Медведев Давид Валерьевич

Стажер Data scientist

КОНТАКТЫ:

Телефон: 89822678287
Email: medvedev03dv@gmail.com
Telegram: @david_medvedev

GitHub: https://github.com/SalLangg

ОБРАЗОВАНИЕ:

2024 — по настоящее *время*
МФТИ, ФПМИ — Магистратура
Факультет: Прикладная математика и информатика
Направление: Современная комбинаторика

ДОПОЛНИТЕЛЬНОЕ ОБУЧЕНИЕ:

- **Центр "Пуск" МФТИ:** Продвинутые методы машинного обучения.
- **Кафедра интеллектуальных систем:** Введение в машинное обучение (Константин Воронцов) и программирование на Python (Мурат Апишев)
- **Stepik: ML & NN (Сергей Балакирев)**
- **Deep Learning School МФТИ часть 1**
- **Тренировки ML Яндекс.** Сейчас прохожу 3-й спринт по RL который стартовал 9-го сентября. До этого прошел тренировки по классическому ML, CV и NLP

НАВЫКИ

- **Языки:** Python, C# (базовый)
- **ML:** sklearn, CatBoost, LightGBM, PyTorch
- **Инструменты:** Git, Docker, FastAPI, MLflow, PySpark(активно изучаю)
- **Базы данных:** PostgreSQL
- **Обработка данных:** pandas, numpy, seaborn, matplotlib
- **Математика:** теория вероятностей, статистика, оптимизация, линейная алгебра

ОБОМНЕ:

Ищу стажировку в области анализа данных и машинного обучения. Имею опыт работы с классическими ML-задачами, пайплайнами и моделями для табличных и не структурированных данных. Хочу применить свои знания в практике и развиваться в направлении прикладного DS и ML.

На данный момент работаю инженером по информационных технологиям проектирования. Фактически курирую техническую сторону работы проектного отдела в части цифровых инструментов. Занимаюсь разработкой систем автоматизации на Python, а также связываю наш отдел с отделов аналитики собирая и выгружая данные из проектных моделей в PostgreSQL.

МОИ ПРОЕКТЫ:

BuildBIMClassify — ML-система классификации BIM-объектов по сметным работам.

Разработал и внедрил пилотную MVP модель для автоматического сопоставления элементов BIM-модели (Revit) с позициями сметного справочника - классическая задача, выполняемая по большей мере вручную.

Результат: уже на этапе MVP смогли значительно сократить ручное сопоставления и смогли уменьшить нагрузку на проектный отдел. Сейчас тестируется расширение выборки с 20-ти классов на 100, где базовая модель без изменений показала качество Micro Precision(70%) Macro Precision(60%)

Описание:

- *Использовал CatBoost для multi-label классификации с фокусом на Precision (90%). Тестировались и другие подходы к классификации, но в виду особенностей данных решил остановиться именно на этом варианте.*
- *Построил пайплайн сбора данных, их обработки и использования результатов модели.*
- *Подготовил пайплайн для дообучения модели при расширении справочника.*

MorseNet — Декодер аудио файлов с кодом Морзе

Построить модель декодирования сигналов морзе, используя технологии, похожие на обработку естественного языка.

Описание:

- *Построена нейросеть CNN → LSTM с CTC Loss для декодирования из аудиофайлов*
- *Для извлечения признаков использовались Mel-спектрограммы и аугментации (time/freq masking)*
- *Логирование метрик обучения происходит через MLflow*
- *Сохранение моделей на сервере*
- *Реализован FastAPI-сервер с возможностью дообучения модели, независимо от инференса*
- *Решение упаковано в Docker*

Результат: 15-е место Kaggle из 124

GitHub: https://github.com/SalLangg/Morse-Decoder_V2

Классификация изображений

Учебная модель для классификации 42 персонажей по JPEG-изображениям. Код был написан самостоятельно, без использования блоков **Deep Learning School**

Описание:

- *Разработана CNN-модель с 3 сверточными блоками*
- *Использованы техники аугментации и расширения тестовой выборки: случайные повороты, изменение яркости/контраста, горизонтальное отражение*
- ***96.56% accuracy** на тренировочной выборке для самой базовой модели.*

GitHub: https://github.com/SalLangg/Image-classification

Предсказание личности

Учебная модель предсказания личности человека на основе данных

Описание:

- *Топ 3% Kaggle.*
- *Проанализирована степень важности пропусков во всех данных*
- *Созданы новые признаки для расширения выборки*
- *Протестированы различные модели - CatBoost, XGBoost, RamdomForest, а также их стейкинг*

GitHub: https://github.com/SalLangg/Personality_Prediction

RAG-LMM помощник инженера(Приостановлен)

Цель: разработать систему умного поиска по внутренней базе знаний компании с выводом найденной информации в качестве контекста LLM.
Технологии: модели компьютерного зрения, ORC, LLM, RAG
Задачи, которые предстоит решить:

- *Решение проблем с неструктурированной документацией с помощью ORC (выполнено)*
- *Интеграция примечаний с изображениями в систему*
- *Проверка актуальности норм*
- *Фильтрация галлюцинаций LLM*

Результат: построил часть модели, которая отвечает за поиск релевантной информации в pdf. Данная модель может определять качество скана pdf для частей документа и если качество низкое, то применяются ORC подходы для получения информации (например таблицы) и конкатенирует их с качественными сканами. После этого происходит поиск и выбор информации на основании запроса через библиотеку Langchain.