Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Метод автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Освітня програма *Комп'ютерні науки*
Назва освітньої програми

Виконав: *студент 4 курсу, група КН-18-1* _____ *О.В. Козенко*
Курс, група виконавця | Підпис | Ініціали, прізвище

Керівник: *к.т.н., доцент кафедри КН* _____ *О.В. Мазурець*
Науковий ступінь, посада | Підпис | Ініціали, прізвище

Нормоконтроль: *к.т.н., доцент кафедри КН* _____ *Р.О. Багрій*
Науковий ступінь, посада | Підпис | Ініціали, прізвище

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор _____ *О.В. Бармак*
Підпис | Ініціали, прізвище

13 червня 2022 р.

Хмельницький 2022

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *бакалавр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

(підпис)
д.т.н., професор О.В. Бармак
«*25*» *березня* *2022* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

1. Тема кваліфікаційної роботи бакалавра: *«Метод автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом»*

2. Завдання видано студенту *Козенко Олександру Васильовичу*
(прізвище, ім'я, по батькові)

3. Керівник роботи *доцент кафедри КН Мазурець Олександр Вікторович*
(посада, прізвище, ім'я, по батькові)

4. Затверджено наказом університету від «*01*» *березня* *2022* р. № *18*

5. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Мета роботи – розробка й апробація методу автоматизованого підбору відповідей на запитання за семантичною подібністю. Основними функціями виступають семантичний аналіз наявних запитань і користувацького запитання, обрахунок оцінок семантичної подібності користувацького запитань, а також знаходження відповіді на користувацьке запитання. Також необхідно створити програмне забезпечення, яке зможе автоматизовано надавати користувачам (на прикладі співвласників багатоквартирних будинків) відповіді шляхом семантичного аналізу користувацьких запитань. Це дозволить користувачам скоротити час очікування відповіді.

Виконавець: *студент 4 курсу, група КН-18-1* _____ *О.В. Козенко*
Курс, група виконавця — Підпис — Ініціали, прізвище

Керівник: *к.т.н., доцент кафедри КН* _____ *О.В. Мазурець*
Науковий ступінь, посада — Підпис — Ініціали, прізвище

Анотація

Тема кваліфікаційної роботи бакалавра: *«Метод автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом»*

Виконавець кваліфікаційної роботи бакалавра: *студент групи КН-18-1 Козенко Олександр Васильович*

Керівник кваліфікаційної роботи бакалавра: *к.т.н., доцент кафедри КН Мазурець Олександр Вікторович*

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерел інформації	
67	30	10	35	4

Метою кваліфікаційної роботи бакалавра є розробка методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом. Для розробки програмного продукту було використано мову програмування C#, а також систему керування базами даних MS SQL Server.

Розроблена система призначена для спеціалізованих соціальних груп, що містять багато варіацій типових запитань, відповіді на які можна автоматизувати з метою запобігання витрачання надмірного людського ресурсу.

Напрямками практичного використання розробленої програми на основі методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом є автоматизація відповідей у типових групах типу водоканалу, обленерго, ОСББ.

Ключові слова: семантичний аналіз, дисперсійна оцінка, ОСББ, ключові слова.

Виконавець: *студент 4 курсу, група КН-18-1* (Курс, група виконавця) _____ (Підпис) *О.В. Козенко* (Ініціали, прізвище)

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
БД	База даних
ІС	Інформаційна система
ІТ	Інформаційні технології
КРБ	Кваліфікаційна робота бакалавра
КН	Комп’ютерні науки
ПЗ	Пояснювальна записка
ПП	Програмний продукт
СКБД	Система керування базами даних
ХНУ	Хмельницький національний університет.
TF	Term Frequency
TF-IDF	Term Frequency – Inverse Document Frequency
ОСББ	Об’єднання співвласників багатоквартирних будинків
КС	Ключові слова
ЦНАП	Центр надання адміністративних послуг
NLP	Natural Language Processing
ОС	Операційна система

Вступ

Спілкування між людьми є невід'ємною частиною буття людини. З розвитком технологій спілкування перейшло з листування у конвертах на рівень електронних листів, а далі і до месенджерів та форумів.

З розвитком ІТ та переходам до цифрового спілкування, обертів набирала і сфера машинної обробки природньої мови. Обробка природньої мови є одним із ключових напрямів ШІ, який працює з аналізом, розумінням та генерацією живих мов для взаємодії з комп'ютером і шляхом усного спілкування, і шляхом письмового замість звичного для комп'ютера способу машинних кодів [1]. Відповідно, текстові дані пошуку в мережі аналізуються з метою надання таргетованого рекламного контенту. Також аналізуються тексти у листах, та навіть просто набрані у певних текстових редакторах із задачею пошуку та виправлення орфографічних помилок, тощо. Обробка природньої мови також і присутня у новітніх гаджетах, таких як розумний будинок, сірі, гул-асистент. І навіть підбірка користувацьких новин також виконується завдяки аналізу пошукових запитів користувача.

Мова є своєрідним фундаментом, адже нею можна передавати не лише факти, а і емоційні стани, за допомогою мови можна отримати нові знання, та синтезувати нові. Технології обробки природньої мови мають дуже широкий потенціал, за яким стоїть майбутнє.

Коріння природної обробки мови сягає 1950-х років, коли відомий англійський учений Алан Тьюрінг опублікував статтю «Обчислювальні машини та розум», запропонувавши так званий «Тест Тьюрінга». Одним з його критеріїв є здатність машини автоматично інтерпретувати та генерувати людську мову [2].

7 січня 1954 року вчені з Джорджтаунського університету продемонстрували можливості машинного перекладу. Інженерами було перекладено понад 60 пропозицій з російської на англійську у повністю автоматичному режимі. Ця подія позитивно вплинула на розвиток машинного перекладу та увійшла в історію під назвою Джорджтаунський експеримент.

1966 року американський інформатик німецького походження Джозеф Вейценбаум у стінах Массачусетського технологічного інституту розробив перший у світі чат-бот «Елізу» [3]. Програма працювала за принципом відтворення діалогу із психотерапевтом, використовуючи техніку активного слухання.

Вже з 2013 року людство познайомилось із більш новітніми чат-ботами, і з плином часу та технологічним прогресом їх якість та функціонал значно розширились. Проте, незважаючи на досить широкі можливості автоматизації, обробка природньої мови все ще є дуже актуальним напрямом, та має практично усі сфери життєдіяльності людини, де можна покращити вже існуючі результати, або зробити щось кардинально нове. Відповідно, однією із сфер яка потребує автоматизації є ОСББ. Зазвичай, спілкування у чатах багатоквартирних будинків нагадує балаган, де хтось запитує щось по суті, а хтось має бажання просто поспілкуватись. І важливі повідомлення можуть значно просісти у стрічці загальних повідомлень і залишитись без відповіді. Оскільки більшість запитань схожі одні на одні або періодично повторюються, виникає потреба згрупувати певні асоціативні запитання та сформувати для них актуальний перелік відповідей. Відповідно, автоматизація процесів отримання відповідей мешканцями багатоквартирних будинків є актуальною задачею

Мета кваліфікаційної роботи бакалавра – створення й програмна реалізація методу автоматизованого підбору відповідей на запитання за семантичною подібністю.

Об’єкт дослідження – процес підбору відповідей на запитання в форматі цифрового тексту.

Предмет дослідження – інформаційні технології, моделі, методи та засоби для автоматизованого підбору відповідей на запитання за семантичною подібністю.

.

Розділ 1 Характеристика предметної області: аналіз моделей, методів та реалізацій

1.1 Аналіз інформаційних моделей

Без спілкування немає суспільства, без суспільства немає людини соціальної, розумної, homo sapiens. Найважливішим шляхом до розуміння інших людей є процес спілкування. Люди мають глибоку й універсальну потребу у взаємодії з іншими, і чим більше їх комунікативні здібності, тим більше задоволення та винагороди буде мати їхнє життя [4]. Вираження почуттів, настроїв, планів, ділові документи – усе це породжує величезні пласти інформації поданої у тексті. Оскільки при роботі з великою кількістю текстової інформації витрачається непомірно багато часу, компанії та організацій покладаються на методи вилучення та опрацювання інформації для автоматизації ручної роботи за допомогою праці інтелектуальних алгоритмів. Такі алгоритми можуть сприяти зменшенню витрати часу, а ще зменшити людські зусилля та зробити різноманітні процеси із меншою кількістю помилок.

Обробка природної мови (NLP) відноситься до галузі інформатики, а точніше, до галузі штучного інтелекту або ШІ, яка займається наданням комп’ютерам здатності розуміти текст і вимовлені слова приблизно так само, як і люди [5].

NLP поєднує у собі обчислювальну лінгвістику (моделювання людської мови на основі правил) зі статистичними моделями, моделями машинного навчання та глибокого навчання. Ці технології разом дозволяють комп’ютерам обробляти людську мову у вигляді тексту або голосових даних, а також у буквальному сенсі «розуміти» її повне значення, враховуючи наміри та настрої мовця чи письменника. Наприклад, NLP

може використовуватись, щоб створювати системи типу розпізнавання мовлення, узагальнення документів, машинного перекладу, виявлення спаму, розпізнавання іменованих сутностей, відповіді питання, автозаповнення, предиктивного введення тексту тощо.

Сьогодні у більшості є смартфони з розпізнаванням мови – для них характерно використання NLP для того, щоб розуміти мову та бути спроможними вести з людиною діалог (асистенти). Також багато людей використовують ноутбуки із вбудованим в ОС розпізнаванням мови.

Windows має віртуальний помічник Cortana, який розпізнає мову [6]. За допомогою Cortana є можливість створювати нагадування, відкривати програми, надсилати листи, грати в ігри, дізнаватися інформацію щодо погоди тощо.

Комп'ютерна лінгвістика представляє собою галузь мовознавства, що вивчає мову за допомогою програм, комп'ютерних технологій з організації та обробки даних та використовується не лише в лінгвістиці, а також у суміжних з нею дисциплінах [7]. Досліджує, як і яким чином людська мова може бути автоматично опрацьована та інтерпретована.

Досягнення у сфері комп'ютерної лінгвістики забезпечують не лише збір, обробку та пошук інформації, накопичення, а також забезпечують [7]:

– міжмовний переклад;

– тематичний пошук текстів;

– аналіз тексту чи розмовної мови за змістом, настроями та іншими якостями;

– можливість отримання відповіді на запитання, включно з такими, що роблять висновки чи опис;

– узагальнення тексту;

– створення чат-ботів, здатних виконувати складні завдання.

Інтернет-спілкування посідає чільне місце у засобах комунікації, налічуючи понад 4 млрд користувачів [8]. Зокрема, особливим пластом інтернет-спілкування є спілкування з громадськими установами. Зокрема, коли людина стикається із моментом оплати комунальних платежів, і є необхідність дізнатись потрібний тариф, або незрозуміло певні правові аспекти різких збільшень тарифів тощо. Ще одним яскравим прикладом даної тематики є групи ОСББ. Якщо почитати чати ОСББ, можна знайти цілі переліки типових запитань або проблем мешканців [9, 10]. Для вирішення нагальних питань людей з побутових проблем існують гарячі лінії, куди можна подзвонити та дізнатися необхідну інформацію, прийомні дні у відповідних установах, або форми подачі заявок, які можна заповнити та отримати відповідь на бажане питання у електронному вигляді. В цілому, гарачі лінії державних установ частіше всього будуть зайняті, а відвідавши установи у прийомні дні доведеться вистояти великі черги, при тому немає гарантії що все ж таки проблему буде вирішено, і людині не доведеться йти стояти чергу знову. Тому для таких питань створюються цілі форуми, на кшталт «Збільшення ціни на газ» [11]. Як правило, на таких форумах ставиться багато типових запитань, які мають різні формулювання, проте мають спільні ключові точки – ключові слова. І відповідно відповіді на такі питання легко групуються та мають у собі ті ж ключові одиниці. Формується певна семантична модель. Семантична модель включає слово, його визначення, поєднання з іншими словами, складання з нього фраз та речень.

Семантичний аналіз важке математичне завдання, вирішення якого застосовується у процесі створення ШІ, що ускладнюється тим, що є необхідність обробки природного мови. Складність також у тому, що комп'ютер не вміє правильно пояснювати образи, які людина передає за

допомогою символів. Дані якісного семантичного аналізу можуть використовуватись у торгівлі для аналізу попиту на товари за отриманими відгуками, у пошукових системах, системах автоматичного перекладу та ін. [12].

При проведенні семантичного аналізу використовуються різноманітні статистичні показники. До статистичних показників належать:

- кількість символів з пробілами та без;
- кількість слів, у тому числі унікальних та значущих;
- стоп-слів;
- кількість води;
- граматичних помилок;
- семантичне ядро;
- відсоток класичної та академічної нудоти.

При підрахунку враховується кількість унікальних слів (без повторень), число значних слів (іменників), стоп-слів (які позбавлені свого сенсу). Відсоток води визначається шляхом розподілу числа значних слів на загальну кількість слів. Кількість води не може вважатись показником якості тексту, проте все ж таки краще, щоб цей показник не перевищував 65%. Якщо в тексті виявлено 75% води або більше того, варто зменшити кількість незначних слів [12]. Нудота у класичному розумінні виражає, скільки разів повторюється в тексті одне й те саме слово. Ще нудота має іншу назву – щільність ключових слів. Тобто відношення кількості ключових слів в тексті до загальної кількості слів у даному тексті. Щільність виражається у відсотковому співвідношенні. Оптимальною щільністю ключових слів вважається 4-6% [13].

Враховуючи вищесказане, розробка методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом є актуальною задачею, яка допоможе одразу двом категоріям:

– Адміністратори (голови ОССБ, працівники комунальних підприємств, працівники ЦНАП тощо). Для цієї категорії доцільно автоматизовувати процеси надання автоматизованих відповідей населенню, оскільки це допоможе розвантажити великі черги та дозволить значно економити робочий час працівників.

– Кінцеві користувачі. Для цієї категорії користувачів також буде доцільно використовувати метод автоматизованого підбору відповідей на запитання, оскільки це дозволить знайти потрібні відповіді та не потребує простою у чергах. Також важливим ефектом є те, що користувачу не потрібно шукати типові запитання.

Отже, напрямок є досить актуальним та має шляхи для використання ІТ з метою автоматизації.

1.2 Огляд теоретичних підходів до розв’язку подібних задач

Ключове слово у аналізі текстової інформації (зокрема, й у пошукових системах) представляє собою деякий набір слів, які мають зміст тексту та одержуються лінгвістичними та математичними методами (наприклад, аналізуючи частоти появи слів у тексті, дисперсійною оцінкою тощо). Це особливо важливі, загальнозрозумілі, ємні та показові для окремо взятої теми слова з тексту, набір яких може дати високорівневий опис змісту для читача, забезпечивши компактне уявлення та збереження у пам'яті сенсу тексту.

У результаті систематизації даних дослідників [14] виділено певний перелік суттєвих властивостей, а також функцій ключових слів у текстах, значущих щодо контексту моделювання та алгоритмізації процесу їх вилучення. Отже, ключові слова характерні такими властивостями:

– є найбільш уживаними (частотними) найменуваннями, позначають ознаку предмета, стан або дію;

– представлені значимою лексикою, є узагальненими щодо своєї семантики, стилістично нейтральні, не оцінні;

– пов'язані один з одним деякою мережею семантичних зв'язків;

– більше 50% слів наповненості ядра тематичного компонента складається з ключових термінів, а найменший набір КС наближається до інваріанту змісту при їхньому логічному впорядкуванні;

– набір КС складається у середньому з 5-15 слів або 8-10 слів, що відповідає обсягу оперативної пам'яті людини, у тексті міститься близько 25-30% КС;

– набір КС визначає контексти слів, що є максимально передбачуваними.

Будь-який алгоритм вилучення КС реалізує одну або кілька систем розпізнавання образів. Вона розбиває вхідну множину слів на два класи: ключові слова та інші.

Можна виділити наступну сукупність ознак (рисунок 1.1) [15]:

– наявність елементів навчання, а також підходи для реалізації процесу навчання;

– вид математичного складу системи розпізнавання, що пояснюється формою подачі інформації ознак КС;

– вид використовуваних реалізації методу лінгвістичних ресурсів.

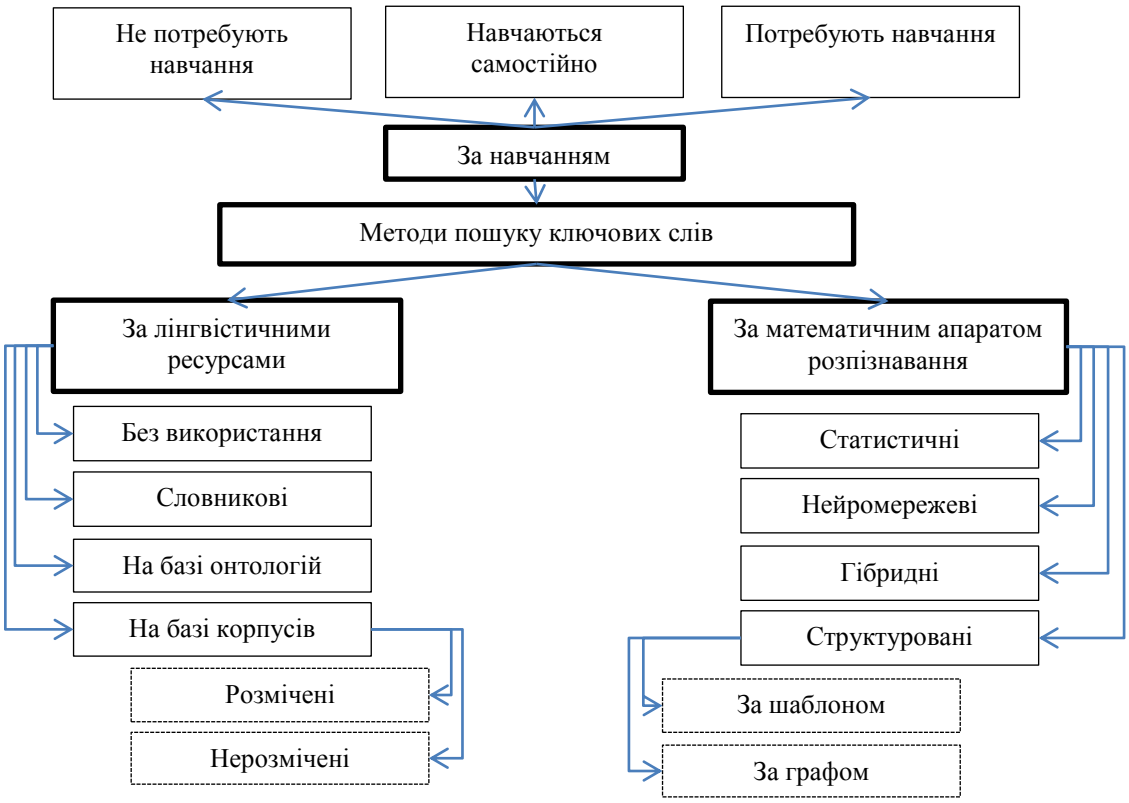


Рисунок 1.1 – Класифікація методів пошуку ключових слів [15]

За присутністю елементів навчання виділяють: ненавчені, які навчаються та навчені методи вилучення КС. Методи, що не потребують навчання, мають на увазі контекстно-незалежне виділення КС з окремого тексту на основі апріорно складених моделей, а також правил. Вони підходять для гомогенні за функціональним стилем корпусів текстів, що збільшуються з часом у обсягах, наприклад наукових праць чи нормативних актів. Методи, які потребують навчання, припускають використання різноманітних лінгвістичних ресурсів для налаштування критеріїв прийняття рішень при детектуванні ключових слів. Тут велике значення має коректне виділення КС у вибірці, що використовується для навчання. Серед методів з навчанням виділяється підклас самонавчання, якщо навчання ведеться без вчителя або з підкріпленням (на основі пасивної адаптації).

