

СИСТЕМА ДЛЯ ГЕНЕРАЦІЇ СТРУКТУРОВАНОГО ТЕКСТУ

Качанов С. А.

Науковий керівник – доцент каф. ПІ, к.т.н., доцент Турута О.П.

Харківський національний університет радіоелектроніки

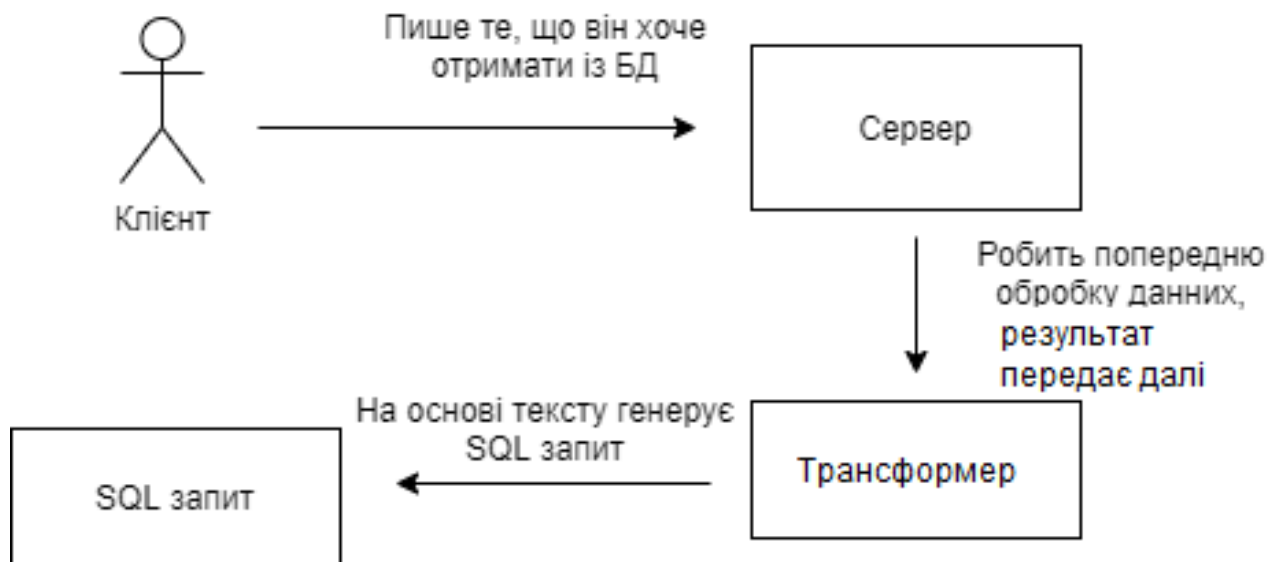
(61166, Харків, просп. Науки, 14, каф. ПІ, тел. (057) 702-14-46)

e-mail: stanislav.kachanov@nure.ua

All information about everything such as all sales in different online-shops, information about people etc. save and base in database. This work is devoted to the problem of understanding and using database for ordinary people. I propose one variant how to solve this problem – transform human-language to SQL query.

