ПРОЕКТ МЭРА МОСКВЫ

ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ

АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

самолет

Сквозь турникеты в ML

Задача 16

Алгоритм для поиска предложенных скидок в телефонных разговорах с клиентами

Задача и цель

ПРОЕКТ МЭРА МОСКВЫ | ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ | АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

Задача: разработка автоматизированного решения для определения скидок в транскрибированных записях телефонных разговоров

Целевой результат: повышение качества анализа влияния скидок на принятие решений клиентами

Решение: классическая задача распознавания сущностей, для которой используются предобученные модели для решения NER-задач

Пример записи:

... где он находится о там есть да уже готовы какая площадь вас интересует секунду так ну вот готовая двухкомнатная пятьдесят квадратных метров с отделкой десять миллионов триста да минимум пятнадцать процентов миллион шестьсот продаж дополнительно могу вам отправить **скидку два процента** она действует в течение двух дней сегодня и завтра удобно сегодня к нам подъехать ...

"{'I-value': [40], 'B-value': [39], 'B-discount': [38]}"

План работы

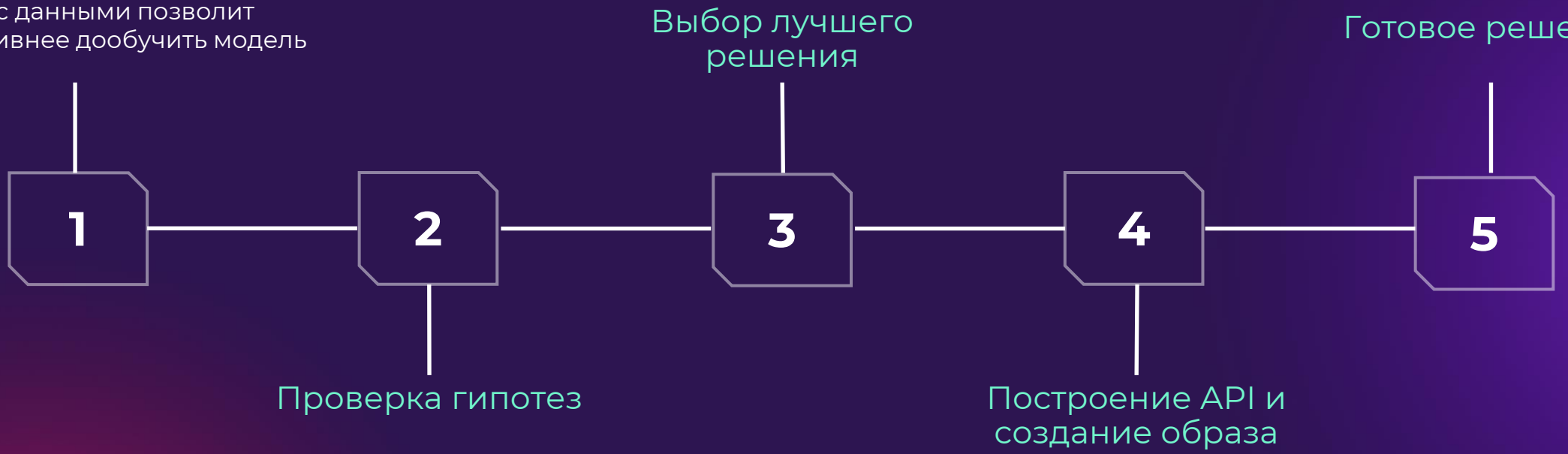
ПРОЕКТ МЭРА МОСКВЫ

ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ

АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

Построение гипотез

- Использование различных NER-моделей трансформеров повысит точность определения необходимых сущностей
- Работа с данными позволит эффективнее дообучить модель



Проверка гипотез и ход работы

ПРОЕКТ МЭРА МОСКВЫ

ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ

АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

503

не пустых значения в target из 3399

344	377	493
I-value	B-value	B-discount

Разведочный анализ данных выявил наличие нескольких записей с большим количеством тегов «I-value» - на этих записях была поправлена разметка