За іншою ознакою класифікації виділяють статистичні та структурні методи вилучення КС.

Статистичні методи враховують відносні частоти морфологічних, лексичних, синтаксичних одиниць, а також їх комбінацій. Це робить створювані на їх основі алгоритми досить простими, проте недостатньо точними, оскільки ознака частотності ключових слів переважає [16]. Одним із класичних методів у даному класі є розрахунок для кожного слова міри TF-IDF (Term Frequency-Inverse Document Frequency) [17]. Метод TF-IDF працює як визначення відносної частоти слів у конкретному документі у порівнянні з оберненою пропорцією цього слова по всьому корпусу документів. Інтуїтивно, це обчислення визначає, наскільки релевантне дане слово у деякому документі. Слова, які зустрічаються в одному або невеликій групі документів, як правило, має вищі номери TF-IDF, ніж звичайні слова, такі як артиклі та прийменники.

В основі структурних методів лежить представлення про текст, як систему семантично і граматично взаємопов'язаних елементів-слів, які характеризуються наборами лінгвістичних ознак. По цій причині ряд дослідників називають цей клас методів лінгвістичним.

Тут у першому наближенні можуть бути виділені два підкласи – графові та синтаксичні (шаблонні) методи. Графові (граф-орієнтовані) методи представляють текст безліччю слів-вершин (або вершин-словосполучень) та ребервідносин між ними [18]. Ці відносини можуть висловлювати для кожної пари слів факти послідовної появи в тексті, наявності у вікні заданого розміру та семантичну близькість. Для вершин отриманого графа обчислюються заходи центральності та пороговому критерію відбираються ключові слова. Відмінності між цими методами

перебувають у особливостях обліку значимості кожної вершини та обчислення відносин між ними.

В основі синтаксичних або шаблонних методів лежить уявлення про регулярні синтаксичні конструкції. Вони містять на певних позиціях ключові слова. У чистому вигляді такі методи слабо застосовні, проте можуть використовуватися у поєднанні з іншими.

Нейромережні методи до завдання вилучення КС стали застосовуватися порівняно нещодавно і засновані на властивості штучних нейронних мереж до узагальнення та виділення прихованих залежностей між вхідними та вихідними даними [19]. Однак для формування наборів даних для навчання та функціонування нейромереж потрібно виділення структурних та статистичних ознак, тому практично методи виділення ключових слів є гібридними, тобто, що поєднують у собі елементи основних розглянутих класів.

Алгоритми вилучення КС, які реалізують зазначені методи, можуть взагалі не використовувати будь-які лінгвістичні ресурси, або використовувати різновиди словників, онтології та тезауруси, а ще корпуси текстів (Без розмітки або з розміткою).

Найбільш класичний метод побудови вектора представлення текстів вважають мішок слів [20]. Ключове припущення цього підходу полягає в тому, що текст може бути виражений за допомогою невпорядкованого набору частот слова (терміни) у тексті. Кількість виділених ознак (слів) часто можна зменшити шляхом перетворення слів у їх родову форму (процес лемматизації). Подання частоти тексту (TF) дуже часто змінюється за допомогою інвертованої частоти оформлення документів (IDF), що дає TF-IDF представлення текстів.

Для реалізації методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом доцільно використовувати статистичні методи пошуку ключових слів, а саме – метод дисперсійної оцінки [21], призначений для оцінки важливості кожного слова в досліджуваному тексті, який є оцінкою дискримінантної сили слів. Метод дисперсійного оцінювання дозволяє відділити із загальної множини широковживаних у тексті слів слова, що розташовані рівномірно.

Для забезпечення збереження типових користувацьких запитань та типових відповідей на них доцільно використовувати БД. Для реалізації користувацького інтерфейсу планується використати ООП, оскільки вищезазначений метод планується для використання для груп осіб – доцільно буде використати класи та принципи ООП.

1.3 Аналіз існуючих програмних рішень

Враховуючи специфіку предметної області – буде потреба аналізувати невеликі текстові повідомлення, схожі за обсягом на твіти. Для аналізу яких потрібно знаходити ключові слова. Відповідно, дана задача схожа із завданнями SEO оптимізації інтернет-контенту, тому далі буде розглянуто існуючі програмні реалізації які містять алгоритми пошуку ключових слів.

Google AdWords: Keyword Planner (Планувальник ключових слів) – безкоштовний інструмент, який дозволяє підбирати різні варіанти слів на основі заданого головного ключа. Крім того, можна отримати дані за статистикою запитів зазначених словосполучень, сезонними коливаннями, рекомендованою ціною кліка в контекстній рекламі та рядом інших корисних параметрів [22].

Планувальник ключових слів Google розроблено, щоб досліджувати ключові слова для використання в кампаніях у пошуковій мережі. Планувальник ключових слів Google Ads використовується для:

– Відкриття нових пошукових ключових слів.

– Перегляд середніх місячних пошукових чисел для ключових слів.

– Визначення затрат.

– Створення нових рекламних кампаній у пошуковій мережі.

Вигляд інтерфейсу зображено на рисунку 1.2.

Рисунок 1.2 – Вебінтерфейс Google AdWords: Keyword Planner [23]

Наступним сервісом буде розглянуто Answer the Public, який витягує запити автозаповнення Google у великій кількості та поділяє їх на різні списки. Це дуже корисно для аналізу питань, які люди ставлять у Google [24]. Вигляд інтерфейсу зображено на рисунку 1.3.

Рисунок 1.3 – Вебінтерфейс Answer the Public [25]

Ще одним вебінтерфейсом для корисним для CEO оптимізації на основі пошуку ключових слів є Ahrefs' Free Backlink Checker, що безкоштовно показує 100 найкращих зворотних посилань на будь-який сайт або веб-сторінку [24].

Він також розкриває п'ять найбільш пов'язаних сторінок, найбільш поширених анкорів, загальну кількість зворотних посилань і доменів, що посилаються. Також можна бачити власні рейтинги рейтингу домену (DR) та рейтингу URL (UR).

Цей інструмент оснащений тією самою базою даних, як і преміальна версія Ahrefs. Деякі статистичні дані:

– 16 трильйонів відомих посилань;

– 176 мільйонів унікальних доменів;

– 6 мільярдів сторінок переглядаються щодня

Індекс зворотних посилань оновлюється свіжими даними кожні 15 хвилин, тому ви можете негайно скористатися новими можливостями. Зовнішній вигляд зображено на рисунку 1.4.

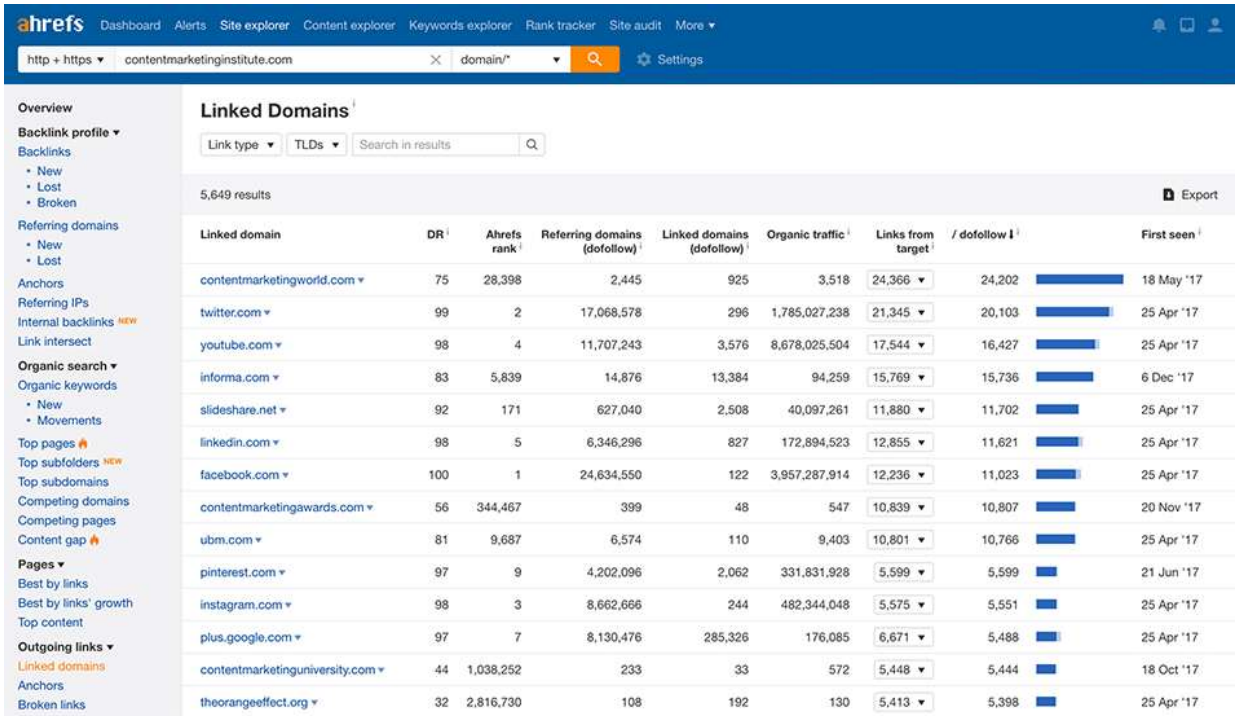


Рисунок 1.4 – Вебінтерфейс Answer the Public [26]

Також є розробка для спеціальної класифікації тексту, яка називається користувальна класифікація тексту. Це одна з функцій, які пропонує Azure Cognitive Service for Language [27]. Це хмарна служба API, яка застосовує інтелект машинного навчання, щоб дозволити створювати власні моделі для завдань класифікації тексту.

Спеціальна класифікація тексту пропонується як частина користувацьких функцій в Azure Cognitive for Language. Ця функція дає змогу користувачам створювати власні моделі штучного інтелекту для класифікації тексту за користувацькими категоріями, попередньо визначеними користувачем. Створюючи власний проєкт класифікації тексту, розробники можуть ітеративно позначати дані, навчати, оцінювати

та покращувати продуктивність моделі, перш ніж зробити її доступною для використання. Якість позначених даних сильно впливає на продуктивність моделі. Щоб спростити створення та налаштування моделі, сервіс пропонує власний веб-портал, доступ до якого можна отримати через мовну студію [28].

З опрацьованих досліджень можна зробити висновок, що розробка програмного забезпечення для вирішення задачі автоматизованого підбору відповідей на запитання за семантичною подібністю є актуальною.

1.4 Аналіз сучасних засобів створення програмного забезпечення

Для реалізації методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом доцільно використати десктопний застосунок, який буде мати інтуїтивно зрозумілий користувацький інтерфейс. Під віконним застосунком розуміється застосунок, розроблений в Windows [29]. Оскільки програмне забезпечення носить дослідницький характер, важлива максимальна інформативність та швидкодія. Мобільний застосунок незручний через маленький екран, тому віконний застосунок для вирішення задачі автоматизованого підбору відповідей на запитання за семантичною подібністю є більш доцільним.

Сучасний ІТ-розробник має чимало способів створення програмного продукту: для кожної унікальної задачі можна підібрати найбільш зручний фреймворк, мову програмування та СКБД.

Найбільш потужні та популярні мови програмування: серед них C# та Java. За індексом TIOBE [30] ці мови програмування займають третє та п’яте місце (рисунок 1.5) серед усіх існуючих, що вказує на їх популярність та потужність.

May 2022	May 2021	Change	Programming Language	Ratings	Change
1	2	^	Python	12.74%	+0.86%
2	1	v	C	11.59%	-1.80%
3	3		Java	10.99%	-0.74%
4	4		C++	8.83%	+1.01%
5	5		C#	6.39%	+1.98%

Рисунок 1.5 – Рейтинг мов програмування за TIOBE [30]

Деталі про мову програмування Java: це об'єктно-орієнтована мова програмування, реліз якої відбувся 1995 року, із 2009 року мову підтримує компанія «Oracle» [31].

Java можна використовувати для створення додатків на низці платформ. Настільні комп’ютери, сервери, мобільні телефони, планшети, та веб-браузери використовують цю мову програмування. Оскільки Java відповідає вимогам WORA, той самий код можна запускати на всіх платформах із середовищем виконання Java (JRE) без перекомпіляції коду [32]

Детальніше про мову програмування C# – об’єктно-орієнтована мова програмування, реліз якої відбувся у 2000 році компанією Microsoft.

Особливості:

– Гнучкість: при створенні додатків будь-якого типу призначення можна використовувати цю мову.

– Статична типізація.

– Легкість сприйняття.

Підтримка поліморфізму.

C# – це мова програмування загального призначення, що використовується для створення різних типів програм і додатків [32].

Розглянемо існуючі СКБД, котрі можемо використати для реалізації програмного продукту. Одними із найпотужніших сьогодні є MS SQL Server та MySQL. Проаналізуємо ці два продукти.

MySQL – це безкоштовна СКБД з відкритим вихідним кодом, Це стабільне, надійне та потужне рішення з такими розширеними функціями, як:

– захищеність даних;

– масштабування;

– висока продуктивність;

– цілодобова підтримка.

MS SQL Server –це СКБД, розроблена Microsoft. Включає в себе мову SQL і Transact-SQL (T-SQL), власну мову Microsoft Механізм баз даних розділений на два сегменти, реляційний механізм, який використовується для обробки команд і запитів. Другий – це механізм зберігання даних, призначений для керування різними функціями бази даних, такими як таблиці, сторінки, файли, індекси та транзакції [33].

Переваги:

– легко встановити;

– покращена продуктивність;

– кілька випусків SQL Server;

– дуже безпечний;

– чудовий механізм відновлення та відновлення даних;

– нижча вартість володіння.

Таким чином, було проаналізовано найпопулярніші існуючі засоби для створення програмних продуктів, їх призначення та переваги та обрано для реалізації методу автоматизованого підбору відповідей за семантичною подібністю віконний застосунок на платформі .NET мовою програмування C# та редактором програмного коду VisualStudio. У якості СКБД було визначено доцільним використання MS SQLServer.

1.5 Мета, задачі та вимоги до реалізації інформаційної системи

Метою кваліфікаційної роботи бакалавра є розробка й апробація методу автоматизованого підбору відповідей на запитання за семантичною подібністю, для чого слід вирішити задачі:

– Провести аналіз предметної області.

– Розробити метод автоматизованого підбору відповідей на запитання за семантичною подібністю.

– Виконати проєктування інформаційної системи на базі методу автоматизованого підбору відповідей на запитання за семантичною подібністю.

– Зробити вибір засобів розробки інформаційної системи.

– Розробити програмну реалізацію методу автоматизованого підбору відповідей на запитання за семантичною подібністю, провести її тестування.

Розроблена програмна реалізація методу автоматизованого підбору відповідей на запитання за семантичною подібністю в вигляді інформаційної системи на платформі.NET має виконувати наступні основні групи функцій:

1. Семантичний аналіз наявних запитань і тестового користувацького запитання:

– Первинна обробка тексту кожного запитання (розділові знаки, регістр тощо).

– Формування вектора слів запитання – впорядкованої множиною слів.

– Формування множини оригінальних слів за впорядкованою множини слів.

– Обрахунок семантичної ваги кожного слова у множині оригінальних слів за методом дисперсного оцінювання.

2. Обрахунок оцінок семантичної подібності користувацького запитання до кожного з наявних у базі запитань з асоційованими відповідями:

– Пошук однакових оригінальних слів у користувацькому запитанні й кожному з наявних запитань з асоційованими відповідями у базі запитань.

– Обрахунок оцінок семантичної подібності користувацького запитання до кожного з наявних у базі запитань з асоційованими відповідями.

3. Знаходження відповіді у базі на користувацьке запитання:

– Знаходження наявного запитання у базі, що має максимальну оцінку семантичної подібності до користувацького запитання (робоче запитання).

– Знаходження відповіді у базі, яка асоційована з робочим запитанням.

– Видача повідомлення користувачу, якщо не вдалося знайти відповіді.

Наведеного функціоналу достатньо для прикладного тестування методу автоматизованого підбору відповідей на запитання за семантичною подібністю.

Розділ 2 Проєктування інформаційної системи

2.1 Метод автоматизованого підбору відповідей на запитання за семантичною подібністю

Загальна схема методу підбору асоціативних відповідей зображена на Рисунку 2.1. Користувач ставить своє запитання, у якому методом дисперсійної оцінки відбувається пошук ключових слів та їх оцінка.

