Решение О! хакатона

от команды FULLSTACK EXCEL

О! Хакатон. Кейс 2: автоматическое тегирование номеров отелей

Описание решения

ЯДРО РЕШЕНИЯ: Использование современных высокопроизводительных подходов в рамках **NLP и машинного обучения**.

ОЧИСТКА ДАННЫХ Очистка от мусорных символов с помощью регулярных выражений. **Эксперименты:** стемминг и лемматизация.

БАЗОВЫЕ МОДЕЛИ

В рамках единого пайплайна для каждого таргета по отдельности выбирались:

1. Векторайзеры для текста **(TF-IDF или count-vectors);**
2. Легкие модели **(логистическая регрессия, SVM)** и их **гиперпараметры** (*optuna*).

Эти модели хорошо закрывают “простые” таргеты на основе подбираемых ими ключевых слов

DEEP-LEARNING МОДЕЛИ

1. Для каждого таргета обучалась легковесная **bert-like** модели на базе **энкодера bge-micro** (*дистилляция* bge-small);
2. Проводились эксперименты по **multilabel классификации** (интересное продуктовое решение);
3. Инференс через **ONNX-runtime** ради графовой оптимизации и роста скорость;
4. **ostrovok-BGE-micro** закрывает более семантически сложные кейсы чем бейслайн.

Архитектура решения

ПОЧЕМУ МЫ?

Помимо самого алгоритма, для его легкой интеграции мы реализовали **API** на **GO (Fiber)** позволяющее работать как с csv файлами, так и со стандартными json-запросами.
А еще мы реализовали **UI-интерфейс**, в теории необходимый для удобной отладки работы алгоритма силами непрофессионалов.

Docker-Compose *(можно поднять весь сервис за один make up)*

ML (python)

- Обрабатывает входные данные и предсказывает 8 классов;
- Гибко настраивается через yaml-конфиг;
- *Bedding, view, capacity* - через **BGE-micro (ONNX-runtime);**
- *Class, club, quality, bathroom*- через легкие модели.

kafka ⟷

Back (go)

- Написан на **fiber**;
- Общается с ML через **kafka;**
- Обрабатывает асинхронные запросы с фронта.

API ⟷

Front (type script)

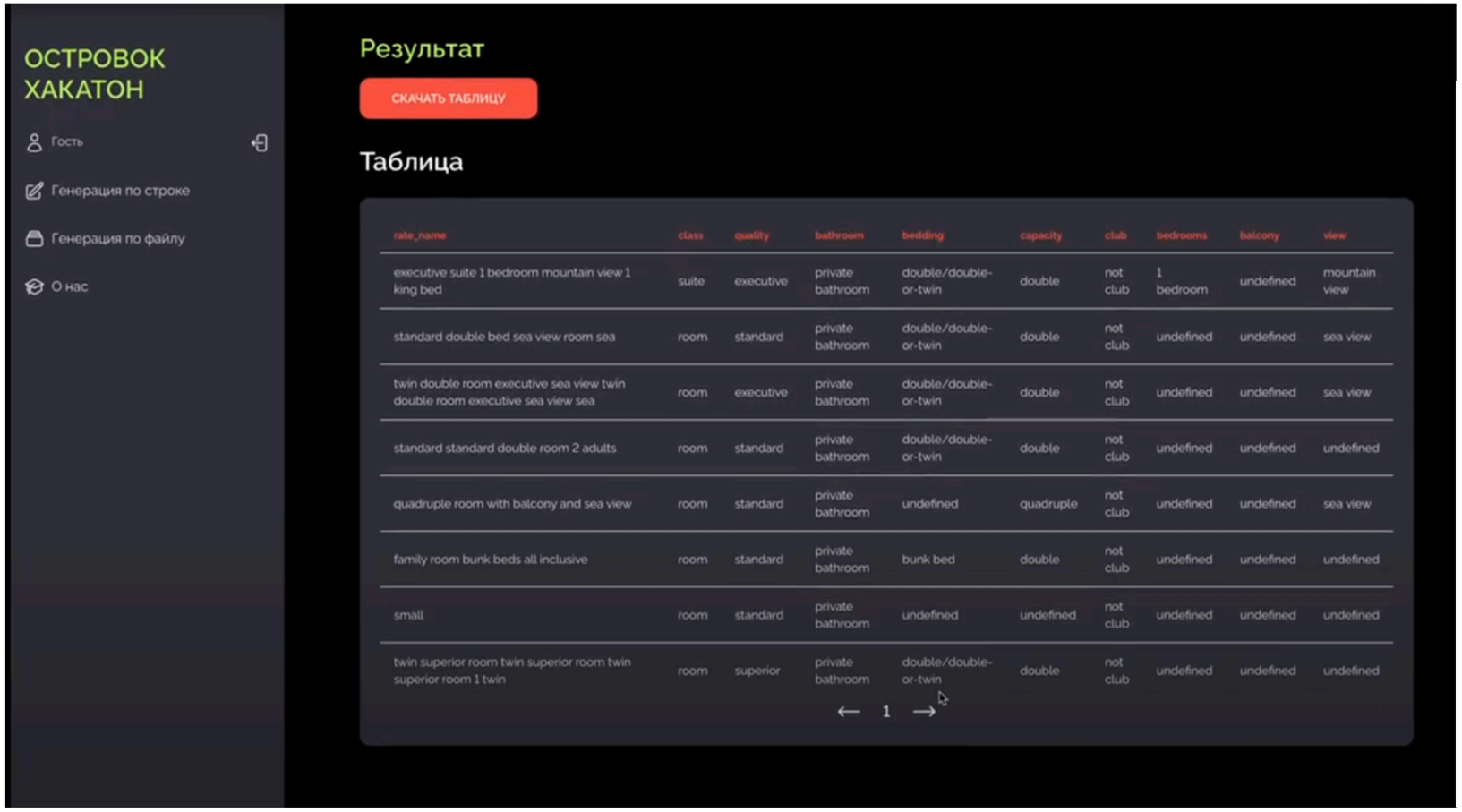
- Написан на **Type Script + React**;
- Позволяет отправить файл и отдельный запрос.
- Позволяет валидировать работу ML глазами.

и еще есть скрипты для переобучения обоих типов моделей по любой разметке.

Примеры интерфейса

ПОМНИТЕ

Мы говорили про реализованный интерфейс. Это он. Он позволяет вам оценить работу модели глазами. В рамках доработок планируется выводить ее.
Полное видео лежит в репозитории и по QR коду.



Примеры работы системы

Показательно Визуально сложные примеры того, как наша модель отработала на грязных данных.

`rate_name`	`dirty['...']`	`predict`
Standard Single Bed in 8-Bed Dormitory Room (En Suite Share Bathroom)	suite (class)	dorm
Double Deluxe Flexible Room Only 200$	run-of-house	room

Пояснение:
En Suite Bathroom - совмещенный санузел
Flexible room - тип зонирования и планировки

Точность **98.5-99.9%** на отложенной выборке в зависимости от колонки

СПАСИБО ЗА ВНИМАНИЕ!

Команда:

Игитов Максим - team-lead, DS. **НИУ ИТМО**

Богодист Всеволод - ML-lead, MLE. **НИУ ИТМО**

Грушевский Георгий - MLE, MLops. **НИУ ИТМО**

Вичук Артем - back-end, dev-ops. **НИУ ИТМО**

Горошко Андрей - front-end, ux-ui. **СПБГУ**