет	Умеренное.	Средняя	Требуется WAV	Нет
Amazon Transcribe	Нет	Умеренное	Требует настройки	Требуется WAV	Нет



Почему LanceDB:



ПРОЕКТ
МЭРА
МОСКВЫ



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ

Сводка результатов по 10 000 случайных запросов.

Case	Elasticsearch (QPS)	LanceDB (QPS)
FTS: Serial	399.8	<u>468.9</u>
FTS: Concurrent	<u>1539.0</u>	528.9
Vector search: Serial	11.9	<u>54.0</u>
Vector search: Concurrent	50.7	<u>71.6</u>

1. Среди всех конкурентов HNSW/RHNSW (PQ), IVF и FLAT **IVF-PQ** дает **максимальный прирост скорости**, но меньший recall при увеличении числа векторов(**). Вычисление этого индекса можно ускорить на **GPU** и **тонко настроить его/поиск**.
2. **Единственная БД с IVF-PQ.**
3. Единственная БД у которой **все индексы disk-based (*) + zero-copy data access(*)**.
4. Встроенная (**бессерверная**) специализированная архитектура, созданная с нуля. Нужна **минимальная** настройка.



LanceDB

Источники: thedataquarry.com/posts/vector-db-3/ (*), thesequence.substack.com/p/guest-post-choosing-the-right-vector (**), github.com/prrao87/lancedb-study



