МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Южно-Уральский государственный университет (национальный исследовательский университет)»
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

Разработка веб-сервиса для моделей машинного обучения по классификации текстовых данных

Рецензент:
доцент кафедры ВМиИТ
ФГБОУ ВО «ЧелГУ», к.ф.-м.н.
А.Ю. Маковецкий

Автор работы:
студент группы КЭ-228
А.Э. Жулев

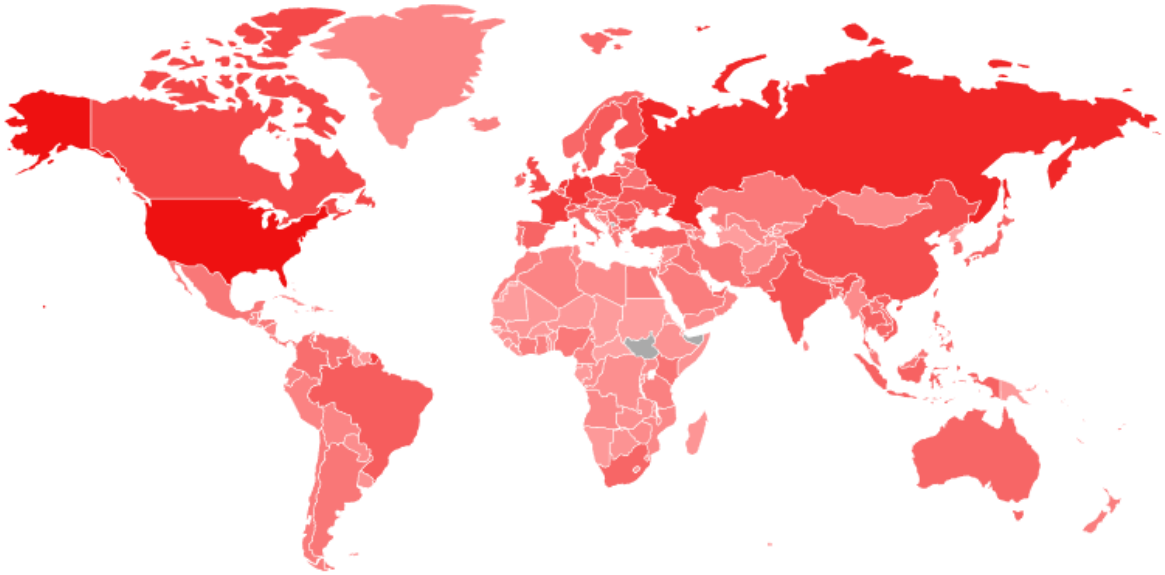
Научные руководители:
доцент кафедры СП, к.ф.-м.н.
С.У. Турлакова

ст. преподаватель кафедры СП
Н.С. Силкина

Челябинск, 2024 г.

АКТУАЛЬНОСТЬ

- Сбор данных для обучения
- Предварительная обработка данных
- Обучение моделей с разными параметрами
- Развертывание моделей
- Интеграция с внешними системами



Июнь 2023 - Апрель 2024

#	Страна	Трафик	
1	США	135 901 037	26.71%
2	Россия	61 538 966	12.10%
3	Германия	32 026 184	6.29%

https://cleantalk.org/ru/spam-stats

ЦЕЛЬ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Цель

Разработка веб-сервиса для моделей машинного обучения по классификации текстовых данных

Задачи

1. Провести анализ предметной области
2. Подготовить набор данных и обучить нейросетевую модель
3. Спроектировать и реализовать веб-сервис
4. Провести тестирование веб-сервиса

АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

Архитектура	Набор данных	F1
FastText + GloVe + BiGRU [1]	OLID	0,790
	посты Twitter	0,670
Semantic LSTM (SLSTM) [2]	SMS	**0,992**
	посты Twitter	0,968
Sequential Stacked CNN-LSTM (SSCL) [3]	SMS	**0,993**
	посты Twitter	0,971

Недостатки рекуррентных нейронных сетей:

1) плохо распараллеливаются
2) теряется контекст в длинных последовательностях

Решение: использование архитектуры «Трансформер» с механизмом внимания

[1] https://doi.org/10.1016/j.procs.2022.09.132
[2] https://doi.org/10.1007/s41870-018-0157-5
[3] https://doi.org/10.1007/s10472-018-9612-z

НАБОР ДАННЫХ

Данные собраны из открытых сообществ соц. сети «ВКонтакте»

Сбор и разметка текста производились с использование приложения «SpamDatasetSaver»

Набор сбалансированный и содержит классы: не спам (0), спам (1)

Величина набора – 2 000 элементов

Предобработка

- Замена омоглифов (α → a) и пробельных символов
- Замена множественных пробелов единичными
- Удаление URL, IP, @username, адресов электронной почты

ИССЛЕДУЕМЫЕ МОДЕЛИ

Модель	Количество параметров (млн.)	Объем (МБ)
Полновесные		
ai-forever/ruRoberta-large	355	1 454
bert-base-multilingual-cased	179	714
ai-forever/ruBert-base	178	716
Дистиллированные		
distilbert/distilbert-base-multilingual-cased	135	542
cointegrated/rubert-tiny2	29,4	118
cointegrated/rubert-tiny	11,9	47,7
DeepPavlov/distilrubert-tiny-cased-conversational-v1	10,4	41,6
DeepPavlov/distilrubert-tiny-cased-conversational-5k	3,6	14,6

Все модели имеют архитектуру «Трансформер»

