

СТРУКТУРОВАНІЙ ОПТИМІЗОВАНИЙ ПОШУК У НЕСТРУКТУРОВАНІХ ДАНИХ ДЛЯ ЗАДАЧІ АНАЛІЗУ МЕНЮ

У статті представлено розроблення пошукової системи для цифрових меню закладів харчування Києва українською мовою. Проєкт реалізовано з використанням сучасних методів обробки природної української мови, зокрема для завдань лематизації, класифікації текстів і фільтрації даних. Основну увагу приділено розробленню алгоритмів, які забезпечують пошук інформації про складники страв, харчові обмеження, алергени та інші ключові характеристики.

У рамках дослідження здійснено аналіз цифрових меню київських закладів, що надало змогу сформувати цілісну картину сучасного стану ресторанного бізнесу в Україні. Результати роботи демонструють можливість структурованого та оптимізованого пошуку в неструктурованих даних меню українською мовою, що дає змогу користувачам легко обирати страви, які розміщено ресторанами.

Ключові слова: обробка природної мови, цифрове меню, заклад харчування, місто Київ, пошуковик, ресторанний бізнес, класифікація текстів, видобування інформації, фільтрація даних, структурований пошук, база даних, база знань, онтологія, Protégé, OWL, SWRL, Pellet, ризонер, Retrieval-Augmented Generation, інформаційні технології, інформаційна система, великі мовні моделі.

Вступ

У сучасному світі зростає обсяг неструктурованої 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. З огляду на це, автоматизація процесу пошуку страв стає ключовою умовою підвищення якості обслуговування клієнтів. Такий підхід дає змогу враховувати харчові вподобання, дієтичні обмеження та алергічні реакції, що робить аналіз меню більш персоналізованим і зручним для споживачів.

Сучасні технології обробки природної мови (NLP) надають нові можливості для роботи з текстами цифрових меню, зокрема щодо їхньої класифікації і фільтрації. У розробленій системі для аналізу та стандартизації текстів запитів користувачів використано великі мовні моделі (LLM), а також для автоматичного розширення запитів технологію RAG (Retrieval-Augmented Generation) [6], що дає змогу підвищити точність і релевантність пошуку.

Актуальність статті обґрунтовано необхідністю удосконалення українськомовних рішень у ресторанному бізнесі. Сучасні застосунки здебільшого орієнтовано на англійськомовних користувачів, тому впровадження методів для аналізу і структурування цифрових меню українською мовою сприяє оптимізації систем обслуговування для користувачів.

Зважаючи на сформовану збірку даних оцінено популярності харчових пропонуваль, 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.

Використання NLP-моделей для оброблення текстів

NLP-модель (модель обробки природної мови) — це комп'ютерна система, яку створено для аналізу людської мови. Такі моделі використовують у 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», подано недоречний текст або ж опису бракує. Також використовують молодіжний сленг і популярні фрази, наприклад «Трушна паста», «Топчик», «Хелсі страва» (див. рис. 1). Помічено, що наявність зазначених фраз ускладнює точне оброблення і класифікацію страв, тому що створює «шум» у даних. Такі випадки оброблено з використанням NLP-моделі, видалено недоречні слова і виділено складники в описах. Використано такі моделі: GPT від OpenAI та LLaMA від Meta AI.

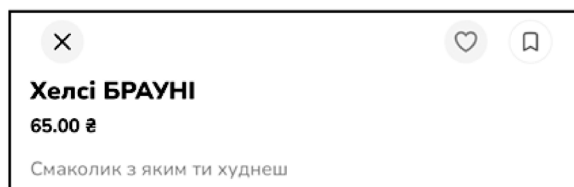


Рис. 1. Приклад недоречного опису

Формування бази даних

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 [2]. Таке рішення зумовлено тим, що на цьому сайті міститься понад шість тисяч меню закладів харчування і всі вони мають однакову структуру.