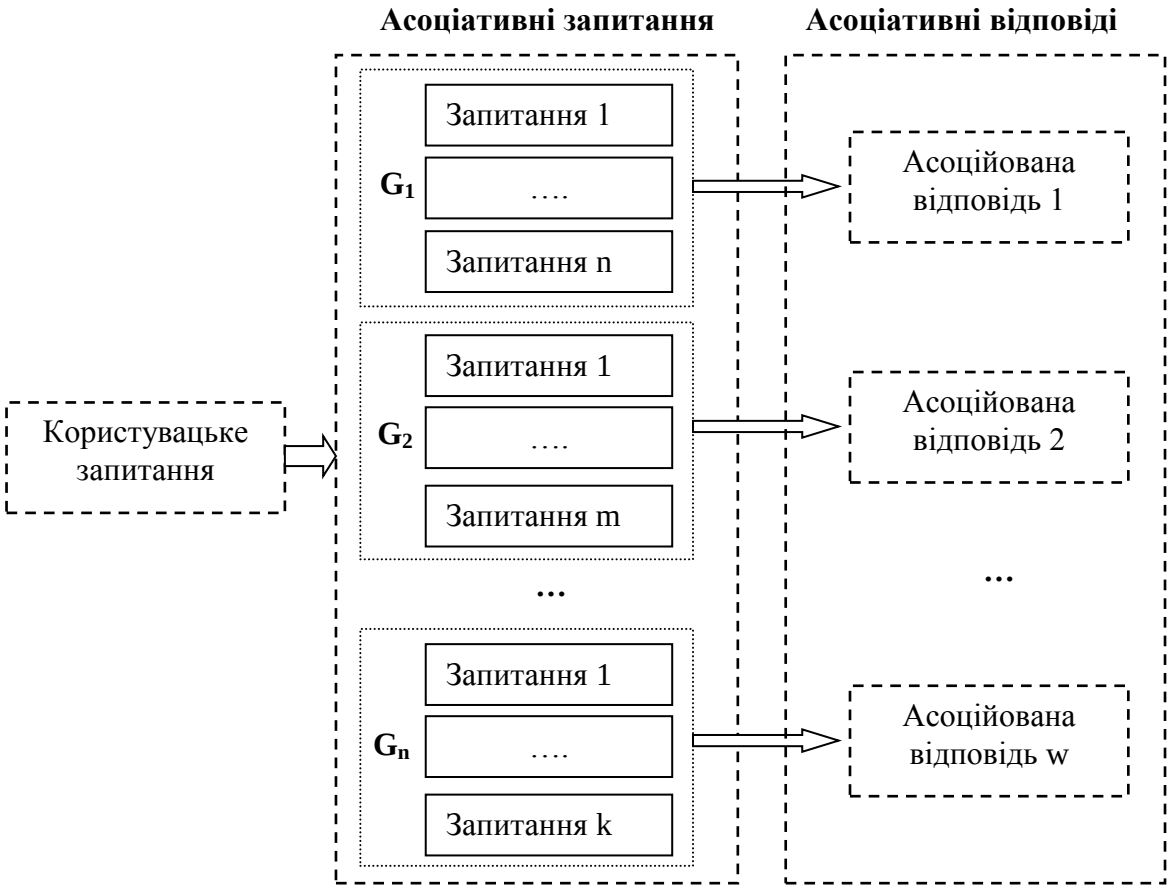


Рисунок 2.1 – Схема автоматизованого підбору відповіді на користувацькі запитання

У базі асоціативних запитань методом дисперсійної оцінки також виконується пошук ключових слів та оцінювання кожного ключового слова

окремого запитання $q_1, q_2, .., q_n$ із груп G_1,.., G_n, де $G_n = \{q_1, q_2, .., q_n\}$. Далі відбувається оцінка схожості користувацького запитання за ключовими словами із існуючими в базі асоціативними запитаннями.

Кожній із груп запитань притаманна певна асоціативна відповідь ans_w, закріплена за кожною групою із множини $ans = \{ans_1, ans_2, .., ans_w\}$.

Вхідними даними методу для автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом є база наявних запитань та притаманних їм асоціативних відповідей та тестове користувацьке запитання, відповідь на яке користувач повинен отримати (рисунок 2.2).

Відповідно, алгоритм у своїй роботі проходить певні кроки. Першим кроком є аналіз наявних запитань і тестового користувацького запитання, що включає в себе первинну обробку тексту кожного запитання. У первинну обробку включено видалення розділових знаків, приведення всіх слів у нижній регістр, видалення стоп-слів тощо. Далі на першому етапі формується вектор слів запитання, що являє собою впорядковану множину слів. Після чого відбувається формування множини оригінальних слів за впорядкованою множиною слів та обрахунок семантичної ваги кожного слова у множині оригінальних слів за методом дисперсного оцінювання.

Другим кроком є обрахунок оцінок семантичної подібності користувацького запитання до кожного з наявних у базі запитань із асоційованими відповідями. Цей етап включає в себе пошук однакових оригінальних слів у користувацькому запитанні й кожному з наявних запитань з асоційованими відповідями у базі запитань та обрахунок оцінок семантичної подібності користувацького запитання до кожного з наявних у базі запитань з асоційованими відповідями.

Рисунок 2.2 – Схема методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом

Третім кроком є знаходження відповіді у базі на користувацьке запитання. На цьому етапі відбувається знаходження наявного запитання у базі, що має максимальну оцінку семантичної подібності до користувацького запитання (робоче запитання), після чого відбувається знаходження відповіді у базі, яка асоційована з робочим запитанням. І на

останок відбувається видача повідомлення користувачу, якщо не вдалося знайти відповіді.

Вихідними даними методу є безпосередньо відповідь на тестове користувацьке запитання.

Оцінка семантичної подібності $P_{a,b}$ користувацького запитання a до наявного запитання b до деякої асоційованої відповіді визначається наступним чином:

$$P_{a,b} = \sum_{i=1}^{n} D_{a,i} D_{b,i}, \tag{2.1}$$

де n – кількість однакових оригінальних слів у користувацькому запитанні a й наявному запитанні b до деякої асоційованої відповіді, $D_{a,i}$ – оцінка семантичної важливості слова i у користувацькому запитанні a, $D_{b,i}$ – оцінка семантичної важливості слова i у наявному запитанні b до деякої асоційованої відповіді.

На базі вищевикладеного матеріалу формується метод автоматизованого підбору відповідей на запитання за семантичною ознакою.

2.2 Інформаційна структура системи

2.2.1 Проектна архітектура системи та взаємозв'язок компонентів

Узагальнена схема програмного комплексу за методом автоматизованого підбору відповідей на запитання за семантичною подібністю зображена на рисунку 2.3.

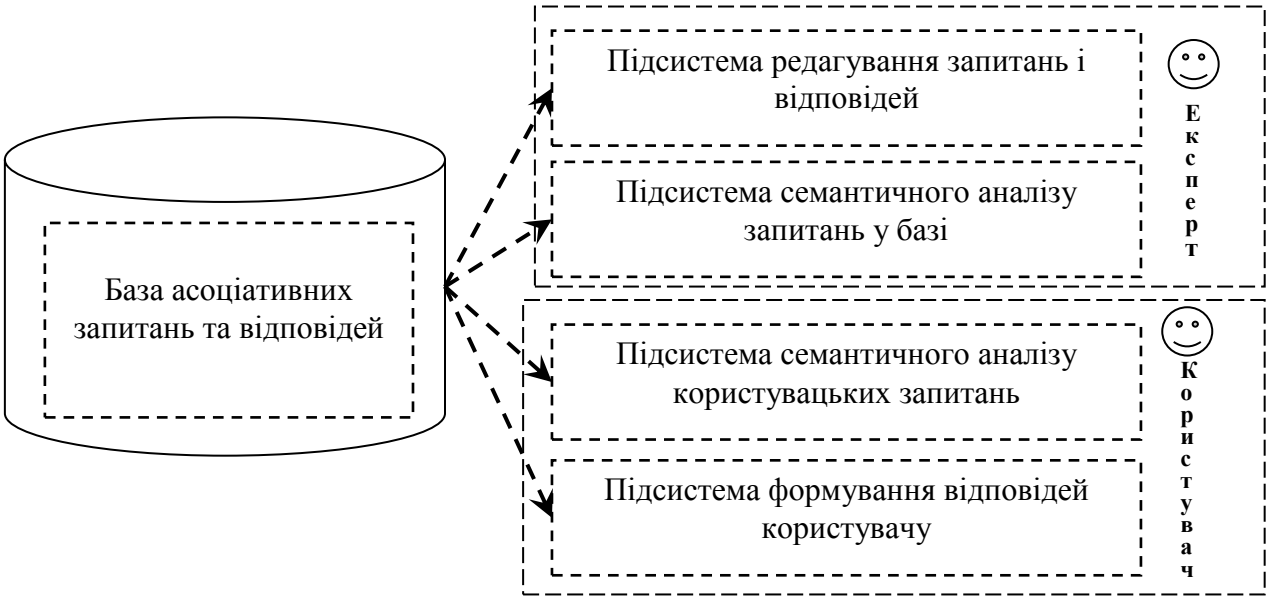


Рисунок 2.3 – Схематична ілюстрація підсистем експерта та користувача

Групи функцій зображених на рисунку 2.3 розділяються для експертів та користувачів. Для експерта доступно дві підсистеми:

– Підсистема редагування запитань і відповідей.

– Підсистема семантичного аналізу запитань у базі.

У першій підсистемі представлено функціонал редагування даних з бази запитань та відповідей: додавання нової відповіді чи запитання у базу, перегляд доступних запитань та відповідей, редагування.

У підсистемі семантичного аналізу запитань у базі доступні функції побудови векторів ключових слів а також оцінка їх семантичної важливості дисперсійним методом.

Група функцій для користувачів також представляє собою дві підсистеми:

– Підсистема семантичного аналізу користувацьких запитань.

– Підсистема формування відповідей користувачу.

Перша підсистема показує користувачу, які слова з його запиту є ключовими, а також їх оцінки важливості. Робить аналіз схожості запитань на наявні у базі методом описаним у 2.1.

Підсистема формування відповідей користувачу дозволяє отримати відповідь на поставлене запитання, а також побачити оцінку на скільки його запитання близьке до наявних у базі за ключовими словами.

2.2.2 Інформаційна модель

Для забезпечення коректної роботи методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом, необхідно створити базу даних, що дозволить зберігати інформацію максимально надійно та забезпечити зв’язки між таблицями. Інформаційна модель зображена на рисунку 2.4

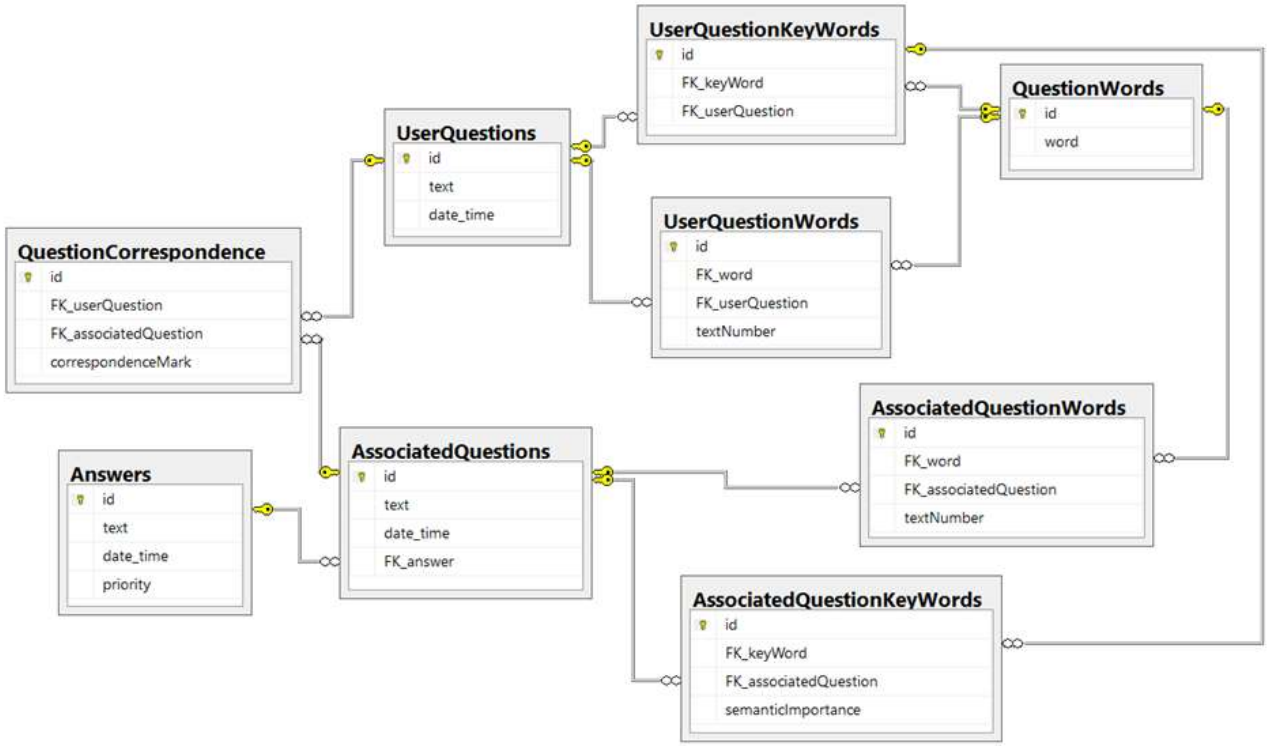


Рисунок 2.4 – Даталогічна модель застосунку для автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом

Таблиця «UserQuestions» (таблиця 2.1) призначена для збереження в базі даних питань, що задають користувачі, а саме їх текст та дату й час додання до БД.

Таблиця 2.1 – Атрибути таблиці «UserQuestions»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці
2.	text	text	Текст запитання, розміщеного на платформі.
3.	date_time	datetime	Дата й час розміщення питання на платформі.

Таблиця «QuestionWords» (таблиця 2.2) призначена для збереження кожного окремого слова, що зустрічається в запитаннях.

Таблиця 2.2 – Атрибути таблиці «QuestionWords»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці
2.	word	varchar(50)	Слово, що використовується у запитанні

Таблиця 2.3 – Атрибути таблиці «UserQuestionWords»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	FK_word	int	Вторинний ключ, посилання на запис таблиці «QuestionWord» для співставленням із відповідним словом.
3.	FK_userQuestion	int	Вторинний ключ, посилання на запис таблиці «UserQuestions» для співставленням із відповідним запитанням користувача.
4.	textNumber	int	Порядковий номер слова в реченні.

Таблиця «UserQuestionWords» (таблиця 2.3) використовується для збереження даних щодо слів, що зустрічаються в запитанні, розміщеному користувачем.

Таблиця «UserQuestionKeyWords» (таблиця 2.4) використовується для збереження даних щодо ключових слів, що містяться в запитанні, розміщеному користувачем.

Таблиця 2.4 – Атрибути таблиці «UserQuestionKeyWords»

№ п/п	**Назва атрибуту**	**Тип даних**	**Опис**
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	FK_keyWord	int	Вторинний ключ, посилання на запис таблиці «AssociatedQuestionKeyWords» для співставленням із відповідним ключовим словом.
3.	FK_userQuestion	int	Вторинний ключ, посилання на запис таблиці «UserQuestions» для співставленням із відповідним запитанням користувача.

Таблиця «Answers» (таблиця 2.5) використовується задля збереження відповідей на запитання, що часто зустрічаються серед поставлених користувачами.

Таблиця 2.5 – Атрибути таблиці «Answers»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	text	text	Текст, що міститься у відповіді на запитання.
3.	datetime	datetime	Дата й час додання відповіді в базу даних.
4.	priority	int	Пріоритет відповіді на запитання.

У таблиці «AssociatedQuestions» (таблиця 3.6) зберігаються питання, подібні до запитань, що розміщують користувачі.

Таблиця 2.6 – Атрибути таблиці «AssociatedQuestions»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	text	text	Текст, що міститься у асоційованому запитанні.
3.	datetime	datetime	Дата й час додання запитання в базу даних.
4.	FK_answer	int	Вторинний ключ, посилання на запис таблиці «Answers»,

Таблиця «AssociatedQuestionsWords» (таблиця 3.7) використовується задля збереження окремих слів кожного асоціативного запитання.

Таблиця 2.7 – Атрибути таблиці «AssociatedQuestionsWords»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	FK_word	int	Вторинний ключ, посилання на запис таблиці «QuestionWords» із співставленням із відповідним записом про слово запитання.
3.	FK_associatedQuestion	int	Вторинний ключ, посилання на запис таблиці «AssociatedQuestions» із співставленням із відповідним асоціативним запитанням.
4.	textNumber	int	Порядковий номер слова в реченні.

У таблиці «AssociatedQuestionsKeyWords» (таблиця 2.8) містяться ключові слова асоціативних запитань.

Таблиця 2.8 – Атрибути таблиці «AssociatedQuestionsWords»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	FK_keyWord	int	Вторинний ключ, посилання на запис таблиці «AssociatedQuestionsWords» із співставленням із відповідним записом про ключове слово запитання.
3.	FK_associatedQuestion	int	Вторинний ключ, посилання на запис таблиці «AssociatedQuestions» із співставленням із відповідним асоціативним запитанням.
4.	semanticImportance	int	Величина семантичної важливості слова.

У таблиці «QuestionCorrespondence» (таблиця 2.9) міститься інформація про відповідність запитань, поставлених користувачем із базою асоціативних запитань. В цій таблиці також зберігається і оцінка відповідності, обчислена методами, описаними в попередніх розділах.

Таблиця 2.9 – Атрибути таблиці «QuestionCorrespondence»

№ п/п	Назва атрибуту	Тип даних	Опис
1.	ID	int	Первинний ключ, числовий ідентифікатор для однозначного визначення запису таблиці.
2.	FK_userQuestion	int	Вторинний ключ, посилання на запис таблиці «UserQuestion» із співставленням із відповідним запитанням користувача.
3.	FK_associatedQuestion	int	Вторинний ключ, посилання на запис таблиці «AssociatedQuestions» із співставленням із відповідним асоціативним запитанням.
4.	correspondenceMark	varchar(50)	Значення оцінки відповідності заданого користувачем та асоційованого запитання.

Виконавши розділ, отримали структуру бази даних застосунку для автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом, забезпечили зв’язки між таблицями та заповнили таблиці початковою інформацією: текстами запитань користувачів, відповідями на поширені запитання та асоціативними запитаннями. Реалізована база даних забезпечує надійне збереження даних та доступ до них.

2.3 Вибір засобів розробки інформаційної системи

2.3.1 Вибір мови програмування

Відповідно по поставленого завдання, було обрано мову програмування C#, адже для розробки методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом саме ця мова програмування дозволить реалізувати програму в повній мірі, виконавши усі поставлені завдання.

