

Факультет вычислительной математики и кибернетики МГУ имени М.В. Ломоносова



Программа профессиональной переподготовки «Искусственный интеллект и машинное обучение»

Система семантического поиска по корпусу российских ГОСТов на основе архитектуры RAG

ФИО слушателя: Оболенский Дмитрий Александрович

Научный руководитель:

Мочалова Ю.Д., сотрудник ФИЦ ИУ РАН, специалист по
машинному обучению на больших данных





Актуальность: поиск по нормативной документации



50 000+

действующих ГОСТов в РФ нуждаются в эффективном поиске



Проблема

- Полнотекстовый поиск не находит синонимы, аббревиатуры и иные формулировки той же нормы — специалисты тратят время на ручной просмотр документов
- Классические LLM «галлюцинируют», что при разработке внутренних нормативов недопустимо



RAG-подход

Retrieval + Generation (Lewis et al., 2020): поиск релевантных фрагментов + генерация ответа. Эволюция подходов: Naive → Advanced → Modular RAG (Gao et al.)



Новизна

- Применение RAG к корпусу российской технической нормативной документации в открытых источниках ранее не описывалось
- Система может быть элементом цепочки ИИ-агентов при автоматизированной разработке или валидации внутренней документации организации



Цель и задачи



Цель: разработка и развёртывание системы семантического поиска по корпусу ГОСТов с гибридным поиском и опциональной генерацией «естественного» ответа на основе найденных фрагментов.

1

Собрать корпус
ГОСТов (allgosts.ru)

2

Реализовать парсер
исходных данных

3

Выполнить векторную
индексацию

4

Построить поисковый
конвейер

5

Реализовать
интерфейс
пользователя

6

Протестировать
качество поиска



Ключевое ограничение: работа без GPU

Ноутбук: 2,3 ГГц CPU, 32 ГБ RAM, 500 ГБ диск. Виртуальной машине выделено 12 ГБ RAM и 4 CPU.



От файла к чанкам: правила парсинга HTML и PDF

Массовая загрузка — `batch_load.py`

Интерфейс — `/admin/upload` (API)

Файл → `parser_dispatch` по «магическим» байтам:
`.pdf / %PDF` → PDF · «<» → HTML · → иначе отказ (нужен PDF)



HTML-парсер — `parser.py`

`<article>`: `div / h2,h3` / только `p` — 3 формата вёрстки
(`_detect_format`)

Правило	Результат
Заголовок: <code><p> «N.N ...» / <h2><h3><h4></code>	новая секция (не чанк)
<code><p> / <div> / </code> — обычный текст	text
<code><table></code> — глоссарий (термин на двух языках с определением в табличной разметке)	term
<code><table></code> — данные (key: value)	table_semantic + table_display
<code><table></code> — указатель / содержание	пропускается