У роботі застосунку для реалізації функціоналу зберігання, оброблення та швидкого доступу до даних використано систему керування базами даних MySQL [7]. У неї попередньо завантажено опис страв, їхні ціну, вагу та особливі характеристики, а саме: гостра; вегетаріанська — без м'яса, риби чи інших морських істот; веганська — без продуктів тваринного походження; без глютену.

Для створення застосунку, у якому можна 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, тому використано базу даних для зберігання потрібної інформації.

restaurant_name
id : integer * dish_name : varchar (255) * description : text o price: decimal (10,2) o weight : varchar (10) * location: varchar (255) o image: varchar (255) o spicy: boolean o vegetarian: boolean o vegan: boolean o gluten_free: boolean

Рис. 2. Схема бази даних цифрових меню закладів харчування

Окрім реляційної бази даних MySQL, для зберігання та оброблення семантичної інформації про страви використано онтологічну базу знань, яку розроблено в Protégé [4], для виявлення прихованих залежностей, а саме: логічний висновок за допомогою Pellet [3] дає змогу ідентифікувати страви, які відповідають складним критеріям, навіть якщо ці залежності не явно зазначені в базі даних.

Оброблення запиту від користувача

У тексті запиту можливі слова в різних граматичних формах, тому їх оброблено і приведено до основної форми (леми) за допомогою UDPipe. Цей інструмент містить функції для аналізу текстів, а саме: токенізацію, лематизацію, розмітку частинами мови (POS tagging) тощо [8]. Отримані лемми

проаналізовано, і слова, які позначено іменниками (NOUN), використано для здійснення пошуку. Текст запиту очищено від зайвих слів і символів, виділено ключові слова та приведено запити до єдиного формату за допомогою великих мовних моделей (LLM). Для забезпечення коректності пошуку та фільтрації страв у застосунку також сформовано спеціалізовані словники, які містять синонімічні назви інгредієнтів, зокрема сирів, м'яса, овочів тощо (див. рис. 3).

Для покращення формування необроблений текст запиту передано до LLM, наприклад, «Хелсі смачненький круасан із сирною начинкою недорогого». Також до моделі надіслано заздалегідь підготовлені словники, як-от словник для сирів, що містить варіативні назви: «маскарпоне», «брі», «cheder», і додано семантичні зв'язки між інгредієнтами із застосуванням бази знань, яку створено у Protégé, наприклад автоматично визначено, що складник із класу «Молочні продукти» містить алерген «лактоза». Крім того, у запропонованій онтологічній базі знань встановлено логічні залежності між об'єктами за допомогою правил SWRL (Semantic Web Rule Language) [9]. Для реалізації логічного висновку використано рушій Pellet, який інтегровано в Protégé.

Після цього запит приведено до єдиного формату, видалено нерелевантні слова та збережено ключові терміни для пошуку за допомогою мовної моделі. Для розширення запиту користувача використано технологію RAG (Retrieval-Augmented Generation). Наприклад, запит «паста» автоматично розширено на терміни «макарони», «спагеті», «італійська кухня», що дає змогу знайти всі відповідні страви.

У застосунку реалізовано комбінований підхід, який поєднує аналіз ключових слів, фільтрацію та семантичний пошук, так створено високий рівень точності пошукових запитів і надано можливість користувачам легко знаходити страви, що задовольняють їхні особливі вимоги й фінансові можливості. Потрібно зазначити, що в розробленій системі використано не лише методи машинного навчання, а й rule-based підхід, зокрема регулярні вирази застосовано для реалізації фільтрації.

