

Построение вопросно-ответной системы с использованием RAG

команда 55

КУРАТОР:

ГЕОРГИЙ ПАНЧУК TG: [@JMZZOMG](#) GITHUB: [JOEIN](#)

КОМАНДА:

ЕВГЕНИЙ ЯКОВЕНКО – [@YAKOVENKO_EVGENII](#) GITHUB: [YAKOVENKO96](#)

ЛЮДМИЛА ТЕПЛОВА – [@LTEPLOVA](#) GITHUB: [TEPLOVA](#)

АЛЬБЕРТ ТАЙЧИНОВ – [@TAYAR902](#) GITHUB: [TAYAR902](#)

АЛЕКСЕЙ ЯТКОВСКИЙ – [@BLACKR_ORIGINAL](#) GITHUB: [ALEKSEI-IA](#)

ЧТО ТАКОЕ RAG СИСТЕМА



ЦЕЛИ И ЗАДАЧИ

Цель:

- Создать вопросно-ответную RAG систему, способную быстро извлекать документы из БД и на их основе генерировать текст ответа




Задачи:

- Реализовать Retrieval часть
- Реализовать генерацию ответа с помощью LLM на основе релевантных документов, полученных из БД
- Разработка backend для взаимодействия с моделями
- Разработка web-приложения для взаимодействия с пользователем



ДАННЫЕ

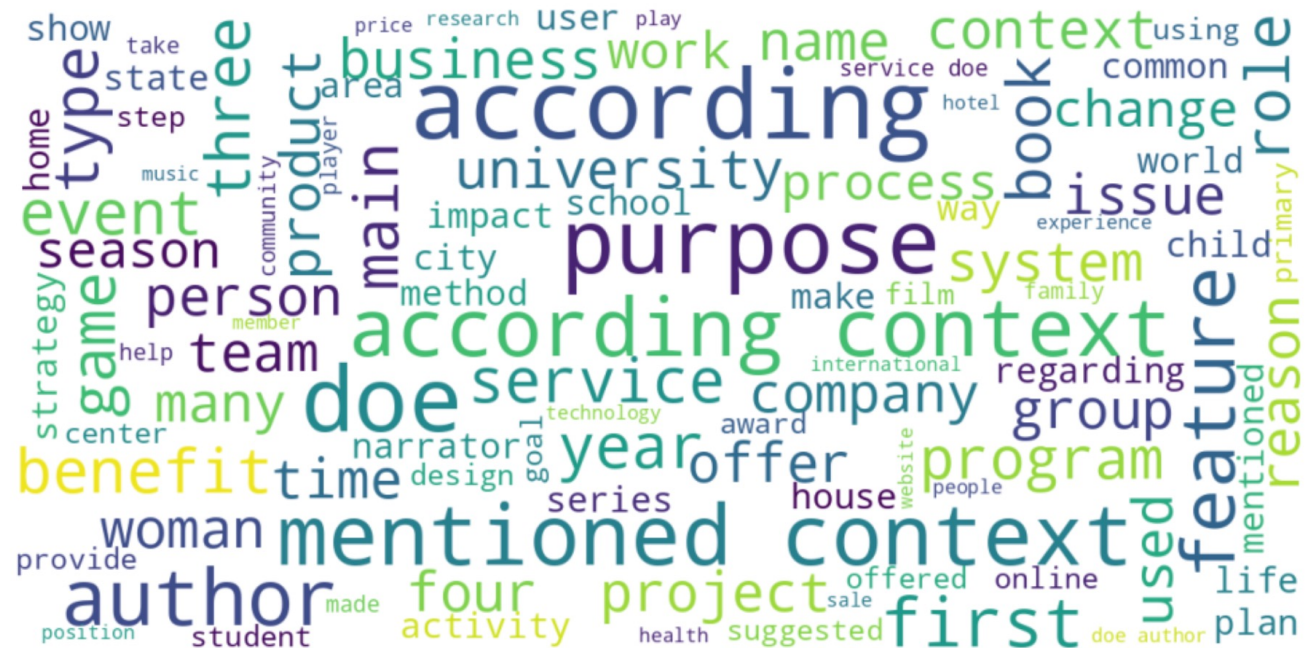
- **Датасет** из huggingface - [neural-bridge/rag-dataset-12000](#)
- **Количество записей:** 12 000
- **Назначение:** обучение/валидация RAG-систем
- **Структура:** вопрос, релевантный документ, ответ
- **Цель использования:** оценка способности находить нужный документ по запросу и правильно генерировать ответ
- **Предобработка:** удаление стоп-слов, дубликатов, текстов на других языках, приведение к нижнему регистру

