

→ самолет

Мобилити

Задача 16. Алгоритм для поиска предложенных скидок в телефонных разговорах с клиентами.



Команда «Мобилити»



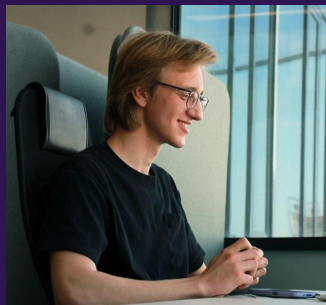
ПРОЕКТ
МЭРА
МОСКВЫ



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ



**Роман
Макаров**

- ML – специалист,
NLP
- @RomanMakarOv
- +79969058641



**Лиана
Марданова**

- ML - специалист
- @liaaaaaana
- +79509469434



**Варвара
Бутенко**

- Дизайнер
- @varbaris
- +79811413788



Введение



Проблема

Отсутствие решения для автоматического распознавания именованных сущностей по транскрипции звонков



Идея

Создание легковесного решения, которое будет распознавать наличие скидок по транскрипту звонка и определять их значение.



Решение

RuBERT-tiny2

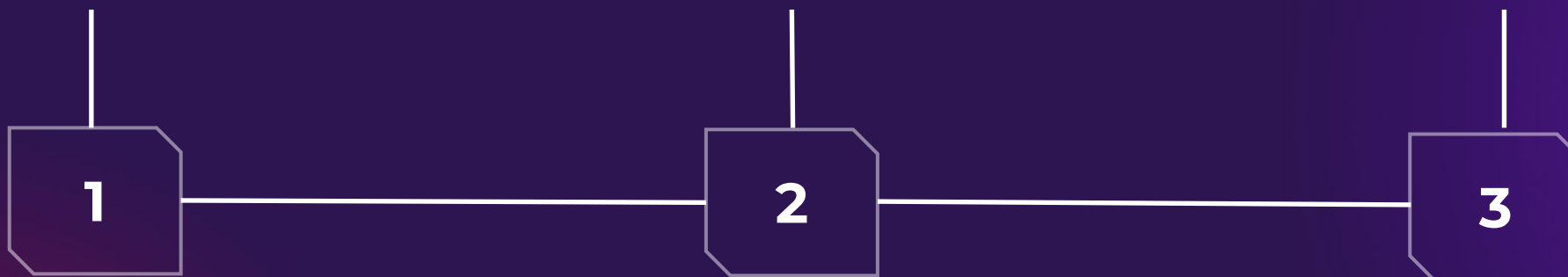
- Тренировка модели на предоставленных данных

FastAPI

Использование фреймворка для создания API для инференса

Usage

Использование - простой GET запрос на страницу '/inference'





Ключевые особенности и преимущества разработанного алгоритма

- ▶ Легковесная модель (RuBERT-tiny2) – веса модели занимают 100 MB. Результат (F1) – **0.72**
- ▶ Быстрый и простой инференс – обработка одного текста (на примере тестовых) через API занимает всего **0.04** секунды. Инференс всего тестового файла занимает 4.2 секунды.
- ▶ Работа строго на CPU, с потреблением всего **400 MB** RAM

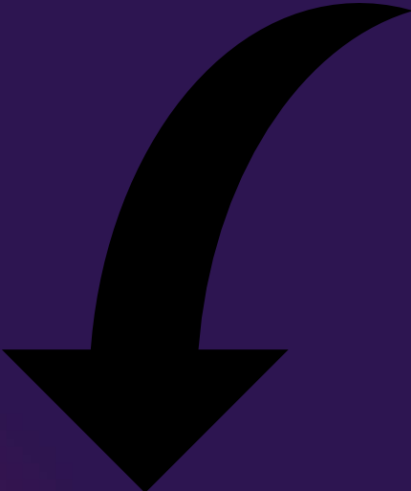


Требовательность к ресурсам и быстродействие

- ▶ Работа строго на CPU, с потреблением всего **400 MB** RAM
- ▶ Инференс одного текста - **0.04** секунды (100 токенов)



Демо



```
import time
import requests

url = "http://0.0.0.0:8000/inference"
data = {"text": "Скидка предоставляется 20 процентов"}

start = time.time()
requests.get(url, json=data)
finish = time.time()

print(f'Time taken: {finish - start:.2f}s\n')
print(response.json())
```

```
arix@arix-ZenBook-UX434DA-UM433DA:~/Desktop/prog/ltc-hack$ python3 test.py
```

```
Time taken: 0.02s
```

```
{'tokens': ['Скидка', 'предоставляется', '20', 'процентов'], 'labels': ['B-discount', '0', 'B-value', 'I-value']}
```

```
arix@arix-ZenBook-UX434DA-UM433DA:~/Desktop/prog/ltc-hack$
```



Планы на будущее

01

Улучшение модели

Fine-tuning модели на новых данных
(и создание синтетических)

02

Доработка прототипа

Создание полноценного веб-приложения, поддерживающего не только API запросы но и прямое взаимодействие через интерфейс

Необходимые ресурсы

- Сервер для обучения модели
- ML инженер
- Full stack Developer



Контакты команды «Мобилити»



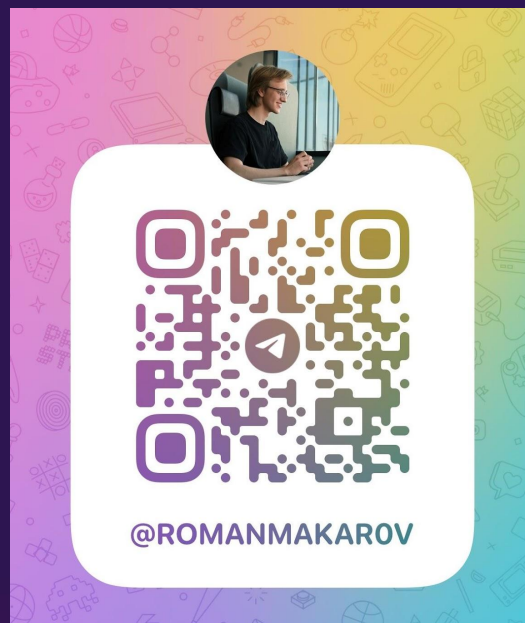
ПРОЕКТ
МЭРА
МОСКВЫ



ДЕПАРТАМЕНТ
ПРЕДПРИНИМАТЕЛЬСТВА
И ИННОВАЦИОННОГО РАЗВИТИЯ
ГОРОДА МОСКВЫ



АГЕНТСТВО
ИННОВАЦИЙ
МОСКВЫ



Роман Макаров

- ML – специалист, NLP
- +79969058641



Лиана Марданова

- ML - специалист
- +79509469434



Варвара Бутенко

- Дизайнер
- +79811413788