



# Люди X5

## Задача 10

Сервис выделения сущностей из поискового запроса клиента в мобильном приложении торговой сети «Пятерочка»

## О команде

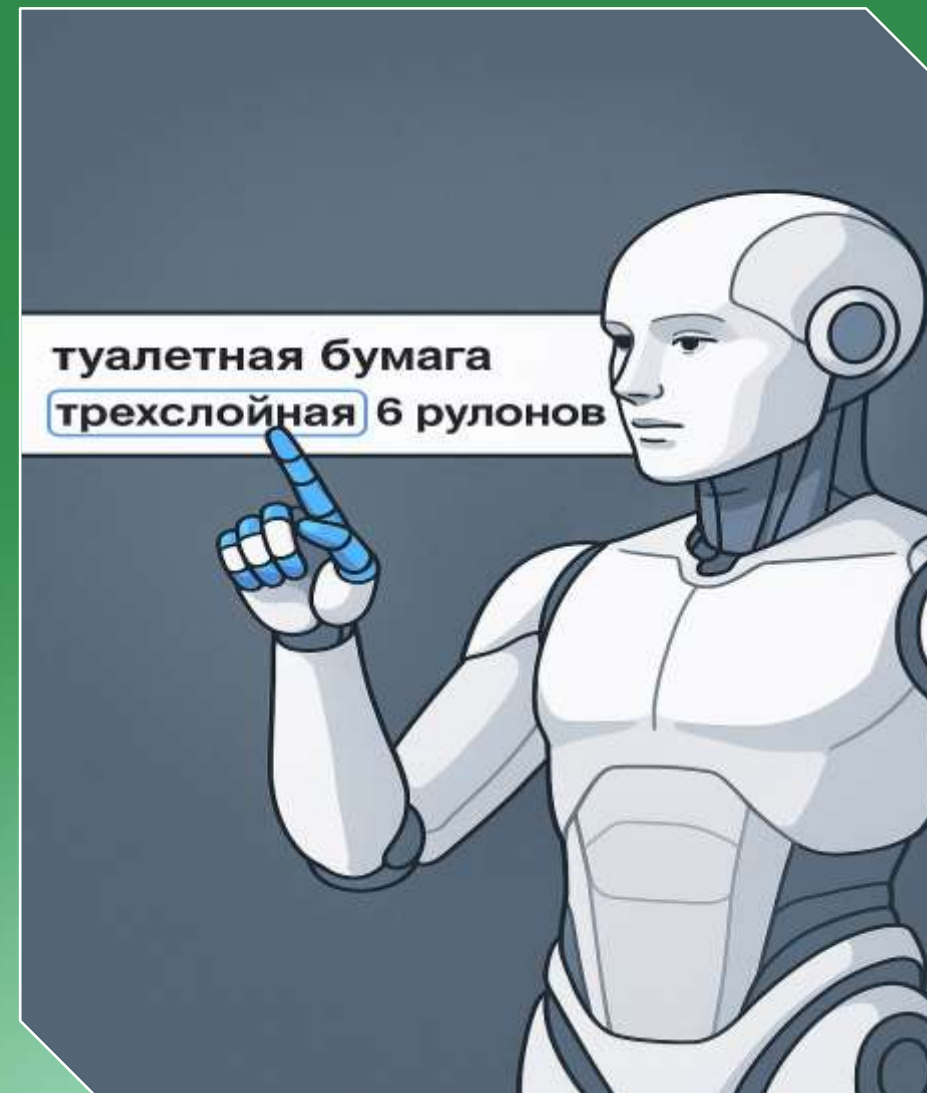
- Город: Москва
- Количество: 4
- Капитан: Шуликов Максим

## Наименование задачи

Сервис выделения сущностей из поискового запроса клиента в мобильном приложении торговой сети «Пятерочка»

## Дальнейшее развитие решения

Добавление новых сущностей: например, упаковка, вес, страна происхождения, чтобы поиск учитывал больше характеристик товара.  
Оптимизация модели: использование более лёгких архитектур, чтобы ускорить работу на мобильных устройствах без потери качества.  
Активное обучение: дообучение модели на реальных пользовательских запросах для постепенного повышения точности.



# КОМАНДА «ЛЮДИ Х5»



ПРОЕКТ  
МЭРА  
МОСКВЫ



ДЕПАРТАМЕНТ  
ПРЕДПРИНИМАТЕЛЬСТВА  
И ИННОВАЦИОННОГО РАЗВИТИЯ  
ГОРОДА МОСКВЫ



РАЗВИТИЕ  
ЧЕЛОВЕЧЕСКОГО  
КАПИТАЛА



ЛИДЕРЫ  
ЦИФРОВОЙ  
ТРАНСФОРМАЦИИ



**Максим  
Шуликов**

- Fullstack
- Tg: @xterris
- 89015974269



**Иван  
Куликов**

- ML
- Tg: @ibroccoli
- 89854259418



**Павел  
Шегай**

- Презентация
- Tg: @shasha\_pegay
- 89266802207



**Раниль  
Хаялиев**

- Tg: ranil\_h



## Краткая история команды

Мы — команда студентов с кафедры «Компьютерные системы и сети» МГТУ им. Н. Э. Баумана.

## Почему выбрали эту задачу?

Мы выбрали задачу выделения сущностей из поисковых запросов, потому что она напрямую связана с реальными потребностями бизнеса и пользователей.

## С какими трудностями мы столкнулись

Самым трудным для нас оказалось создание качественного датасета для обучения, поскольку от полноты и корректности разметки напрямую зависит результат. Также было непросто подобрать оптимальные гиперпараметры и выбрать подходящую модель среди доступных вариантов. Мы преодолели эти трудности за счёт экспериментов, анализа результатов и постепенного улучшения подхода, что в итоге позволило добиться высокой точности.



## Краткое описание решения

Основная задача — выделять сущности из поисковых запросов клиентов в мобильном приложении «Пятёрочка», чтобы улучшить точность поиска товаров. В запросах встречаются такие сущности, как тип продукта, бренд, объём и процентное содержание.

Для решения задачи выбрана модель **sberbank-ai/ruRoberta-large**.

Обучение велось 5 эпох с оптимизатором **AdamW**. Итоговое решение позволяет автоматически выделять сущности в запросах и тем самым повышает релевантность поисковой выдачи.





## Подготовка данных

Для обучения был использован предоставленный организаторами датасет. А также его дополненная версия, сформированная на основе открытого датасета с Kaggle: **Russian Supermarket Prices** (2019-2022). Каждое предложение размечено на уровне символов: выделены начала и концы сущностей и их тип. Далее мы преобразуем разметку:

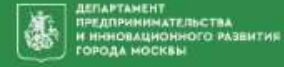
- аннотации приводятся к словесному уровню,
- затем выравниваются с токенами, полученными от токенизатора **HuggingFace**.

## Выбор модели

В качестве базовой модели была выбрана предобученная модель от Сбера: **sberbank-ai/ruRoberta-large**. Она была выбрана, потому что она основана на архитектуре трансформер, которая показывает хорошие результаты в классификации слов с учётом контекста, а также предобучена на корпусе текстов на русском языке.

## Обучение модели

- Архитектура: RoBERTa (Robustly Optimized BERT Pretraining Approach)
- Размер: Large (24 слоя, 1024 hidden size, 16 attention heads)
- Специализация: русский язык
- Параметры: ~355M параметров
- Размер словаря: 50265 токенов



# КОМАНДА «ЛЮДИ X5»