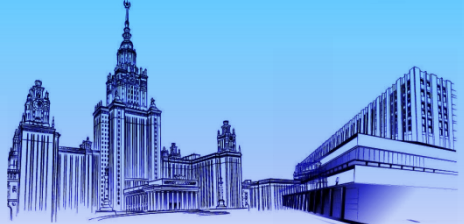
PDF-парсер — `parser_pdf.py`

`pdfplumber`: текст вне таблиц + `extract_tables()`; картинки — `pytupdf`

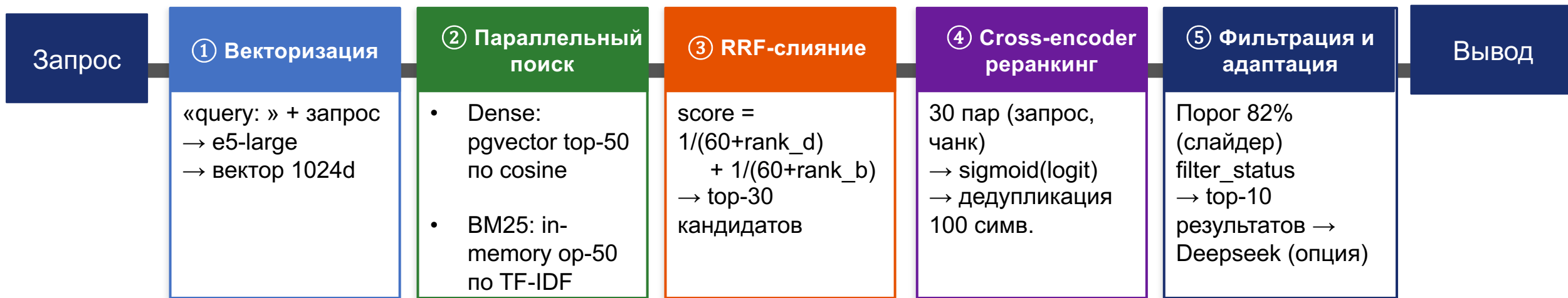
Правило	Результат
Заголовок: regex «Глава/Приложение/Раздел/N.N ...»	новая секция (не чанк)
Текст вне границ таблиц	text
<code>extract_tables()</code> — данные таблицы	table_semantic + table_display
Изображения (<code>pytupdf</code> , >5 КБ)	привязка к чанку (<i>images</i>)

Эмбеddинг — `intfloat/multilingual-e5-large (1024d)`

`text / term / table_semantic` → префикс «`passage:`» и кодирование · `table_display` → нулевой вектор (безэмбеddинга, только для отображения)



Поисковый пайплайн



Метод	Преимущество	Ограничение	Решение
Dense (E5-large)	Понимает смысл, синонимы	Размывает точные термины	✓ Применяется
BM25	Точные термины, обозначения	Не понимает смысла	✓ Применяется
RRF (слияние)	Не требует настройки	Только ранги, не оценки	✓ Применяется
Cross-encoder	Высокая точность реранкинга	Медленно на всём корпусе	✓ Топ-30 кандидатов
HyDE	Лучший recall на неточных запросах	2× вызовов LLM, дорого, возможно искажение моделью сути вопроса	✗ Отклонён
ColBERT	Токен-уровень сопоставления	Требует GPU, большой индекс	✗ Отклонён
Graph RAG	Связи между сущностями	ГОСТы слабо связаны графом	✗ Отклонён
RAPTOR	Иерархические резюме	Искажает нормативные формулировки	✗ Отклонён



Архитектура системы



Хост-машина

Браузер → <http://localhost:8080>

VirtualBox NAT: 8080→80 | 2222→22

Виртуальная машина Ubuntu 24 | 4 CPU 12 ГБ RAM без GPU

Nginx
(порт 80)

FastAPI / uvicorn (порт 8000)

PostgreSQL 17 + pgvector
(Docker, порт 5433)

Модели загружаются в при старте API автоматически

- Эмбединг: `intfloat/multilingual-e5-large` (1024d)
- Реранкер: `ambroad/bert-multilingual-passage-reranking-msmarco`
- BM25-индекс (in-memory, rank-bm25): строится при запуске API

users
123 id
A-Z username
A-Z password_hash
A-Z role
🕒 created_at

images
123 id
123 chunk_id
A-Z filename
A-Z filepath
A-Z caption
🕒 created_at

documents
123 id
A-Z designation
A-Z title
A-Z oks_code
A-Z source_file
A-Z status
🕒 created_at

chunks
123 id
123 doc_id
A-Z chunk_type
A-Z content
📄 embedding
A-Z section
A-Z table_display
123 chunk_index
🕒 created_at

feedback
123 id
A-Z query
A-Z designation
A-Z chunk_type
123 score_vector
123 score_reranker
123 chunk_id
123 position
123 rating
🕒 created_at



Интерфейс системы



Страница поиска

ГОСТ-поиск

admin

Выйти

#3 требования к болтовым соединения в металлургии

ОТВЕТ ИИ

На основе предоставленных выдержек из стандартов, требования к болтовым соединениям, которые могут быть применимы в металлургии (в контексте конструкционных высокопрочных соединений), содержатся в **ГОСТ 32484.3-2013** и **ГОСТ 32484.6-2013**. Прямых требований, специфичных именно для металлургической отрасли, в тексте нет, но указаны общие требования к высокопрочным болтокомплектam:

1. **Резьба и покрытие:** Поле допуска резьбы болтов (6g) действительно до горячего цинкования. Болты с горячим цинковым покрытием комплектуются гайками с увеличенным полем допуска резьбы (ГОСТ 32484.3-2013, Таблица 3).