Модели взяты из репозитория Hugging Face Hub

https://huggingface.co/

ЭКСПЕРИМЕНТЫ

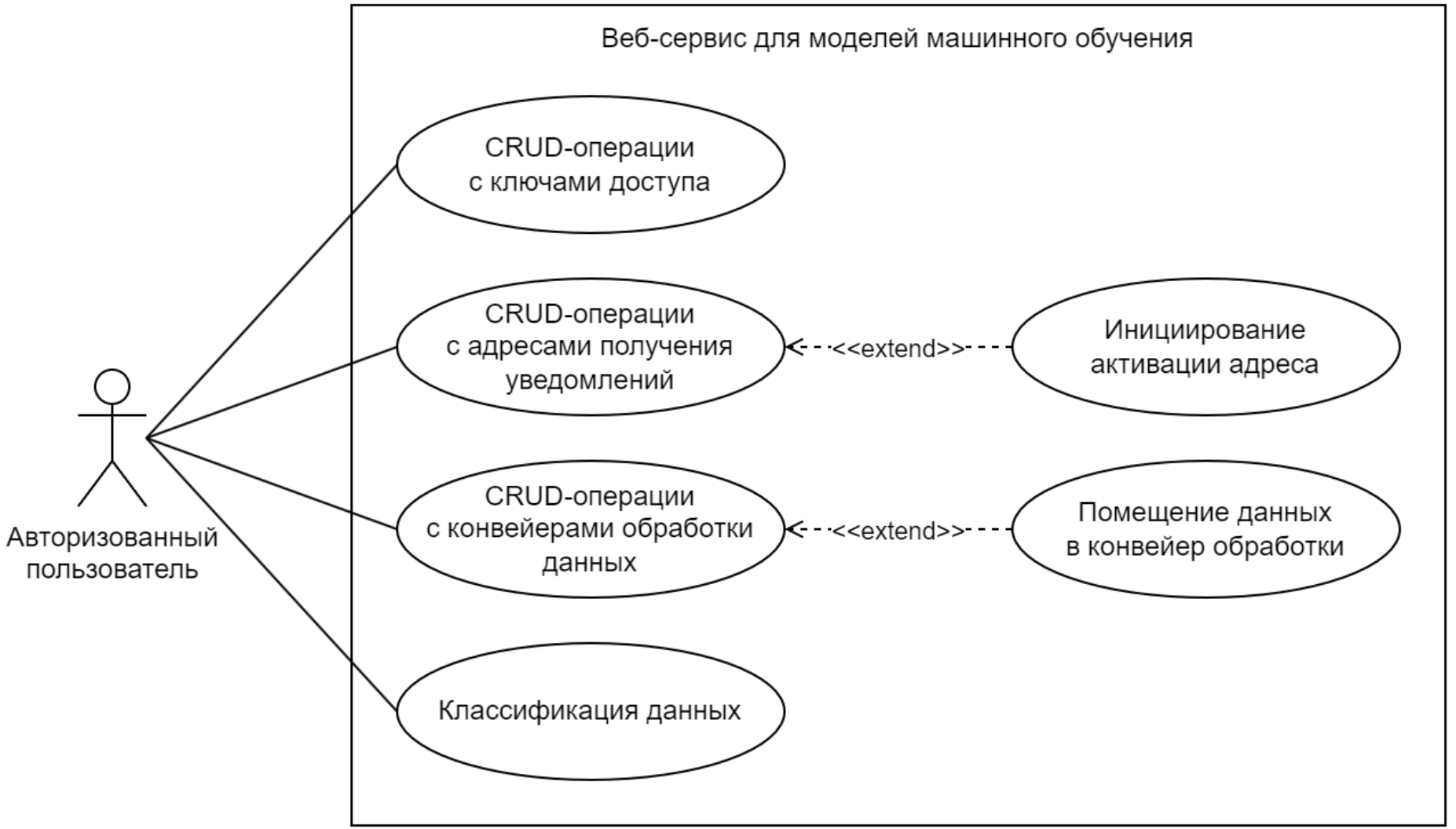
Модель	Время обучения (с)	Количество GPU	F1
Полновесные			
ai-forever/ruRoberta-large	**103**	4	**0,990**
bert-base-multilingual-cased	45	4	0,978
ai-forever/ruBert-base	39	4	0,978
Дистиллированные			
distilbert/distilbert-base-multilingual-cased	**32**	4	**0,960**
cointegrated/rubert-tiny2	**10**	**1**	**0,956**
cointegrated/rubert-tiny	10	1	0,944
DeepPavlov/distilrubert-tiny-cased-conversational-v1	9	1	0,952
DeepPavlov/distilrubert-tiny-cased-conversational-5k	9	1	0,949

Обучающая выборка – 80%, тестовая – 20%

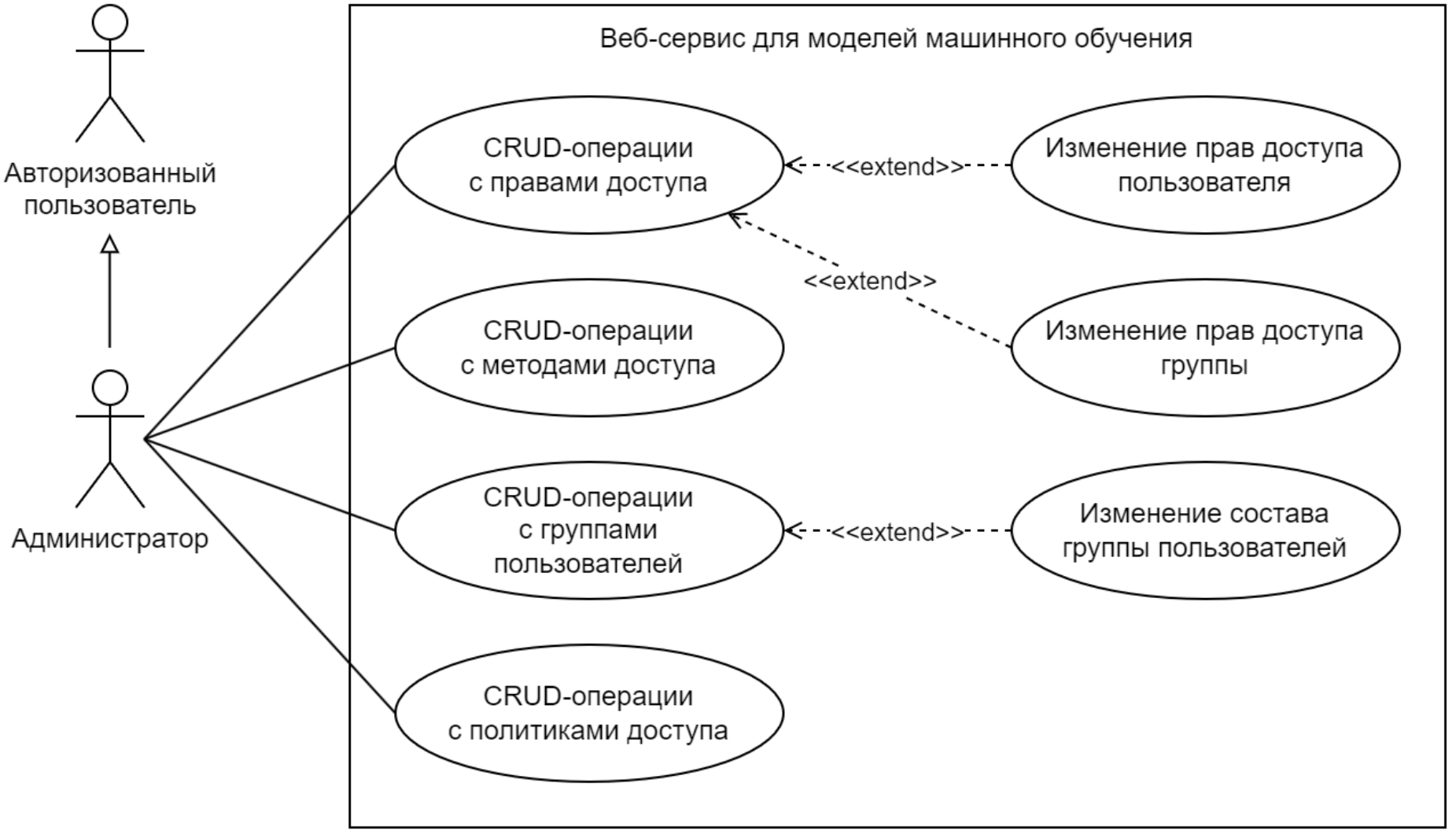
Комплекс «Нейрокомпьютер ЮУрГУ»