context	question	answer
string · lengths	string · lengths	string · lengths
		
2.94k↔3.83k 17.5%	48↔79 47.5%	243↔484 25.6%
<p>Caption: Tasmanian berry grower Nic Hansen showing Macau chef Antimo Merone around his property as part of export engagement activities.</p> <p>THE RISE and rise of the Australian strawberry, raspberry and blackberry industries has seen the sectors redouble their international trade focus, with the release of a dedicated export plan to grow their global presence over the next 10 years.</p> <p>Driven by significant grower input, the Berry Export Summary 2028 maps the</p>	<p>What is the Berry Export Summary 2028 and what is its purpose?</p>	<p>The Berry Export Summary 2028 is a dedicated export plan for the Australian strawberry, raspberry, and blackberry industries. It maps the sectors' current position, where they want to be, high-opportunity markets, and next steps. The purpose of this plan is to grow their global presence over the next 10 years.</p>

EDA

- Проверены размеры и структура данных, выявлены и устранены дубликаты и пропуски
- Выявлено равномерное покрытие различных тем и отсутствие существенного перекоса в выборке
- Результаты EDA подтверждают пригодность данных для последующего обучения моделей на основе RAG

Облако слов колонка question



ML ПОДХОД

Выбор способа векторизации

Протестированы разные эмбеддинги и замерена точность (для каждого вопроса выполняется поиск по близости и сравнение с оригинальным ответом по индексу)

Выбор хранилища данных

Сравнение Pandas и Qdrant (датасет был обогащен до 100 тыс строк и произведены замеры извлечения ответов), лучшее время (0.007с) с использованием Qdrant

Выбор метрики

Dot product / Cosine similarity / Euclidean distance

Базовая модель

Tf-Idf, БД Qdrant, Dot product, Precision - 85.9%

Модель	Точность
Word2Vec	28.5
Fasttext	36.6
glove-twitter-100	13
all-MiniLM-L6-v2	75
tf-idf	85.9