2. **Механические свойства:** Классы прочности болтов — 8.8 или 10.9. Ударная вязкость — не менее 27 Дж (ГОСТ 32484.3-2013, раздел "Материал").

3. **Покрытия:** Запрещены покрытия из кадмия и его сплавов. При выборе обработки поверхности необходимо учитывать риск водородного охрупчивания (ГОСТ 32484.3-2013, Таблица 3).

4. **Шайбы:** Для комплектов применяются плоские шайбы с фаской по ГОСТ 32484.6-2013.

5. **Защита от коррозии:** В контексте заземления (ГОСТ Р 70303-2022, п. 5.3) указано, что соединения должны быть защищены от коррозии и механических повреждений, а для болтовых соединений предусматриваются меры против ослабления контакта. Однако этот стандарт относится к слаботочным системам и заземлению, а не к металлургическому оборудованию.

Если требуется информация о болтовых соединениях именно в металлургических установках (например, печах, прокатных станах), в предоставленных выдержках она отсутствует.

Показано 5 из 27 247 фрагментов

20586.5 мс

скрыто 5 ниже порога

№	ОБОЗНАЧЕНИЕ	НАЗВАНИЕ	РЕЛЕВАНТНЫЙ ФРАГМЕНТ	РЕЙТИНГ
1	ГОСТ 32484.3-2013 21.060.10, 21.060.20 Документ	Болтокомплекты высокопрочные для предварительного натяжения конструкционные. Система HR - комплекты шестигранных болтов и гаек	Таблица 3 - Требования к болтам и ссылочные стандарты: Поле допуска действительно до горячего цинкования. Болты с горячим цинковым покрытием предназначены для комплектации с гайками, имеющими увеличенное поле допуска резьбы. Расположение V-образного надреза (Шарпи) в испытательных образцах из болтов принимают по ISO 898-1. При выборе соответствующего процесса обработки поверхности (например, очистка и покрытие) необходимо учитывать риск... ▼ Показать фрагмент полностью Рисунок 3 - Размеры гаек ► Показать таблицу	97.2% 👍👍
2	ГОСТ Р 70303-2022 33.040.20	Слаботочные системы. Кабельные системы.	Рисунок 1 - Соединение цепей защитного заземления с устройством функционального заземления 5.3 Соединения заземляющих проводников и заземлителей должны быть надежными и обеспечивать непрерывность электрической цепи в соответствии с ГОСТ Р 55982. Соединения стальных проводников выполняются в соответствии с требованиями...	97.2% 👍👍

Порог релевантности

80%

☒ Ответ ИИ

Введите поисковый запрос... (Enter — поиск, Shift+Enter — новая строка)

→

Загрузка нового документа

Загрузить файл

Загрузить по ссылке

ССЫЛКА НА СТРАНИЦУ ALLGOSTS.RU

https://allgosts.ru/01/020/gost_r_iso_30042-2016

Файл будет скачан по адресу: https://allgosts.ru/01/020/gost_r_iso_30042-2016.doc

Проверить

Загрузить

● Покрытие хорошее (≥70%)

Покрытие: 98.6%

Обозначение **ГОСТ Р ИСО 30042-2016**

Название Системы управления терминологией, базами знаний и контентом. Обмен терминологическими базами [TermBase eXchange (TBX)]

Статус Отменен

Код ОК 01.020, 35.240.60

Чанков всего: 204

text: 175

table_semantic: 15

table_display: 14

Макс. размер: 1498 симв.

Символов в файле: 207 127 | Символов в чанках: 204 304 | Покрытие: 98.6%

▼ Первые 3 чанка



Результаты тестирования



Тип поиска	Средние значения по результату тестирования				
	Пример запроса	Место искомого документа	Скор искомого документа	Скор топ-1	Выше порога 82%
Естественный	как обозначаются подшипники по диаметру и ширине	4	72%	75%	5,25
Контрольный	правила дорожного движения для велосипедистов			65%	1
Размытый	классификация подшипников	6,5	53%	95%	5,75
Точный	обозначение размерных серий подшипников	5,5	97%	94%	5,5
Цитата	Подшипники роликовые конические однорядные	2,25	97%	97%	5,75



Производительность
(CPU)

14,8 с

1 пользователь

46,5 с

3 параллельных запроса

9 / 9 запросов -
без ошибок и таймаутов

Всего 18 запросов = 4 × 4 эталонных ГОСТа + 2 Контрольных — полная таблица в Приложении Ж



Куда развивать систему



Сейчас

Рабочий гибридный поиск: Dense (e5-large) + BM25 + RRF + cross-encoder реранкинг, развёрнут без GPU.