Мова програмування C# – це об’єктно-орієнтована мова програмування, яка була створена спеціально для платформи .NET. Вона має ряд переваг, порівняно з іншими мовами програмування:

- висока швидкодія – швидка компіляція коду та робота застосунку;
- строга типізація, яка підвищує рівень безпеки даних;
- проста для вивчення та роботи;
- високий рівень підтримки від Microsoft;
- велика кількість офіційної та сторонньої документації;
- підтримка сторонніх бібліотек, що дуже спрощує роботу розробникам.

Отже, як видно з переваг даної мови програмування, вона найкращим чином допоможе виконати поставлені завдання.

2.3.2 Вибір фреймворку

Серед існуючих платформ для розробки програмного забезпечення, для швидкої розробки та розгортання програми найкраще підходить платформа .NET [34].

Дана платформа розроблена компанією Microsoft та дозволяє створювати різного типу застосунки – від веб сайту до віконного застосунку. Платформа має потужну підтримку від розробника, адже отримує постійні оновлення. Проте в деяких випадках це може негативно відобразитись на вже існуючих застосунках в тому випадку, якщо бібліотеки, які застосовувались для розробки отримують серйозні зміни при оновленні [34].

Платформа .NET вважається однією із надійніших та безпечніших в плані створення застосунків платформою. Отож, у рамках задачі реалізації методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом вона цілком задовольнить потреби у реалізації функцій програмного застосунку.

2.3.3 Вибір редактора програмного коду

У якості редактора програмного коду було обрано Visual Studio, оскільки вона надає безліч можливостей, що полегшують написання коду, та керування ним [35]. Використовуючи структуру, можна розгортати та згортати різні блоки коду. Додаткову інформацію про код є можливість подати за допомогою технології IntelliSense. Для пошуку в коді застосовні такі функції, як «Перейти», «Перейти до визначення» та «Знайти усі посилання». Вставляти блоки коду можна за допомогою фрагментів коду. Код також можна створювати за допомогою таких функцій, як «Створення в результаті використання».

Також за допомогою Visual Studio можна створювати користувацький інтерфейс – даний редактор коду підтримує створення різноманітних компонентів для віконних застосунків [35].

Великою перевагою цього редактора є те, що він підтримує створення діаграм класів для наочного відображення об’єктно-орієнтованих зв’язків.

Visual Studio є продуктом Microsoft, як і .NET та мова програмування C#, що забезпечить максимально тісний зв’язок між цими компонентами для створення застосунків.

Це далеко не весь перелік сильних сторін цього сучасного редактора, отож для написання програмного коду було обрано саме його.

2.3.4 Вибір СКБД

Для реалізації методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом необхідним є збереження даних для роботи застосунку у базі даних, тому необхідним є застосування системи керування баз даних, яка допоможе швидко створити та забезпечувати зручність роботи, саме тому обрано СКБД SQL Server. Серед її переваг варто виділити наступні:

- захищеність даних;
- потужність системи та якість збереження інформації ;
- контроль доступу до даних;
- зручний інтерфейс роботи;
- інтеграція з продуктами Microsoft.

З огляду цих переваг СКБД SQL Server є найоптимальнішим варіантом для роботи з базою даних в рамках даної задачі.

Підсумовуючи, в цьому розділі обрали усі необхідні інструменти для розробки додатку: мову програмування та СКБД. Як результат обрано наступний комплекс: мова програмування C#, .NET Framework, СКБД SQL Server.

Розділ 3 Програмна реалізація інформаційної системи

3.1 Структура та функціональне призначення програмних складових системи

Згідно з поставленим завданням КРБ, була спроектована інформаційна система, діаграма класів якої зображена на рисунку 3.1.

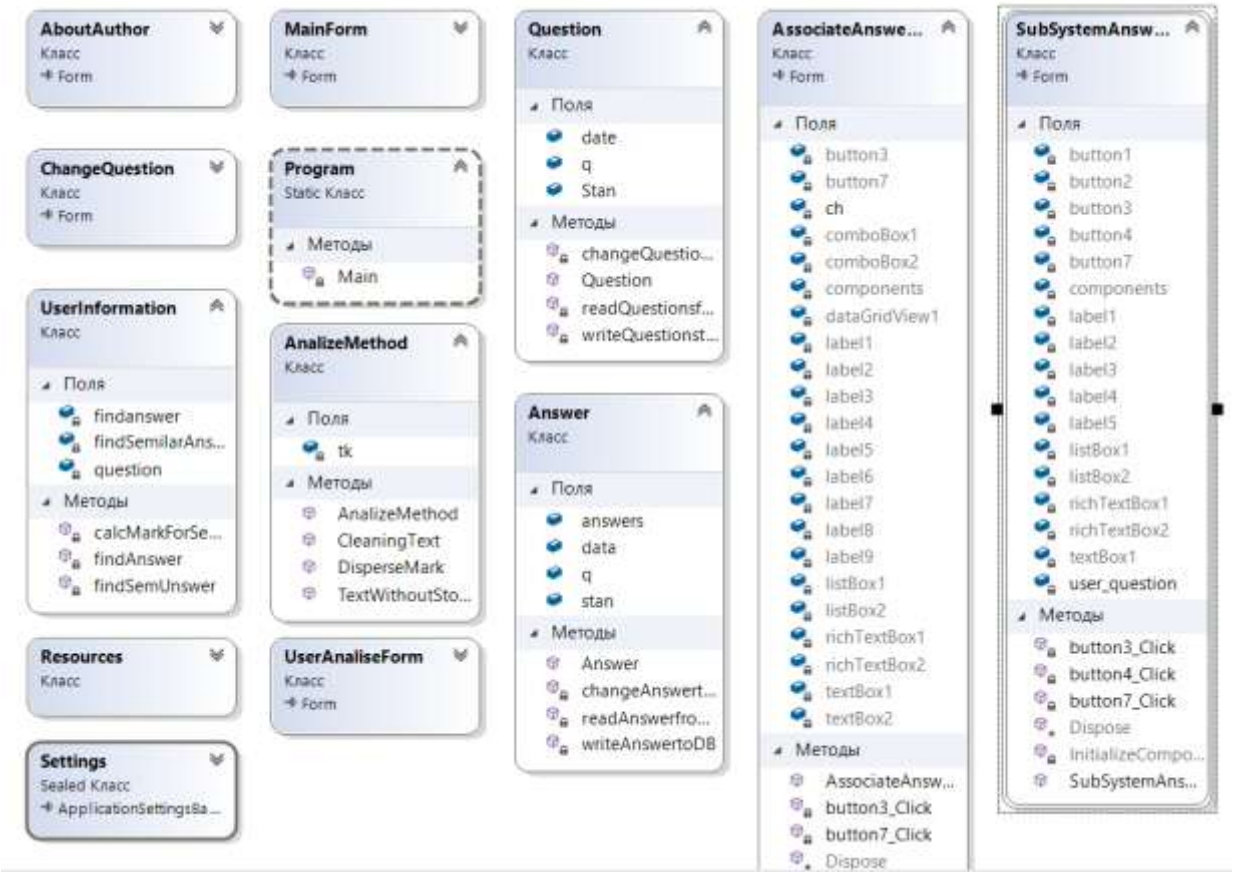


Рисунок 3.1 – Діаграма класів програмної реалізації методу автоматизованого підбору асоціативної відповіді за семантичною подібністю

Клас Answer реалізує логіку роботи з запитаннями з бази даних. Є методи для читання, додавання нового запису та зміни існуючого запису з

відповідними параметрами. Подібний йому за функціоналом є клас Question з відповідними методами для читання, запису та зміни даних запитань.

Клас AnalizeMethod виконує функціонал по роботі з запитаннями стосовно попередньої обробки запитань та векторизації. CleaningText() виконує очистку повідомлення від розділових знаків, а метод TextWithoutStops() виконує очистку від зайвих стоп-слів. Метод DisperseMark() шукає дисперсійну оцінку для кожного ключового слова.

Клас UserInformation призначений для виведення користувачу відповіді на задане ним питання та виводу відповідних схожих на його запитань за семантичною подібністю.

Відповідно до поставленої задачі, реалізації описаних складових буде достатньо для забезпечення базового функціоналу.

3.2 Особливості реалізації програмних складових системи

Згідно з темою КРБ, та описаним методом автоматизованого підбору асоціативної відповіді за семантичними ознаками, для очистки асоціативних запитань від розділових знаків та переведення усіх слів у нижній регістр відбувається у методі CleaningText(), який на вхід отримує запитання, та видаляє з нього все лишнє і повертає користувачу перерахування слів у нижньому регістрі. Код методу подано нижче:

```
public IEnumerable <String> CleaningText(String text)
        {
            var words = text.Split(new char[] { ' ', '.', ',', '!',
'?' })
```

```
            .Where(x => !string.IsNullOrWhiteSpace(x))
            .Select(x => x.ToLower());
              return words;
        }
```

Для очистки тексту також потрібно по видаляти стоп-слова. Оскільки повідомлення досить маленькі, то стоп-словами у рамках реалізації вважаються слова, довжина яких менше двох символів. На вхід метод TextWithoutStops(IEnumerable<String> t) отримує перелік слів, а результатом буде також перелік слів, проте вже без слів з однієї та двох літер. Код методу очистки від стоп-слів подано нижче:

```
        public IEnumerable<String>
TextWithoutStops(IEnumerable<String> t)
        {
            IEnumerable<String> words= t;

             words = words.Where(x =>x.Length>2);

            return words;
        }
```

Відображення вектору ключових слів та частоти їх появи після видалення стоп-слів та очистки даних зображено на рисунку 3.2.

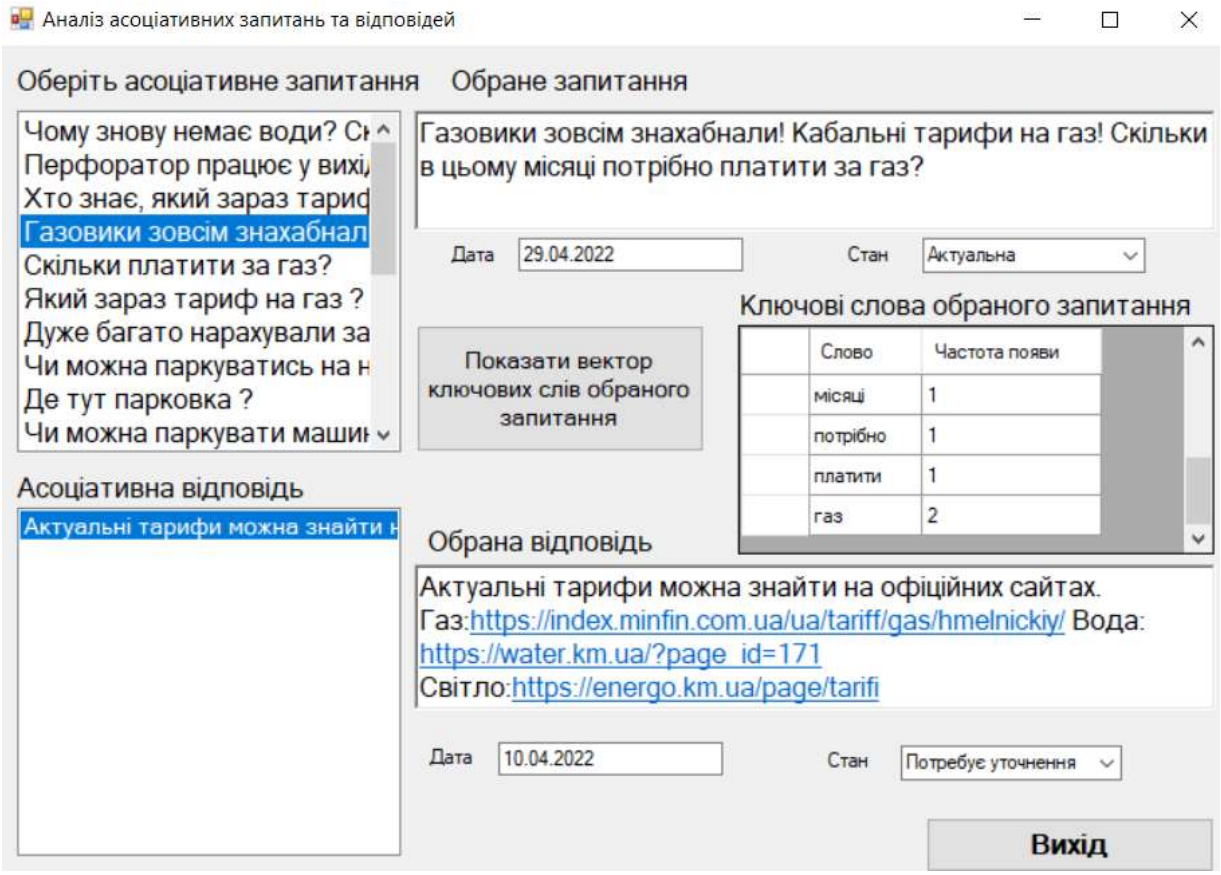


Рисунок 3.2 – Слова та їх частоти зустрічання у асоціативному запитанні

Фрагмент методу, що відповідає за функціонал зміни асоціативної відповіді у базі асоціативних відповідей наведено нижче:

```
int index = listBox1.SelectedIndex;
ans[index].answers = richTextBox1.Text;
ans[index].stan = comboBox1.SelectedIndex;
ans[index].data = textBox1.Text;
string sql = "Update Answers set text = @ans[index].answers
date_time=@ans[index].data where Id = @index";

MessageBox.Show("Дані було змінено! ");
```

Початкова форма редагування зображена на рисунку 3.3.

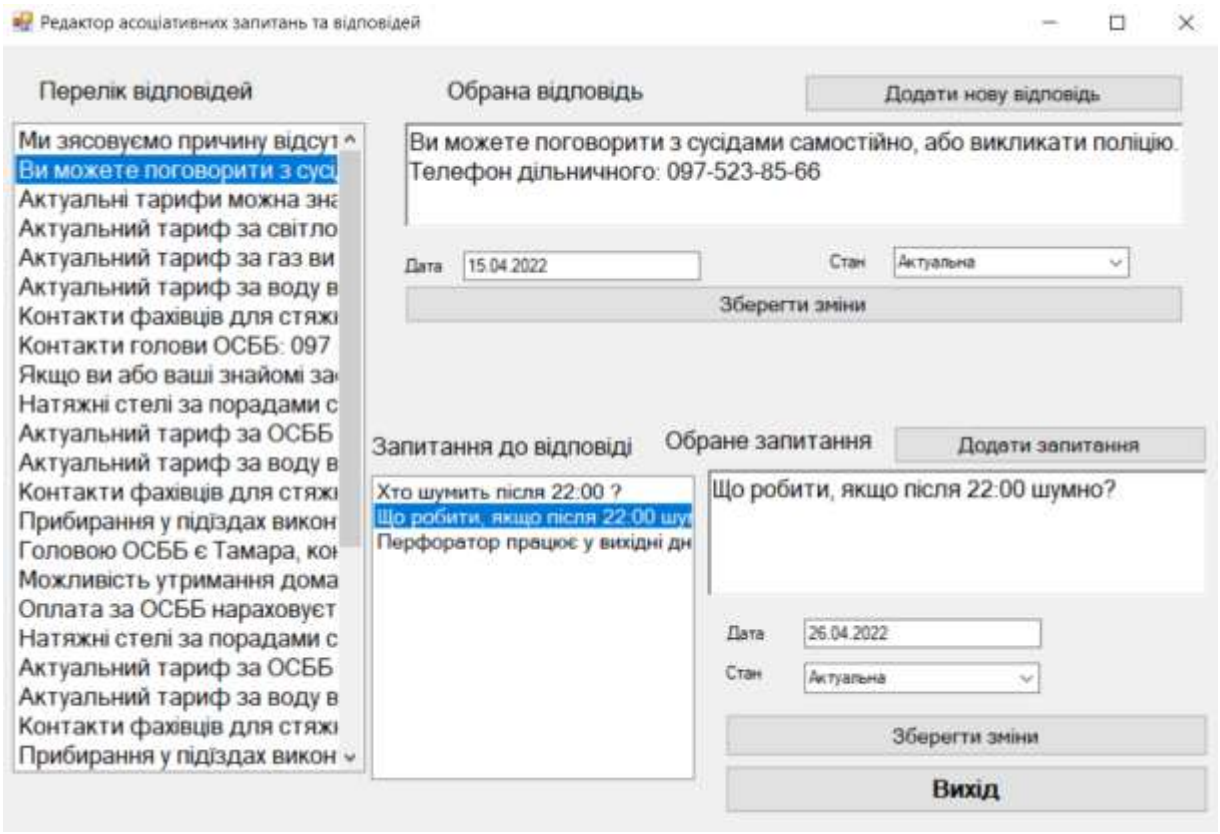


Рисунок 3.3 – Форма редагування

У обраному рядку із відповіддю було змінено дату та текст відповіді (рисунок 3.4).

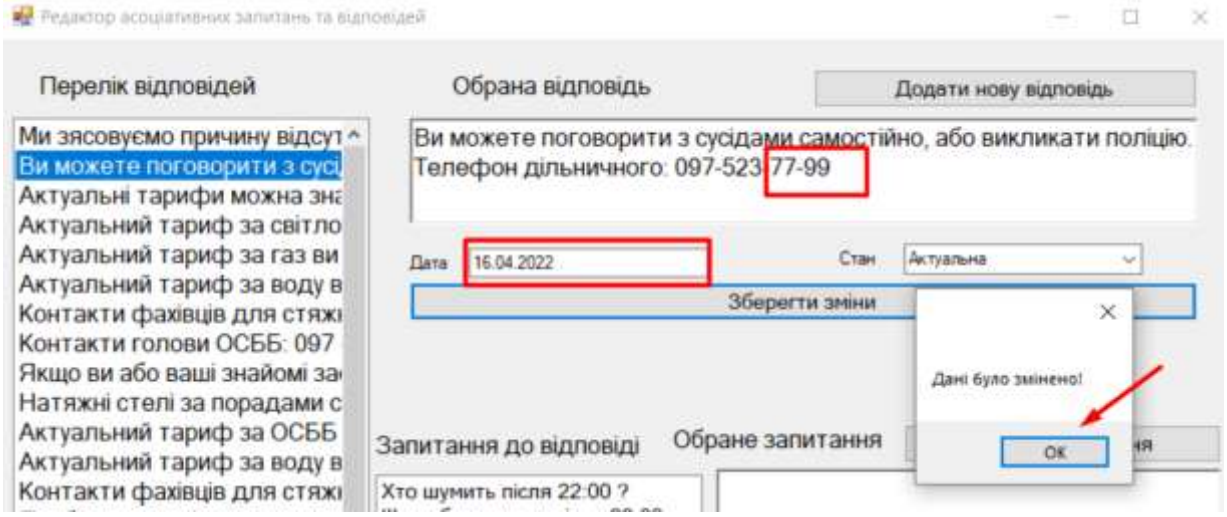


Рисунок 3.4 – Редагування даних асоціативної відповіді

Таким чином, в результаті розробки було реалізовано інформаційну систему на основі методу автоматизованого підбору відповідей на запитання за семантичною подібністю. У подальшому планується удосконалення розробленої системи.