ВЫБОР АРХИТЕКТУРЫ ДЛЯ RETRIEVAL, ТОЧНОСТЬ

Бенчмарк

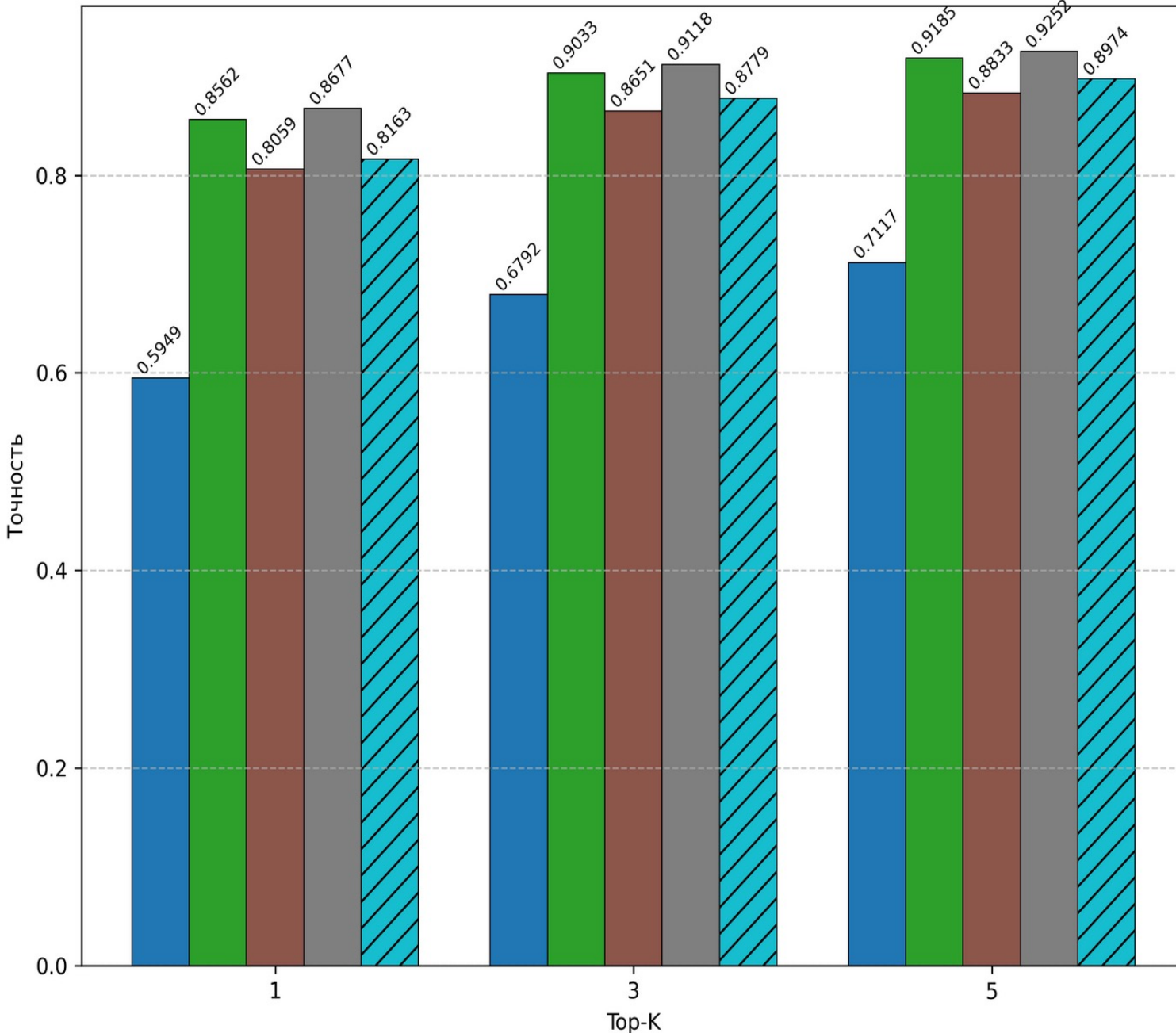
- релевантность извлеченных текстов

$$\text{HitRate}@k = \frac{H_k}{N}$$

H_k - число ответов с релевантным документом в topk

N — общее число запросов

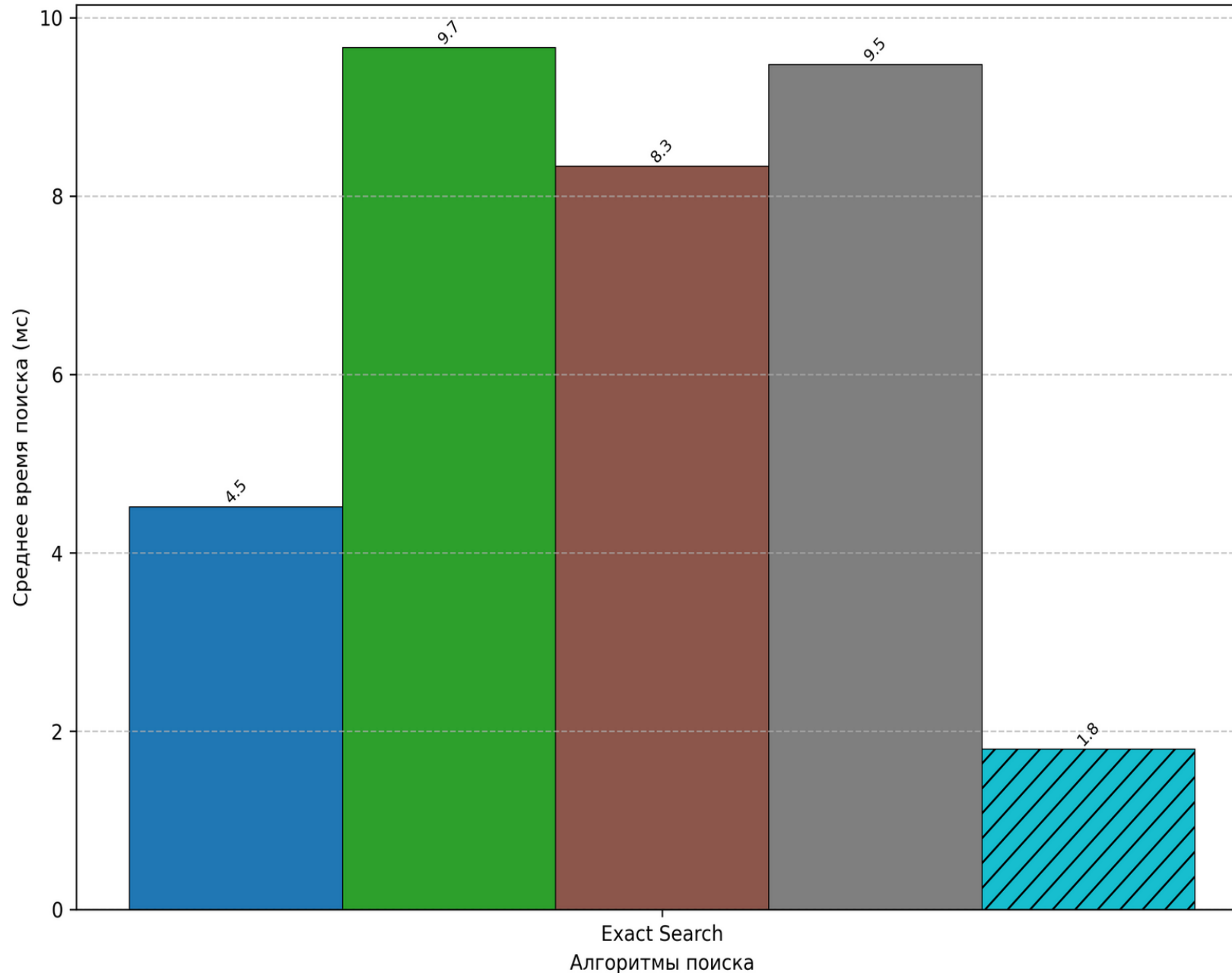
Сравнение производительности RAG системы: Точность поиска