Краткосрочно (без GPU)

- Query expansion через DeepSeek — лёгкий вариант HyDE
- Дообучение cross-encoder реранкера на накопленных оценках пользователей (feedback, hard negatives)
- Очистка вопроса от неинформативных слов/знаков



При появлении GPU

- ColBERT и полноценный HyDE
- Расширение корпуса: СНИП, СП, ISO/IEC.
- ONNX-квантизация моделей



Корпоративный сценарий

- VirtualBox: режим NAT → «Сетевой мост» (Bridged Adapter).
- В белый список CORS добавляется адрес организации.
- Подключается TLS-сертификат.
- Изменения — только конфигурация, без правки кода приложения.



Обратная связь



Оценки пользователей по каждому результату сохраняются в feedback и формируют датасет для дообучения cross-encoder реранкера



Итоги работы



Корпус и парсеры

703 ГОСТа, рост числа чанков на документ $\times 9$ раз (11 $\rightarrow$ 100). Всего 70 676 чанков + 3 012 изображений



Гибридный конвейер

Dense (e5-large) + BM25 + RRF + cross-encoder перанкинг — методы взаимно компенсируют слабые стороны друг друга.



Интерфейс и безопасность

React-интерфейс; JWT-авторизация, CORS, rate-limiting, защита от SSRF.



Автономность

Развёртывание без GPU на обычном компьютере; все компоненты запускаются автоматически при старте ВМ.



Список литературы



1. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. — 2020. — Vol. 33. — P. 9459–9474.
2. Gao Y., Xiong Y., Gao X. et al. Retrieval-Augmented Generation for Large Language Models: A Survey // arXiv:2312.10997. — 2023.
3. Wang L., Yang N., Huang X. et al. Text Embeddings by Weakly-Supervised Contrastive Pre-training // arXiv:2212.03533. — 2022.
4. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. — 2019. — P. 4171–4186.
5. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of EMNLP. — 2019. — P. 3982–3992.
6. Johnson J., Douze M., Jégou H. Billion-scale similarity search with GPUs // IEEE Transactions on Big Data. — 2021. — Vol. 7, No. 3. — P. 535–547.
7. Robertson S., Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond // Foundations and Trends in Information Retrieval. — 2009. — Vol. 3, No. 4. — P. 333–389.
8. Cormack G., Clarke C., Buettcher S. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods // Proceedings of SIGIR. — 2009. — P. 758–759.
9. Gao L., Ma X., Lin J., Callan J. Precise Zero-Shot Dense Retrieval without Relevance Labels (HyDE) // Proceedings of ACL. — 2023. — P. 1762–1777.
10. Khattab O., Zaharia M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT // Proceedings of SIGIR. — 2020. — P. 39–48.
11. Sarthi P., Abdullah S., Tuli A. et al. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval // arXiv:2401.18059. — 2024.
12. Edge J., Trinh H., Cheng N. et al. From Local to Global: A Graph RAG Approach to Query-Focused Summarization // arXiv:2404.16130. — 2024.
13. Wang L., Yang N. et al. Multilingual E5 Text Embeddings: A Technical Report // arXiv:2402.05672. — 2024
14. Документация pgvector [Электронный ресурс]. — URL: <https://github.com/pgvector/pgvector> (дата обращения: 01.06.2026).
15. Документация FastAPI [Электронный ресурс]. — URL: <https://fastapi.tiangolo.com> (дата обращения: 01.06.2026).
16. Документация sentence-transformers [Электронный ресурс]. — URL: <https://www.sbert.net> (дата обращения: 01.06.2026).
17. Портал allgosts.ru — база российских ГОСТов [Электронный ресурс]. — URL: <https://allgosts.ru> (дата обращения: 01.06.2026).