Щоби здійснювати пошук, створено список алергенів [1], які можна вибрати за допомогою порця, наприклад горіхи, гриби, глютен, лактоза, яйця, риба, соя, селера. Окрім алергій, люди можуть мати певні харчові звички, наприклад веганство, вегетаріанство, безглютенова дієта.

```
"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", "Грюер", "Фетакса", "Маасдам", "Едам",  
        "Проволоне", "Буррата", "Манчего", "Стілсон", "Радамер",  
        "Панір", "Скаморца", "Асьяго", "Раклет", "Шевр", "Ансьяго"]
```

Рис. 3. Приклад словника для сортів сиру

Дослідження на основі вхідних даних

Зважаючи на сформовану збірку даних цифрових меню закладів харчування Києва з усіма назвами, складниками, ціною та вагою страв, проведено кількісні аналізи. Спершу проведено аналіз для оцінювання популярності страв у меню ресторанів українською мовою. Результати візуалізовано у вигляді графіка для наочності (рис. 4). За отриманими результатами найпопулярнішою стравою є салат, який трапляється 899 разів. Це дослідження також є прикладом використання підходів комп'ютерної соціальної науки (computational social science), що дає змогу вивчати гастрономічні переваги споживачів та взаємозв'язок між географічним розташуванням закладів харчування і ціноутворенням. Проаналізовано великі обсяги даних і отримано соціально значущу інформацію для вдосконалення ресторанного бізнесу з використанням такого підходу.

Також проведено кількісний аналіз складників страв (див. рис. 5). 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 болоньезе або бешамель. Другим за популярністю інгредієнтом є сир, звідси можна емпірично припустити, що найчастіший алерген, який трапляється у стравах, це лактоза, що підтверджено далі в роботі.