- snowflake-arctic-embed-s - Exact Search
- mxbai-embed-large-v1 - Exact Search
- jina-embeddings-v2-base-en - Exact Search
- multilingual-e5-large - Exact Search
- BM25 - Exact Search

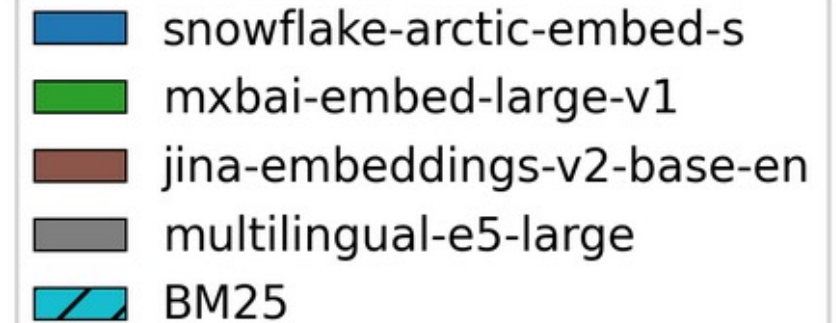
ВЫБОР АРХИТЕКТУРЫ ДЛЯ RETRIEVAL, СКОРОСТЬ

Сравнение производительности RAG системы: Скорость поиска



Бенчмарк

- скорость извлечения ответа из БД



ВЫБОР АРХИТЕКТУРЫ ДЛЯ RETRIEVAL, HYBRID SEARCH & RERANKER

Бенчмарк

- Hybrid search + Reranker:TextCrossEncoder (jina-reranker-v1-turbo-en)

Vectors Configuration (Name, Size, Distance)

