МИНИСТЕРСТВО ЦИФРОВОГО РАЗВИТИЯ СВЯЗИ И МАССОВЫХ КОММУНИКАЦИЙ
Ордена Трудового Красного Знамени
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский технический университет связи и информатики»

Кафедра «Математическая кибернетика и информационные технологии»

Лабораторная работа №10

по дисциплине «Математические основы баз данных»

по теме:

«Обращение к Базе данных на естественном языке.»

Выполнил: Студент группы
БПИ2403
Сон Владимир Сергеевич
Проверил:
Старший преподаватель
Фатхулин Тимур Джалилевич

Москва
2025

Цель работы

Изучить механизмы работы с созданием модели, посредством которой реализуется обращение к базе данных на естественном языке.

Основные теоретические сведения

- NLP (Natural Language Processing) - обработка естественного языка;
- TAPAS - нейросетевая модель Google AI для ответов на вопросы по табличным данным;
- Токенизация - разбиение текста на отдельные элементы (токены);
- TensorFlow/Transformers - библиотеки для машинного обучения.

Задачи

1. Импортировать таблицу из базы данных;
2. Подключить базу данных к среде Google Collab;
3. Создать модель и функцию, обучение которых приводит к нужным результатам;
4. Задать вопросы базе данных.

Ход работы

1) Импортируем таблицу из базы данных

```
> sudo mariadb -D TradingArea -e "SELECT ClientID, REPLACE(CompanyName, ',', ';') as
 CompanyName, REPLACE(Address, ',', ';') as Address, Phone, BankDetails, REPLACE(Con
tactPerson, ',', ';') as ContactPerson FROM Client" | sed 's/\t/,/g' > ~/TradingArea
Data/client.csv
> sudo mariadb -D TradingArea -e "SELECT * FROM Store" | sed 's/\t/,/g' > ~/TradingA
reaData/store.csv
> sudo mariadb -D TradingArea -e "SELECT LeaseID, ContractID, StoreID, REPLACE(Lease
Period, ',', ';') as LeasePeriod FROM Lease" | sed 's/\t/,/g' > ~/TradingAreaData/le
ase.csv
```

2) Подключаем базу данных к среде Google Collab. Создаём модель и функцию, обучение которых приводит к нужным результатам

```python
from transformers import AutoModelForTableQuestionAnswering, AutoTokenizer, pipeline
import pandas as pd
from google.colab import drive

drive.mount('/content/drive')

client_data = pd.read_csv('/content/drive/MyDrive/TradingAreaData/client.csv')
print(client_data.head())

store_data = pd.read_csv('/content/drive/MyDrive/TradingAreaData/store.csv')
print(store_data.head())


lease_data = pd.read_csv('/content/drive/MyDrive/TradingAreaData/lease.csv')
print(lease_data.head())

client_data = client_data.astype(str)
store_data = store_data.astype(str)
lease_data = lease_data.astype(str)

model_name = 'google/tapas-base-finetuned-wtq'
tapas_model = AutoModelForTableQuestionAnswering.from_pretrained(model_name)
tapas_tokenizer = AutoTokenizer.from_pretrained(model_name)

nlp = pipeline('table-question-answering',
               model=tapas_model,
               tokenizer=tapas_tokenizer)

def ask_question(query, data):
    result = nlp({'table': data, 'query': query})
    answer = result['cells']
    return answer

ask_question(
    "What is the maximum DailyRentCost?",
    store_data
)
```

Импортированны нужные библиотеки для работы с моделью. Подключён Google Диск для экспорта БД и конвертированны данные из CSV. Загружена модель TAPAS и созданна функция для задавания вопросов.

3) Задаём вопросы базе данных

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
   ClientID               CompanyName                       Address  \
0         1             ООО "Силсонг"     г. Москва; ул. Мтуси; д. 1
1         2             ЗАО "KDECALL"    г. Москва; ул. Мтуси; д. 34а
2         3             ООО "Гномики"    г. Москва; ул. Мтуси; д. 89
3         4  ЗАО "Sonderkraftfahrzeug"  г. Байконур; ул. Янгеля; д. 6
4         6          ОАО "МамДайДеняг"    г. Москва; ул. Мтуси; д. 25

                Phone           BankDetails                ContactPerson
0  8 (999) 111-22-33  40702810580000001111  Дзикевич Максим Вячеславович
1  8 (999) 444-55-66  40704577890000001111       Лучшев Вадим Алексеевич
2  8 (999) 777-88-99  40708983440000001111   Голубенко Максим Алексеевич
3  8 (916) 675-75-75  40700542010000001111        Сон Владимир Сергеевич
4  8 (999) 535-67-89  40702810222222222222          Сон Андрей Сергеевич
   StoreID  Floor    Area  HasAirConditioning  DailyRentCost  ShoppingCenterID
0        8      1   50.00                   1         2500.0                 4
1        9      1  120.50                   1         5500.0                 4
2       10      2   75.25                   0         3200.0                 4
3       11      1  200.00                   1         9000.0                 5
4       12      3   45.75                   1         1800.0                 5
   LeaseID  ContractID  StoreID LeasePeriod
0        1           1        8  12 месяцев
1        2           1        9   6 месяцев
2        3           1       10   6 месяцев
3        4           2       11  18 месяцев
4        5           2       12   6 месяцев
Device set to use cpu
/usr/local/lib/python3.12/dist-packages/transformers/models/tapas/tokenization_tapas.py:2700: FutureWarning: Series.__getitem__ t
  text = normalize_for_match(row[col_index].text)
/usr/local/lib/python3.12/dist-packages/transformers/models/tapas/tokenization_tapas.py:1494: FutureWarning: Series.__getitem__ t
  cell = row[col_index]
['9000.0']
```

Вывод

В ходе работы была изучена и успешно протестирована система для интерактивного взаимодействия БД с использованием технологий обработки естественного языка (NLP).