Міністерство освіти й науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра мультимедійних систем факультету інформатики

NLP: АНАЛІЗ МЕНЮ ЗАКЛАДІВ ХАРЧУВАННЯ КИЄВА ДЛЯ ПОШУКОВИКА СТРАВ

Текстова частина до курсової роботи
за спеціальністю «Комп'ютерні науки» 122

Керівник курсової роботи
ст.в. Смиш О.Р.

_____
(підпис)
«____» _____ 2024 р.

Виконала студентка КН-3
Чижова А.О.
«____» _____ 2024 р.

Київ 2024

Міністерство освіти й науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра мультимедійних систем факультету інформатики

ЗАТВЕРДЖУЮ

Зав.кафедри мультимедійних систем,

Доцент., к. ф.-м. н. О.П. Жежерун

_____(підпис)

«____»_____2024 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

на курсову роботу

студентці 3 року навчання БП «Комп'ютерні науки» факультету інформатики

Чижовій Анастасії Олексіївні

ТЕМА: NLP: Аналіз меню закладів харчування Києва для пошуковика страв

Зміст ТЧ до курсової роботи:

Індивідуальне завдання
Вступ
1. Аналіз наявних рішень
2. Формування бази даних
3. Створення застосунку
4. Аналіз зібраних даних
5. Апробація результатів дослідження
Висновки
Список літератури
Додаток

Дата видачі „___» _____ 2023 р. Керівник _____
(підпис)

Завдання отримав _____
(підпис)

Тема:_____

Календарний план виконання роботи:

о /п	Назва етапу курсової роботи	Термін виконання етапу	Примітка
1.	Отримання завдання на курсову роботу.	30.07.2023	
2.	Огляд технічної літератури за темою роботи	22.08.2023 — 15.10.2023	
3.	Аналіз сучасних методів обробки природної мови	10.08.2023 — 24.04.2024	
4.	Збір меню закладів харчування Києва	06.08.2023 — 29.03.2024	
5.	Оброблення зібраних даних	17.10.2023 — 30.03.2024	
6.	Розроблення вебзастосунку для пошуку страв	15.11.2023 — 07.05.2024	
7.	Написання текстової частини роботи.	01.04.2024 — 10.05.2024	
8.	Створення слайдів для доповіді та написання доповіді.	02.05.2024 — 11.05.2024	
9.	Аналіз отриманих результатів із керівником.	11.05.2024	

10.	Коригування роботи за результатами перевірки керівником.	12.05.2024	
11.	Подання роботи на перевірку на плагіат.	13.05.2024	
12.	Захист курсової роботи.	22.05.2024	

Студентка Чижова Анастасія Олексіївна

Керівник Смиш Олег Русланович

«_____»_____

АНОТАЦІЯ

У роботі описано створення пошуковика страв для українськомовних цифрових меню закладів харчування Києва з використанням сучасних можливостей обробки природної української мови, які застосовано для лематизації, класифікації текстів і оптимізованої фільтрації даних за допомогою алгоритмів для пошуку інформації про складники, харчові обмеження, алергени та інші характеристики страв.

У дослідженні проведено аналіз страв цифрових меню закладів харчування Києва, з використанням NLP-моделей, задля унаочнення та формування цілісної картини сучасного стану ресторанного бізнесу у воєнний період в Україні.

Кінцевий програмний продукт дає змогу здійснювати оптимізований і структурований пошук у неструктурованих цифрових меню закладів харчування українською мовою й обирати страви, які розміщено ресторанами.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

NLP – Natural Language Processing (обробка природної мови)

PoS-tagging – Part-of-Speech tagging (розмічування частин мови)

API – Application Programming Interface (прикладний програмний інтерфейс)

JSON – JavaScript Object Notation (запис об'єктів JavaScript)

HTML – Hypertext Markup Language (мова розмітки гіпертекстових документів)

XML – EXtensible Markup Language (розширювана мова розмітки)

QR-код – Quick Response code (код швидкої відповіді)

GPT – Generative Pre-trained Transformer (породжувальний попередньо натренований трансформер)

LLaMA – Large Language Model Meta AI (велика модель мови (LLM), випущена Meta AI)

XPath – XML Path Language (мова запитів, для вибору вузлів з XML документів)

SQL – Structured Query Language (мова структурованих запитів)

ВСТУП

Обґрунтування вибору теми дослідження. У сучасному світі зростає обсяг не структурованої текстової інформації, тому здатність ефективно обробляти та аналізувати великі масиви даних є важливою для різноманітних галузей. Індустрія ресторанного бізнесу в Києві стрімко розширюється, кожен рік відкриваються нові заклади, які розміщують меню українською мовою в цифровому форматі. Автоматизація процесу пошуку страв є неодмінною умовою вдосконалення інструментів для аналізу меню та їхніх складників, зважаючи на дієтичні обмеження, алергічні реакції та особливі харчові вподобання людей.

Покращення технологій методів обробки природної мови (NLP) надає нові можливості для роботи з текстами меню закладів харчування та їхнього оброблення. Лематизація, токенізація, визначення частин мови — методи NLP, за допомогою яких можливо підвищити точність і швидкість пошуку страв.

Вибір теми дослідження також обґрунтовано необхідністю удосконалення україномовних рішень у ресторанному бізнесі. Сучасні застосування здебільшого зосереджено для англомовних користувачів, тому впровадження методів для аналізу і структурування цифрових меню українською мовою сприятиме оптимізації систем обслуговування для користувачів у воєнний період в Україні.

Мета та завдання роботи. Метою запропонованого дослідження є надання можливості користувачам здійснювати оптимізований і структурований пошук у неструктурованих цифрових меню закладів харчування українською мовою й обирати страви, які розміщено ресторанами.

Для досягнення поставленої мети виконано такі завдання:

- Проаналізовано методи обробки природної української мови для дослідження різних аспектів меню закладів харчування.

- Досліджено структуру меню, їхню різноманітність, вміст алергенів у стравах, відповідність харчовим звичкам, а також варіативність опису складників страв.
- Створено застосунок, який надає змогу здійснювати пошук і сортування страв.

Об'єкт дослідження. Процеси пошуку й аналізу інформації меню закладів харчування з використанням технологій обробки природної української мови.

Предмет дослідження. Методи NLP, які використано для лематизації, класифікації текстів і оптимізованої фільтрації даних за допомогою алгоритмів для пошуку інформації про складники, харчові обмеження, алергени та інші характеристики страв у меню закладів харчування Києва.

Методи дослідження. У роботі застосовано такі методи дослідження: абстрактно-логічний аналіз, експерименти, моделювання, аналіз даних, порівняння. Зазначені методи обрано, зважаючи на поставлену мету та завдання дослідження.

Наукова новизна отриманих результатів.

Уперше реалізовано метод, що уможливлює пошук страв українськомовних цифрових меню закладів харчування Києва за допомогою фільтрації складників страв, які можуть спричиняти алергічну реакцію, аби оптимізувати процес вибору харчових пропонувань для користувача.

Уперше проведено аналіз страв українськомовних цифрових меню закладів харчування Києва, з використанням NLP-моделей, задля унаочнення та формування цілісної картини сучасного стану ресторанного бізнесу у воєнний період в Україні.

Удосконалено метод автоматичного видобування значущої інформації зі страв українськомовних цифрових меню закладів харчування задля підвищення точності визначення інгредієнтів і назв страв, що уможливлює коректне їхнє оброблення.

Практичне значення отриманих результатів. У результатах роботи продемонстровано сучасні рішення для аналізу текстів цифрових меню українською мовою, що допомогло оптимізувати структурування описів страв для кращого розуміння наявності певних інгредієнтів.

Розроблену систему використано для покращення пошуку страв у меню ресторанів з урахуванням дієтичних обмежень, спеціальних харчових звичок і потреб. Із застосуванням методів сортування покращено взаємодію з користувачем, що сприяє збільшенню попиту до закладів харчування в Києві та уможливлює оптимальний підхід до здійснення вибору.

Зважаючи на сформовану збірку даних проведено оцінку популярності харчових пропонувань, їхніх складників і ціноутворення. На основі цих даних ресторани можуть визначати актуальність страв і враховувати поточні кулінарні тренди, покращувати клієнтоорієнтованість і конкурентоспроможність у галузі громадського харчування.

Структура й обсяг курсової роботи. Робота складається зі вступу, п’яти розділів, загальних висновків роботи, використаних джерел і додатків. Загальний обсяг курсової роботи 43 сторінки.

РОЗДІЛ 1

АНАЛІЗ НАЯВНИХ РІШЕНЬ

Індустрія закладів харчування активно збільшується, лише за 5 місяців у 2023 році відкрито 172 кафе й ресторани [1]. З розвитком цифрових технологій у сфері ресторанного бізнесу почали використовувати мобільні та вебзастосунки. Вони полегшують взаємодію клієнта із закладом харчування і покращують сервіс персоналізованого обслуговування.

У цьому розділі оглянуто існуючі мобільні й вебзастосунки, які використовуються в харчовій індустрії. Також зазначено методи обробки природної мови (NLP), які використано для оброблення текстової інформації цифрових меню закладів харчування Києва. Описано основні методи NLP, а саме: лематизацію, токенізацію, стемінг, видобування інформації і аналіз частин мови.

1.1. Аналіз застосунків

На платформах розповсюдження прикладних мобільних програм App Store[2] і Google Play[3] знайдено застосунки, як-от: Zupa[4], що працює для закладів у місті Львів, WeWest[5] і Expirenza by mono[6], що працюють у всіх великих містах України. Зазначені застосунки пов’язані з меню різних закладів харчування. Вони мають подібний функціонал, наприклад, можна переглянути розташування закладів на мапі, як проілюстровано на рисунку 1.1.