- NVIDIA Tesla V100 SXM2 x 4

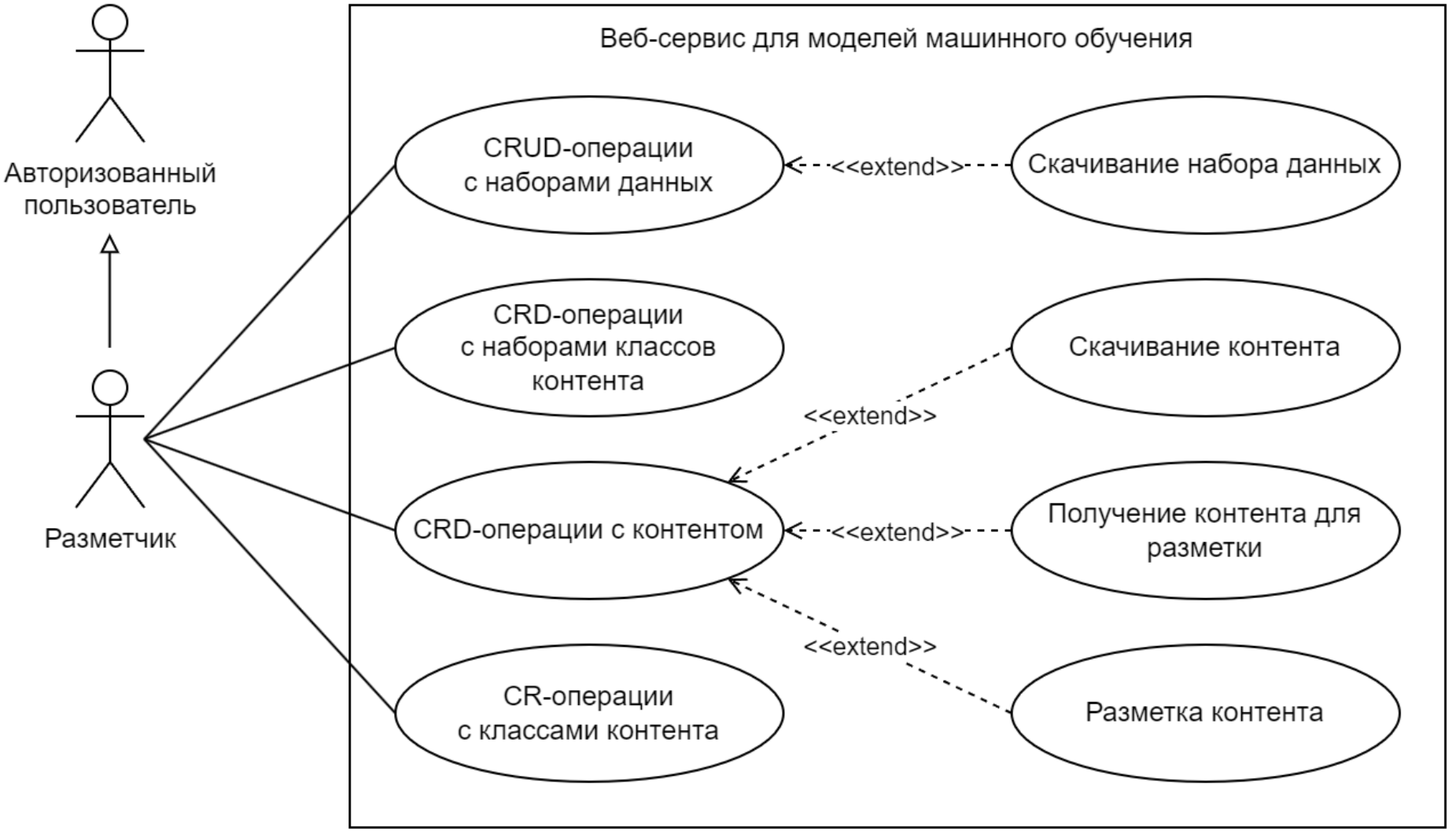
АВТОРИЗОВАННЫЙ ПОЛЬЗОВАТЕЛЬ



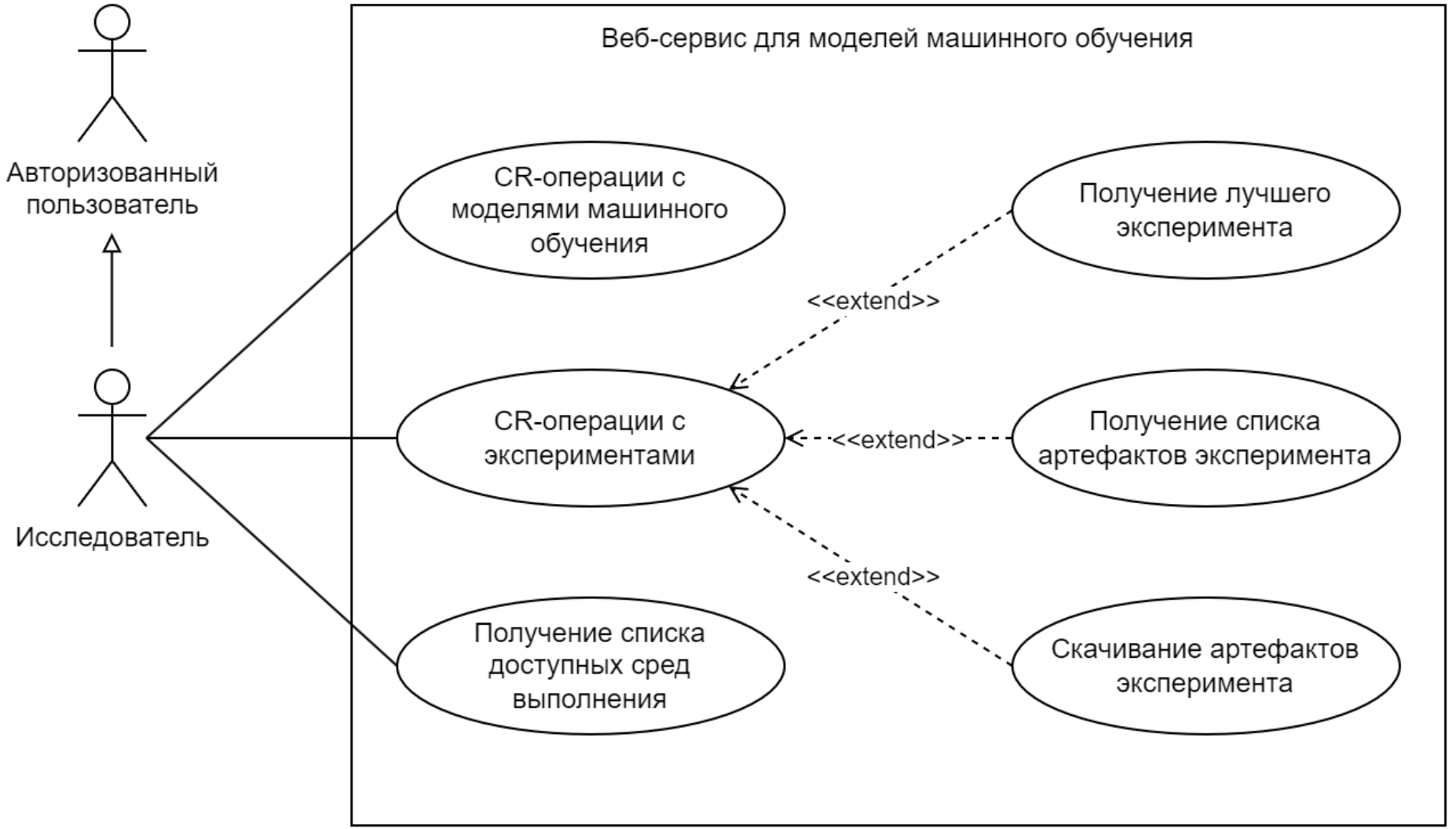
АДМИНИСТРАТОР



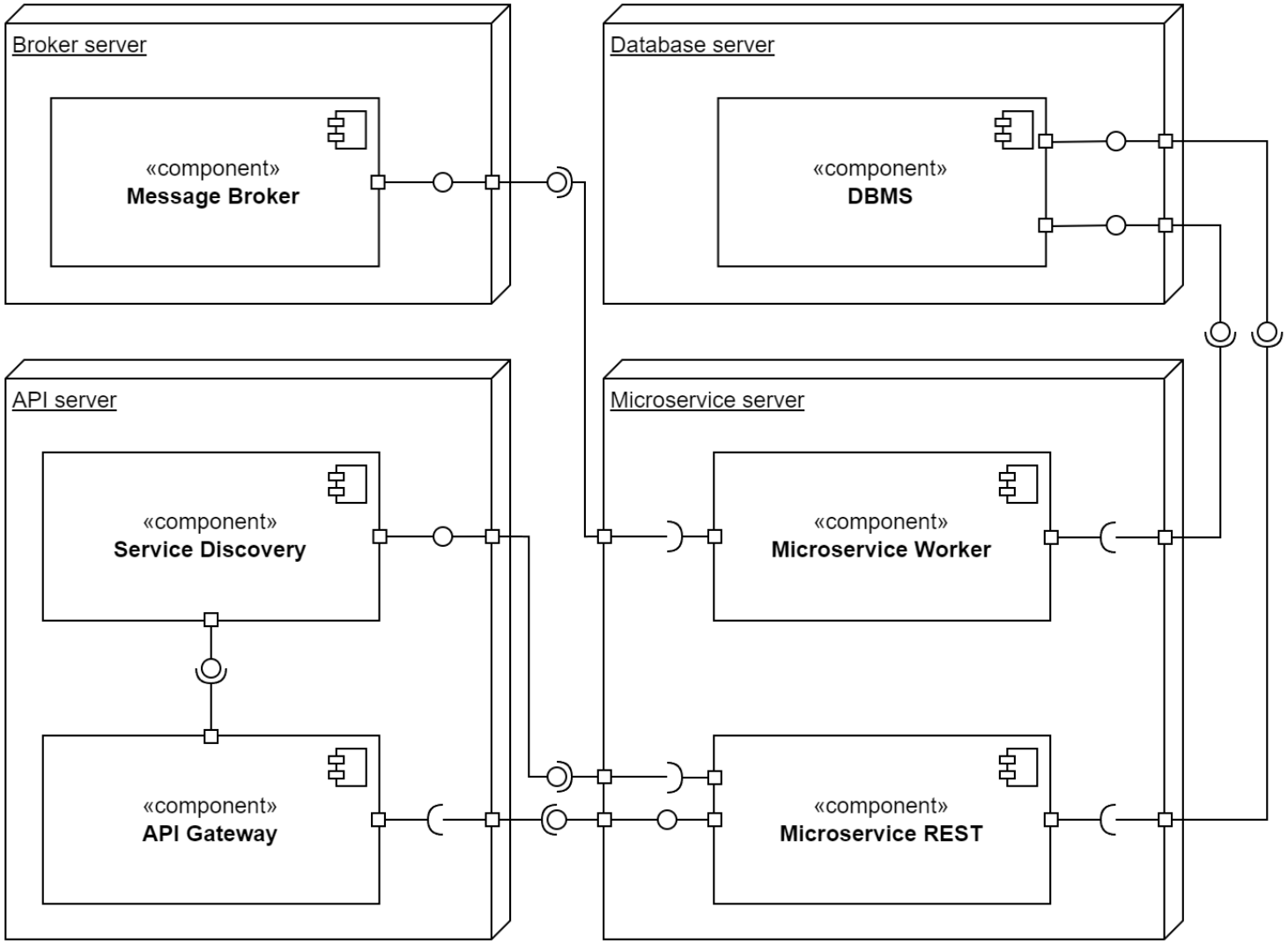
РАЗМЕТЧИК


Веб-сервис для моделей машинного обучения
Авторизованный пользователь
Разметчик
CRUD-операции с наборами данных
CRD-операции с наборами классов контента
CRD-операции с контентом
CR-операции с классами контента
Скачивание набора данных
Скачивание контента
Получение контента для разметки
Разметка контента
<<extend>>
<<extend>>
<<extend>>
<<extend>>

ИССЛЕДОВАТЕЛЬ



АРХИТЕКТУРА СИСТЕМЫ



МИКРОСЕРВИСЫ

- Auth – аутентификация и авторизация пользователей