colbertv2.0 128 Cosine

dense 1024 Cosine

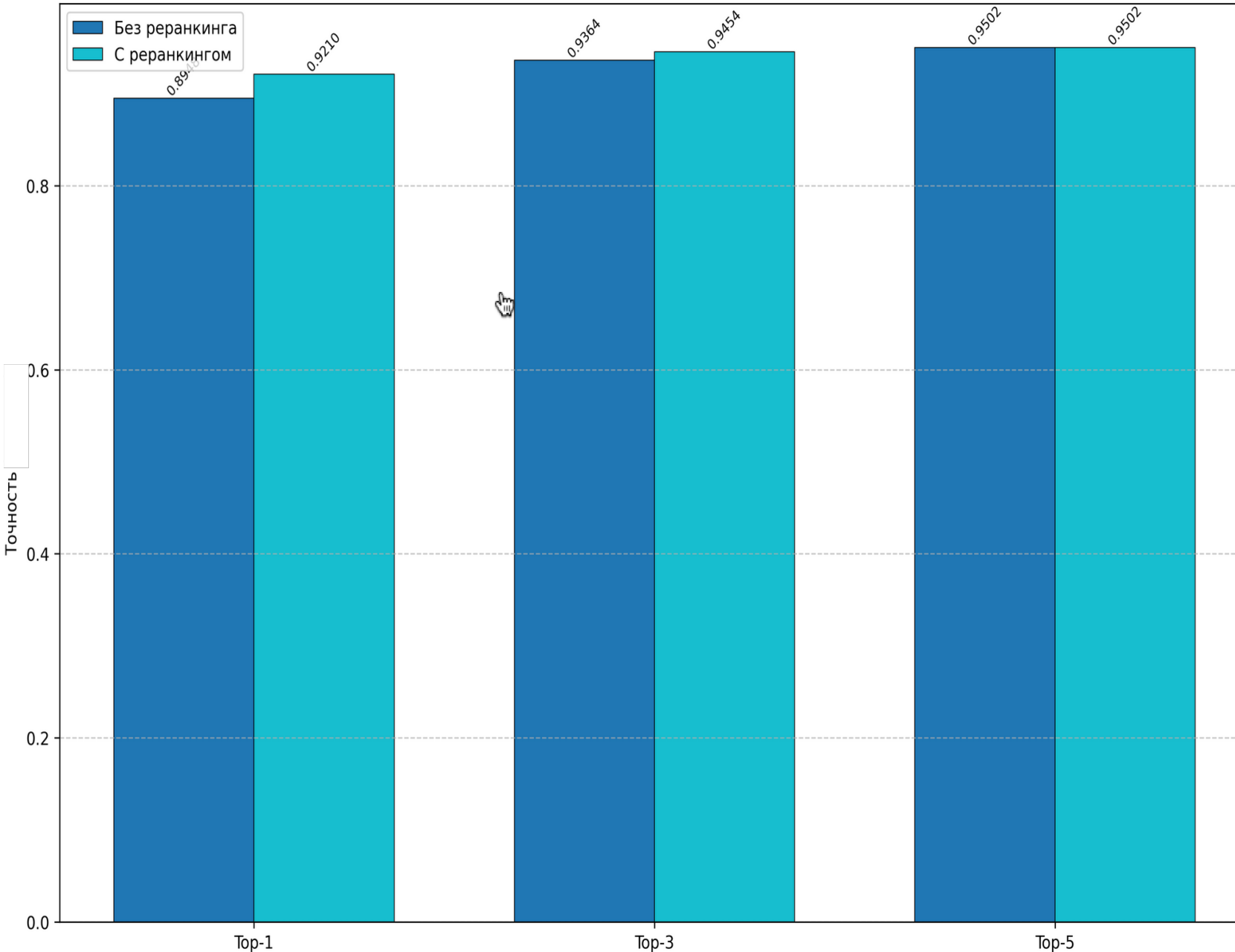
bm25 Sparse

Время поиска ответа:

с реранкингом: 17.878 мс

Без реранкинга : 17.034 мс

Сравнение для гибридного поиска с реранкингом и без: Точность поиска



Name	Vectors Configuration (Name, Size, Distance)
hybrid_collection	colbertv2.0 128 Cosine dense 1024 Cosine bm25 Sparse

Модель

- ☐ Настроить
- ☒ Инференс

Текст отправлен в модель

ПРИЛОЖЕНИЕ: АРХИТЕКТУРА

rag_system 0.1.0 OAS 3.1
[/api/openapi.json](#)

rag_system

GET	/ Root
POST	/api/v1/models/load_dataset Fit
POST	/api/v1/models/fit_save Fit Save
POST	/api/v1/models/load_model Load Model
POST	/api/v1/models/unload_model Unload Model
POST	/api/v1/models/find_context Find Context
POST	/api/v1/models/find_answer Find Answer
POST	/api/v1/models/quality_test Quality Test
GET	/api/v1/models/get_datasets Get Datasets
GET	/api/v1/models/list_models List Models
DELETE	/api/v1/models/remove/{model_id} Remove
DELETE	/api/v1/models/remove_all Remove All

What are the unique features of the Coolands for Twitter app?