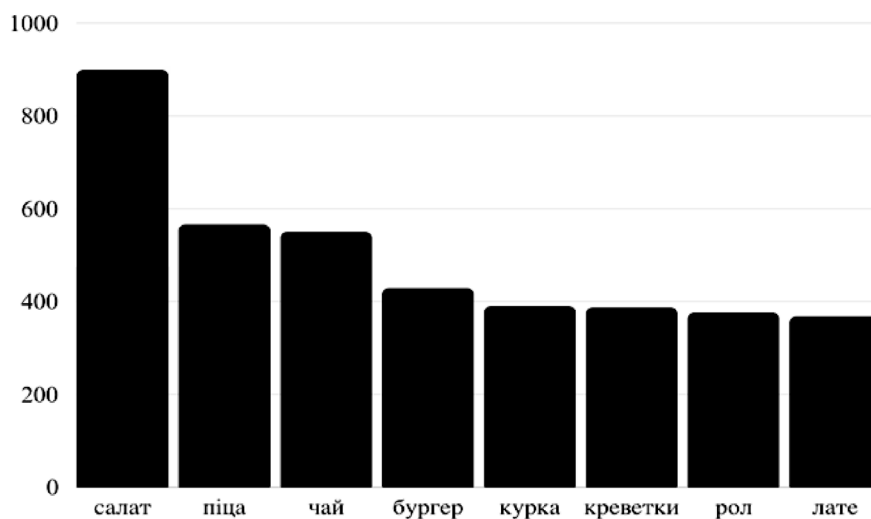


Рис. 4. Перелік найпопулярніших страв

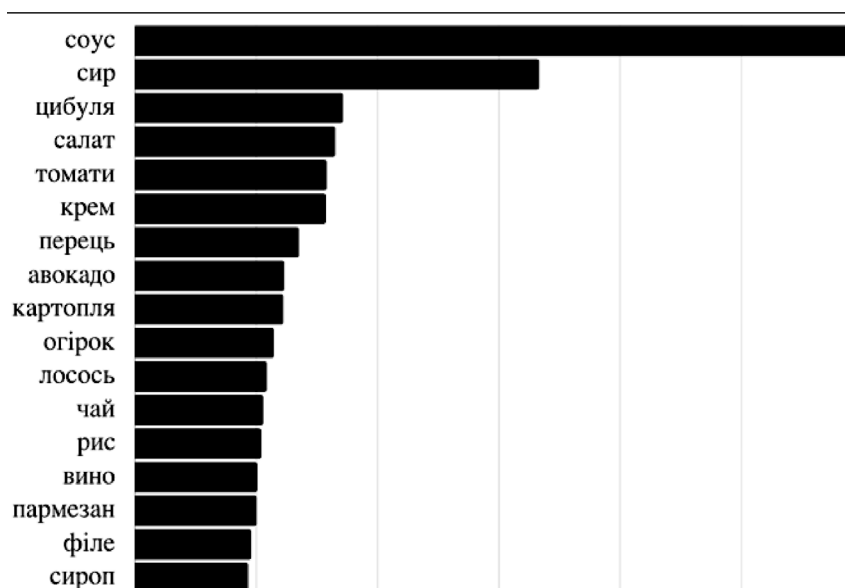


Рис. 5. Перелік найчастіших інгредієнтів у стравах

Додатково проведено аналіз найчастіших алергенів, де безперечним лідером є лактоза, яка міститься у 7652 стравах, що становить 21 % усіх страв. Кількість решти алергенів у меню показано на рис. 6.

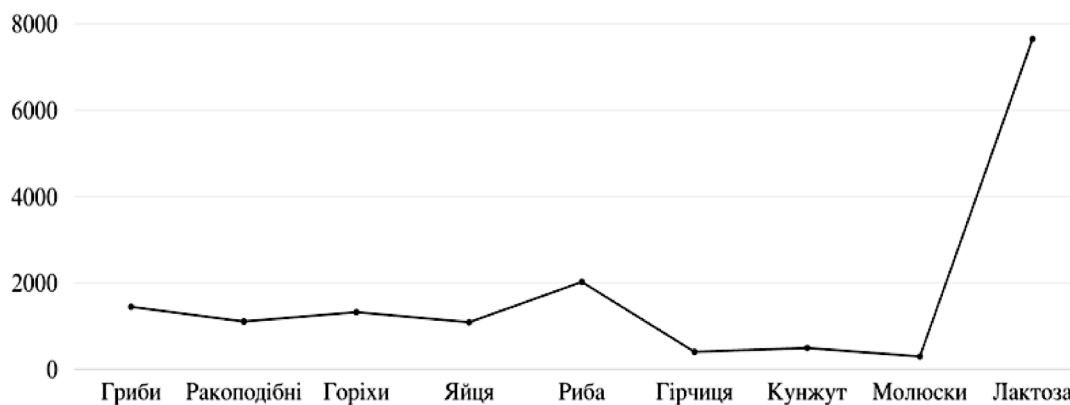


Рис. 6. Перелік найчастіших алергенів

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 становить 291,25 грн. 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. 7).