- **Dataset** – хранение и разметка наборов данных
- **Learning** – обучение моделей
- **Pipeline** – управление конвейерами обработки данных
- Classification* – классификация данных
- Notification* – отправка уведомлений

* группа, каждый миросервис которой специализирован

СРЕДСТВА РЕАЛИЗАЦИИ

Язык программирования: Python 3.11.3

Основные библиотеки:

1) FastAPI (0.100.0)
2) Pydantic (2.7.0)
3) SQLAlchemy (2.0.19)
4) Celery (5.3.1)

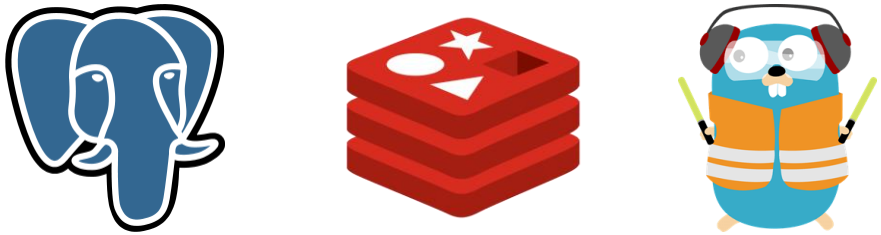
СУБД: PostgreSQL (15.6)

Брокер сообщений: Redis (5.0.14)

API-шлюз: Traefik (2.11.0)

Обнаружение сервисов: Consul (1.18.1)

Контейнеризация: Docker (26.1)