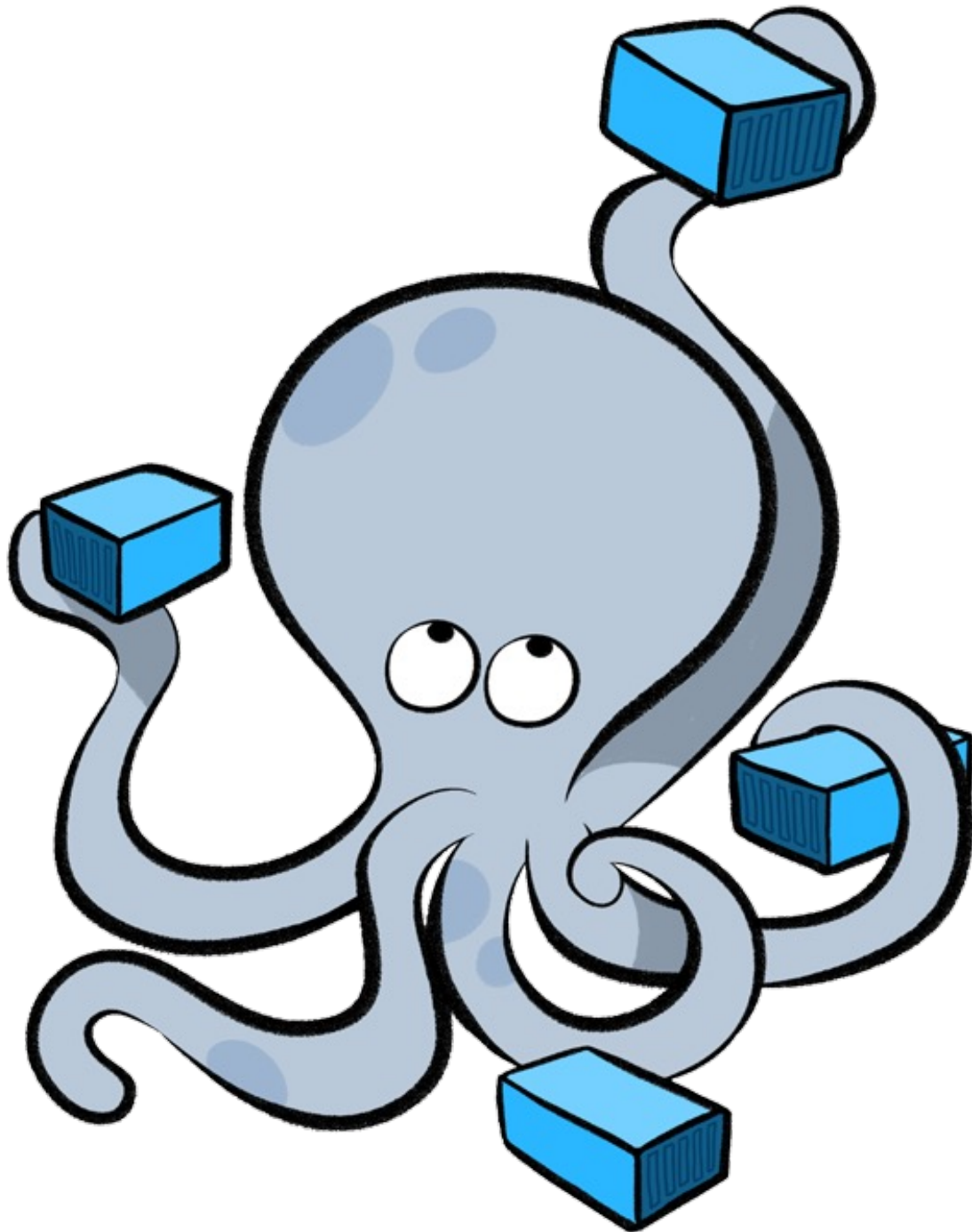
Ответ: The unique features of the Coolands for Twitter app are:

- Real-time find and refresh, eliminating the need for a refresh button.
- Avatar indicator, which shows small avatars in the title to indicate messages or tweets that mention you in real-time.
- Smart bookmark, which allows you to easily go back to a previous reading point in your timeline.

Additionally, the app has other features such as user-level notification settings, basic Twitter client features, and support for multiple accounts.

Оценка качества ответа

Релевантность: 4/5 **Точность:** 3/5 **Грамотность:** 4/5 **Описание:** Ответ соответствует вопросу, правильно интерпретируя контекст и перечисляя ключевые функции приложения. Однако он пропускает некоторые детали из контекста, такие как прямые ссылки и другие особенности, что снижает точность. Текст ясный и связный, что делает его грамотным и понятным.