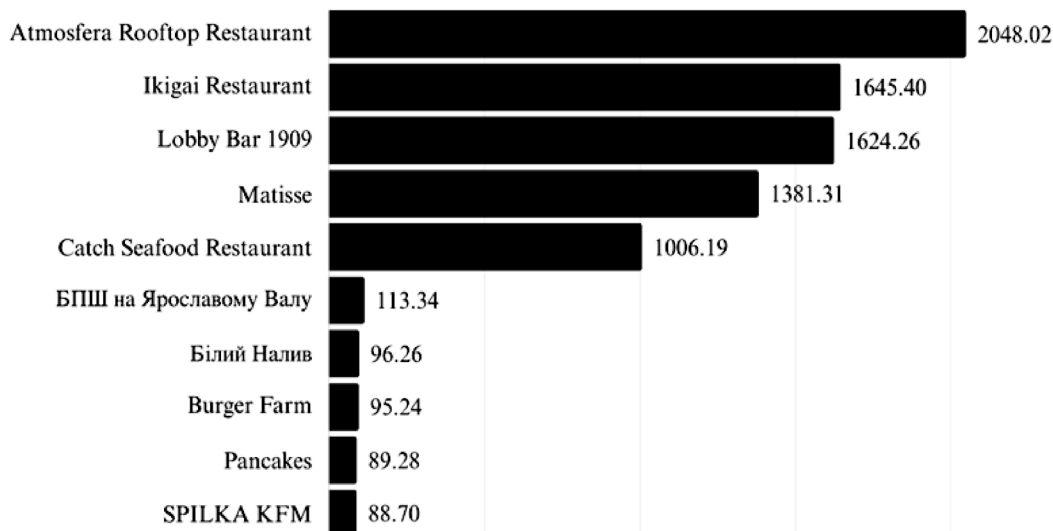


Рис. 7. Перелік п'яти найдорожчих і п'яти найдешевших ресторанів

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 рис. 8 показано розташування на мапі закладів харчування, що використано при створенні пошуковика. Щоби продемонструвати зміну цін залежно від розташування, використано теплокарту: теплі кольори (червоний, оранжевий) показують високі значення, холодні кольори (синій, зелений) — низькі значення.

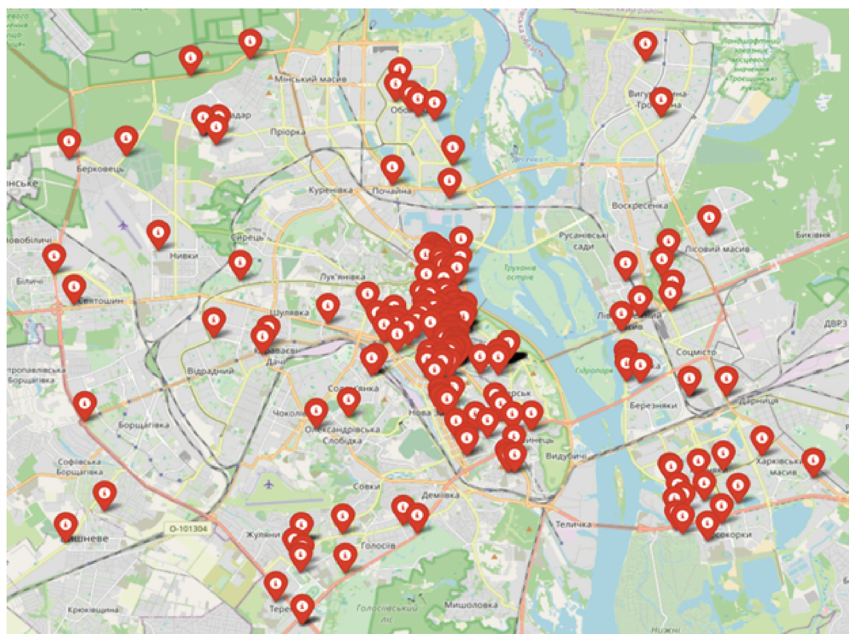


Рис. 8. Розташування всіх проаналізованих закладів харчування в місті Київ

На рисунку 9 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.

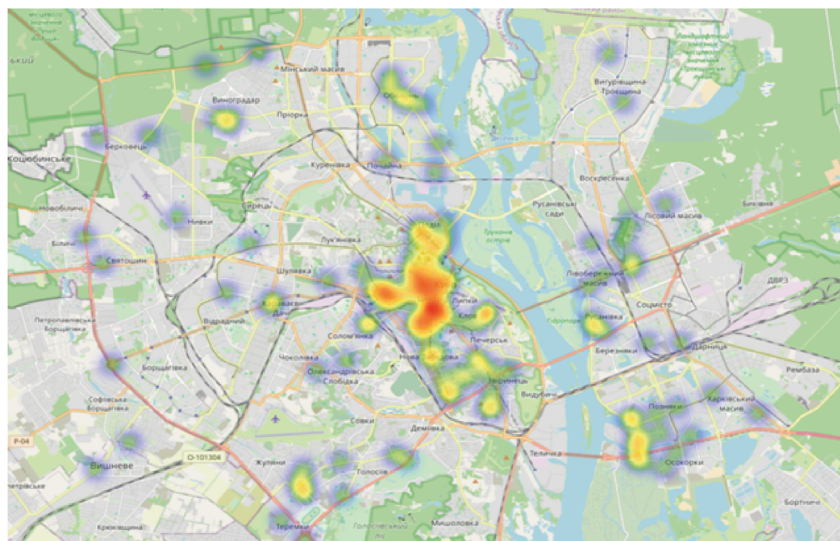


Рис. 9. Тенденція зміни середніх цін у закладах харчування Києва залежно від розташування

