СЕРВИС поиска параграфов

ГАЗПРОМ НЕФТЬ

Selected Distincts

СЕРВИС
поиска
параграфов

Щекинов Михаил Алексеевич

капитан

архитектура и взаимодействие

Худицкий Василий Олегович

разработчик

работа с моделью

Иванов Иван Геннадьевич

разработчик

структура и реализация базы данных

Кривых Наталья Викторовна

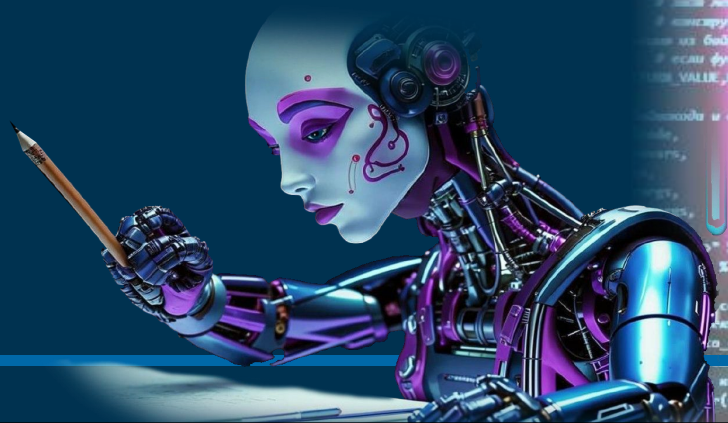
разработчик

идентификация векторных структур в базе, дизайн

§ Постановка задачи

Требуется реализовать **сервис поиска параграфов**,
как технической документации, так и художественной литературы.
Упаковать решение в docker image.

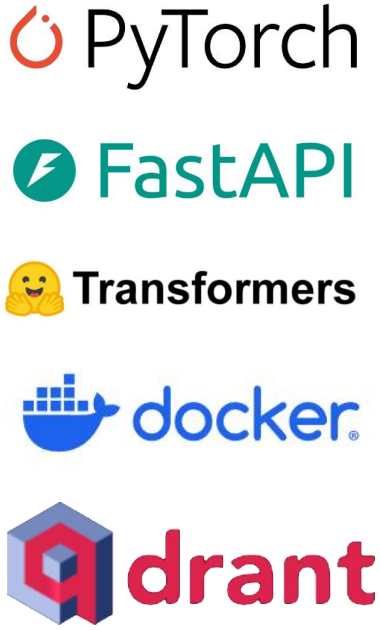
Решение должно реализовывать **две «ручки» - API endpoints**,
одна из которых будет осуществлять **индексацию**,
а другая делать **поиск**.



§ Техническая реализация

Технологии и инструменты:

- Transformers
- PyTorch
- FastAPI
- Qdrant
- docker

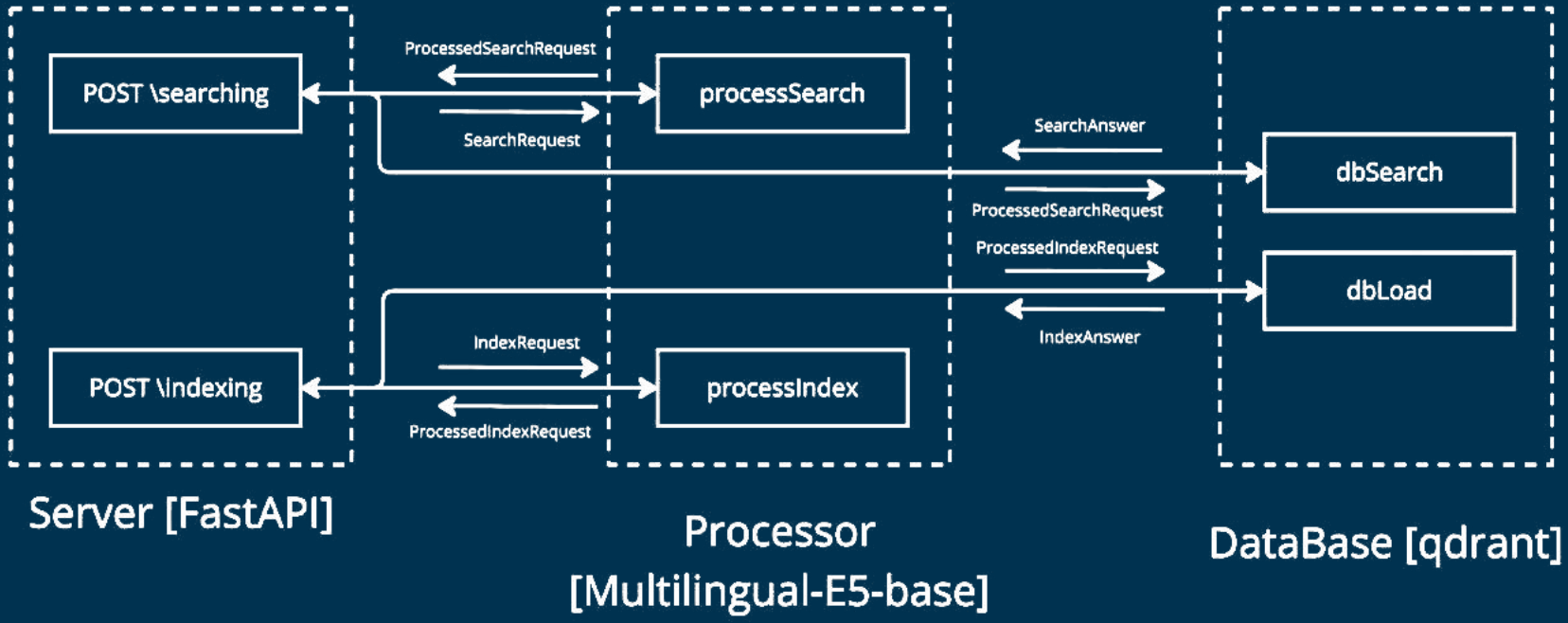


§ Решение

Решение составлено из трех основных блоков, взаимодействующих между собой:

- сервера;
- процессора: для преобразования текста в векторы;
- хранилища данных.

§ Архитектура



СЕРВИС
поиска
параграфов

§ API

POST /searching Search

search in DataBase with given json

Parameters

No parameters

Request body required — application/json

Example Value | Schema

```
{
  "text": "string",
  "top_k": 3,
  "filter_by": "string",
  "keywords": []
}
```

Responses

Code	Description	Links
200	Successful Response Media type application/json Controls Accept header. Example Value | Schema `{ "success": true, "content": [ "string" ], "count": 0 }`	*No links*
422	Validation Error	*No links*

```
// Search request example [POST /searching]
{
  "text"     : "piece of wood",
  "top_k"    : int = 2,
  "filter_by" : null,
  "keywords"  : ["wood"]
}
// Search answer example
{
  "success"   : true,
  "content"   : [
    "How it happened that Mastro Cherry, carpenter, found a piece of wood that wept and laughed like a child.",
    "The piece of wood lay on the forest floor, half-buried in the leaves, its surface rough and weathered. It had been shaped by years of rain and sun, telling a story of resilience and endurance. As the light filtered through the trees, it caught the grains of the wood, revealing intricate patterns that whispered of the life it once had as part of a mighty oak."
  ],
  "count"    : 2
}
```

СЕРВИС
поиска
параграфов

§ API

POST /indexing Index

Loads given json to DataBase

Parameters

No parameters

Request body required — application/json

Example Value | Schema

```
{
  "content": "string",
  "queries": [],
  "keywords_or_phrases": [],
  "chunk_id": "string"
}
```