В режиме кросс-валидации на 5-ти фолдах были обучены 3 модели.
Оценка производилась по метрикам f1_score и classification_report из sklearn.metrics

0.5637 F-weight
bert-base-cased

0.6650 F-weight
sberbank-ai/ruBert-base

0.7461 F-weight
Babelscape/wikineural-multilingual-ner

Анализ найденных моделью сущностей показал, что модель находит сущностей в тексте, больше, чем есть в разметке тренировочной выборки

Нужно дополнить разметку

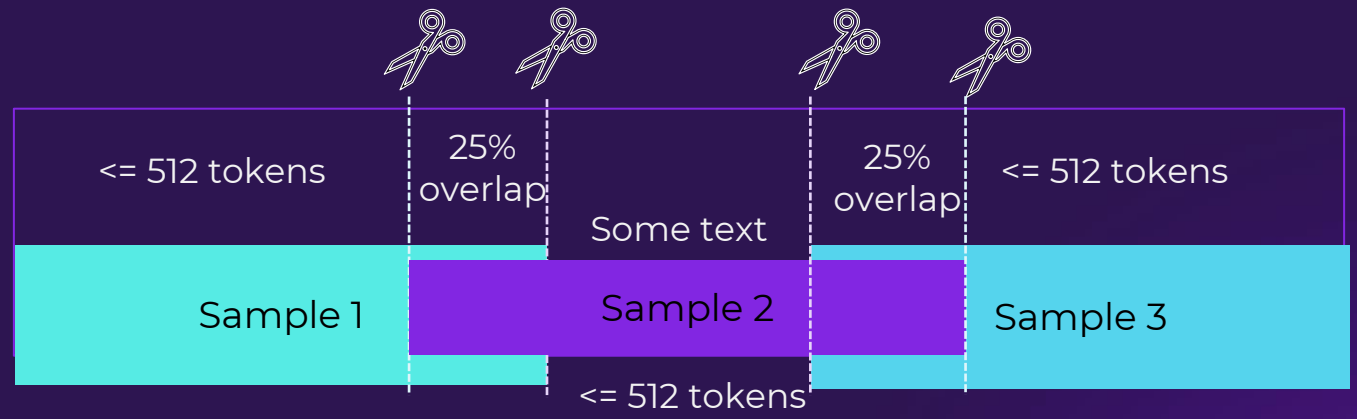
Проверка гипотез и ход работы

503
не пустых значений в target из 3399

4 итерации по дополнению разметки

886
не пустых значений в target из 3399

Тексты в обучающей выборке разбиты на части с перекрытием 25%, чтобы каждая часть не превышала 512 токенов



На обновленном датасете были обучены 3 модели:

модель	Взвешенная F-мера
Babelscape/wikineural-multilingual-ner	0,78136
DeepPavlov/rubert-base-cased-conversational	0,83367
microsoft/mdeberta-v3-base2	0,83379

Результаты последних двух моделей примерно одинаковые, решено остановиться на модели от DeepPavlov, т.к. она более легковесная: в полтора раза меньше по размеру и быстрее обучается, но не забываем про DeBERTa...

Проверка гипотез и ход работы

ПРОЕКТ МЭРА МОСКВЫ
ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ
АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

Сгенерированы синтетические данные:

токенизатор + head GPT2

pretrained sberbank-ai/rugpt3medium_based_on_gpt2

модуль Faker

*max_length случайное количество слов в тексте от 100 до 500 + Faker для добавления случайного количества случайных слов

5500

синтетических записей с разметкой

случайное числительное от 1 до 30 / числительное с фразами от «сто тысяч» до «миллиона»

+

Случайное написание слов «процент», «рубль» и «скидка»

Исходный датасет

3399

Обучающая выборка

400

тестовых значений

Синтетические данные

5000

Обучающая выборка

500

тестовых значений

Проверка гипотез и ход работы

ПРОЕКТ МЭРА МОСКВЫ
ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ
АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

После генерации набора данных выяснилось, что отсутствуют сущности вида «**скидка примерно два три процента**».
Создан второй набор синтетических данных.
Каждый набор по 5000 строк был добавлен к исходному датасету из 3000 строк, обучены 2 модели:

модель	набор данных	F-weight
DeepPavlov/rubert-base-cased-conversational3	первый	0,86485
DeepPavlov/rubert-base-cased-conversational3	второй	0,85205

На втором наборе скор упал.