3.3 Тестування інформаційної системи

Для достовірності роботи створеної інформаційної системи згідно з темою КРБ, було проведено такі види тестування: Unit Test, Test Case та функціональне тестування. Першими були зроблені Unit Test-и.

Перший тест показує коректність роботи методу CleaningText(), де на вхід подається текст, у якому потрібно перевести всі символи у нижній регістр та повідкидати розділові знаки. На виході повинна бути множина слів. Код юніт тесту буде таким:

```
[TestMethod()]
 public void CleaningTextTest()
 {
     AnalizeMethod a = new AnalizeMethod();
     var text = "Терміново потрібен сантехнік, надайте, будь ласка, контакти";
     var foreig = new List<string>()
     { "терміново", "потрібен","сантехнік", "надайте", "будь", "ласка", "контакти"  };

     var wordsrez = a.CleaningText(text);
```

```
            Assert.IsTrue(foreig.SequenceEqual(wordsrez));
        }
```

Результат успішного виконання юніт-тесту проілюстровано на рисунку 3.5.

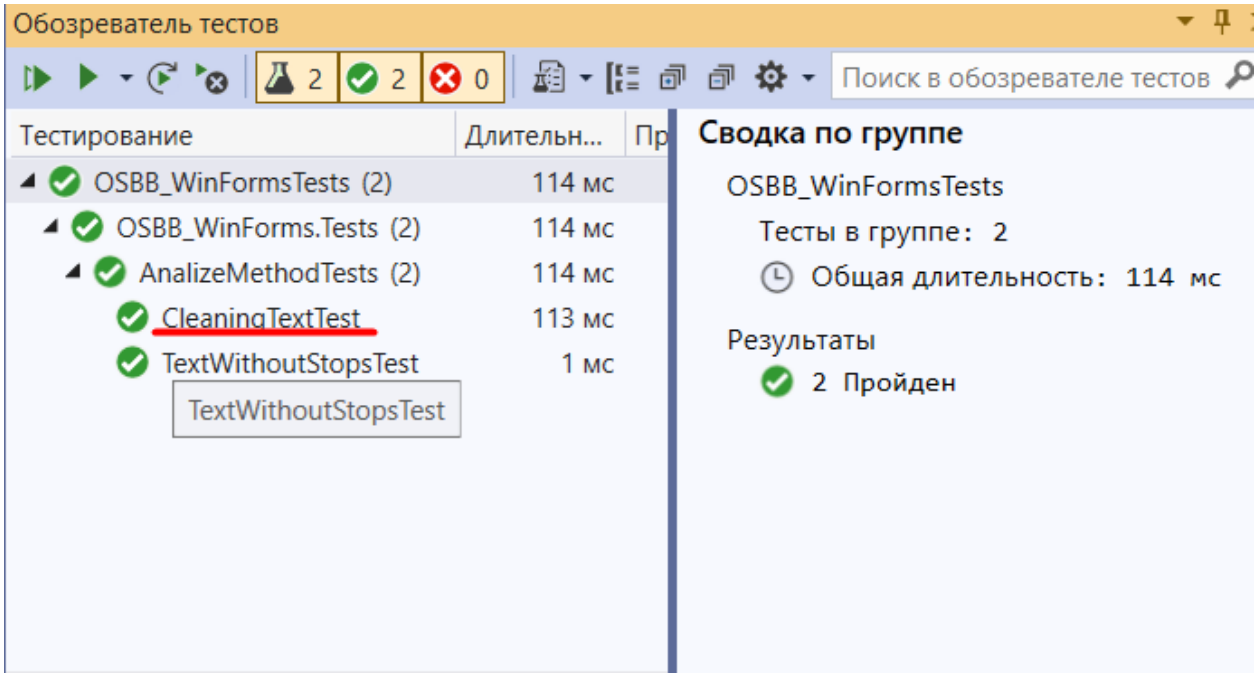


Рисунок 3.5 – Проходження юніт-тесту для методу CleaningText()

Це підтверджує, що дана система відповідає поставленій задачі та може проходити подальшу перевірку на вказаний функціонал.

Другим юніт-тестом буде досліджено роботу методу TextWithoutStop(). Його код проілюстровано нижче.

```
[TestClass()]
    public class AnalizeMethodTests
    {
        [TestMethod()]
```

```
        public void TextWithoutStopsTest()
        {
            AnalizeMethod a = new AnalizeMethod();
            var words = new List<string>()
            { "газовики", "зовсім","знахабнали", "кабальні",
"тарифи", "на", "газ"  };
            var foreig = new List<string>()
            { "газовики", "зовсім","знахабнали", "кабальні",
"тарифи", "газ"  };

            var wordsrez = a.TextWithoutStops(words);

            Assert.IsTrue(foreig.SequenceEqual(wordsrez));

        }
    }
```

При запуску системи тестування на екрані відображено успішне проходження тесту (рисунок 3.6)

Одразу було накладено функціональну перевірку через графічний інтерфейс користувача. При запуску застосування та запуску потрібної підсистеми аналізу асоціативних запитань та відповідей та натисненні кнопки «Показати вектор ключових слів обраного запитання» результат відображено на рисунку 3.7.

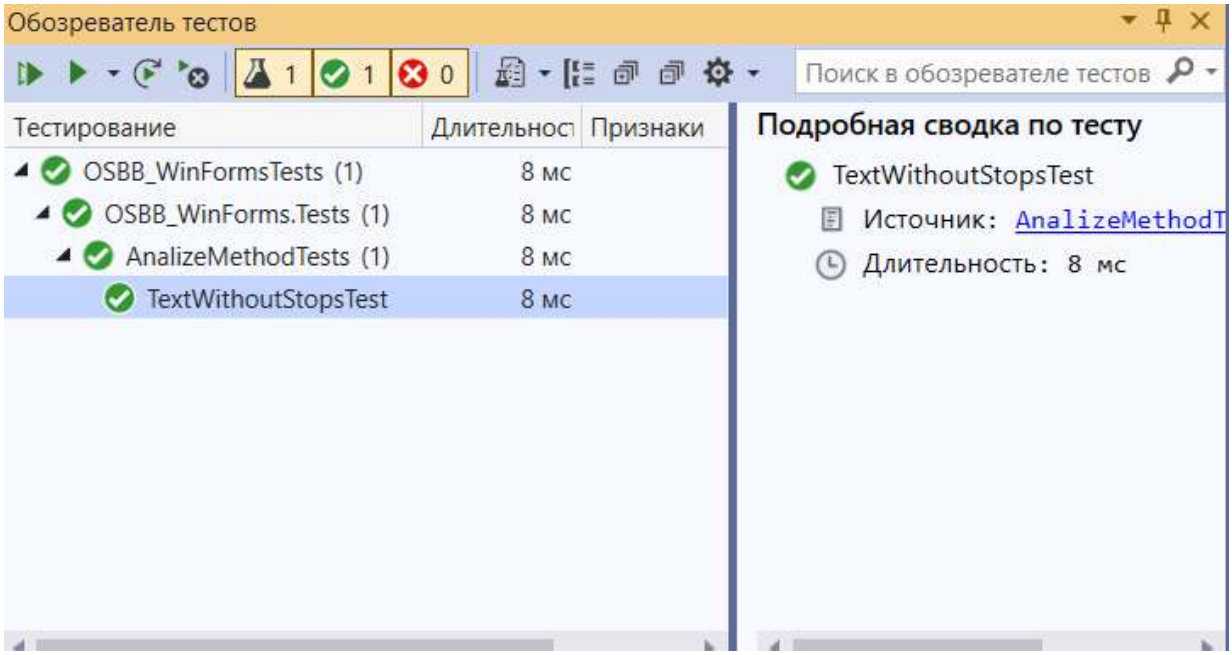


Рисунок 3.6 – Проходження юніт-тесту для методу TextWithoutStopsTest()

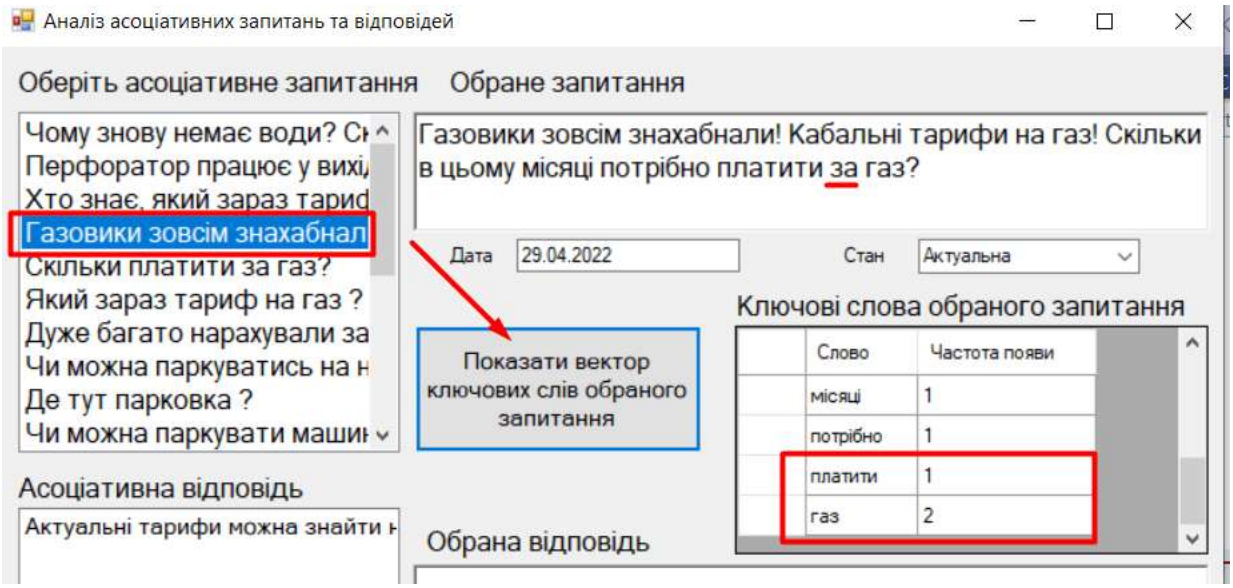


Рисунок 3.7 – Функціональне тестування методу TextWithoutStopsTest()

Також було проведено тестування за допомогою тест-кейсів. Перший тест-кейс перевіряє коректність зміни існуючих асоціативних відповідей та відображений у таблиці 3.1.

Таблиця 3.1 Тест-кейс TS0001

<table>
<tr><td>Тест-кейс ID: TS0001</td><td>Приорітет: 1</td><td>Створено: 1.05.2022, Козенко О.В.</td></tr>
<tr><td colspan="3">Назва: перевірка функціоналу зміни існуючої обраної асоціативної відповіді.
Вхідні дані: Замінити тестову відповідь з «Ви можете поговорити з сусідами самостійно, або викликати поліцію. Телефон дільничного: 097-523-85-66» на «Викликайте поліцію. Телефон дільничного: 097-523-85-66»</td></tr>
<tr><td colspan="2">Кроки</td><td>Очікуваний результат</td></tr>
<tr><td colspan="2">1. Запустити застосування.
2. На головній формі обрати пункт «Редактор асоціативних запитань та відповідей»
3. Обрати у переліку відображених відповідей «Ви можете поговорити з сусідами самостійно, або викликати поліцію. Телефон дільничного: 097-523-85-66»
4. Замінити її на таку «Викликайте поліцію. Телефон дільничного: 097-523-85-66»
5. Натиснути кнопку «Зберегти зміни»</td><td>Запуск за стосунку

Відповідь повинна відобразитись у вікні «Обрана відповідь»

Заміна відповіді на вказану

Відображення повідомлення про успішне внесення змін.</td></tr>
<tr><td colspan="3">Результат виконання тест-кейсу: перевірку пройдено успішно.</td></tr>
</table>

Після запуску програми необхідно виконати вказані у таблиці 3.1 кроки. Після чого у програмі користувач побачить результат про успішні зміни (рисунок 3.8).

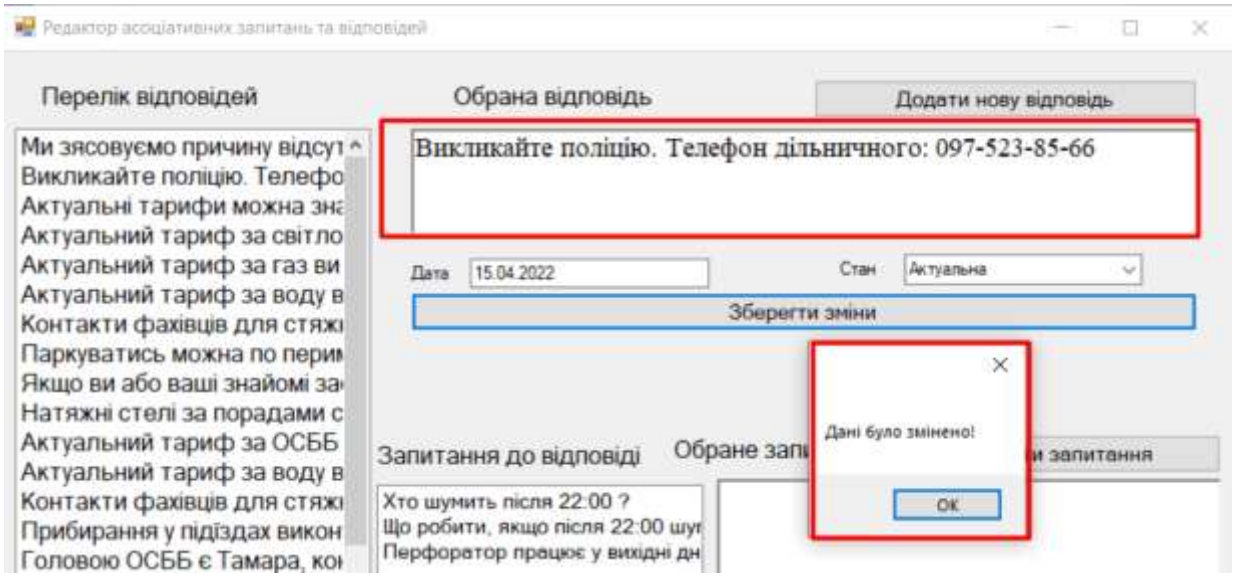


Рисунок 3.8 – Результат тестування збереження змін

При тестуванні даної програми некоректно працюючих функцій виявлено не було. У результаті проведеного тестування можна зробити висновок, що програмна реалізація працює коректно згідно поставленої задачі.

3.4 Інструкція користувача

При запуску розробленого застосування користувачу відображається стартова форма, з якої можна перейти на підсистеми, описані в 2.1. Вигляд стартової форми зображено на рисунку 3.9.

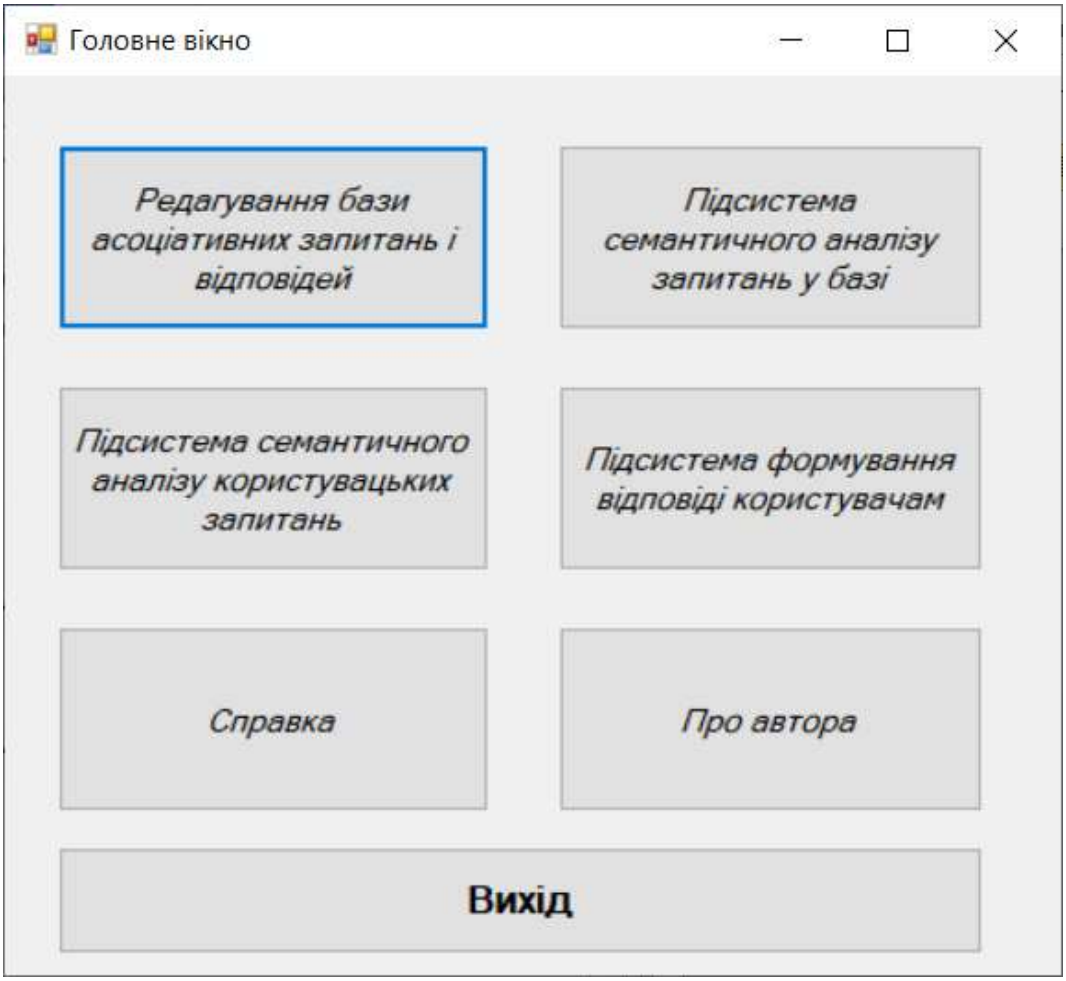


Рисунок 3.9 – Стартовий екран програми

З головного вікна користувач може перейти на одну з реалізованих підсистем. Для переходу до підсистеми «Редагування бази асоціативних запитань і відповідей» користувачу потрібно натиснути на однойменну кнопку. Після переходу відкриється відповідна форма (рисунок 3.10), з якої стане доступною редагування запитань та відповідей.