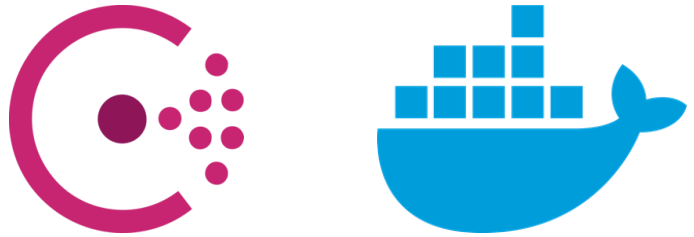
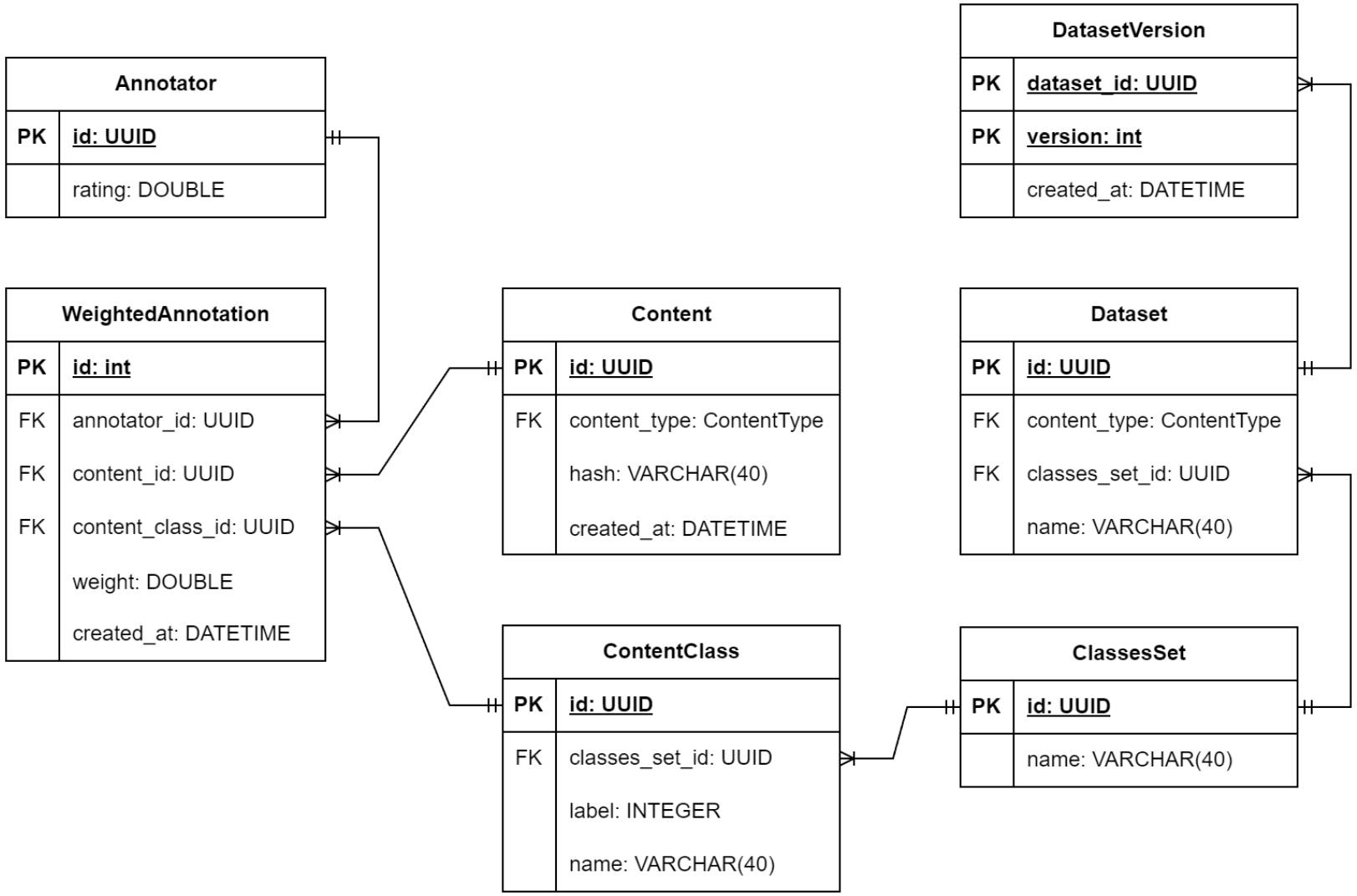
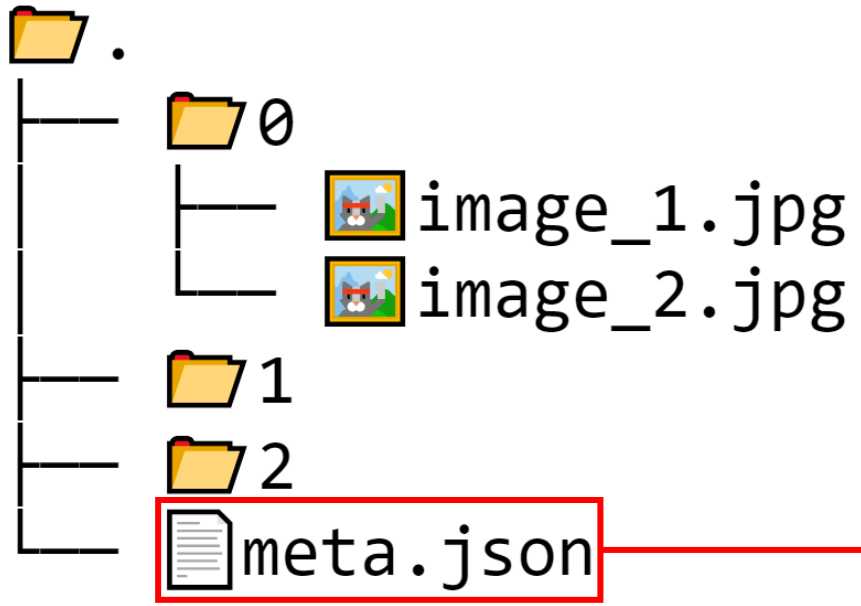