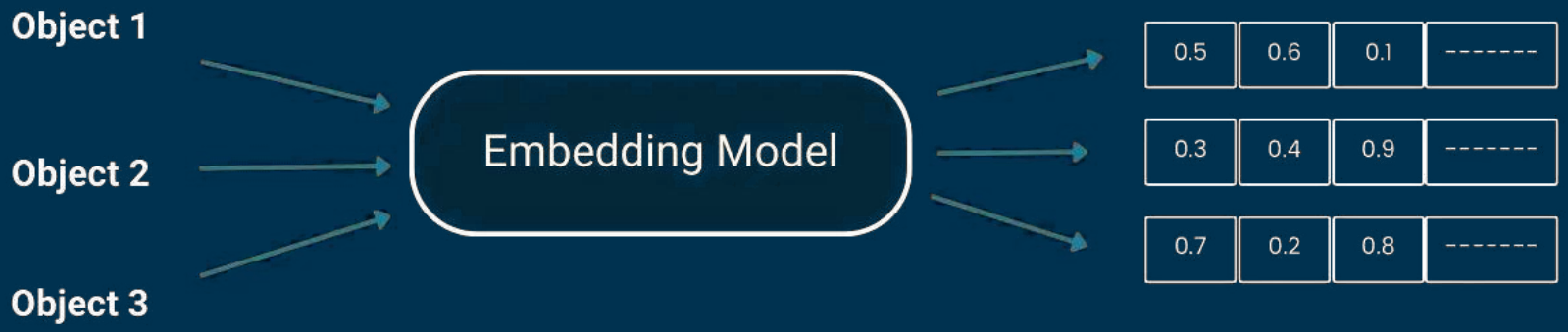
Responses

Code	Description	Links
200	Successful Response Media type application/json Controls Accept header. **Example Value** | Schema `{ "success": true }`	*No links*

```
// Index request example [POST /indexing]
{
  "content"  : "How it happened that Mastro Cherry, carpenter, found a piece of wood that wept and laughed like a child.",
  "queries"  : [],
  "keywords"  : [
    {
      "keyword_or_phrase": "Mastro Cherry",
      "explanation": "Mastro Ciliegia is the main antagonist of Pinocchio and Friends"
    }
  ],
  "chunk_id"  : "e60b46b6-849b-5527-bb13-88e58c2bb2f9"
}
// Index answer example
{
  "success": true
}
```

СЕРВИС
поиска
параграфов

§ Векторизация данных



§ Qdrant

Открытая облачно-ориентированная векторная база данных, разработанная для приложений искусственного интеллекта следующего поколения

- Поддерживает широкий спектр критериев запросов и типов данных, таких как числовые диапазоны, сопоставление текстов, геолокации и многое другое
- Предназначен для высокопроизводительного векторного поиска, отличается скоростью и эффективностью извлечения релевантных документов из больших объёмов данных

СЕРВИС
поиска
параграфов

§ Демонстрация

Реализованное решение осуществляет хранение и быстрый поиск по параграфам.
Работает как независимый сервер.

§ Демонстрация

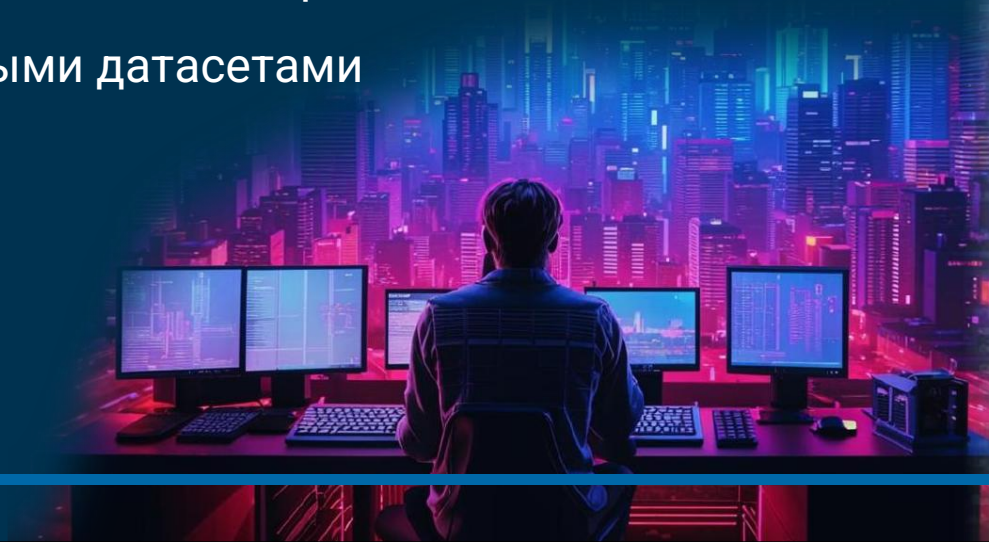
Реализованное решение осуществляет хранение и быстрый поиск по параграфам.
Работает как независимый сервер.

§ Области применения

- **при аудите** проектной документации для поиска нормативных определений строительных терминов;
- в сервисных центрах **для быстрого нахождения** ответов на часто задаваемые вопросы;
- в исследовательских центрах **для поиска информации** в научных статьях, исследованиях и технических отчетах, для работы с внутренними документами в компаниях и т.д.

§ План развития

- Эксперименты с алгоритмами поиска
- fine-tuning модели e5
- Усовершенствование системы векторизации
- Работа с более обширными датасетами



СЕРВИС
поиска
параграфов

§ Выводы

- **Реализована** запланированная **модель**
- Реализованы требуемые **API-points**
- Выполнено **требование по времени отклика**

СЕРВИС
поиска
параграфов

§ Преимущества модели

- Высокая скорость поиска за счет использования **векторной базы данных**
- **Расширяемость и гибкость решения**, что достигнуто разделением на компоненты
- Возможность **распределённого монтирования**