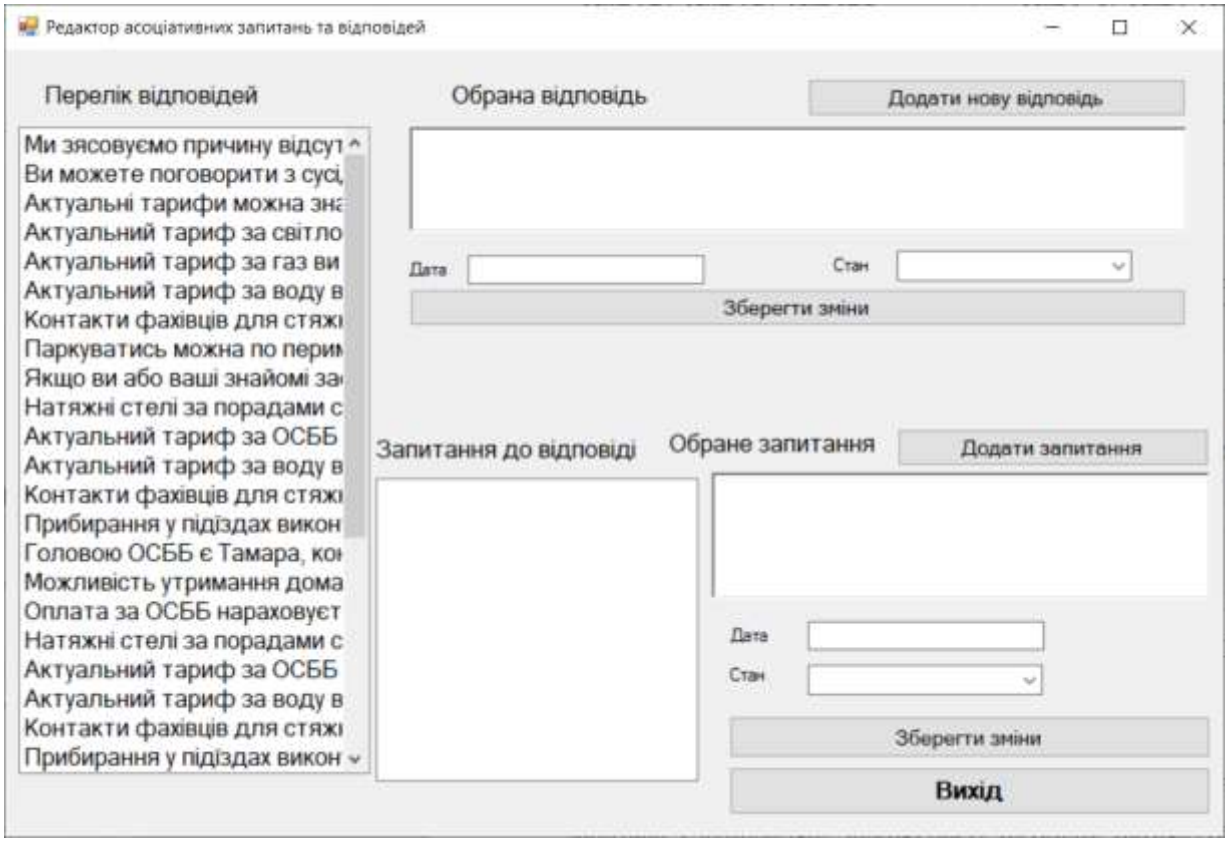


Рисунок 3.10 – Екран «Редагування бази асоціативних запитань і відповідей»

Користувач відповідь яка його цікавить з переліку, і відповідна відповідь відобразиться у вікні «Обрана відповідь». Також відобразяться автоматично відповідні їй запитання у вікні «Запитання до відповіді» (рисунок 3.11).

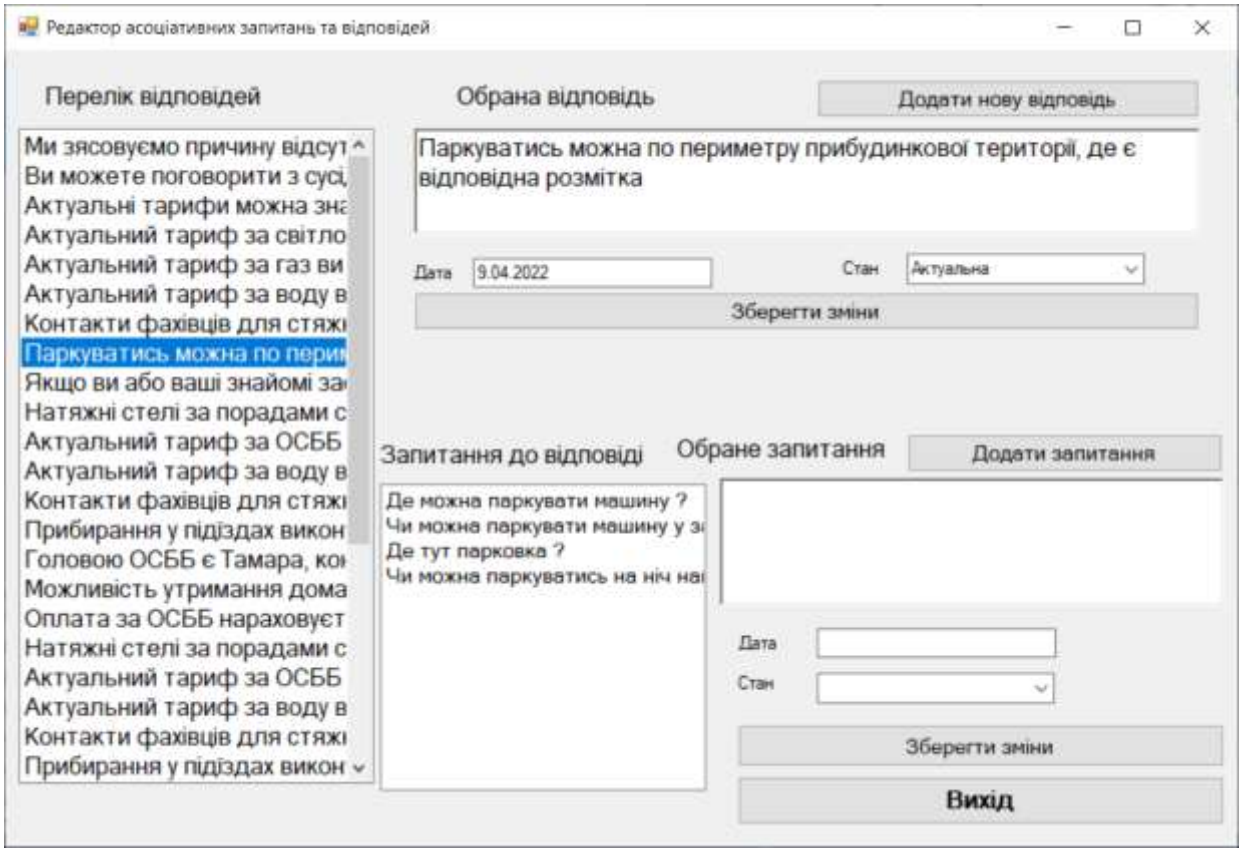


Рисунок 3.11 – Екран «Редагування бази асоціативних запитань і відповідей». Заповнення запитань

Натиснувши на запитання, яке цікавить користувача у вікні «Обране запитання» відобразиться відповідь на нього (рисунок 3.12).

Також для цієї форми доступно функціонал додавання та зміни запитань та відповідей. Якщо внести зміни у відповідну обрану відповідь, відповідні зміни будуть внесені до бази даних та користувач побачить повідомлення про те що дані було змінено (рисунок 3.13).

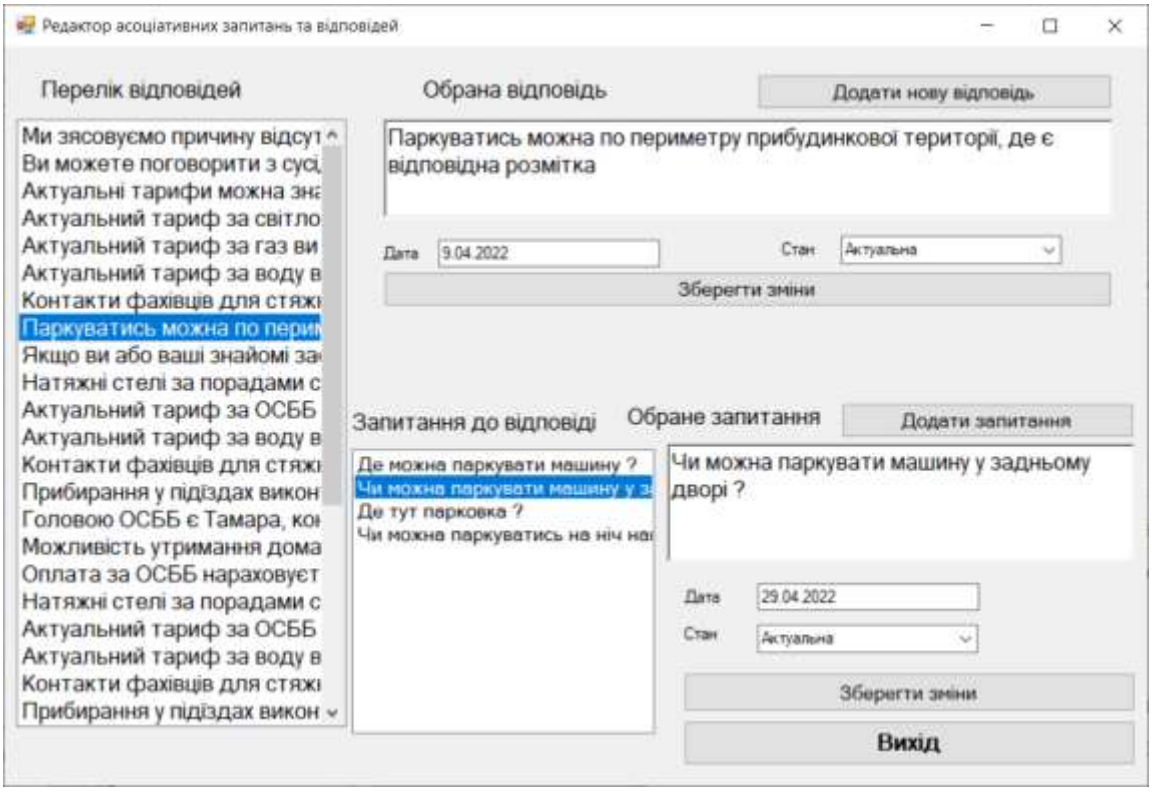


Рисунок 3.12 – Екран «Редагування бази асоціативних запитань і відповідей». Обране запитання

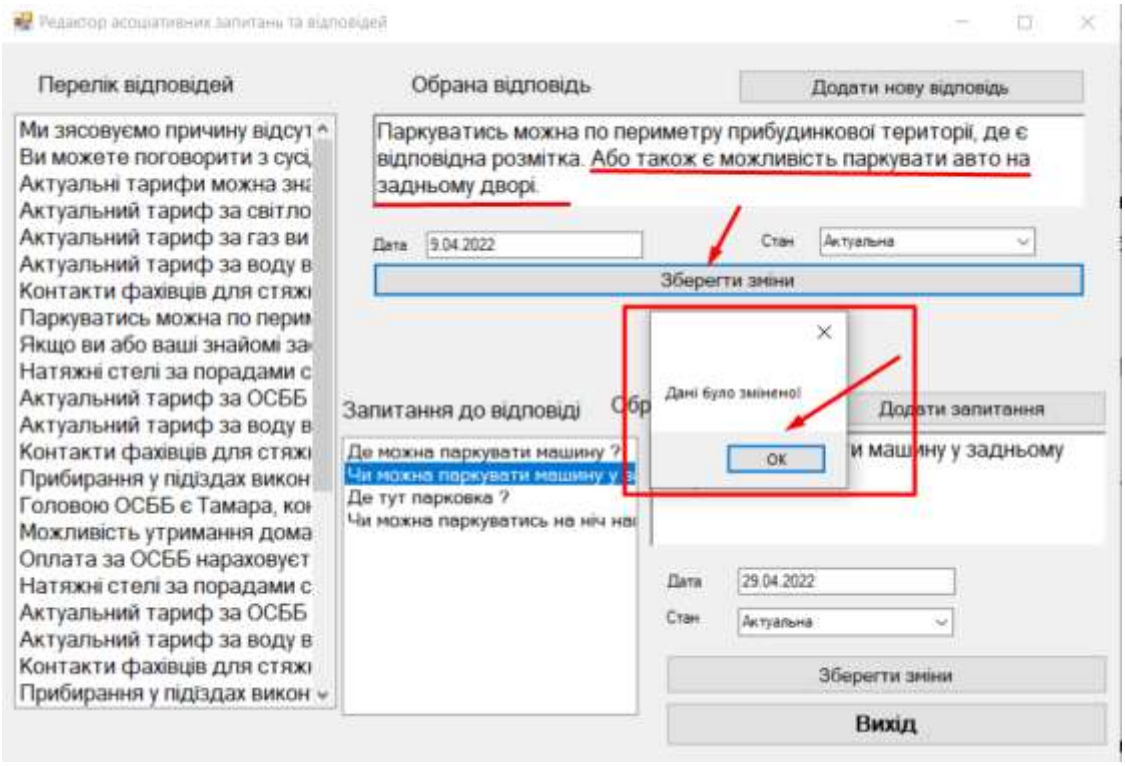


Рисунок 3.13 – Екран «Редагування бази асоціативних запитань і відповідей». Редагування відповіді

Аналогічним чином відбувається зміна у базі відповідних запитань. Також є можливість додати нове запитання до відповідної відповіді. Для цього потрібно ввести запитання та натиснути на кнопку «Додати запитання» (рисунок 3.14).

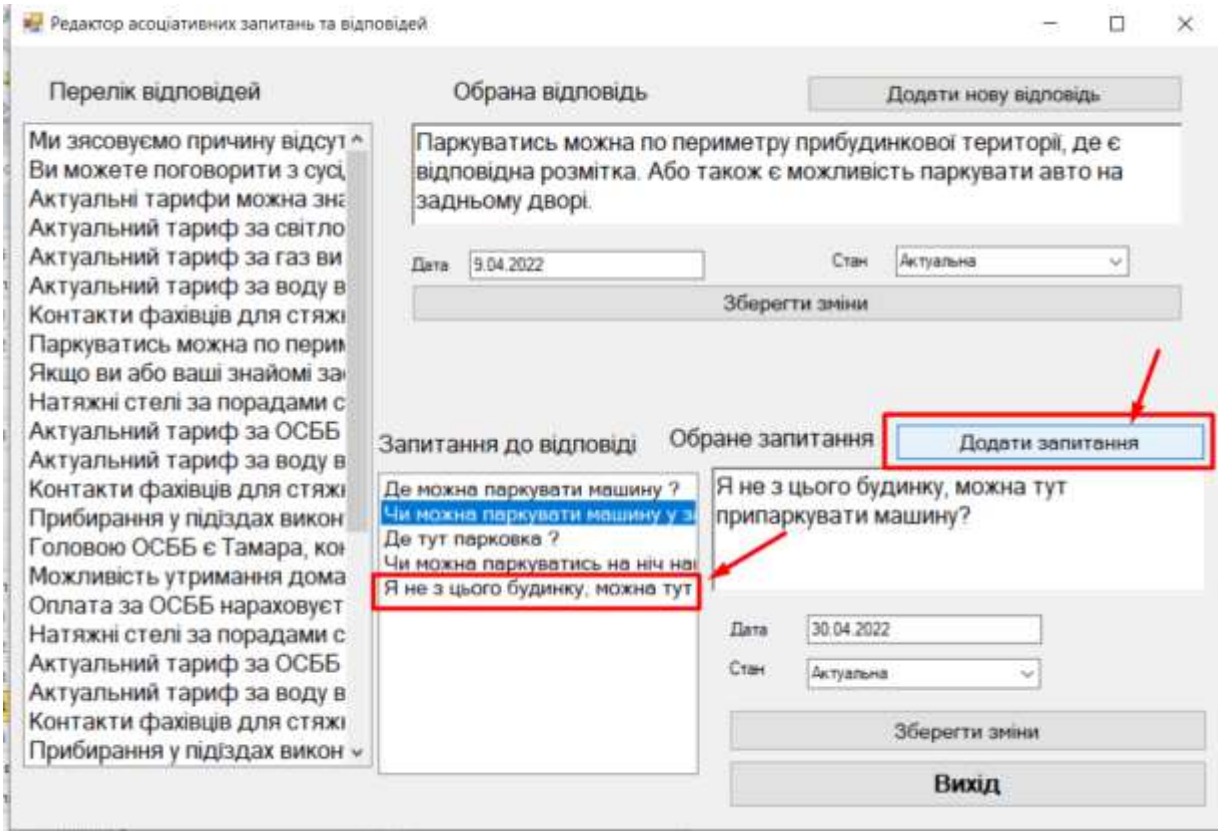


Рисунок 3.14 – Екран «Редагування бази асоціативних запитань і відповідей». Додавання нового запитання

Також на цій формі є кнопка виходу, після натиснення на яку відбудеться вихід з програми.

Натиснувши на головному екрані кнопку «Про автора» буде відображено сторінку з відомостями про автора (рисунок 3.15).