СХЕМА БАЗЫ ДАННЫХ СЕРВИСА DATASET



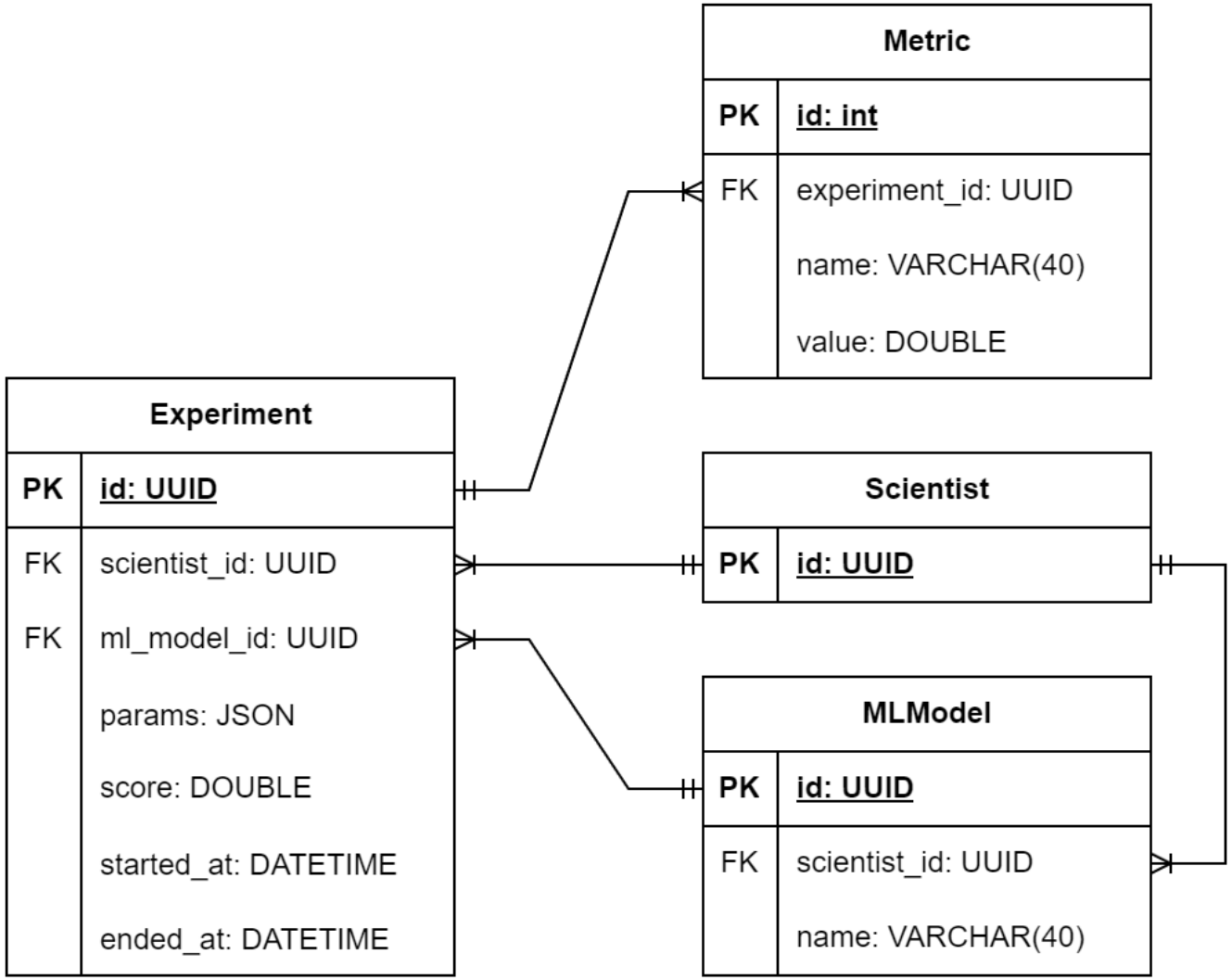
БИБЛИОТЕКА SIMPLE_FILE_DATASET




```
{
  "name": "pets",
  "version": 1,
  "labels": {
    "0": "cat",
    "1": "dog",
    "2": "snake"
  }
}
```

1. Поддержка любых файлов
2. Минимальный объем мета-данных
3. Поддержка ZIP-архивов:
   - упаковка/распаковка наборов
   - чтение наборов без распаковки

СХЕМА БАЗЫ ДАННЫХ СЕРВИСА LEARNING



БИБЛИОТЕКА REMOTE_ENVIRONMENTS

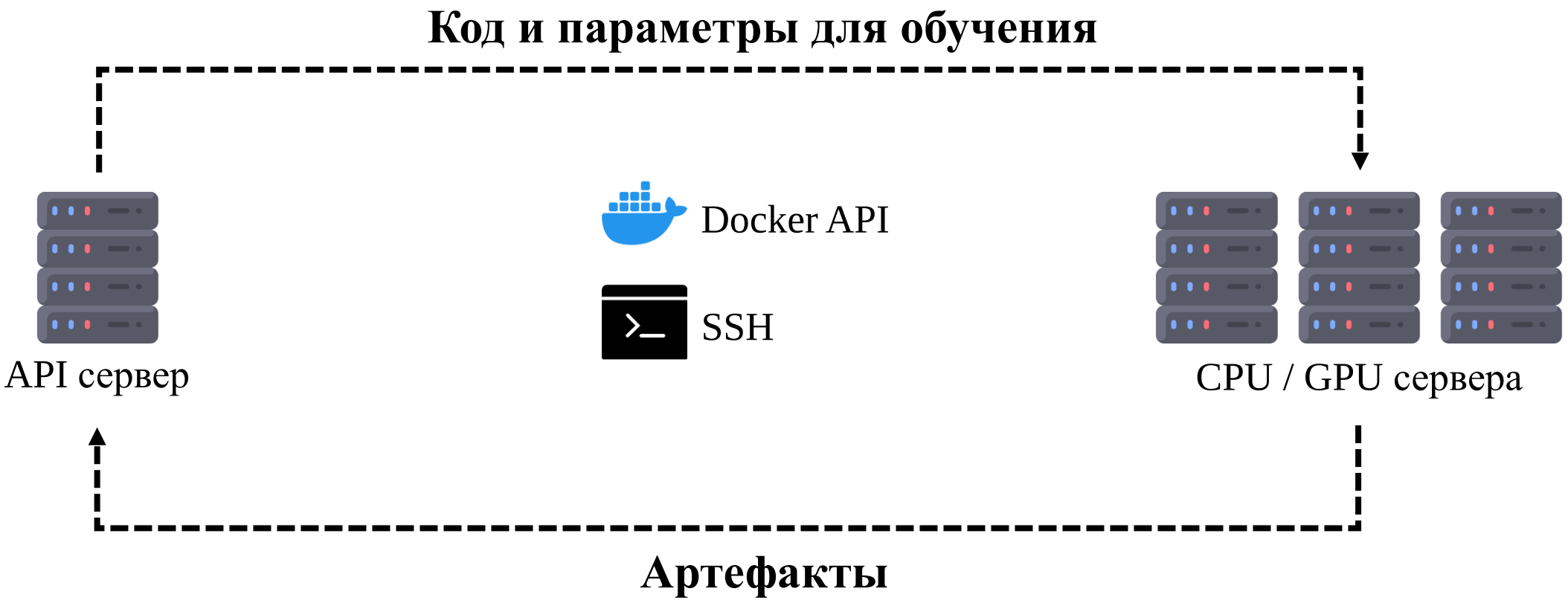
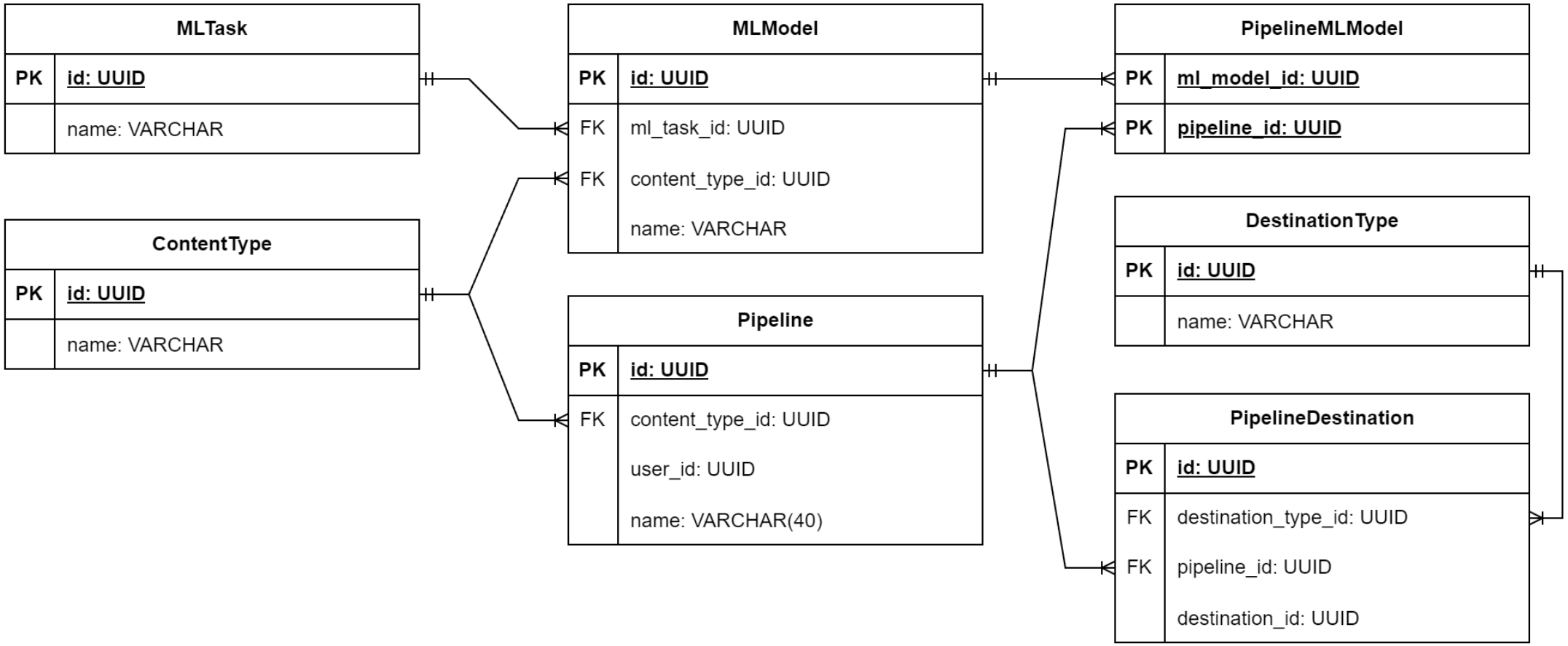
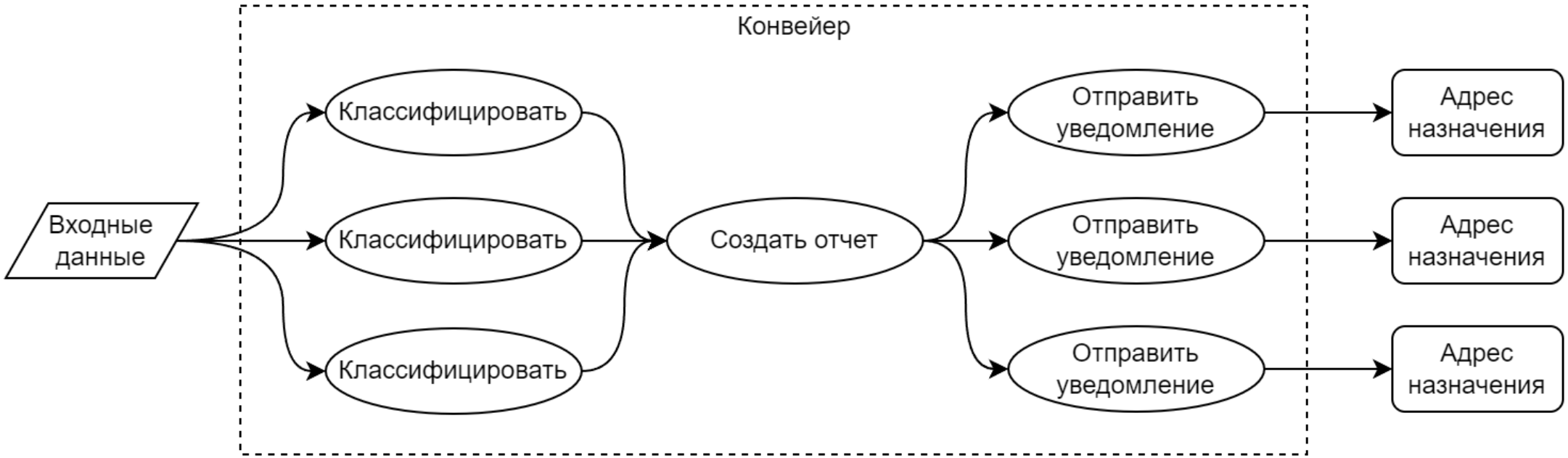