Скорость поиска на 1М датасете



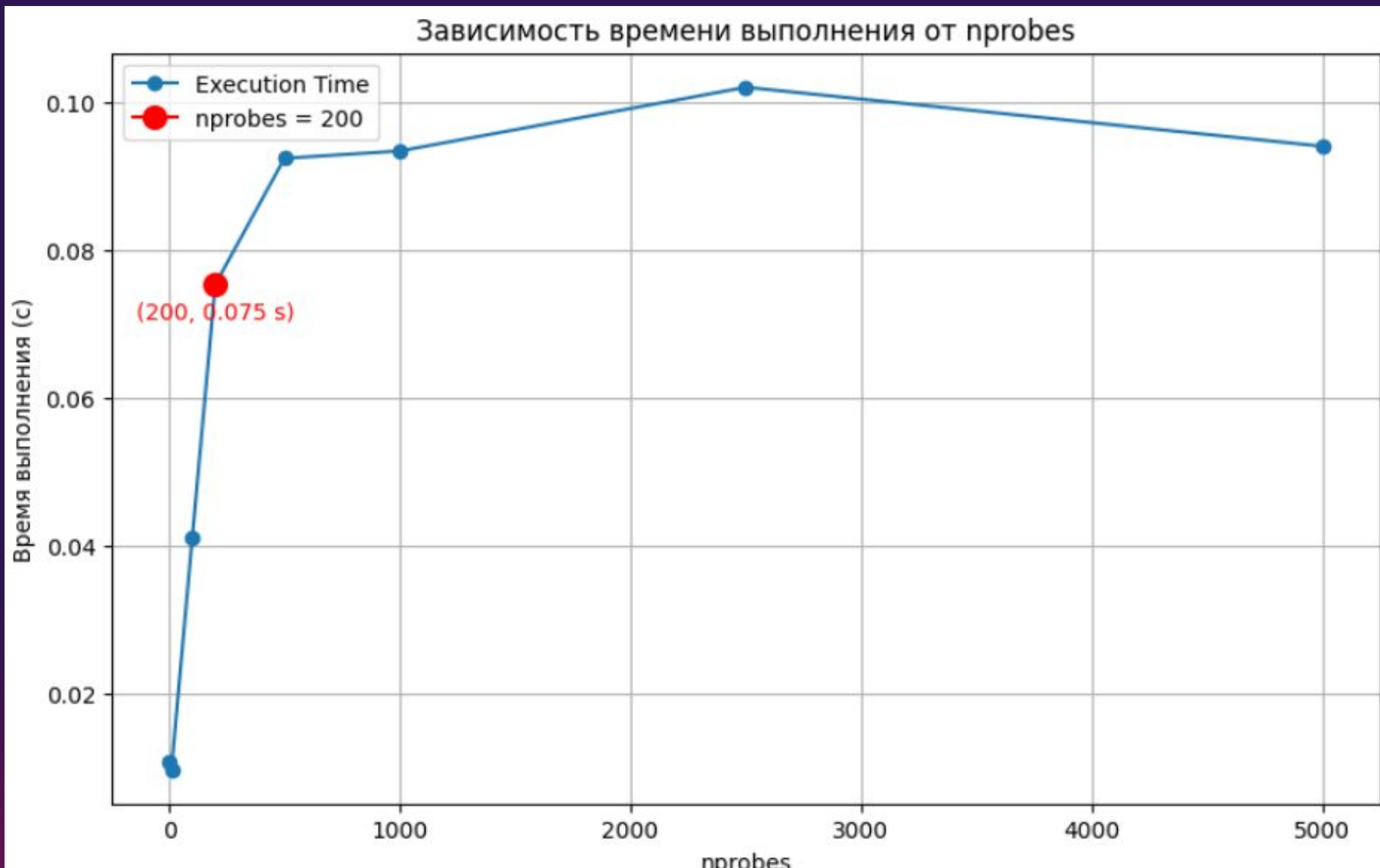
ПРОЕКТ
МЭРА
МОСКВЫ



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ



Без индекса -
1.98s

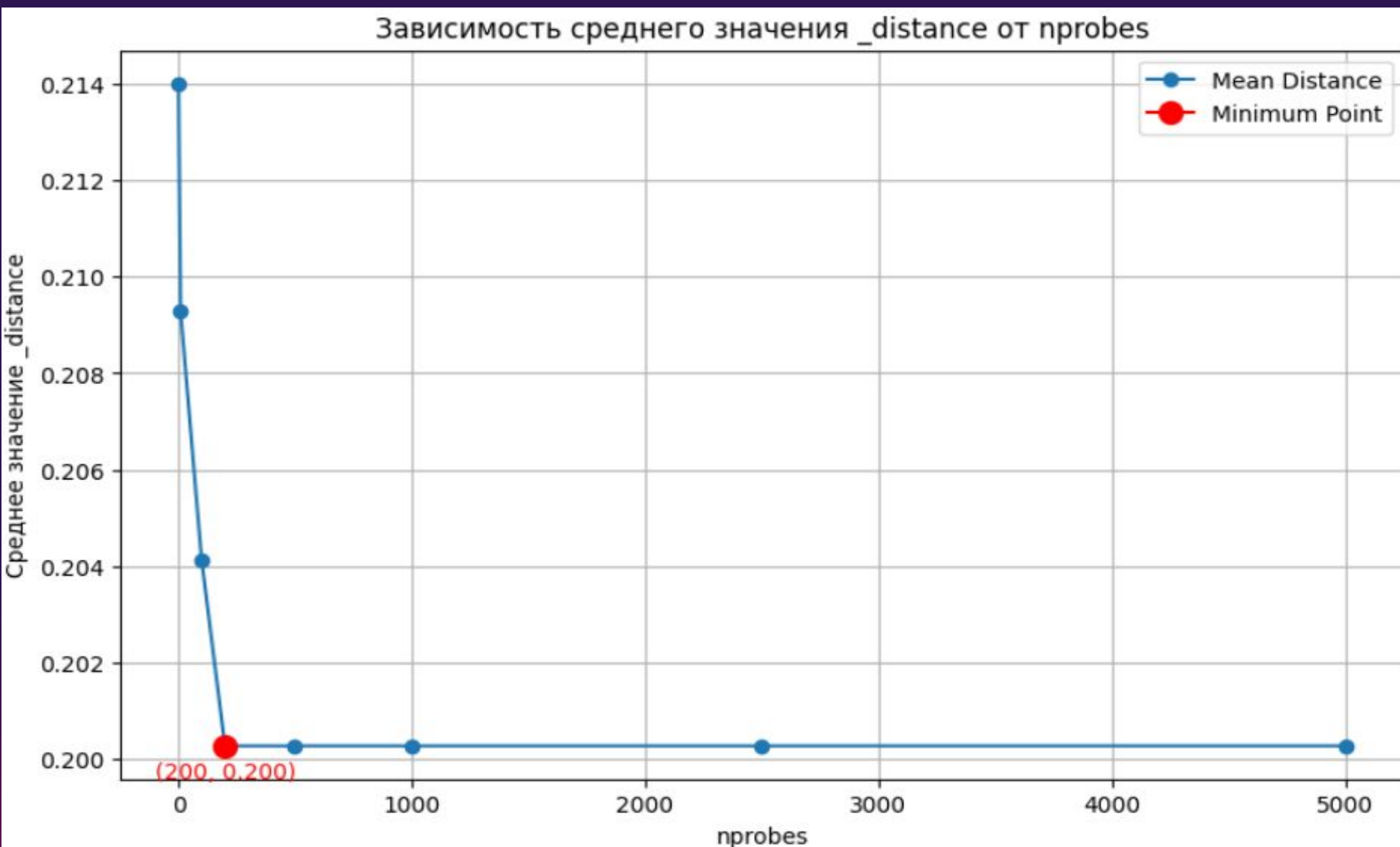
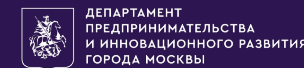
С индексом -
200 nprobes 0.075s