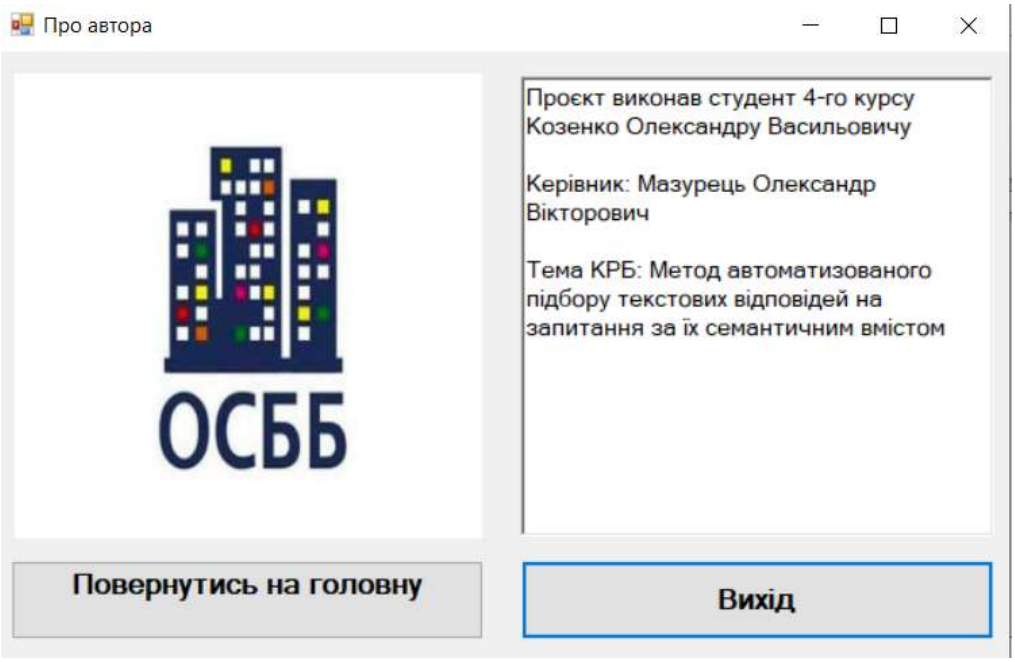


Рисунок 3.15 – Екран «Про автора»

Натиснувши на кнопку «Підсистема семантичного аналізу» користувачу відкриється форма для проведення семантичного аналізу запитань (рисунок 3.16).

Тут як і на попередній формі необхідно обрати запитання з переліку для проведення аналізу, після чого обране запитання деталізується та відобразиться у полі «Обране запитання». Для знаходження ключових слів та обрахунку їх частоти і дисперсійної оцінки потрібно натиснути на кнопку «Показати вектор ключових слів обраного запитання» (рисунок 3.17)

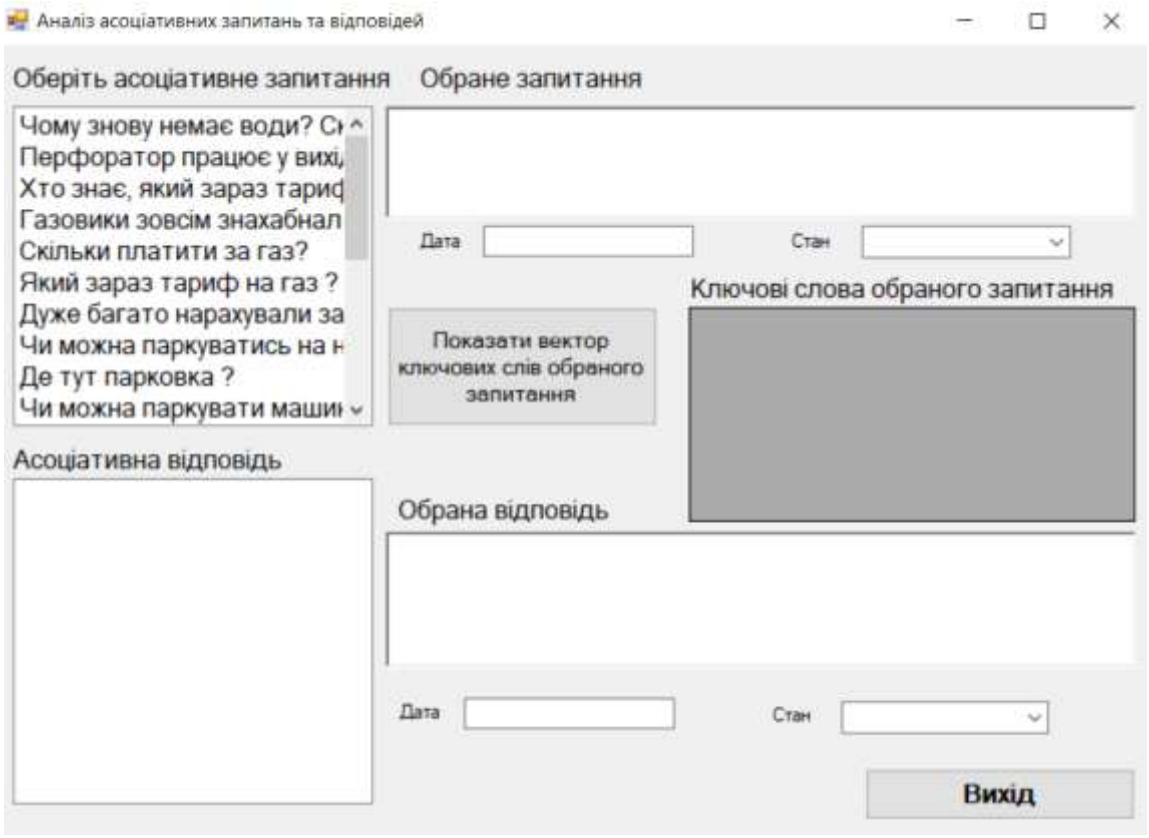


Рисунок 3.16 – Екран «Аналіз асоціативних запитань та відповідей»

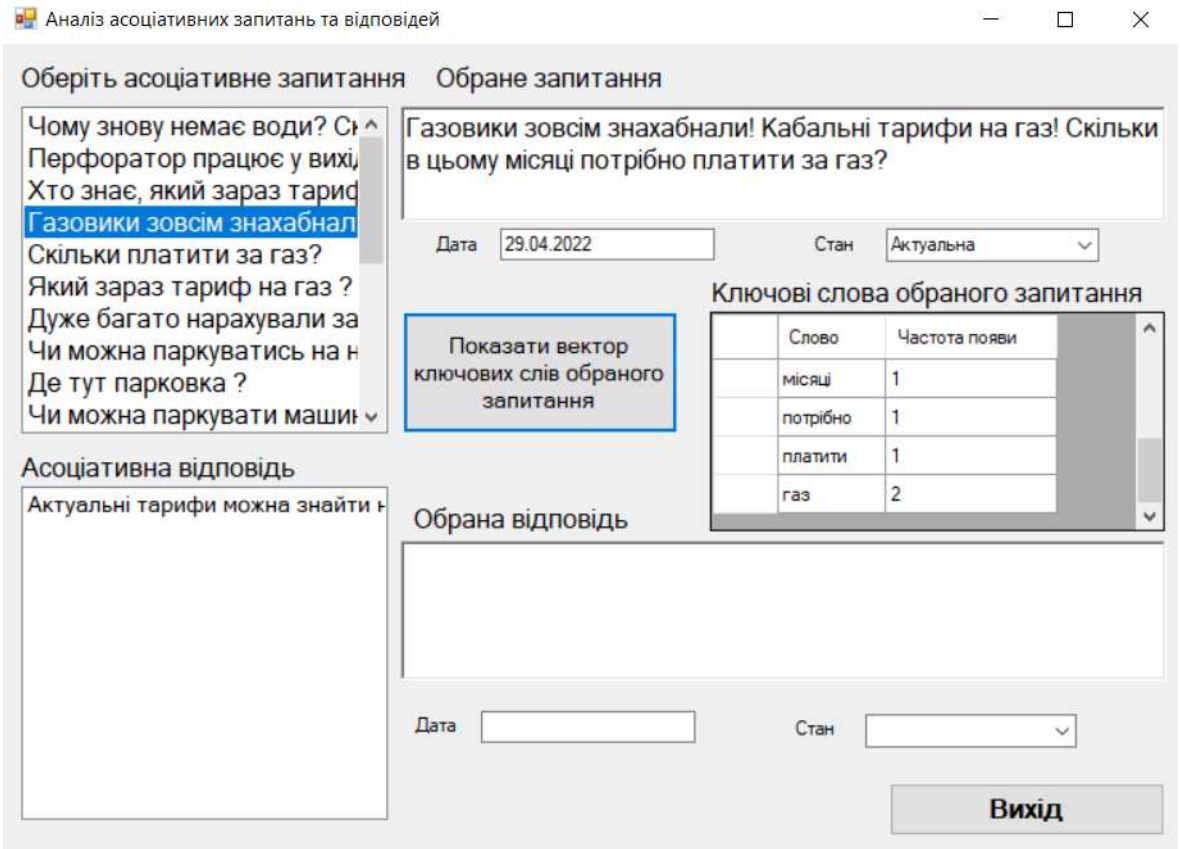


Рисунок 3.17 – Екран «Аналіз асоціативних запитань та відповідей»

Для відображення асоціативної відповіді потрібно натиснути на неї і вона буде відображена разом з актуальними для неї датою та станом (рисунок 3.18).

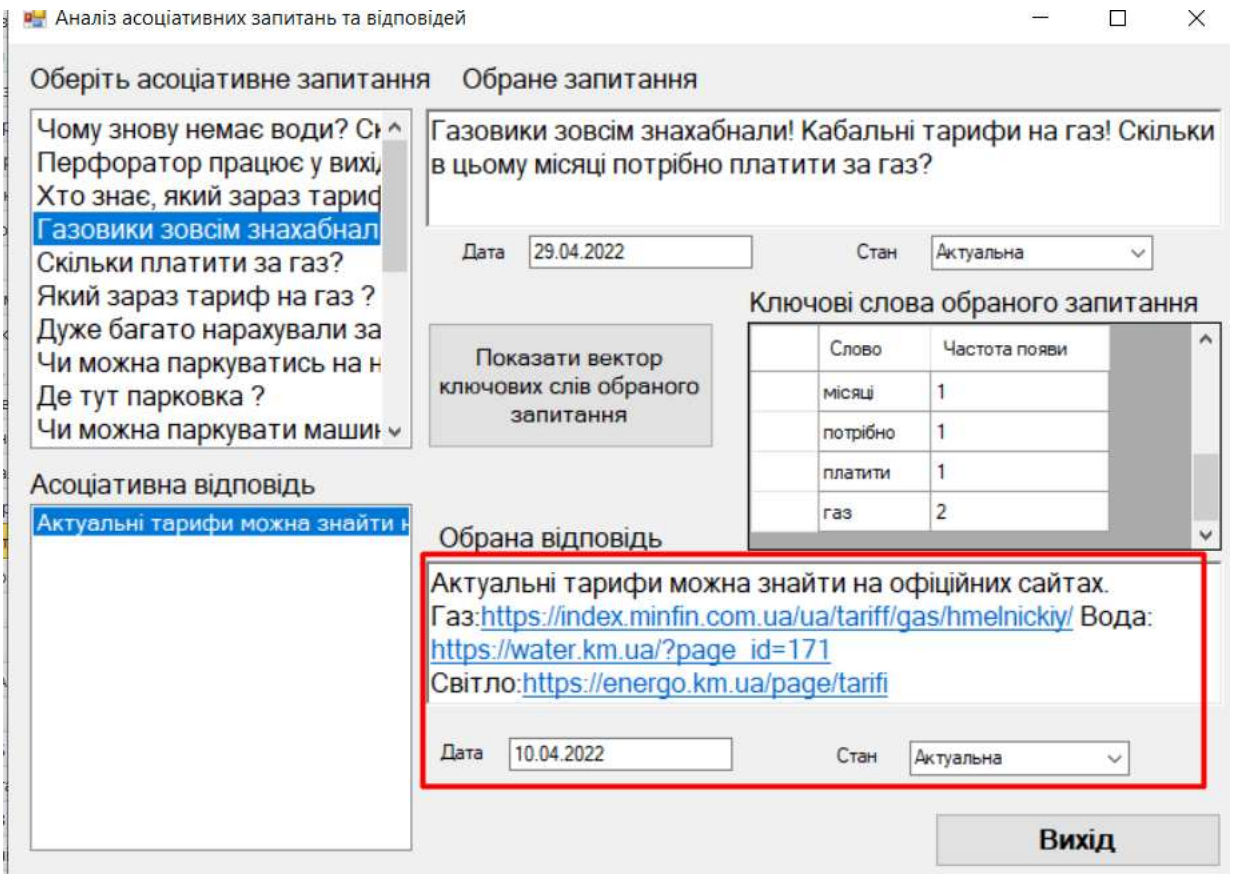


Рисунок 3.18 – Екран «Аналіз асоціативних запитань та відповідей».

Асоціативна відповідь

При натисненні на кнопку «Підсистема формування відповідей користувача» користувач перейде до форми отримання відповіді на введене питання. Питання вводиться у поле «Введіть ваше запитання» (рисунок 3.19).

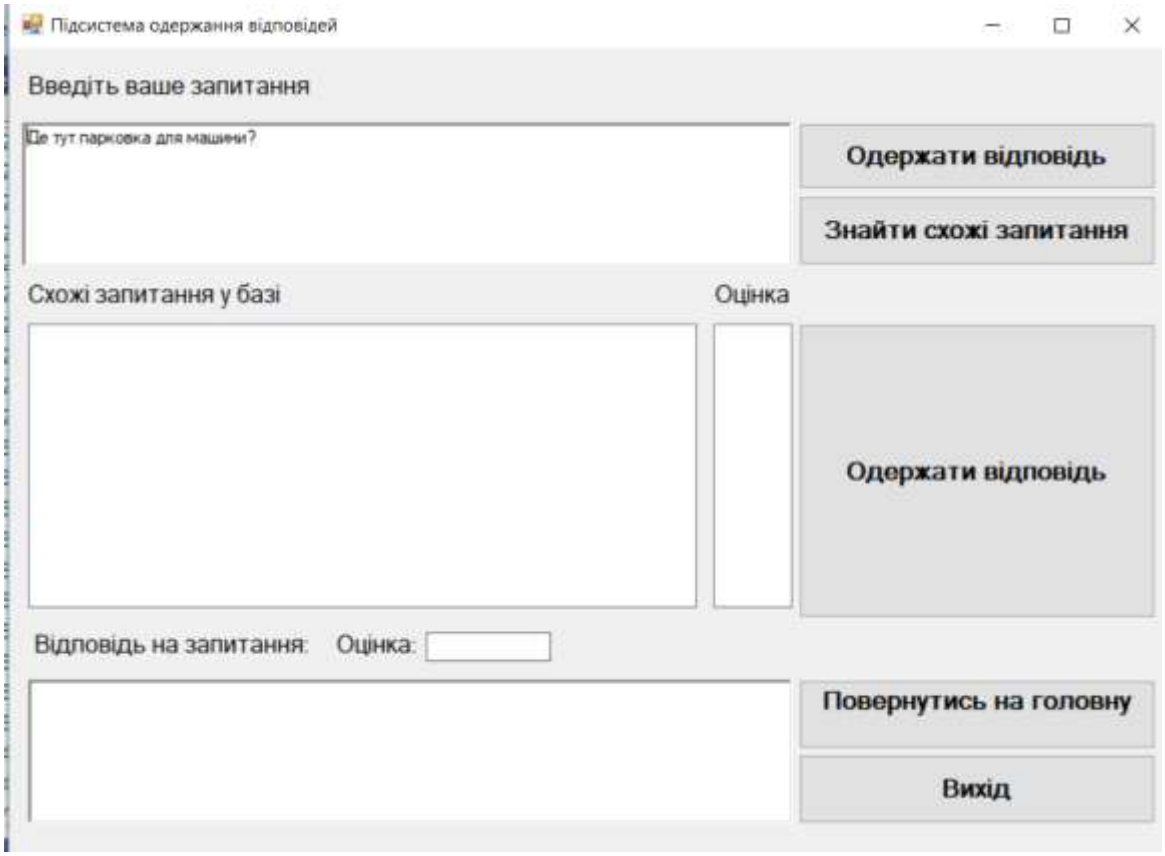


Рисунок 3.19 – Екран «Підсистема одержання відповідей»

Можна одержати близькі запитання до введеного, натиснувши на кнопку «Знайти схожі запитання». Натиснувши на кнопку у полі «Схожі запитання у базі» та «Оцінки» відобразиться перелік схожих запитань з відповідними оцінками (рисунок 3.20).

Натиснувши на кнопку «Одержати відповідь» користувачу буде відображено найвірогіднішу відповідь на поставлене запитання за дисперсною оцінкою (рисунок 3.21).