Висновки

У статті описано та продемонстровано роботу пошукової системи для цифрових меню закладів харчування Києва. Розроблену систему інтегровано із сучасними методами обробки природної мови (NLP), зокрема з використанням морфосинтаксичного аналізу, семантичного аналізу із застосуванням онтологічної бази знань, розширення запитів за допомогою технології RAG (Retrieval-Augmented Generation), логічних висновків із правилами SWRL та рушієм Pellet. У системі використано реляційну базу даних MySQL для зберігання інформації про страви та онтологічну базу знань, створену в Protégé, для збагачення семантичної інформації та підвищення точності пошуку. Це надає змогу здійснювати структурований пошук страв з урахуванням складників, дієтичних обмежень, алергенів, цінкових параметрів та географічного розташування закладів.

Система поєднує результати роботи великих мовних моделей (LLM), з використанням яких очищено і стандартизовано текст запиту, а також його розширено за допомогою заздалегідь сформованих словників та правил, для забезпечення повноти пошукових результатів. Для точного виділення інгредієнтів та фільтрації страв за заданими критеріями застосовано rule-based підхід, зокрема регулярні вирази, і підвищено точність і релевантність пошуку. Інтегровано онтологічну базу знань із правилами SWRL і рушієм Pellet, автоматично визначено логічні залежності між інгредієнтами та стравами, а також виявлено нові знання.

На прикладах цифрових меню київських закладів харчування проілюстровано можливості системи щодо автоматичного оброблення даних, фільтрації страв за алергенами та дієтичними обмеженнями, а також кількісного аналізу популярності страв і їхніх складників, оцінено вплив географічного розташування закладів на цінову політику, що є прикладом використання підходів комп'ютерної соціальної науки (computational social science) і може слугувати основою для оптимізації меню ресторанів.

Отже, запропоновані рішення є значущим кроком для вдосконалення клієнтоорієнтованих послуг у галузі ресторанного бізнесу.

Список літератури

1. Харчові алергени в меню ресторану: як зробити ваш заклад безпечним? — Блог ChoiceQR [Електронний ресурс] // ChoiceQR — Smart solutions for modern restaurants. — Режим доступу: <https://choiceqr.com/uk/news/harchovi-alergeny-v-menyu-restoranu-yak-zrobyty-vash-zaklad-bezpechnym/>.
2. ChoiceQR — онлайн QR меню для закладів [Електронний ресурс] // ChoiceQR — Smart solutions for modern restaurants. — Режим доступу: <https://choiceqr.com/uk/>.
3. Pellet — Semantic Web Standards [Electronic resource] // W3C. — Mode of access: <https://www.w3.org/2001/sw/wiki/Pellet>.
4. protégé [Electronic resource] // protégé. — <https://protege.stanford.edu/>.
5. Sabry F. Information Extraction [Electronic resource] / Fouad Sabry. — [S. l.]: One Billion Knowledgeable. — 78 p.
6. Siddharth L. Retrieval Augmented Generation using Engineering Design Knowledge [Electronic resource] / L. Siddharth, Jianxi Luo // Knowledge-Based Systems. — 2024. — P. 112410. — <https://doi.org/10.1016/j.knosys.2024.112410>.
7. SQL: Learn SQL (using MySQL) in One Day and Learn It Well. SQL for Beginners with Hands-on Project. (Learn Coding Fast with Hands-On Project): LCF Publishing, Chan, Jamie: 9781731039668.