ПРИЛОЖЕНИЕ: АРХИТЕКТУРА

DOCKER-COMPOSE -P APP_RAG UP -D:

- FASTAPI_BACK
- STREAMLIT
- QDRANT
- LOKI
- PROMTAIL
- GRAFANA

ПРИЛОЖЕНИЕ РАЗВЕРНУТО НА VPS

[HTTP://178.130.43.233:8501/](http://178.130.43.233:8501/)

МОНИТОРИНГ:

[HTTP://178.130.43.233:3000/](http://178.130.43.233:3000/)



РЕЗУЛЬТАТЫ И ВЫВОДЫ:

Общее время работы пайплайна:

Retrieval (hybrid+rerank) - 17.034 мс

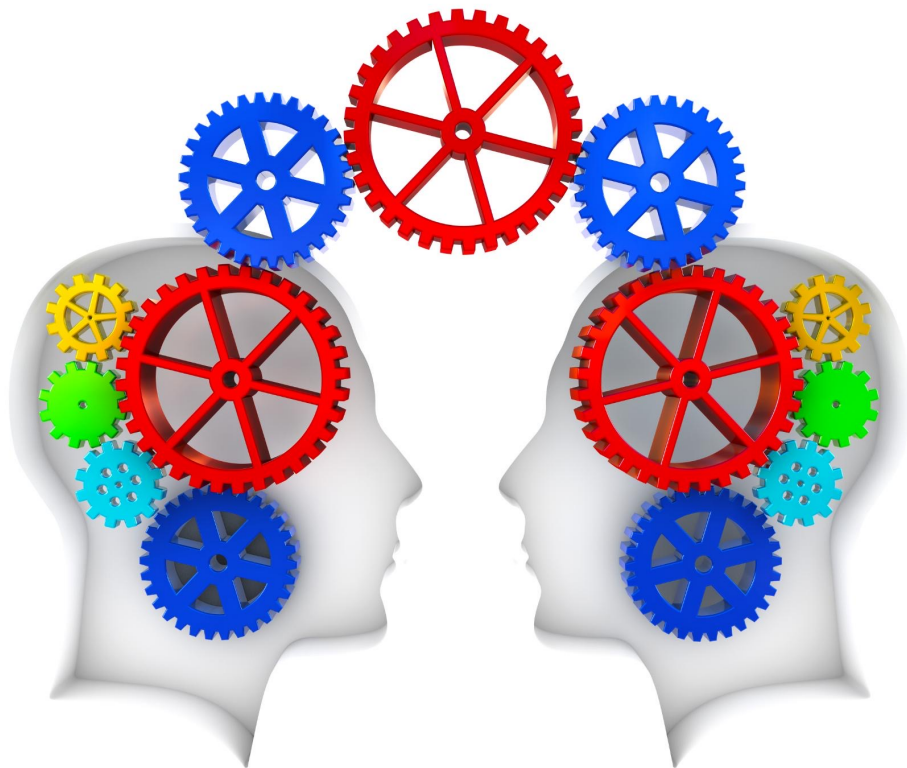
Augmented Generation – 0.5 с – 2 с,
в зависимости от занятости сервера



Подготовка источников знаний
**ColBERT Multivectors with
FastEmbed**



РАБОТА В КОМАНДЕ



- EDA - каждый участник провел EDA
- Baseline:
 - Fasttext - Евгений Яковенко
 - Word2vec/Sentence Transformers - Людмила Теплова
 - Tf-idf - Альберт Тайчинов
 - BM25 with ANN/all-MiniLM-L6-v2/Hybrid Search (BM25 + Sentence Transformers) - Алексей Ятковский
- Backend (FastAPI) - Альберт Тайчинов
- Frontend (Streamlit) - Людмила Теплова, Евгений Яковенко
- Docker + VPS + Grafana - Евгений Яковенко
- Pre-commiter - Алексей Ятковский
- Бенчмарк для выбора архитектуры Retrieval - Людмила Теплова, Альберт Тайчинов
- LLM для генерации - Евгений Яковенко

ИТОГ



ЧТО ПОЛУЧИЛОСЬ?

- ЧЕКПОИНТЫ ВЫПОЛНЕННЫ
- ПОСТАВЛЕННАЯ ЗАДАЧА ВЫПОЛНЕНА



ЧТО НЕ ПОЛУЧИЛОСЬ?

- ТЕСТИРОВАНИЕ ДРУГИХ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ
- ИСПОЛЬЗОВАНИЕ ДАТАСЕТА БОЛЬШЕГО ОБЪЁМА.