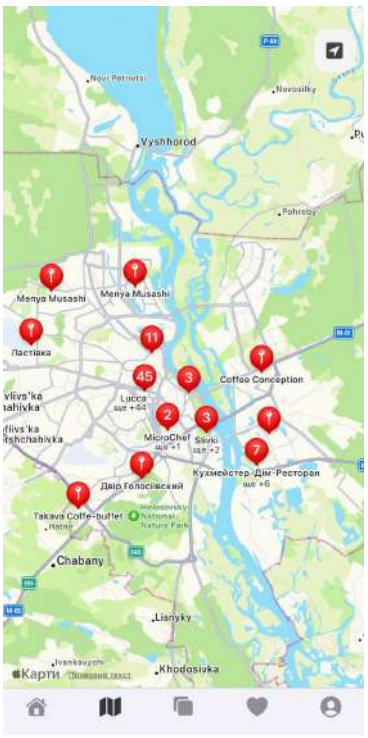

Рисунок 1.1. Приклад мапи з розташуванням закладів у застосунку WeWest

Також, у застосунках можна відсортувати заклади за категоріями, національними кухнями й іншими особливостями, як показано на рисунку 1.2. Наприклад, чи є в закладі дитяче меню, чи можна приходити із тваринами, чи є тераса тощо. Пошук у застосунках WeWest і Zupa можна здійснювати лише за назвою закладу. На головній сторінці розміщено список закладів із їхнім рейтингом, покликанням на меню і номером телефону. В Expirenza, окрім пошуку за назвою закладу, можна здійснювати також пошук за найменуванням страви, до того ж застосунок дає змогу користувачам сканувати QR-код у закладі, щоби переглядати меню і деталізовані описи страв, робити замовлення і оплачувати рахунок зі смартфона.

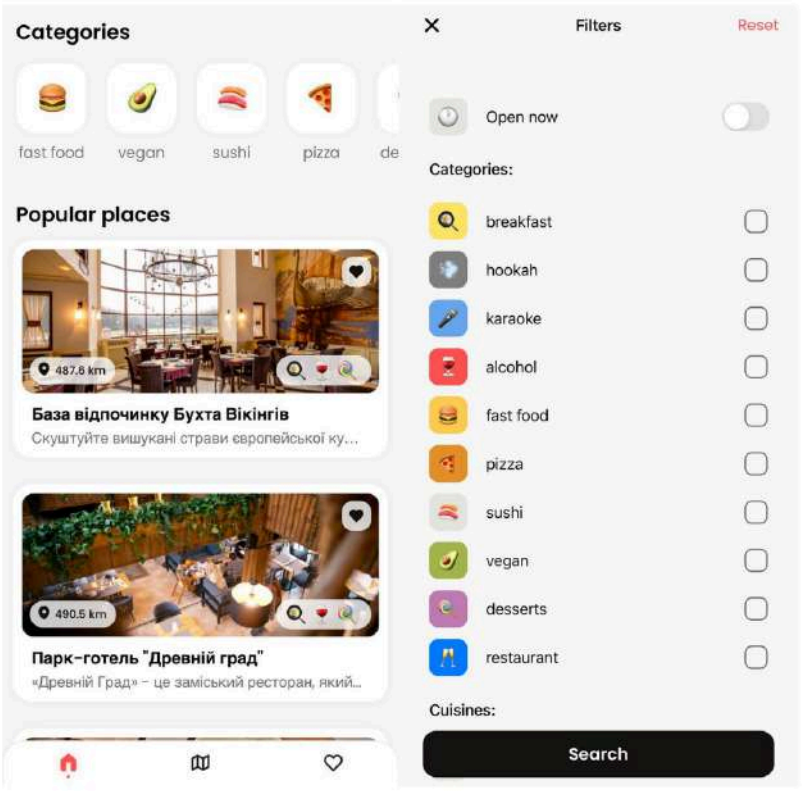

Рисунок 1.2. Опції для сортування закладів харчування у Zupa

QR-код — це різновид двовимірного штрихкоду, який містить певні дані. QR — скорочення від «Quick Response» (швидка відповідь). QR-код кодує цифрову інформацію, що складається з літер та цифр у піксельному форматі.[8]

Окрім мобільних застосунків є також і веб. Наприклад, сайт ChoiceQR [7], він, як і Expirenza надає можливість клієнту сканувати QR-код і має схожий функціонал до вищезгаданих застосунків, проте на відміну від них, у ChoiceQR кожен заклад має окрему вебсторінку, де можна переглянути повне меню, ознайомитися з деталями страв, замовити доставлення їжі, залишити відгук тощо (див.рисунок 1.3).

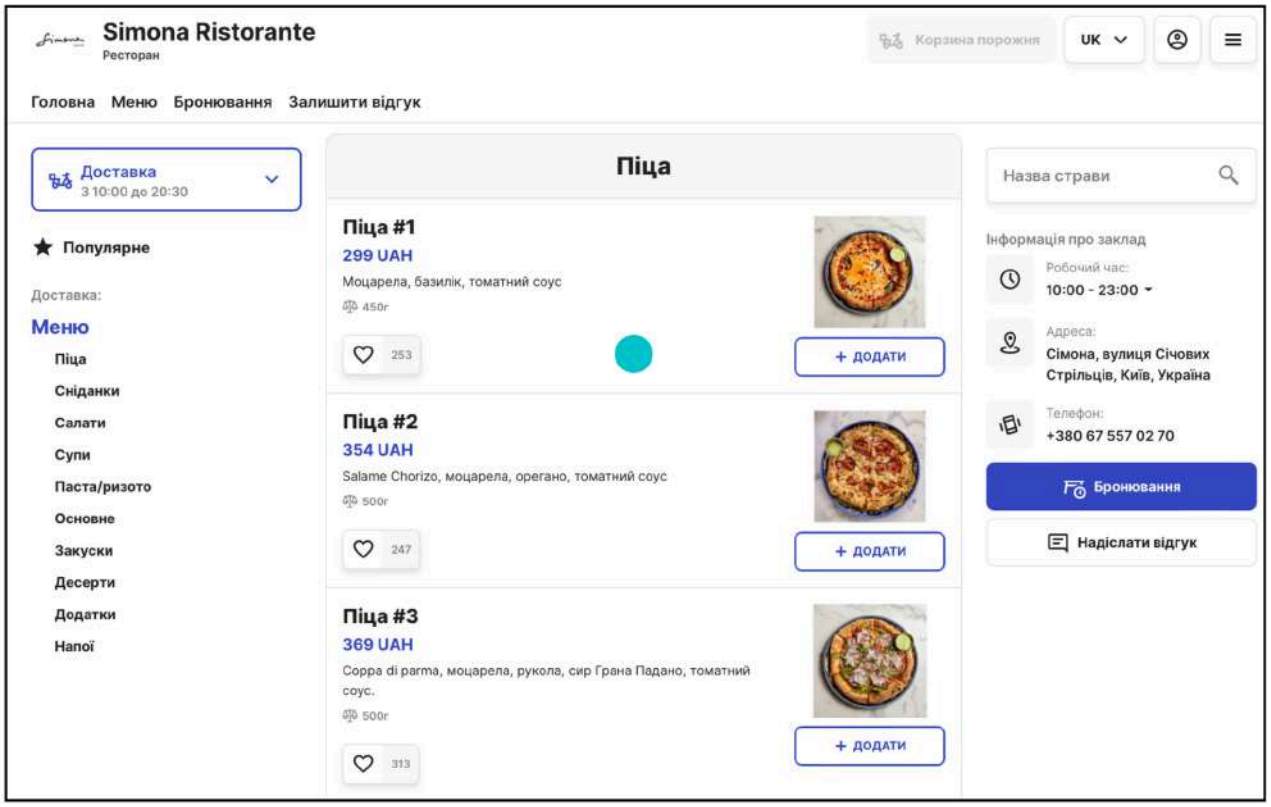

Рисунок 1.3. Загальний вигляд вебзастосунку ChoiceQR

1.2. Обробка природної мови

Мова — це спосіб спілкування, за допомогою якого люди можуть ділитися інформацією, думками, ідеями. Однак, люди не завжди дотримуються правил використання мови, що може створювати непорозуміння не тільки під час спроби передати інформацію комп’ютеру, а й навіть у живому спілкуванні. У розмові притаманно використовувати слова в переносному значенні, описувати думки із різними емоціями, і мову, яку використовують у такій комунікації називають живою. Для того, щоб комп’ютер міг її зрозуміти використовують технології обробки природної мови.

NLP (Natural Language Processing) або обробка природної мови — галузь комп’ютерної лінгвістики, що спрямована на створення технологій, які дають змогу машинам розуміти, аналізувати й відтворювати людську мову. Цю галузь використовують для розуміння змісту та контексту повідомлення. У

запропонованому дослідженні використано декілька ключових методів NLP, які описано далі.

Токенізація — це поділ тексту на менші частини, які називаються токенами. Токенами можуть бути слова, символи або фрази. Отримані токени зазвичай використовуються як вхідні дані для подальших етапів обробки, зокрема векторизація, де токени перетворюються на числові представлення для використання в моделях машинного навчання.

Лематизація (Lemmatization) — процес перетворення слова (токена) до початкової форми. В українській мові для іменників початковою формою є називний відмінок однини, для прикметників — називний відмінок однини чоловічого роду, для дієслова — інфінітив, наприклад, «стрибнув» → «стрибати».

Стемінг (Stemming) — це метод зведення слів із різними відмінками або формами до їхнього основного кореня. Цей процес полегшує впорядкування слів за семантичним значенням і аналіз тексту, тому що зменшує кількість форм слів, які потрібно обробити. Проте в роботі його не використано, тому що порівнюючи зі стемінгом у процесі лематизації враховано контекст слова і визначено його частину мову, тому отримано більш точну базову форму слова.

Видобування інформації (Information Extraction) — процес механічного вилучення структурованої інформації, часто передбачає певну обробку природної мови [10]. Методи видобування інформації використано, наприклад, для визначення назв страв, їхніх характеристик, даних про меню закладів харчування, тому що з їхнім застосуванням можливо зручно організовувати і структурувати великі обсяги неструктурованого тексту, який є в цифрових меню.

Аналіз частин мови (POS tagging) — визначення того, якою частиною мови є кожне слово в реченні (іменник, дієслово, прикметник, прислівник тощо). POS tagging використано для з’ясування ключових слів у запиті від користувача в пошуковику. Наприклад, у реченні «Дешева паста з куркою», «дешева» є

прикметником, що вказує на те, що користувач хоче отримати страву із ціною нижче середнього значення цін у місті Київ. А іменник «паста» свідчить про те, що користувач очікує знайти конкретний вид страви.