8. Straka Milan. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task / Milan Straka // *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning*. — Association for Computational Linguistics, Stroudsburg, PA, USA, 2018. — Pp. 197–207.
9. SWRL: A Semantic Web Rule Language Combining OWL and RuleML [Electronic resource] // W3C. — Mode of access: <https://www.w3.org/submissions/SWRL/>.

References

- Chan, J. (2018). *SQL: Learn SQL (using MySQL) in One Day and Learn It Well. SQL for Beginners with Hands-on Project. (Learn Coding Fast with Hands-On Project)*. LCF Publishing.
- ChoiceQR — online QR menu dlia zakladiv. (n.d.). ChoiceQR — Smart solutions for modern restaurants. <https://choiceqr.com/uk/>.
- Kharchovi alerheny v menu restoranu: yak zrobyty vash zaklad bezpechnym? — Bloh ChoiceQR. (2023, 23 lystopada). ChoiceQR — Smart solutions for modern restaurants. <https://choiceqr.com/uk/news/harchovi-alergeny-v-menu-restoranu-yak-zrobyty-vash-zaklad-bezpechnym>.
- Pellet — Semantic Web Standards. (2011, 9 sichnia). W3C. <https://www.w3.org/2001/sw/wiki/Pellet>.
- protégé. (n.d.). protégé. <https://protege.stanford.edu/>.
- Sabry, F. (2023). *Information Extraction*. One Billion Knowledgeable.
- Siddharth, L., & Jianxi, L. (2024). *Retrieval augmented generation using engineering design knowledge*. Knowledge-Based Systems. <https://doi.org/10.1016/j.knosys.2024.112410>.
- Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning*, 197–207.
- SWRL: A Semantic Web Rule Language Combining OWL and RuleML. (2004, 21 травня). W3C. <https://www.w3.org/submissions/SWRL/>.

O. Smysh, A. Chyzhova

STRUCTURED OPTIMIZED SEARCH IN UNSTRUCTURED DATA FOR MENU ANALYSIS TASKS

The article describes the development of a dish search engine for digital restaurant menus in Kyiv, focusing on Ukrainian-speaking users. The system integrates modern Natural Language Processing (NLP) methods such as lemmatization, text classification, and data filtering, alongside Retrieval-Augmented Generation (RAG), specialized ingredient dictionaries, a database, and an ontological knowledge base designed in Protégé. Using rules from the Semantic Web Rule Language (SWRL) and logical inference through the Pellet reasoning engine, the system performs semantic analysis of user queries, automatically identifying relationships between dish components, and improving search relevance.

The search algorithm utilizes a multi-layered approach that combines machine learning, logical reasoning, and rule-based filtering. User queries, often containing informal or varied phrasing, are first processed by a large language model (LLM) to identify and standardize key terms. The LLM is enhanced with pre-defined dictionaries (e.g., for cheese types like “mascarpone”, “brie,” or “cheddar”) and connected to an ontological knowledge base, which enriches the query with semantic relationships. RAG extends this functionality by automatically expanding search terms to include synonyms or related concepts, such as linking “pasta” to “macaroni” or “spaghetti.”

The study incorporates principles of computational social science to analyze semi-structured data from digital restaurant menus, such as the popularity of dishes and their ingredients, as well as the impact of restaurant location on pricing. The data highlights trends in customer preferences and provides actionable insights for optimizing restaurant menus.

The developed system successfully integrates NLP techniques, logical reasoning, and structured data storage, achieving high accuracy and relevance in search results. By incorporating an LLM, RAG, and ontological reasoning, the system demonstrates the potential for significantly enhancing customer-oriented services in the restaurant industry through advanced data analysis and semantic search tools.

Keywords: natural language processing, digital menu, restaurant, Kyiv city, search engine, restaurant business, text classification, information extraction, data filtration, structured search, database, knowledge base, ontology, Protégé, OWL, SWRL, Pellet, reasoner, Retrieval-Augmented Generation, information technologies, information system, large language models.

Матеріал надійшов 02.12.2024