СХЕМА БАЗЫ ДАННЫХ СЕРВИСА PIPELINE



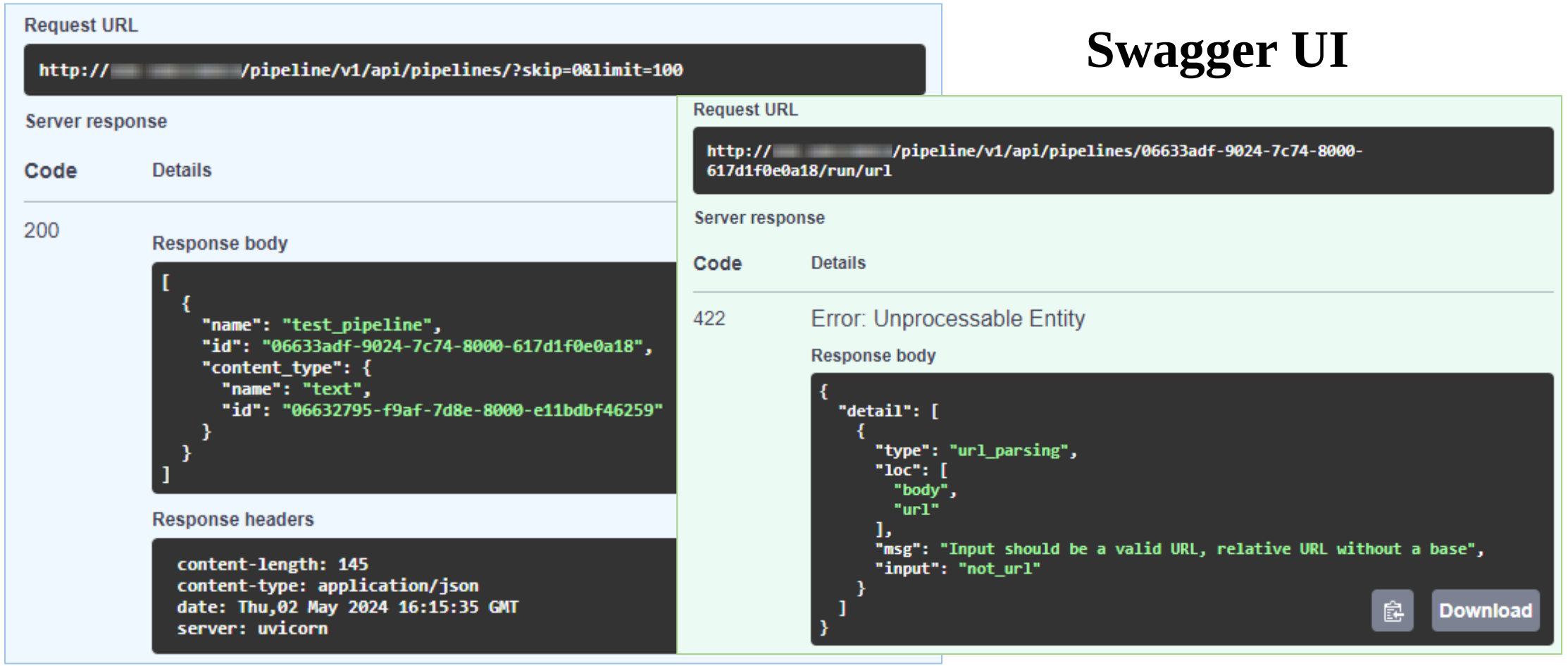
СТРУКТУРА КОНВЕЙЕРА



ТЕСТИРОВАНИЕ

>200 функциональных тестов для 84 конечных точек доступа

Swagger UI



ПУБЛИКАЦИЯ

- Жулев А. Применение нейронных сетей для выявления спама. // Параллельные вычислительные технологии: Короткие статьи и описания плакатов XVIII всерос. науч. конф. с международным участием (2 - 4 апреля 2024 г., Челябинск) – Челябинск: Издательский центр ЮУрГУ, 2024. – С. 88–93.
  DOI: 10.14529/pct2024