1.3. Огляд наявних NLP-моделей

NLP-модель (модель обробки природної мови) — це комп’ютерна система, яка створена для аналізу, розуміння, інтерпретації і генерації людської мови. Такі моделі використовуються в застосунках для розпізнавання мовлення, машинного перекладу, автоматичного підсумовування текстів, розуміння природної мови тощо.

Зазвичай на сайтах закладів можна помітити, що в описі страв написано: «Гарного дня», «Смачного», недоречний текст, або опису бракує. Також використовувався молодіжний сленг і популярні фрази, наприклад, «Трушна паста», «Топчик», «Хелсі страва» (див.рис. 1.4). Помічено, що наявність зазначених фраз ускладнює точне оброблення і класифікацію страв, тому що створює «шум» у даних. Також, брак детального опису додає труднощі в процесі видобування необхідної інформації, а саме: інгредієнти, алергени тощо. Такі випадки оброблено з використанням NLP-моделі, видалено недоречні слова і виділено складники в описах.

Рисунок 1.4. Приклад недоречного опису

Для аналізу обрано моделі: GPT від OpenAI, LLaMA від Meta AI, та Gemini від Google.

LLaMA 2 70B (Large Language Model Meta AI) — це велика мовна модель, розроблена компанією Meta AI. Вебінтерфейс для роботи з LLaMA має вигляд чату, який можна налаштовувати, де System Prompt це повідомлення, яке надається як підказка для керування відповідями системи, Temperature — регулювання випадковості відповіді (чим менша температура, тим точніша відповідь). Кількість токенів, з яких сформовано відповідь теж можна обмежити. У моделі LLaMA обраховується кількість токенів, а не кількість слів.

ChatGPT 3.5 є однією з версій мовних моделей Generative Pre-trained Transformer від OpenAI. Інтерфейс для роботи з ChatGPT 3.5 також має форму чату, де користувач може вводити свої запитання або коментарі, а модель генерує відповіді. Ця модель не має доступу в інтернет і за вказаною інформацією її дані актуальні до січня 2022 року.

Gemini, або Google Bard, — це велика мовна модель, розроблена компанією Google AI. На відміну від інших моделей має не лише інтерфейс чату, але й інтерфейс Notebook, який дає змогу писати код мовою Python і використовувати Gemini для генерації тексту, перекладу різних мов, інформативних відповідей на запитання тощо. А також проти інших моделей має доступ в інтернет і може надавати актуальну інформацію.

UDPipe — це сучасний багатомовний інструмент обробки природної мови. Цей інструмент містить функції для аналізу текстів, а саме: токенізацію, лематизацію, розмітку частинами мови (POS tagging) тощо.

РОЗДІЛ 2

ФОРМУВАННЯ БАЗИ ДАНИХ

2.1. Збір цифрових меню закладів харчування

Для реалізації застосунку пошуку за стравами розглянуто меню закладів харчування в місті Київ, які розміщено в інтернеті в цифровому форматі. Для пошуку й завантаження використано сайт Choice.qr [7]. Таке рішення зумовлено тим, що на цьому сайті міститься понад шість тисяч меню закладів харчування і всі вони мають однакову структуру.

На головній сторінці сайту Choice.qr немає змоги здійснити пошук закладів. Дізнатися адресу ресторану можливо безпосередньо в його меню, тому це враховано під час створення програми для автоматизованого завантаження меню. На вебсторінках закладів містяться переліки страв, для кожної з них вказано назву, проте виявлено, що в деяких випадках не зазначено опис, ціну, вагу.

Створено й використано метод на мові Python і завантажено меню ресторанів з усіма сторінками на локальне сховище в HTML-форматі. Оскільки, зазвичай у закладах категорії меню розміщено за різними покликаннями, у програмі створено функцію для їхнього об’єднання в одну сторінку для зручного оброблення надалі.

Категорії в меню закладів харчування використовуються для організації страв за певними групами, наприклад, основні страви, десерти, напої тощо.

2.2. Бібліотека «Selenium» для автоматизації оброблення даних

Для роботи із мовними моделями, які зазначено в першому розділі, потрібно вводити повідомлення у форматі: «Випиши в один рядок із тексту лише перелік усіх інгредієнтів через кому, українською мовою без зайвого

тексту. Інгредієнти записати через кому в один рядок. Текст: {description}». Модель ChatGPT 3.5 впоралася із цим завданням найкраще, тому для подальшої роботи обрано її. Далі отриманий результат потрібно скопіювати і вставити на місце попереднього опису страви. Така обробка вручну була б часозатратною і неефективною. Для автоматизації цього процесу використано бібліотеку Selenium[19], що є інструментом для автоматизації роботи з веббраузерами, який дає змогу імітувати дії користувача, наприклад, клацання на кнопки, введення даних у форми, навігація між сторінками та інші взаємодії. Тому дані відправлено у формі запитів до моделі й зібрано відповіді від неї із застосуванням зазначеної бібліотеки.

Щоби така автоматизація працювала, необхідно вказати вебелементи сторінки й описати поведінку для програми. Для оброблення описів меню використано: стрічку для введення запиту; кнопку для надсилання і елемент, куди виведено результати роботи моделі.

Для доступу до елементів інтерфейсу ChatGPT 3.5 використано ID або ж XPath. Наприклад, поле для введення запитів у моделі має ID «prompt-textarea». XPath, що розшифровується як XML Path Language, є мовою запитів, яку використано для вибору вузлів з XML-документів. В автоматизації тестування з Selenium, XPath використано для визначення елементів, з якими потрібно взаємодіяти на вебсторінці. Приклад розробленого коду для розпізнавання показано на рисунку 2.1.

```
try:
ingredients_element = WebDriverWait(driver, 10).until(
EC.presence_of_element_located((By.XPATH, "//div[@data-message-author-role='assistant']//p"))
)
```

Рисунок 2.1. Демонстрація коду для пошуку елементу відповіді моделі

Спочатку здійснено пошук елементу для введення запиту, далі автоматизовано введено текст для виокремлення саме інгредієнтів страв. Кожна страва має власні складники, у меню опис розміщено під тегом «styles_CollapsedText__blArN styles_collapsed__3d5qr styles_line_lampShort__JCqRm». Враховано всі можливі зазначені інгредієнти страви, тому що в запит додано не лише текст із вказаного вище тегу, а й із її назви, якій присвоєно тег «styles_menu-item-title__92eAl». Розроблено функцію для формування запитів до моделі і збереження отриманих відповідей із використанням відповідних HTML-елементів.

2.3. Використання сформованої бази даних

У роботі застосунку використано систему керування базами даних, що є комплексом програмних засобів, які використовуються для створення нової бази даних — структури, призначеної для зберігання та обробки великого, взаємопов’язаного обсягу інформації і її редагування. У неї попередньо завантажено страви із їхнім описом, ціною, вагою та особливими характеристиками, а саме: гостра; вегетаріанська — без м’яса, риби чи інших морських істот; веганська — без продуктів тваринного походження; без глютену.

Для створення застосунку, у якому можливо здійснювати пошук страв, потрібен швидкий доступ до структурованих меню закладів. Для користувача вагомою є інформація про назву страви, її складники, ціну, вагу, можливі алергени та доступність страви, тобто розташування закладу, де її можна замовити. У завантажених HTML-файлах меню містилася інформація, що не стосується страв та їхніх складників, наприклад, покликання на соціальні мережі, тому використано базу даних для зберігання потрібної інформації.

Встановлено застосунок MAMP[13], який застосовано для роботи з MySQL. Усі інструменти для розроблення встановлюється одним пакетом без

необхідності окремого налаштування Apache[15], PHP[14] або інших компонентів. У проєкті під’єднано та заповнено базу даних у PhpMyAdmin[17] з використання програми, яку створено на Python з бібліотекою BeautifulSoup[16]. BeautifulSoup[16] призначена для парсингу HTML-документів. Вона дає змогу збирати інформацію з вебсторінок, у ній надано інструменти для навігації, пошуку та зміни структури даних. Вебсторінки зазвичай неправильно структуровані, тому зазначену бібліотеку адаптовано до різних особливостей документів для обробки та аналізу вебданих.

PhpMyAdmin[17] — це інструмент для адміністрування баз даних MySQL через вебінтерфейс. З використанням його зручного інтерфейсу написано SQL-запити, здійснено керування таблицями, стовпцями, зв’язками, індексами, користувачами, дозволами та іншими об’єктами бази.

Для кожного закладу створено окрему таблицю, де її назва відповідає назві закладу. У першому стовпчику містяться назви всіх страв, що є в меню, у другому стовпчику розміщено складники кожної страви, у наступних стовпчиках зазначено розташування закладу в Києві, покликання на зображення страви й за допомогою змінної Boolean заповнено стовпчики, що відповідають характеристикам страви як: гостра, вегетаріанська, веганська, без глютену (див. рисунок 2.2). Також кожній страві присвоєно унікальний ID для того, щоб забезпечити єдиний, неповторний ідентифікатор. У базі даних створено та налаштовано керування зв’язками між таблицями і її використано для уникнення дублікатів записів.