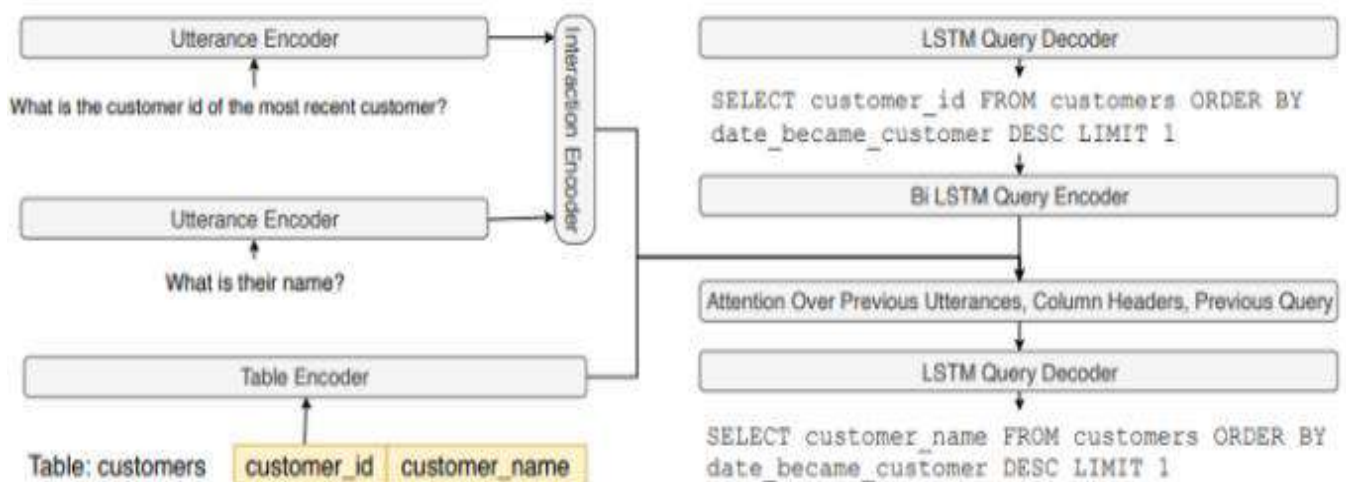
Використання новітніх технологій у сучасному світі, для оптимізації роботи з документами набуває все більшого поширення у світовій практиці. Злагоджена організація роботи з документами є важливою складовою частиною процесів і прийняття управлінських рішень, яка впливає на економічність, оперативність і надійність функціонування апарату управління установи, культуру праці управлінського персоналу і якість управління. Універсальним інструментом створення систем управління документами є системи управління базами даних (СУБД). Проте, на превеликий жаль, більшість людей просто не знають як користуватися базами даних, тому що для цього потрібно знати спеціальний мовний запитів - SQL, яким ти будеш взаємодіяти з базою даних (БД). Для цього потрібні Data Base Analytics. Але ж це досить незручно, та досить дорого.

Саме тому я пропоную варіант розв'язання цієї проблеми – зробити програмну систему, яка буде трансформувати звичайний текст до SQL-запиту. Ця проблема нова і, насправді, існує не так багато варіантів її розв'язку. Усі вони використовують штучний інтелект (AI) та машинне навчання (ML), що досить логічно. Наприклад, за допомогою трансформерів. Трансформер надає тисячі попередньо навчених моделей для виконання таких завдань над текстами, як класифікація, витяг інформації, відповіді на запитання, узагальнення, переклад, генерація тексту тощо на понад 100 мовах. Його мета - зробити сучасну обробку природної мови (NLP) простішою у використанні для всіх. У той же час кожен модуль python, що визначає архітектуру, може бути використаний як автономний і модифікований для швидких дослідницьких експериментів. Мабуть, найбільш сучасним трансформером є GPT-3 – авторегресійна модель мови, яка використовує глибинне навчання для генерації тексту. Повна версія GPT-3 має ємність у 175 мільярдів параметрів машинного навчання, остання версія GPT-3 є частиною тенденції попереднього тренування представлень мови в системах обробки природної мови.



Діаграма 1 – Концепція роботи програми.

На Діаграмі 1 змальована загальна концепція програмної системи, тобто людина звертається з тим, що вона хоче отримати, далі після попередньої обробки та звертання до трансформеру в результаті отримується SQL-запит.



Діаграма 2 – Загальна схема генерації запиту.

На Діаграмі 2 описан концептуальний алгоритм роботи генерації запиту.

Отже, ця робота може, дійсно, стати досить корисною у повсякденному житті у різних сферах та використовуватися навіть у нашому навчальному закладі. Проте очевидно, що стовідсоткової правильності та точності при генерації запитів отримати не вдасться, але треба зробити все, аби максимізувати її навіть при важких запитах з підзапитами, різними аналітичними функціями та створенням допоміжних таблиць та уявлення.