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Проведен анализ предметной области
2. Подготовлен набор данных и обучена нейросетевая модель
3. Спроектирован и реализован веб-сервис
4. Проведено тестирование веб-сервиса

РАСПРЕДЕЛЕННОЕ ОБУЧЕНИЕ

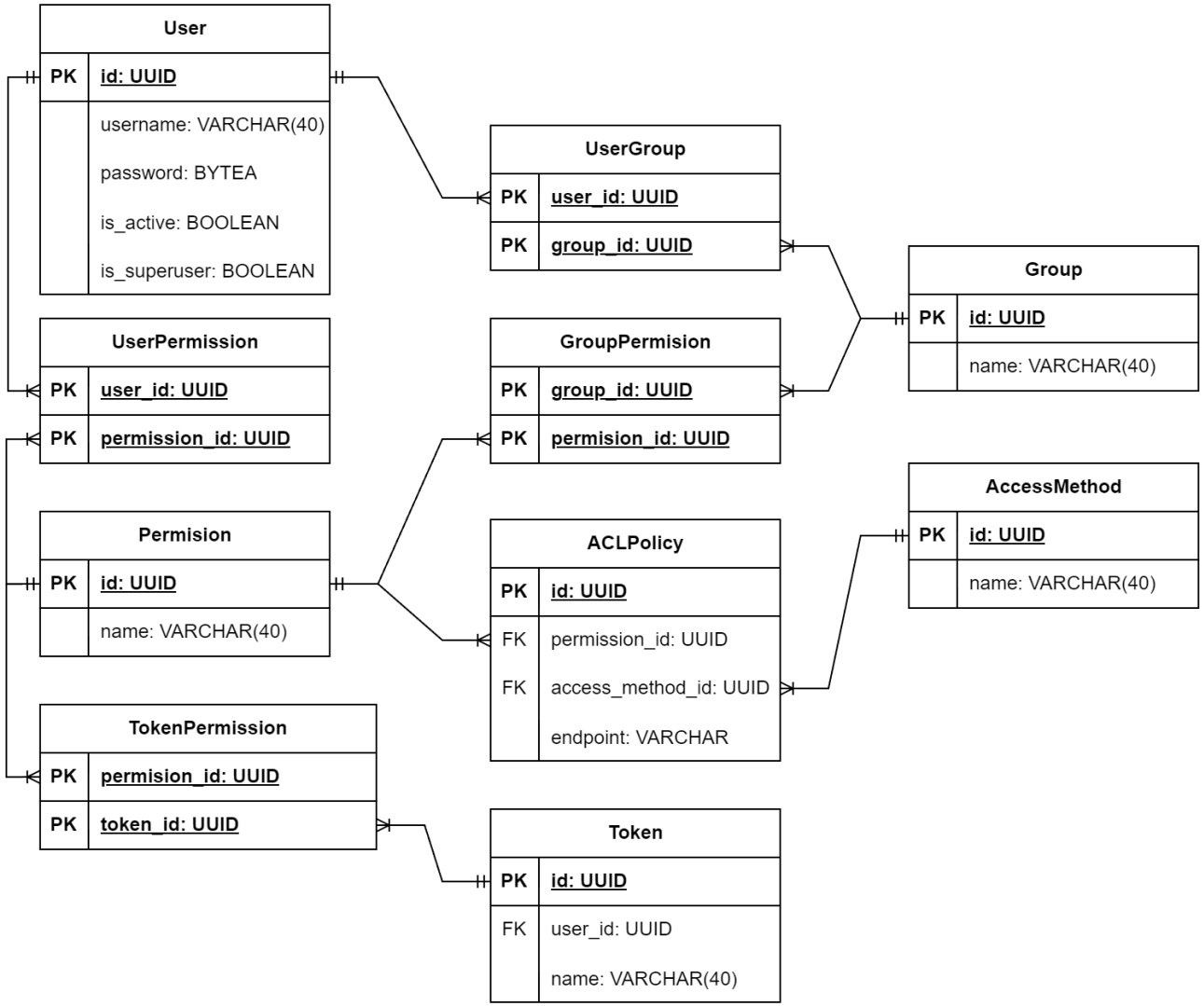
Модель	Время обучения (с)			Ускорение (среднее)
	GPU x 1	GPU x 2	GPU x 4	
Полновесные				
ai-forever/ruRoberta-large	263	162	103	**1,598**
bert-base-multilingual-cased	86	62	45	1,382
ai-forever/ruBert-base	66	51	39	1,301
Дистиллированные				
distilbert/distilbert-base-multilingual-cased	47	40	32	1,213
cointegrated/rubert-tiny2	10	17	19	**0,741**
cointegrated/rubert-tiny	10	16	19	**0,734**
DeepPavlov/distilrubert-tiny-cased-conversational-v1	9	15	17	**0,741**
DeepPavlov/distilrubert-tiny-cased-conversational-5k	9	16	18	**0,726**

ПОИСК ГИПЕРПАРАМЕТРОВ

Модель	Accuracy	F1	batch size	lr
Полновесные				
ai-forever/ruRoberta-large	0,991	**0,990**	32	3e-5
bert-base-multilingual-cased	0,978	0,978	16	3e-5
ai-forever/ruBert-base	0,980	0,978	32	3e-5
Дистиллированные				
distilbert/distilbert-base-multilingual-cased	0,958	**0,960**	16	5e-5
cointegrated/rubert-tiny2	0,955	0,956	16	5e-5
cointegrated/rubert-tiny	0,944	0,944	32	5e-5
DeepPavlov/distilrubert-tiny-cased-conversational-v1	0,950	0,952	32	3e-5
DeepPavlov/distilrubert-tiny-cased-conversational-5k	0,950	0,949	32	5e-5

batch size: 16, 32, 64, 128
lr: 1e-5, 3e-5, 5e-5

СХЕМА БАЗЫ ДАННЫХ СЕРВИСА AUTH



КЛЮЧИ ДОСТУПА

Виды ключей:

1) пользователя (user)
2) выборочного доступа (custom)

Содержимое ключа:

1) тип ключа
2) идентификатор пользователя
3) идентификатор ключа

Содержимое зашифровано криптографическим алгоритмом симметричного шифрования Fernet