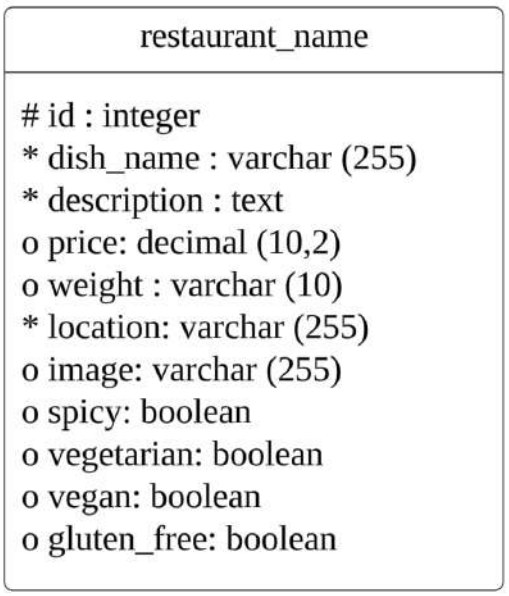

Рисунок 2.2. Схема бази даних цифрових меню закладів харчування

2.4. Висновки до розділу

У цьому розділі розглянуто процес створення та налаштування бази даних для розробленого вебзастосунка. Зібрано цифрові меню закладів харчування українською мовою міста Київ із сайту ChoiceQR. Застосовано метод автоматизованого завантаження HTML-сторінок для забезпечення повного охоплення всіх меню. Застосовано бібліотеку Selenium для автоматизації введення запитів до моделі ChatGPT 3.5 та отримання відповідей із переліком інгредієнтів страв. Це дало змогу значно скоротити час для оброблення великих обсягів даних із цифрових меню ресторанів. Для кожного закладу створено окрему таблицю в базі даних, де записано необхідні дані про страви.

Отже, у такий спосіб організовано значний обсяг інформації з неструктурованих цифрових меню закладів харчування Києва для подальшого застосування.

РОЗДІЛ 3

СТВОРЕННЯ ЗАСТОСУНКУ

Третій розділ дослідження присвячено написанню вебзастосунку, який надає користувачам детальну інформацію про страви в цифрових меню закладів харчування Києва за допомогою оптимального пошуку та сортування, враховуючи різні критерії, наприклад, інгредієнти, ціну, вагу, наявність алергенів та харчові вподобання.

У розділі описано процес розроблення застосунку, який складається з декількох етапів. Спочатку здійснено оброблення запиту від користувача, лематизовано і проаналізовано частини мови, а також виокремлено ключові слова задля пошуку, для цього використано технології обробки природної мови (NLP). Далі написано алгоритми, які враховують спеціальні запити, наприклад, сортування за ціною.

Детально описано, як реалізовано пошук страв у базі даних, яку сформовано з окремих таблиць для кожного закладу. Створено та виконано SQL-запити через Express-сервер для цього розглянуто і використано технології Node.js та MySQL.

3.1. Оброблення запиту від користувача

На головній сторінці розміщено поле для введення запиту, яке використано для пошуку страв (див. рис. 3.1). У тексті запиту можливі слова в різних граматичних формах, тому їх оброблено і приведено до основної форми (леми) у функції, яку написано для надсилання запиту на локальний сервер UDPipe. Отримані леми проаналізовано, і слова, які позначено іменниками (NOUN), використано для здійснення пошуку. Застосовано такий алгоритм, тому що в запиті найбільш важливою частиною є іменники, наприклад, у тексті «Паста з грибами» виокремлено «паста» і «гриб».

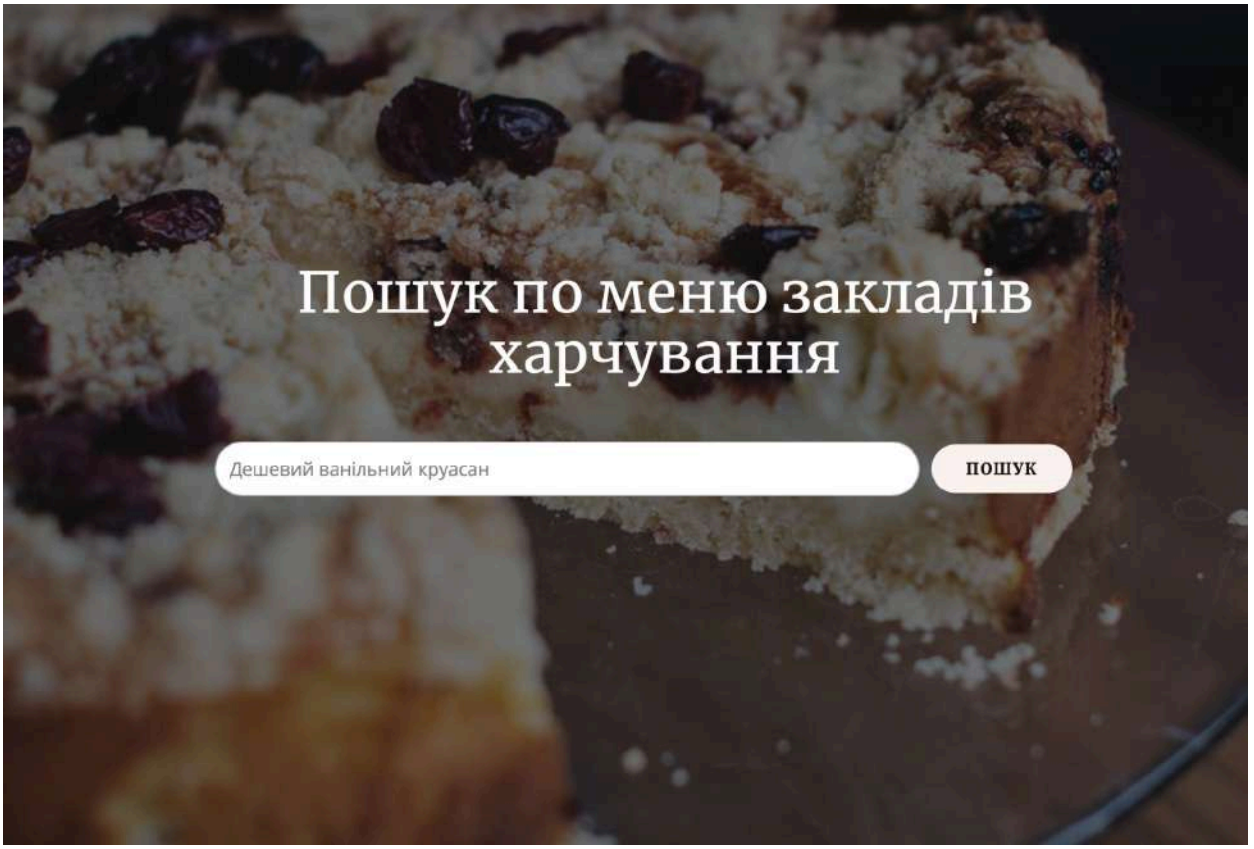

Рисунок 3.1. Демонстрація головної сторінки застосунку з пошуковою стрічкою

Після першого кроку, де отримано лематизовані іменники, запит оброблено далі для визначення, чи містяться спеціальні ключові слова для фільтрації. Наприклад, користувач може шукати страви, що не містять певні алергени або вегетаріанські страви. У застосунку проаналізовано такі запити та відповідно налаштовано пошук, щоби відобразити точний результат.

Також у застосунку розроблено алгоритм, який реагує на слово «дешевий» у запиті, наприклад, «Дешевий ванільний круасан». У такому випадку результати пошуку посортовано за зростанням ціни, і першими відображено найдешевші страви. Поєднано аналіз ключових слів і фільтрацію, й у такий спосіб створено високий рівень точності пошукових запитів, надаючи користувачам можливість легко знаходити страви, що задовольняють їхні особливі вимоги і фінансові можливості.

3.2. Відображення страв у застосунку

Як зазначено вище, усі страви розміщено в базі даних, де кожному закладу створено окрему таблицю. Для того, щоби виводити в застосунку потрібні страви, пошук здійснено серед усіх наявних таблиць. Під час створення застосунку для об'єднання таблиць використано SQL запити, які сформовано і виконано через Express-сервер із застосуванням технологій Node.js[18] — середовище виконання програм, яке надає можливість працювати з кодом мови програмування JavaScript на сервері та MySQL.

Спочатку виконано SQL запит «SHOW TABLES;'» і визначено всі наявні таблиці в базі даних MySQL. Отримано перелік таблиць і в застосунку сформовано окремі SQL запити для кожної з них. Запити застосовано для вибору потрібних полів, наприклад, «dish_name», «description», «price», «weight». Результати об'єднано за допомогою оператора «UNION ALL», який дає можливість зібрати всі відповідні записи з різних таблиць в один результат. (див. рис. 3.2.) Для користувача отримані дані оброблено й відображено у вигляді переліку страв, де кожна страва містить назву, опис, ціну.

```
const queries = tables.map(table => {
    const cleanName = table.replace(/^\d+\./, '');
    console.log(`Table: ${table}, Clean Name: ${cleanName}`);
    return `SELECT dish_name,
        description,
        price,
        weight,
        '${table}' AS table_name FROM \`${table}\``;
});

const query = queries.join(' UNION ALL ');
```

Рисунок 3.2. Приклад коду, де сформовано SQL запит

Також, створено вікно з деталями страви, де окрім назви, опису й ціни, відображено вагу або об'єм, назву закладу харчування, у якому можливо замовити обрану страву, також мапу з розташуванням ресторану. Для того, щоб відкрити це вікно, написано функцію і оброблено жест клацання на страву. У базі даних розташування закладу в Києві вказано покликанням, тому спочатку покликання перетворено на координати формату «41.40338, 2.17403», а потім нанесено на мапу, приклад коду зображено на рисунку 3.3. Географічне розташування ресторану відображено з використанням ключа до Google Maps API.

```
function initMap(mapUrl) {
    const queryURL = new URL(mapUrl);
    const address = decodeURIComponent(queryURL.searchParams.get('q'));

    const geocoder = new google.maps.Geocoder();
    geocoder.geocode({ address: address }, function(results, status) {
        if (status == 'OK') {
            const map = new google.maps.Map(document.getElementById('map'), {
                zoom: 16,
                center: results[0].geometry.location
            });

            new google.maps.Marker({
                map: map,
                position: results[0].geometry.location
            });
        } else {
            console.error('Geocode was not successful for
                           the following reason: ' + status);
        }
    });
}
```

Рисунок 3.3. Код, де перетворено покликання в географічні координати

3.3. Фільтрація страв

Певні харчові продукти, наприклад, молоко, яйця, горіхи, соя, пшениця, є алергенами, тобто можуть спровокувати алергічну реакцію в людей, що чутливі до них. У розробленні застосунку для меню закладів харчування враховано цю інформацію, тому що це надає змогу користувачам заздалегідь дізнатися, які страви містять потенційно небезпечні для них інгредієнти.