Идея: смешать оба набора данных, но в какой пропорции?

модель	набор 1	набор 2	F-weight
DeepPavlov/rubert-base-cased-conversational3	3500	1500	0,85453
DeepPavlov/rubert-base-cased-conversational3	2000	1000	0,86684
DeepPavlov/rubert-base-cased-conversational3	4000	2000	0,88433
DeepPavlov/rubert-base-cased-conversational3	5000	2000	0,86573

Лучшее решение

ПРОЕКТ МЭРА МОСКВЫ
ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ
АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

9000
Обучающий датасет

3000
исходный датасет

4000
первый набор синтетики

2000
второй набор синтетики

1200
Валидация

400
исходный датасет

400
первый набор синтетики

400
второй набор синтетики

Устройство	Данные	Количество записей	Время	Текст/сек
Kaggle GPU P100	valid	1200	0:45	25-30
Kaggle CPU	valid	1200	8:52	2-3
RTX 3060	valid	1200	0:23	50-55
i5-13500	valid	1200	1:49	10-12
Kaggle GPU P100	gt_test.csv	482	0:30	15.99
Kaggle CPU	gt_test.csv	482	3:25	2.35
RTX 3060	gt_test.csv	482	0:28	16.67
i5-13500	gt_test.csv	482	2:03	3.89

Тексты размечены на 6 классов:
наличие меток 'B-value', 'I-value', 'B-discount' + синтетика это или нет
В этих моделях использовалась кросс-валидация на 7 фолдах

модель	набор данных	F-weight
DeepPavlov/rubert-base-cased-conversational3	3000+4000+2000	0,93324
microsoft/mdeberta-v3-base2	3000+4000+2000	0,94434

Используемые технологии и ресурсы

ПРОЕКТ МЭРА МОСКВЫ | ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ | АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

01

Python:

PyTorch, Pandas, Numpy, Scikit-learn

02

FAST API, Docker

- Сборка докер-контейнера (15 минут):

docker build -t aeroplane_app

- запуск контейнера:

docker run -d -p 8000:8000 --name aeroplane aeroplane_app

- Размер:

4.88GB в оперативной памяти

03

Обучение модели:

kaggle

- Обучение моделей семейства BERT на данных v1 и v2 - 4.5 часа на GPU P100 (обучение на одном фолде - 50-55 минут). Модель семейства DeBERTa обучается примерно 8 часов (по 1.5 часа на фолд).
- Обучение моделей семейства BERT на данных v3 - 6.5 часов на GPU P100 (обучение на одном фолде - 75-80 минут). Модели семейства DeBERTa не хватило выделенного лимита на сессию.
- в более 80% случаев наибольший скор получался у моделей, обученных на первом фолде.
- Для дальнейшего обучения использовался первый фолд, время обучения составило чуть более 4-х часов.

04

Получение синтетических данных:

kaggle

Генерация 5500 текстов (> 11 часов)