Создание индекса
04 min 17 sec
("CUDA")

Код: gitlab.com/fizzzgen/video-indexer/-/blob/main/test/DB_testing.ipynb



Качество поиска на 1М датасете



refine_factor = 10
Датасет 1М
Размерность 512
T4 GPU
Индекс IVF-PQ



Достоинства веб-сервиса:



ПРОЕКТ
МЭРА
МОСКВЫ



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ

Сэмплирование:

Выбираем самый репрезентативный фрейм по косинусному расстоянию.

В итоге обрабатываем всего 1 кадр, слабо теряя в качестве.

Высокая скорость:

Индексация в ~10x раз быстрее, поиск в ~2x раз быстрее регламента на посредственном железе без GPU.

Раздвоенный поиск:

Ищем одновременно по AI-тексту и содержимому в видео.

Это дало лучшие результаты по сравнению с комбинированными векторами.

Повышенная точность:

Суммаризируем весь текст, помимо препроцессинга для однородности. Получившиеся эмбединги получаются когерентными из-за одной CLIP-модели.

Автономность:

Redis для кэширование эмбедингов запросов.

Независим от внешних API.

Все модели установлены локально и их легко заменить на аналоги в пайплайне.

Масштабируемость:

Целевой язык - русский. Модели зафайнтюнены на нем. Также модели хорошо работают с английским и других языках.

Решение можно значительно ускорить используя GPU: все модели и индекс IVF-PQ перенести на GPU.



Недостатки веб-сервиса:



ПРОЕКТ
МЭРА
МОСКВЫ



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ

Отсутствие ранжирования:

Отсутствие шаффла между визуальными и текстовыми эмбедами. Мы **предлагаем добавить в пайплайн еще одну модель в начало для классификации запроса** - визуального или текстового. Потенциально это может ускорить сервис.

Похожим решением будет добавить модель в самый конец пайплайна для шаффла результатов.

Хранение без репликации:

В LanceDB оптимизирована для работы на диске и без сервера. **Предлагаем использовать S3 для хранения** и на операции write. Для операций read предлагаем использовать **реплику** LanceDB в файловой системе. LanceDB **легко интегрируется** в облако, но нет собственных блокировок.

Сложный контент:

Мы берем 1 кадр из содержания видео и 7 для OCR. “Тяжелые” мультимодальные модели выдадут лучший результат для видео, в котором **много разнородных эвентов и они все важны для понимания происходящего на видео.**

Взвешенный поиск:

Размывается смысл в итоговом векторе, если складывать AI-текст и содержимое воедино. **Веса не помогли. Предлагаем оставить в продакшене FTS поиск по описаниям и включить его в шаффл-модель/модель классификации.**



Спасибо за внимание.
~Команда **GO_HACK.**