Для того, аби здійснювати пошук, створено список алергенів, які можна обрати за допомогою прапорця (див. рис. 3.4). Відповідно до Закону Про інформацію для споживачів щодо харчових продуктів 2019 року, ресторатори в Україні зобов'язані інформувати гостей про інгредієнти, які використано в стравах. Зокрема, про продукти, які входять до офіційного списку харчових алергенів і можуть бути фатальними в меню при алергії. (див. додаток А)

У списку для сортування страв зазначено перелік найбільш поширених харчових алергенів, які надано сайтом ChoiceQR.

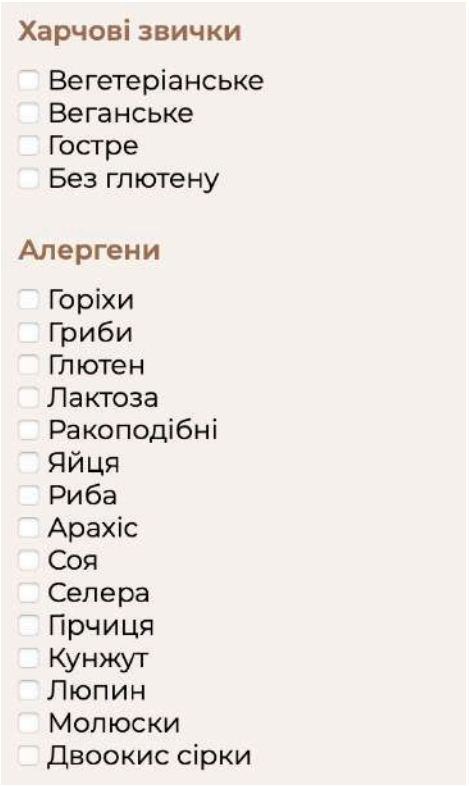

Рисунок 3.4. Перелік харчових звичок і алергенів у застосунку

Якщо прапорець активовано, то всі страви, що містять певний алерген приховано. Їх визначено за допомогою відповідності їхніх складників словникам, у яких вміщено повний перелік продуктів, що є алергенами, адже, наприклад, лактоза міститься в молоці, сирі, кефірі, айрані тощо. Приклад словника показано на рисунку 3.5.

Окрім алергій, люди можуть мати певні харчові звички, наприклад, веганство, вегетаріанство, безглютенова дієта тощо. Для зручності користувачів у застосунку також передбачено можливість фільтрування страв із такими харчовими побажаннями або обмеженнями. У застосунку для пошуку страв вказано чотири варіанти харчових звичок, які показано на рисунку 3.4. У базі даних створено окремі Boolean стовпці, які відповідають цим параметрам. На сайті ChoiceQR їх також було виділено окремо, тому це застосовано під час реалізації сортування страв за харчовими звичками. Якщо активний прапорець хоча б однієї з харчових звичок, то на відміну від страв з алергенами, які приховано за такої умови, ці страви навпаки відображено.

```
"Сир": ["Брі", "Камамбер", "Чеддер", "Гауда",
        "Горгонзола", "Рокфор", "Пармезан", "Моцарела",
        "Фета", "Бринза", "Грюйер", "Емменталь", "Рікотта",
        "Козячий сир", "Блакитний сир", "Пекоріно", "Качокавалло",
        "Халлумі", "Чечіль", "Шевр", "Горгонзола", "Горгондзола",
        "Філадельфія", "Плавлений сир", "Маскарпоне", "Рікотта",
        "Сулугуні", "Грюєр", "Фетакса", "Маасдам", "Едам",
        "Проволоне", "Буррата", "Манчего", "Стілсон", "Радамер",
        "Панір", "Скаморца", "Асьяго", "Раклет", "Шевр", "Ансьяго"],
```

Рисунок 3.5. Приклад словника для сортів сиру

Також є параметри, де надано змогу встановлювати діапазон (від числа до числа) для ваги й ціни. У деяких меню вагу для страв не вказано, тому в застосунку додано прапорець, який можна активувати, якщо потрібно відобразити такі страви.

РОЗДІЛ 4

АНАЛІЗ ЗІБРАНИХ ДАНИХ

4.1. Дослідження на основі вхідних даних

Зважаючи на сформовану збірку даних цифрових меню закладів харчування Києва з усіма назвами, складниками, ціною та вагою страв проведено кількісний аналіз. Це метод обробки даних, який використовується для визначення частоти виникнення певних елементів у наборі даних. Зібрані дані підготовлено до аналізу, видалено дублікати записів і виправлено помилки в написанні найменувань. У контексті цієї роботи спершу проведено аналіз для оцінювання популярності страв у меню ресторанів українською мовою.

Результати візуалізовано у вигляді графіку для наочності, як показано на рисунку 4.1. За отриманими результатами найпопулярнішою стравою є салат, який трапляється 899 разів. Проведений аналіз популярності харчових пропонувань може надати можливість ресторанам оптимізувати свої меню і врахувати страви, які мають найбільший попит серед споживачів.

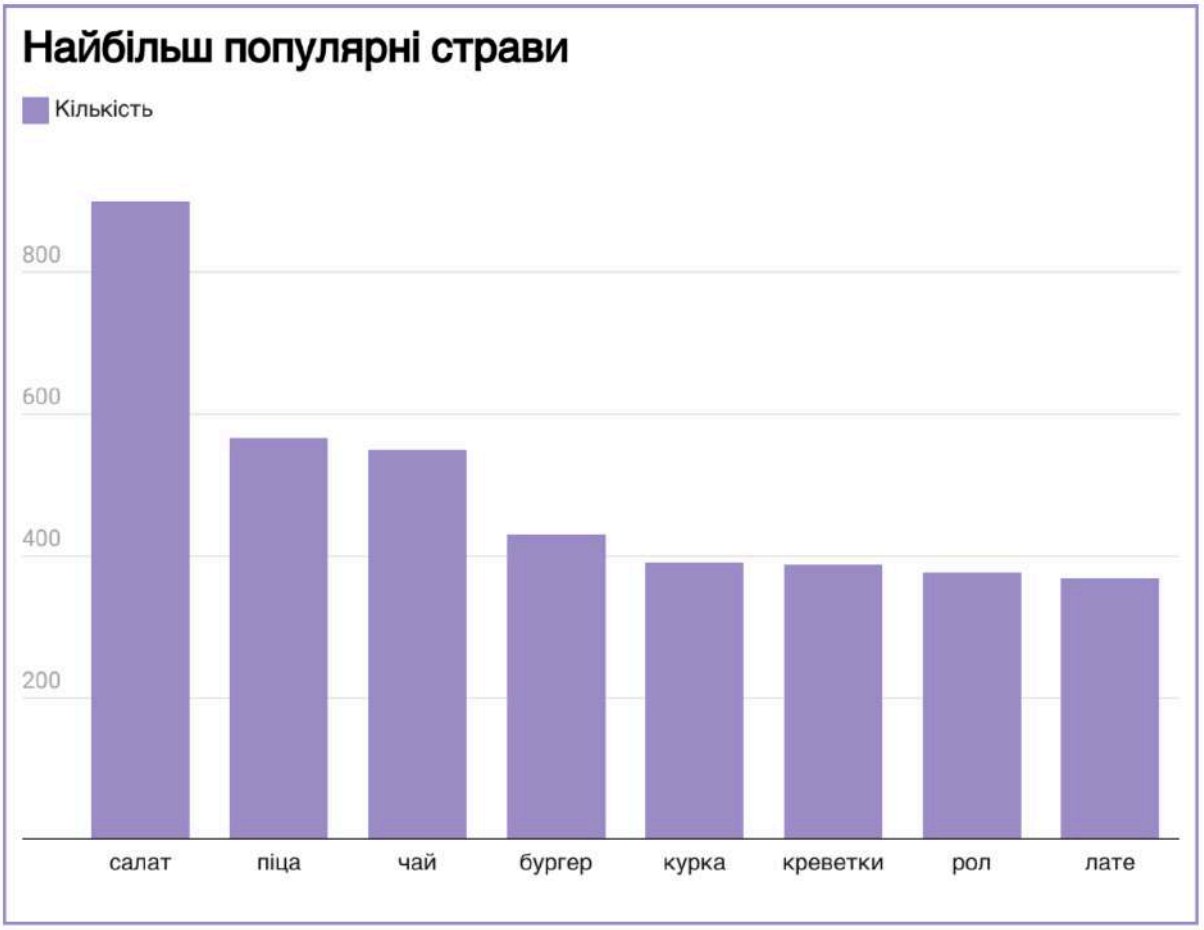

Рисунок 4.1. Перелік найпопулярніших страв

Також проведено кількісний аналіз складників страв (див.рис. 4.2). Найчастішим інгредієнтом виявлено соус, що є досить неоднозначно, оскільки є прості соуси, наприклад, майонез або кетчуп, проте знайдено й такі, що містять багато складників, наприклад, болоньєзе або бешамель. Через те, що цей інгредієнт може містити різноманітні компоненти, які є алергенами, а саме: горіхи, соєві продукти або глютен, то можна припустити, що в такому випадку при фільтруванні страв, яке описано в третьому розділі, неможливо приховати всі страви, які можуть спричиняти алергічну реакцію.

Другим за популярністю інгредієнтом виявлено сир, звідси можна емпірично припустити, що найчастіший алерген, який трапляється в стравах, це лактоза, що підтверджено далі в роботі.