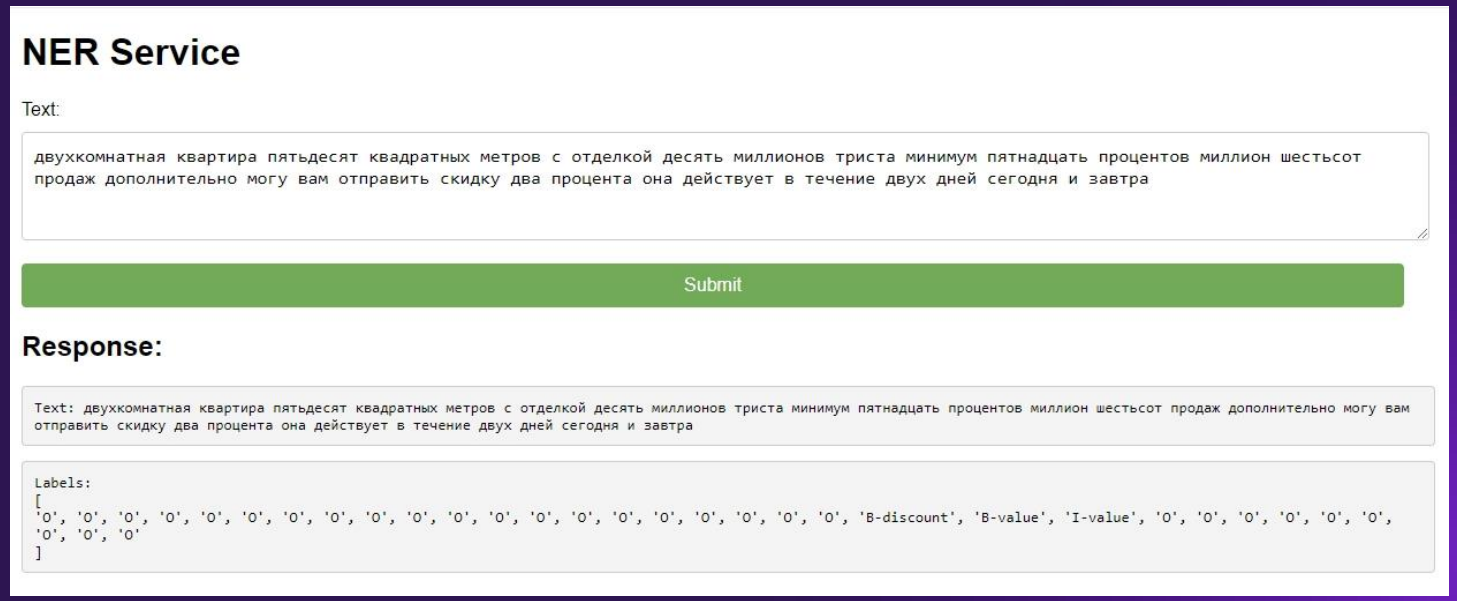
Развертывание сервиса

ПРОЕКТ МЭРА МОСКВЫ
ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ
АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

Отправка запроса к API на Python

```python
import requests
import json

url = "http://127.0.0.1:8000/ner"

text = (
    "двухкомнатная квартира пятьдесят квадратных метров с отделкой десять миллионов триста "
    "минимум пятнадцать процентов миллион шестьсот продаж дополнительно могу вам отправить "
    "скидку два процента она действует в течение двух дней сегодня и завтра")

payload = {"text": text}
headers = {"Content-Type": "application/json"}

response = requests.post(url, headers=headers, data=json.dumps(payload))

if response.status_code == 200:
    result = response.json()
    print("Text:", result['text'])
    print("Labels:", result['labels'])
else:
    print(f"Error: {response.status_code}")
    print(response.text)
```

Отправка запроса к API из браузера




```
Text: двухкомнатная квартира пятьдесят квадратных метров с отделкой десять миллионов триста минимум
 пятнадцать процентов миллион шестьсот продаж дополнительно могу вам отправить скидку два процента
 она действует в течение двух дней сегодня и завтра
Labels: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O',
 'O', 'O', 'B-discount', 'B-value', 'I-value', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
```

Предложения и развитие проекта

ПРОЕКТ МЭРА МОСКВЫ
ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ
АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

Использование более точной модели транскрибации и диаризации, что позволит обрабатывать более структурированный текст.

Расширение датасета и повторная проверка разметки.

Использование более сложной модели.

Использование более сложной схемы разметок. Вместо использования простых меток сущностей, таких как B-value, I-value и O, возможно использование вложенной схемы разметок, которая различает перекрывающиеся сущности и другие более сложные отношения.

Использование предварительно обученных вложений. Вместо обучения вложений слов с нуля, использовать предварительно обученные вложения, такие как GloVe или ELMo, которые уже содержат семантическую информацию о словах.

Применение ансамблевого обучения. Обучение нескольких моделей NER с разными архитектурами и гиперпараметрами с объединением их прогнозов для повышения качества.

Сквозь турникеты в ML