

Балобин Дмитрий

Data Scientist, NLP

Москва, Россия

hanaan@internet.ru

t.me/d0rj3228

github.com/d0rj

Опыт работы

Сбер

Июнь 2022 – По настоящее время

Junior Data Scientist

10 мес.

Команда NLP B2C

- Добавление в пайплайн обучения LLM логирования в ClearML и расчёта метрик качества;
- Разработка модели Intent Recognition для диалоговой системы (классификация текста, similarity search, rule-based NER). Увеличил количество интенгов одновременно с улучшением метрик на 3.5%;
- Сбор и валидация датасетов для модели классификации (Toloka, ПРОМ данные, opensource данные). Объём данных увеличился в 4 раза с 10000 до 40000 строк;
- Оптимизация моделей для использования в ПРОМе (машины с CPU ресурсами). Ускорил общий инференс на целевом железе в 2 раза с использованием ONNX, Faiss, PyTorch;
- Речевая аналитика телефонных звонков (классификация / кластеризация текста);
- Внедрение системы версионирования данных и отслеживания экспериментов ClearML в команде

Сбер

Декабрь 2021 – Май 2022

Стажёр Data Scientist

6 мес.

Команда NLP B2C

- Речевая аналитика телефонных звонков (кластеризация, визуализация);
- Улучшение пайплайна обучения модели intent recognition (перенос на PyTorch Lightning), адаптация его под внутренние требования качества;
- Увеличение скорости и качества на ПРОМе путём анализа существующих решений;
- Разработка Telegram ботов для оптимизации рабочих задач (проверка диалоговой логики, тестирование моделей)

Образование

РТУ МИРЭА

2019 – По настоящее время

Прикладная математика, Анализ данных

СберУниверситет

Ноябрь 2022 — Январь 2023

Курс - Архитектура нейронных сетей: Трансформеры

СберУниверситет

Январь 2022 — Май 2022

Курс - NLP (углубленный курс)

Tinkoff Fintech

Февраль 2021 — Май 2021

Курс - Разработка на Scala

Используемые технологии

Языки программирования: Python

Фреймворки: PyTorch, Sklearn, Transformers, PyTorch Lightning, Natasha, Streamlit

Инструменты: Git, DVC, Docker, MLFlow, Tensorboard, ClearML, ONNX, Poetry