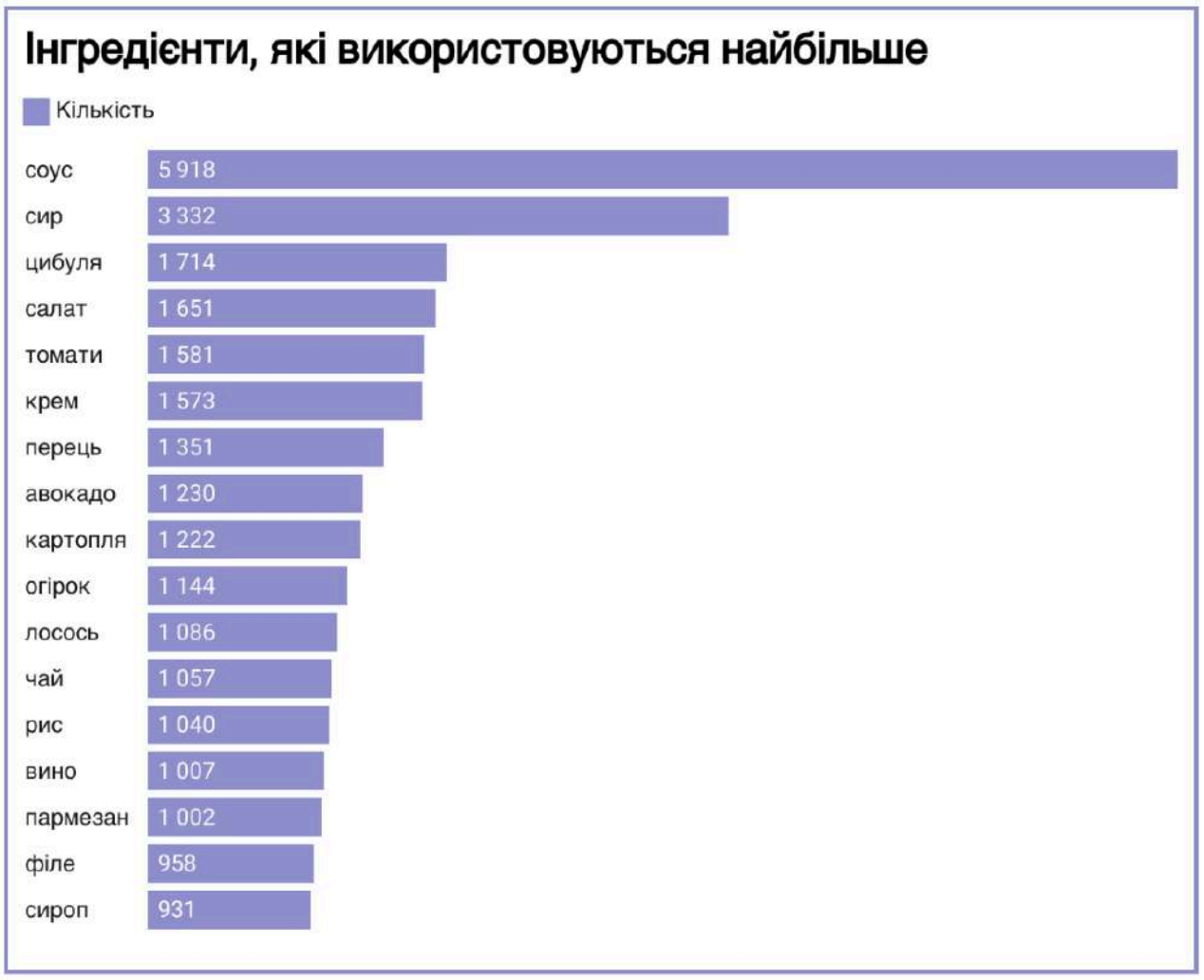

Рисунок 4.2. Перелік найчастіших інгредієнтів у стравах

Додатково проведено аналіз найчастіших алергенів, де безперечним лідером є лактоза, яка міститься в 7652 стравах, що складає 21 % від усіх страв. Зважаючи на отримані результати доведено, що фільтрація алергенів, яку описано в третьому розділі є вкрай важливою, щоб уникнути потенційних алергічних проявів у споживачів. Кількість решти алергенів у меню показано на рисунку 4.3.

Лактоза	7652
Риба	2 024
Гриби	1 448
Горіхи	1 325
Ракоподібні	1 108
Яйця	1 091
Кунжут	495
Гірчиця	405
Соя	382
Молюски	298
Селера	177
Арахіс	121
Люпин	32
Двоокис сірки	20

Рисунок 4.3 Перелік найчастіших алергенів

4.2. Цінова політика закладів харчування Києва

Оскільки в базі даних описаного пошуковика страв міститься окремий стовпець, де вказано ціни всіх страв, то їх також проаналізовано. За отриманими результатами середня ціна страви в Києві складає 291,25 гривні. Для того, щоби переконатися, що такий результат не виник через можливу появу одного ресторану з великою середньою ціною, виведено п’ять найдорожчих закладів і п’ять найдешевших (див.рис. 4.4).

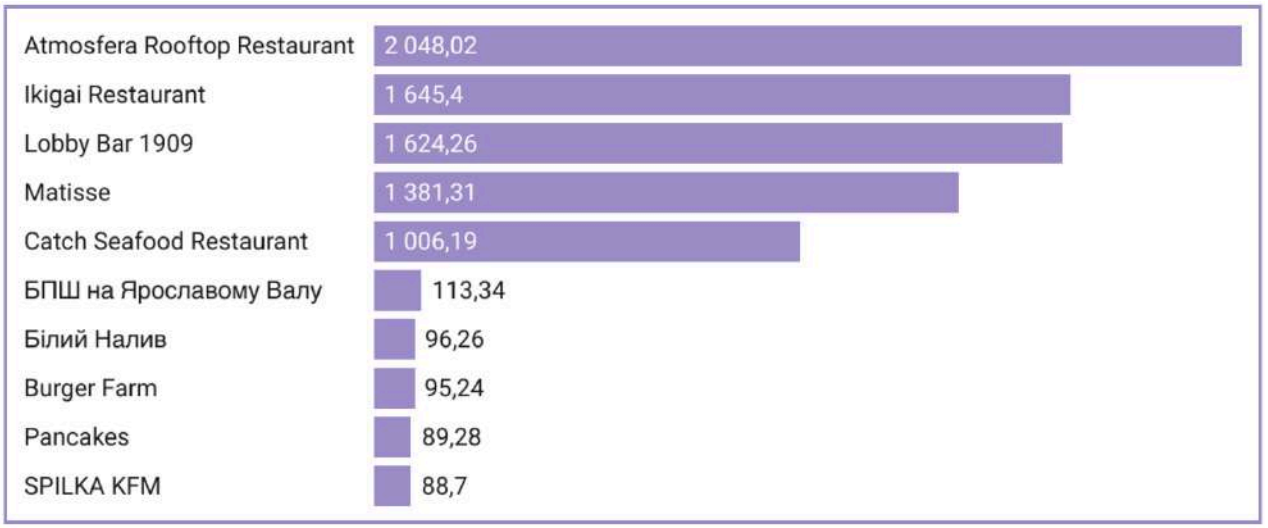

Рисунок 4.4. Перелік п’яти найдорожчих і п’яти найдешевших ресторанів

Наступним кроком проведено дослідження варіацій у ціноутворенні страв залежно від географічного розташування ресторанів у місті Києві. Гіпотеза полягає в тому, що заклади з розташуванням ближче до центру міста, мають вищі ціни порівняно з тими, які розташовуються на периферії. На рисунку 4.5. показано розташування на мапі закладів харчування, що використано при створенні пошуковика. Щоби продемонструвати зміну цін залежно від розташування використано HeatMap — це графічне відображення даних, де значення представленні кольорами. Теплі кольори (червоний, оранжевий) використовуються для показу високих значень, тоді як холодні кольори (синій, зелений) показують низькі значення.

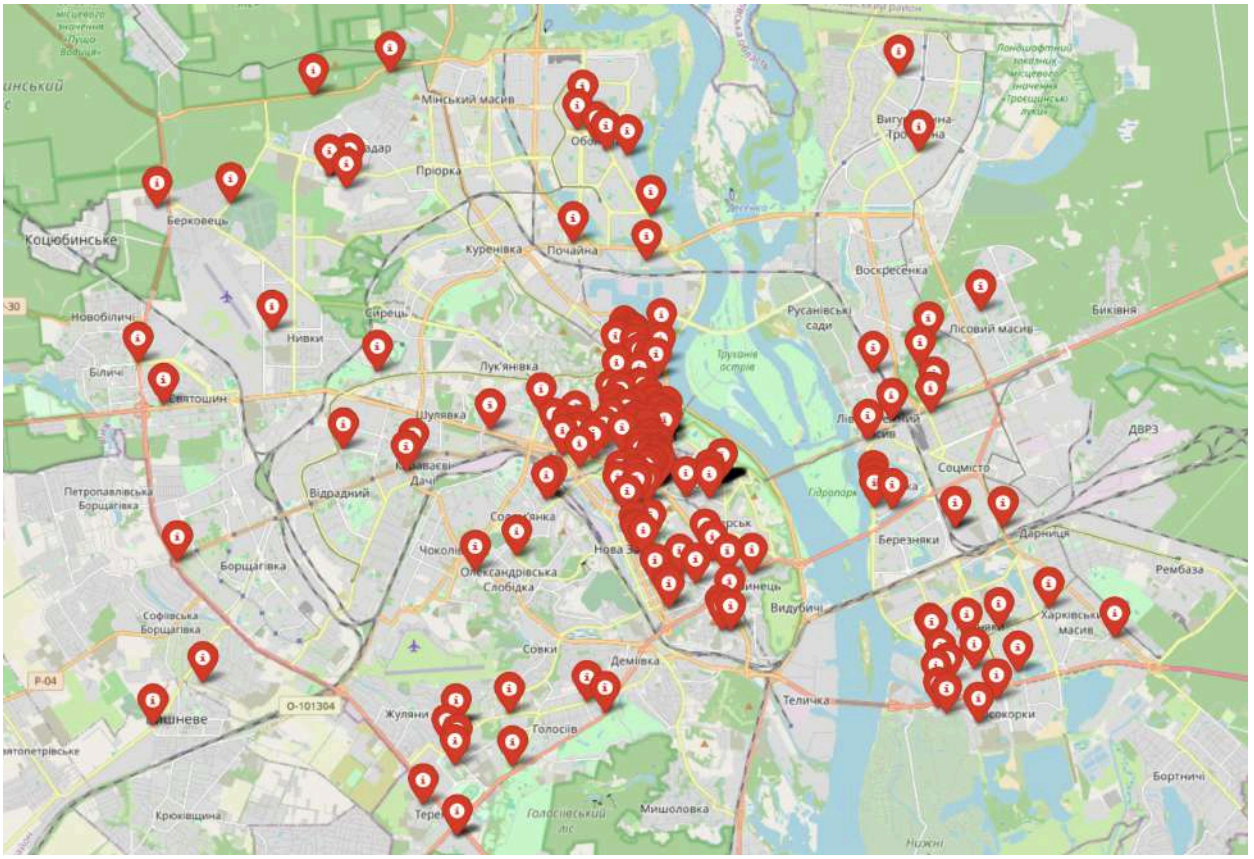

Рисунок 4.5. Розташування всіх проаналізованих закладів харчування в місті Київ

Для побудови HeatMap взято значення з бази даних про назви ресторанів і їхнє розташування у форматі покликання, а також використано попередньо обраховані середні ціни на страви. Покликання перетворено на географічні координати для нанесення на карту за допомогою написаного коду мовою програмування Python.

На рисунку 4.6. можна побачити, що гіпотезу підтверджено. Географічне розташування має значний вплив на ціноутворення в ресторанах. Найдорожчі заклади харчування розміщено в центрі міста, а заклади на периферії мають нижчу цінову політику.

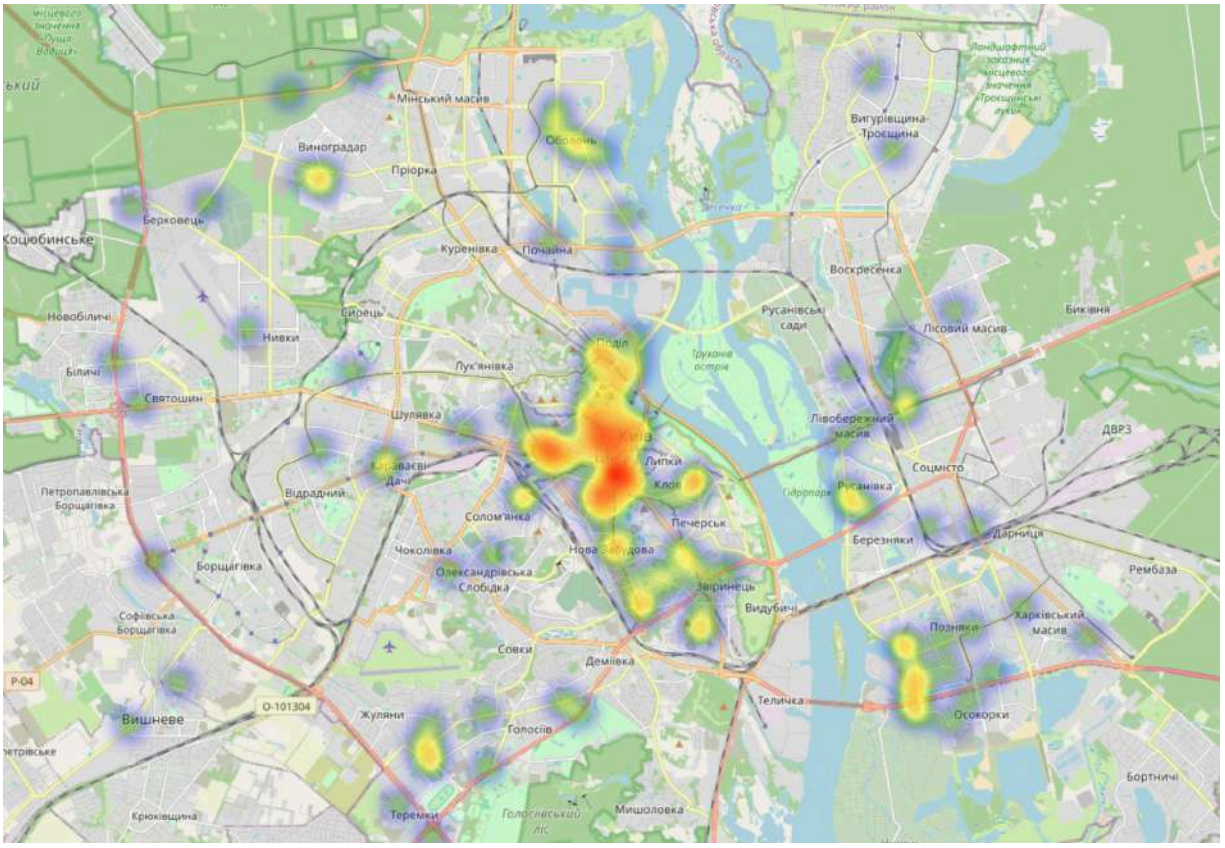

Рисунок 4.6. Тенденція зміни середніх цін у закладах харчування Києва залежно від розташування

4.3. Висновки до розділу

За допомогою проведеного аналізу даних цифрових меню закладів харчування Києва зібрано інформацію про популярність страв, частоту використання інгредієнтів та розподіл алергенів у стравах. За результатами дослідження варіацій у ціноутворенні страв виявлено, що середня ціна страви в Києві складає 291,25 гривні. Продемонстровано, що на ціни страв значно впливає географічне розташування з використанням аналізу найдорожчих та найдешевших закладів харчування. Цю інформацію можливо використати для підвищення якості обслуговування споживачів та конкурентоспроможності закладів харчування.

РОЗДІЛ 5

АПРОБАЦІЯ РЕЗУЛЬТАТІВ ДОСЛІДЖЕННЯ

5.1. Приклад використання пошуковика страв

Для демонстрації роботи застосунку вибрано запит «Дешева піца з куркою». Після того, як запит надіслано, отримано результат показаний на рисунку 5.1. Оскільки введено слово «дешева», то страви відсортовано за зростанням ціни. Біля кожної страви вказано її назву, опис, ціну та зображення.

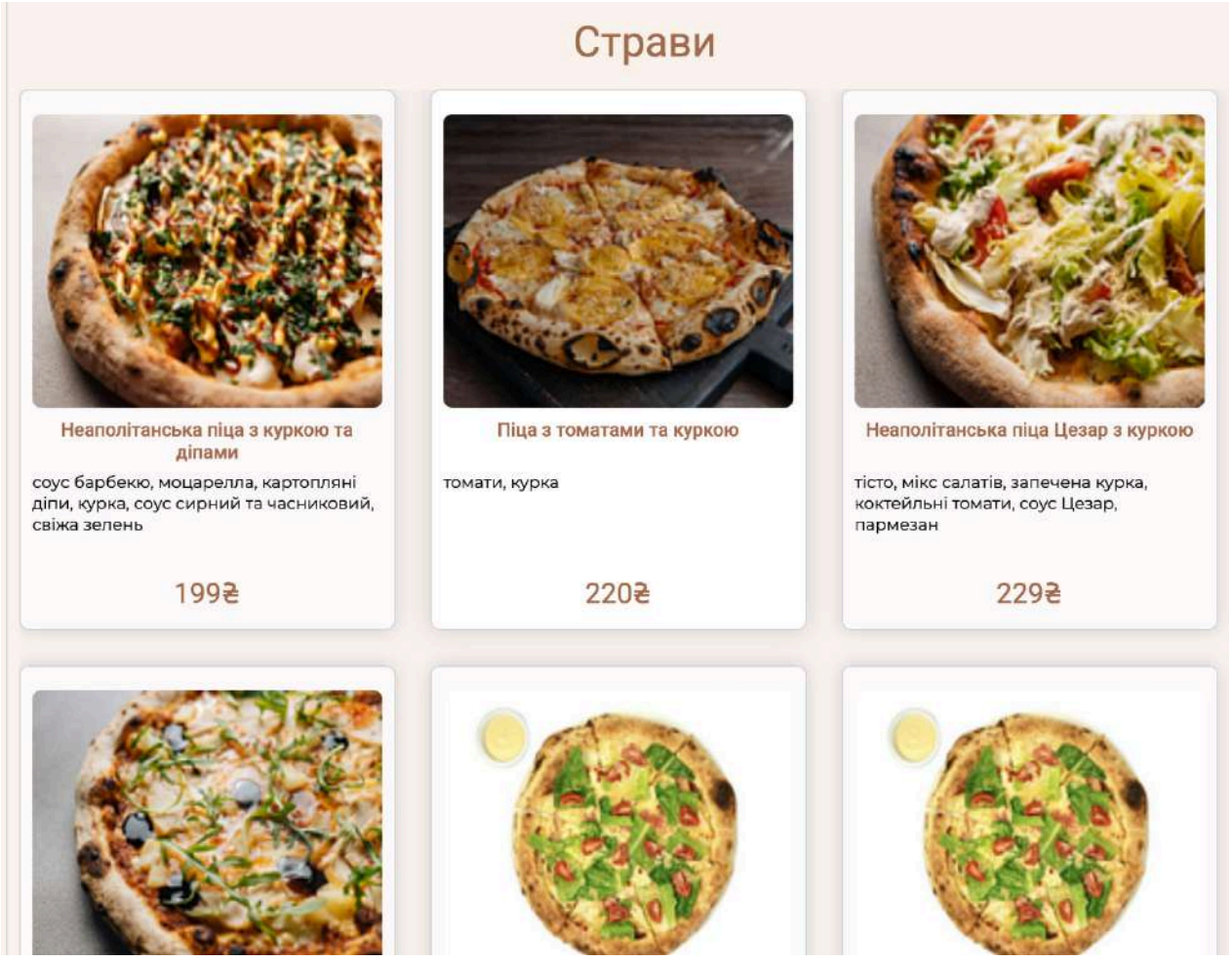

Рисунок 5.1 — Демонстрація результатів пошуку за запитом «Дешева піца з куркою»

При натисканні на обрану страву з’явиться вікно з детальним описом страви, її вагою, ціною та розташуванням на мапі (див.рис. 5.2).

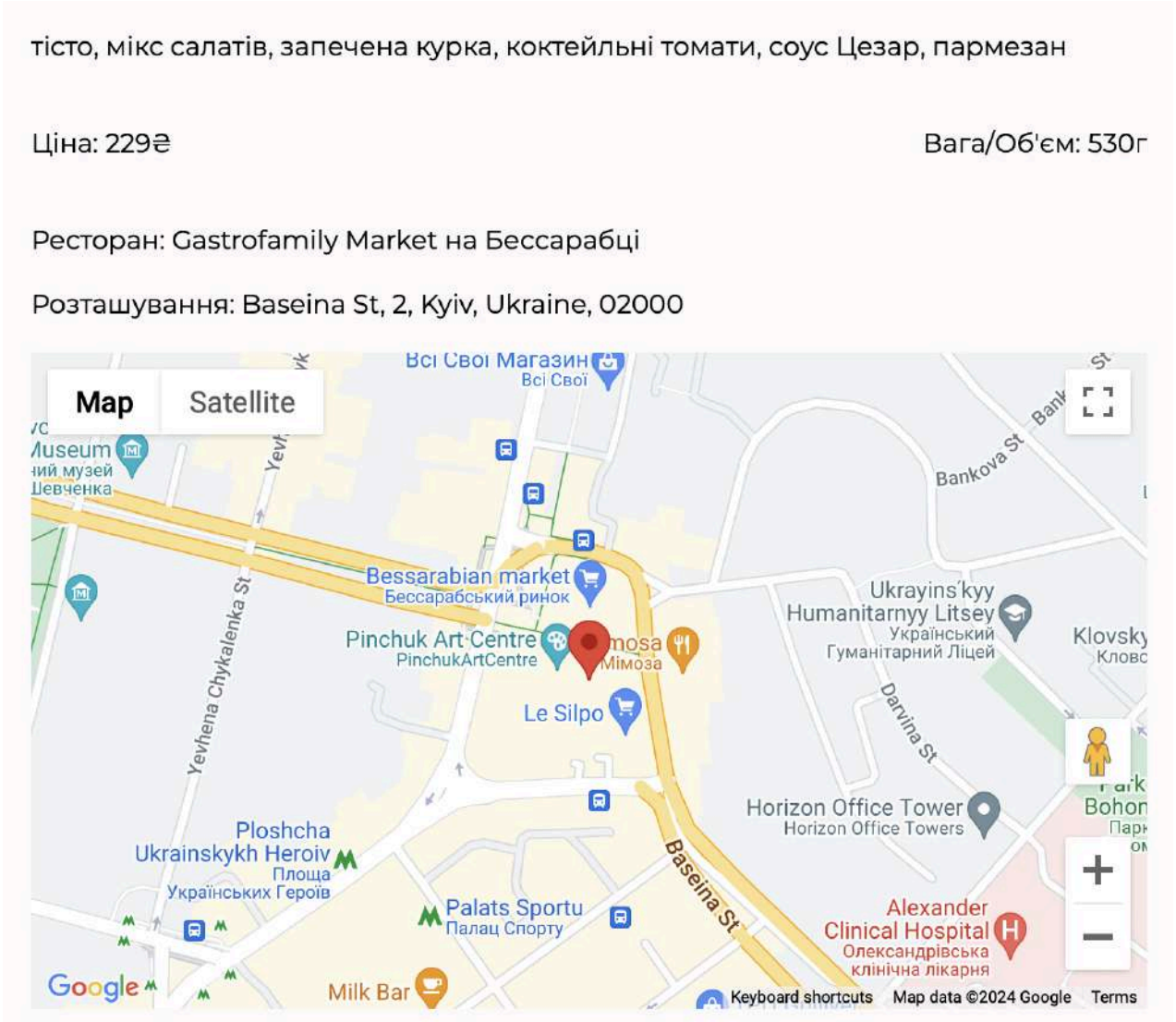

Рисунок 5.2. Деталі страви й розташування закладу на мапі

5.2. Порівняння отриманих результатів із наявними рішеннями

У першому розділі цієї роботи зазначено наявні рішення, які знайдено в інтернеті, проте всі вони в більшості спрямовані на пошук саме ресторанів, і хоча коли здійснено пошуку в цих застосунках, відображено меню різних

закладів харчування, вони не структуровані й зазвичай розміщені не в зручному для користувача форматі.

На відміну від цих рішень, пошуковик, що описано і розроблено в цій роботі спрямовано на роботу з меню та стравами закладів харчування. Основні переваги створеного застосунку полягають у тому, що реалізовано фільтрацію за алергенами, харчовими звичками та цінами для страв. Завдяки можливості здійснювати запити одночасно до всіх меню, користувачу надано повний огляд доступних страв і уможливлено пошук бажаної страви за оптимальною ціною та з відповідними характеристиками.

ВИСНОВКИ

Мету досягнуто, тому що створено пошуковик, у якому надано можливість користувачам здійснювати оптимізований і структурований пошук у неструктурованих українськомовних цифрових меню закладів харчування Києва й обирати страви, які розміщено ресторанами.

Завдання виконано, зокрема проаналізовано методи обробки природної української мови й досліджено різні аспекти цифрових меню закладів харчування, а саме: опрацьовано різноманітне подання структури меню, вміст алергенів у стравах, відповідність харчовим звичкам, а також варіативність опису складників страв.

Уперше реалізовано метод, що уможливлює пошук страв українськомовних цифрових меню закладів харчування Києва за допомогою фільтрації складників страв, які можуть спричиняти алергічну реакцію, аби оптимізувати процес вибору харчових пропонувань для користувача.

Уперше проведено аналіз страв українськомовних цифрових меню закладів харчування Києва, з використанням NLP-моделей, задля унаочнення та формування цілісної картини сучасного стану ресторанного бізнесу у воєнний період в Україні.

Удосконалено метод автоматичного видобування значущої інформації зі страв українськомовних цифрових меню закладів харчування задля підвищення точності визначення інгредієнтів і назв страв, що уможливлює коректне їхнє оброблення.

Розроблено методи для ефективного збору меню закладів харчування Києва із сайту ChoiceQR. Створено базу даних із 300 меню, які сумарно вміщують 37335 страв. Здійснено кількісний аналіз серед страв та їхніх складників. Проаналізовано вплив географічного розташування закладів на ціноутворення страв.

Продемонстровано методи розроблення пошуковика, створену систему використано для покращення пошуку страв у меню ресторанів з урахуванням дієтичних обмежень, спеціальних харчових звичок і потреб, застосовано методи сортування страв для покращення взаємодії з користувачем, що сприяє збільшенню попиту до закладів харчування Києва та уможливлює оптимальний підхід до здійснення вибору.

СПИСОК ЛІТЕРАТУРИ

1. Interfax-Ukraine. У Києві за 5 місяців 2023 р. відкрилися 172 кафе і ресторанів - експерт. *Інтерфакс-Україна*. URL: https://interfax.com.ua/news/economic/921556.html.

2. App Store. *Apple*. URL: https://www.apple.com/app-store/.

3. Додатки Android у Google Play. *Android Apps on Google Play*. URL: https://play.google.com/store/apps?hl=uk&gl=US.

4. ZUPA. *Clixby.app*. URL: https://clxb.ee/zupa.

5. *Головна — wewest*. URL: https://wewest.pro/

6. *expirenza by mono: єдиний QR-код для ресторанів*. Monobank — Мобільний Банк. URL: https://expz.monobank.ua/?lang=uk.

7. ChoiceQR — онлайн QR меню для закладів. *ChoiceQR — Smart solutions for modern restaurants*. URL: https://choiceqr.com/uk/?utm_source=google&utm_medium=cpc&utm_campaign=brand&keyword=choiceqr&gad_source=1&gclid=Cj0KCQjwir2xBhC_ARIsAMTXk87D0Y4maGQxk-sYgpTMiqrtE89D1xaXVuqhQnkfo64bhOogJmRhqpkaApouEALw_wcB.

8. QR-коди: все, що вам потрібно знати. *Acer Corner*. URL: https://blog.acer.com/ua/discussion/1043/qr-kodi-vse-scho-vam-potribno-znati.

9. Довідник по HTML тегам. *Український веб-довідник*. URL: https://css.in.ua/html/tags.

10. Information Extraction: Fundamentals and Applications. (2023). (n.p.): One Billion Knowledgeable.

11. UDPipe. *LINDAT/CLARIAH-CZ*. URL: https://lindat.mff.cuni.cz/services/udpipe/.

12. Харчові алергени в меню ресторану: як зробити ваш заклад безпечним? – Блог ChoiceQR. *ChoiceQR – Smart solutions for modern restaurants*. URL:

https://choiceqr.com/uk/news/harchovi-alergeny-v-menyu-restoranu-yak-zrobyty-vash-zaklad-bezpechnym/.

13. MAMP & MAMP PRO - your local web development solution for PHP and WordPress development. MAMP & MAMP PRO - Your local web development solution. URL: https://www.mamp.info/en/mac/.

14. PHP: Hypertext Preprocessor. PHP: Hypertext Preprocessor. URL: https://www.php.net.

15. Welcome to The Apache Software Foundation!. Welcome to The Apache Software Foundation!. URL: https://www.apache.org.

16. beautifulsoup4. PyPI. URL: https://pypi.org/project/beautifulsoup4/.

17. phpMyAdmin. phpMyAdmin. URL: https://www.phpmyadmin.net.

18. Node.js – Run JavaScript Everywhere. Node.js – Run JavaScript Everywhere. URL: https://nodejs.org/en.

19. Selenium. Selenium. URL: https://www.selenium.dev.

ДОДАТОК А. СПИСОК НАЙПОШИРЕНІШИХ ХАРЧОВИХ АЛЕРГЕНІВ

1. зернові, що містять глютен;
2. ракоподібні та продукти з них;
3. яйця та яєчні продукти;
4. риба та рибні продукти;
5. арахіс та продукти з арахісу;
6. соя та соєві продукти;
7. молоко та молочні продукти;
8. горіхи (включно з мигдалем, фундуком, волоськими горіхами, кеш'ю, горіхами пекан, бразильськими горіхами, горіхами макадамія та фісташками);
9. селера та продукти із селери;
10. насіння гірчиці та гірчичні продукти;
11. насіння кунжуту та продукти з кунжуту;
12. двоокис сірки та сульфіти в концентрації понад 10 мг/кг або 10 мг/л;
13. люпин та продукти з люпину;
14. молюски та вироби з молюсків.