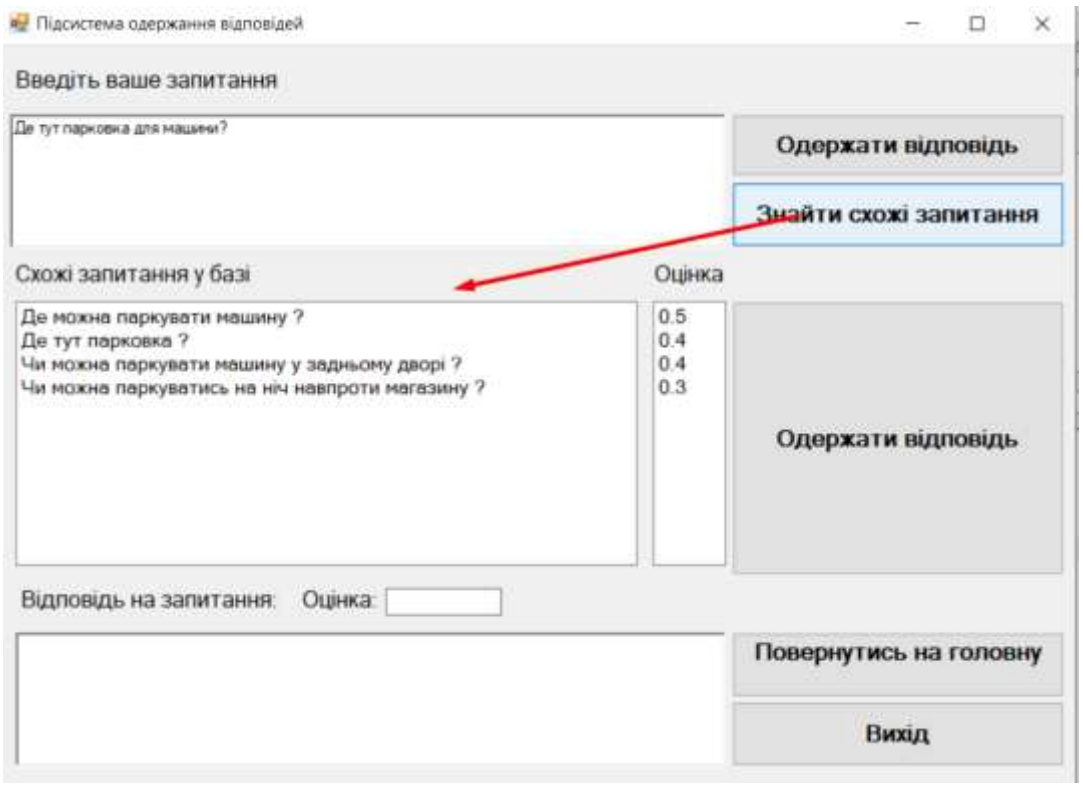


Рисунок 3.20 – Екран «Підсистема одержання відповідей». Відображення схожих запитань

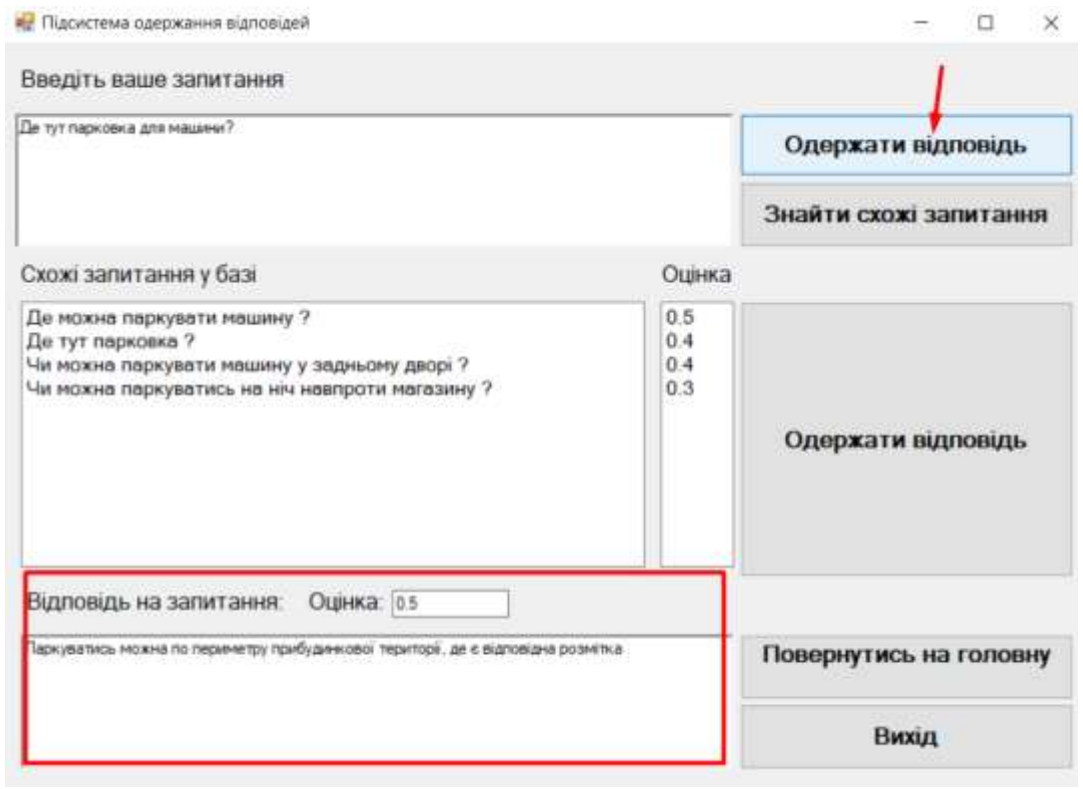


Рисунок 3.21 – Екран «Підсистема одержання відповідей». Одержати відповідь

Також на формі є кнопки «Повернутись на головну» та «Вихід», які відповідно переходять на стартову форму та виходять з програми. У даній програмній реалізації виконані всі поставлені задачі.

3.5 Вимоги до розгортання інформаційної системи

Мінімальними апаратними вимогами для використання розробленої програмної реалізації є:

- частота процесора 800 МГц;
- 3 Гб оперативної пам'яті;
- 1000 Мб вільного місця на жорсткому диску;
- відео-карта;
- мишка;
- клавіатура;
- монітор.

Вимоги до програмного забезпечення:

- Операційна система Windows 2007 та вище;
- .NET Framework.

Висновки

У рамках виконання кваліфікаційної роботи бакалавра було виконано розробку й апробацію методу автоматизованого підбору відповідей на запитання за семантичною подібністю. Зокрема, було проведено аналіз предметної області й досліджено сучасні підходи до автоматизованого пошуку ключових слів, розглянуто існуючі програмні реалізації за цим напрямком.

Було спроектовано інформаційну систему та створено і описано відповідний метод автоматизованого підбору відповідей на запитання за семантичною подібністю. У якості засобів розробки було обрано фреймворк .NET Framework, мову програмування C#, редактор програмного коду VisualStudio та СКБД MS SQL Server. Для зручності користування застосунком було створено інструкцію користувача та описані мінімальні апаратні вимоги до розгортання інформаційної системи.

Для забезпечення коректної роботи програмної реалізації методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом, було створено базу даних, що дозволяє, забезпечуючи зв’язки між таблицями, зберігати інформацію: тексти запитань користувачів, відповіді на поширені запитання та асоціативні запитання. Реалізована база даних забезпечує збереження даних та доступ до них.

Розроблена програмна реалізація методу автоматизованого підбору відповідей на запитання за семантичною подібністю на платформі .NET виконує наступні основні функції:

1. Семантичний аналіз наявних запитань і тестового користувацького запитання.

2. Обрахунок оцінок семантичної подібності користувацького запитання до кожного з наявних у базі запитань з асоційованими відповідями.

3. Знаходження відповіді у базі на користувацьке запитання.

При цьому забезпечується виконання наступних дій:

– первинна обробка тексту кожного запитання (розділові знаки, регістр тощо);

– формування вектора слів запитання – впорядкованої множиною слів;

– формування множини оригінальних слів за впорядкованою множини слів;

– обрахунок семантичної ваги кожного слова у множині оригінальних слів за методом дисперсного оцінювання;

– пошук однакових оригінальних слів у користувацькому запитанні й кожному з наявних запитань з асоційованими відповідями у базі запитань;

– обрахунок оцінок семантичної подібності користувацького запитання до кожного з наявних у базі запитань з асоційованими відповідями;

– знаходження наявного запитання у базі, що має максимальну оцінку семантичної подібності до користувацького запитання (робоче запитання);

– знаходження відповіді у базі, яка асоційована з робочим запитанням;

– видача повідомлення користувачу, якщо не вдалося знайти відповіді.

Даного базового функціоналу виявилось достатньо для прикладного тестування методу автоматизованого підбору відповідей на запитання за семантичною подібністю; функціональність застосунку може бути розширена у подальших дослідженнях.

Перелік посилань

1. Habr. Основи Natural Language Processing для тексту. URL: https://habr.com/ru/company/Voximplant/blog/446738/

2. Що таке обробка природної мови? URL: https://forklog.com/chto-takoe-obrabotka-estestvennogo-yazyka/

3. Що таке чат боти та кому вони потрібні? URL: https://forklog.com/chto-takoe-obrabotka-estestvennogo-yazyka/

4. Skilled Interpersonal Communication URL: https://www.taylorfrancis.com/books/mono/10.4324/9781003182269/skilled-interpersonal-communication-owen-hargie

5. Natural Language Processing (NLP). URL: https://www.ibm.com/cloud/learn/natural-language-processing

6. Голосовий помічник Cortana та його особливості. URL: https://mentamore.com/covremennye-texnologii/golosovoj-pomoshhnik-cortana.html

7. Computatioonal linguistics and artificial intelligence. URL: https://molodyivchenyi.ua/index.php/journal/article/view/390/379

8. Соцмережі-2021. URL: https://hromadske.ua/posts/socmerezhi-2021-tiktok-starshaye-facebook-perevazhno-zhinochij-a-strichku-mi-gortayemo-400-miljoniv-rokiv

9. Відповіді на найбільш частіші питання. URL: https://osbb.work/documentations/faq

10. ОСББ Кондратюка 3. URL: https://www.osbbkondratuka3.com/типові-питання-про-осбб

11. Збільшення ціни на газ. URL: https://infobox-forum.prozorro.org/d/558-zbilshennya-tsini-na-gaz

12. Семантичний аналіз. URL: https://cropas.by/seo-slovar/semanticheskij-analiz/

13. Ключові слова. URL: https://igroup.com.ua/seo-articles/keywords/

14. Автоматичний пошук ключових слів у корпусі масової літератури URL: http://aprus.khpi.edu.ua/article/view/2227-6890.2018.4.13

15. Методи та алгоритми вилучення ключових слів. URL: https://cyberleninka.ru/article/n/metody-i-algoritmy-izvlecheniya-klyuchevyh-slov

16. Корпусна та когнітивна лінгвістика. URL: https://www.ulif.org.ua/system/files/ling_inf_studio_tom_4_umif_b5.pdf

17. Automatic Keyword Extraction on Twitter. URL: https://aclanthology.org/P15-2105.pdf

18. Методи лінгвістичних досліджень. URL: http://discourse.com.ua/lekcii/metodi-lingvistichnih-doslidzhen/

19. Semantic search: Issues and technologies. URL: https://cyberleninka.ru/article/n/semanticheskiy-poisk-problemy-i-tehnologii/viewer

20. Text Document Clustering: Wordnet vs. TF-IDF vs. Word Embeddings. URL: https://aclanthology.org/2021.gwc-1.24.pdf

21. Advanced Informatics for Computing Research. URL: https://books.google.com.ua/books?id=5j18DwAAQBAJ

22. What is Google Keyword Planner Used For? URL: https://raddinteractive.com/what-is-google-keyword-planner-used-for/

23. How to Use Google Ads' Keyword Planner (2022 Edition). URL: https://compose.ly/content-strategy/use-googles-adwords-keyword-planner-free

24. 25 кращих безкоштовних інструментів SEO (випробуваних та протестованих). URL: https://evo.business/25-luchshix-besplatnyx-instrumentov-seo/

25. Answer the public. URL: https://kparser.com/answer-the-public-alternative/

26. Backlink Checker. URL: https://ahrefs.com/ru/backlink-checker

27. What is custom text classification (preview)? URL: https://docs.microsoft.com/en-us/azure/cognitive-services/language-service/custom-classification/overview

28. Welcome to Language Studio. URL: https://language.cognitive.azure.com/

29. Розробка віконних додатків C#. URL: https://ci-sharp.ru/ukr/Teaching/razrabotka-okonnih-prilozheniy-C.html

30. TIOBE Index for May 2022. URL: https://www.tiobe.com/tiobe-index/

31. James Gosling; Bill Joy, Guy Steele, Gilad Bracha (2005). The Java Language Specification, Third Edition. Addison-Wesley. ISBN 0-321-24678-0.

32. What is Java Used For? URL: https://www.aternity.com/blogs/what-is-java-used-for/

33. All About SQL Server: Advantages, Best Practices, and Tools. URL: https://www.tek-tools.com/database/sql-server-best-practices-and-tools

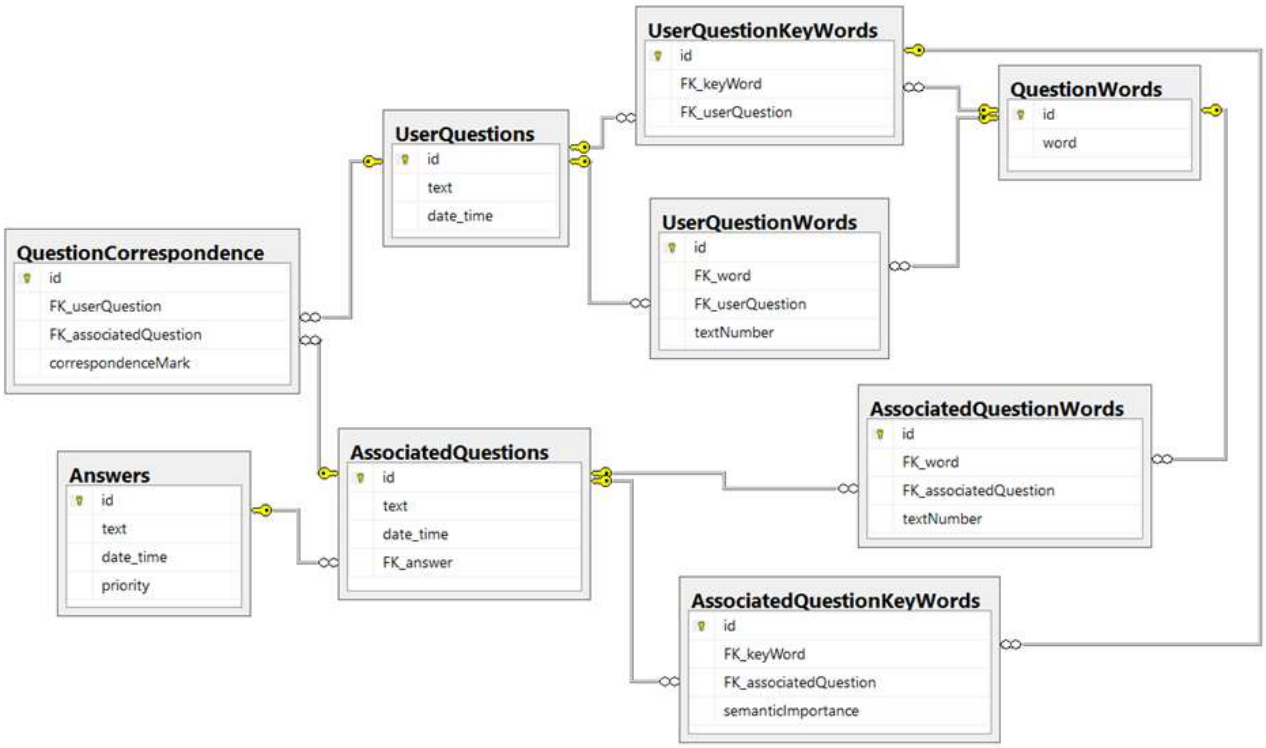
34. Огляд .NET Framework. URL: http://www.williamspublishing.com/PDF/978-5-6040043-7-1/part.pdf

35. Features of the code editor. URL: https://docs.microsoft.com/en-us/visualstudio/ide/writing-code-in-the-code-and-text-editor?view=vs-2022

ДОДАТКИ

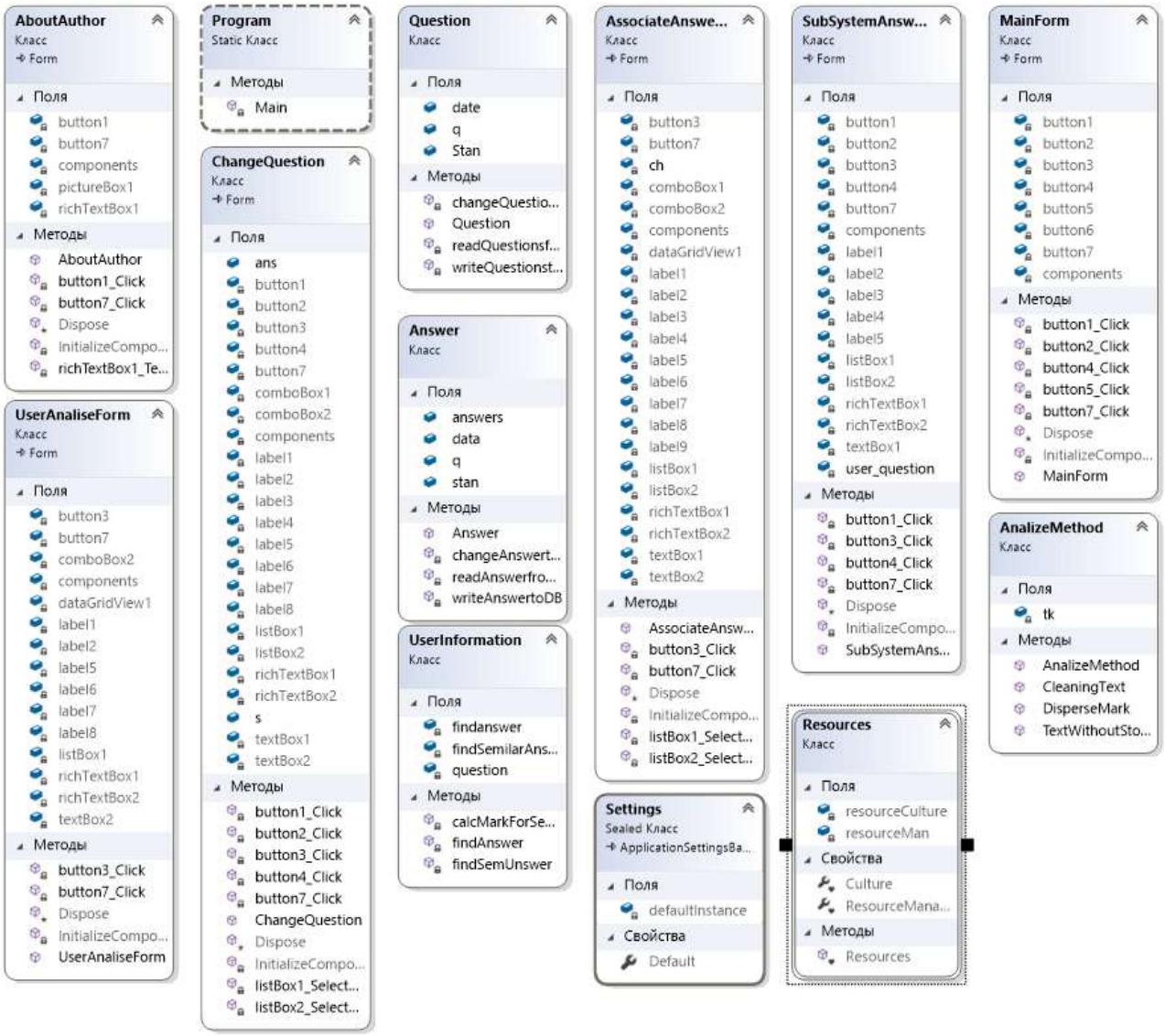
Додаток А

Структура бази даних для прикладної реалізації методу автоматизованого підбору відповідей за семантичною подібністю



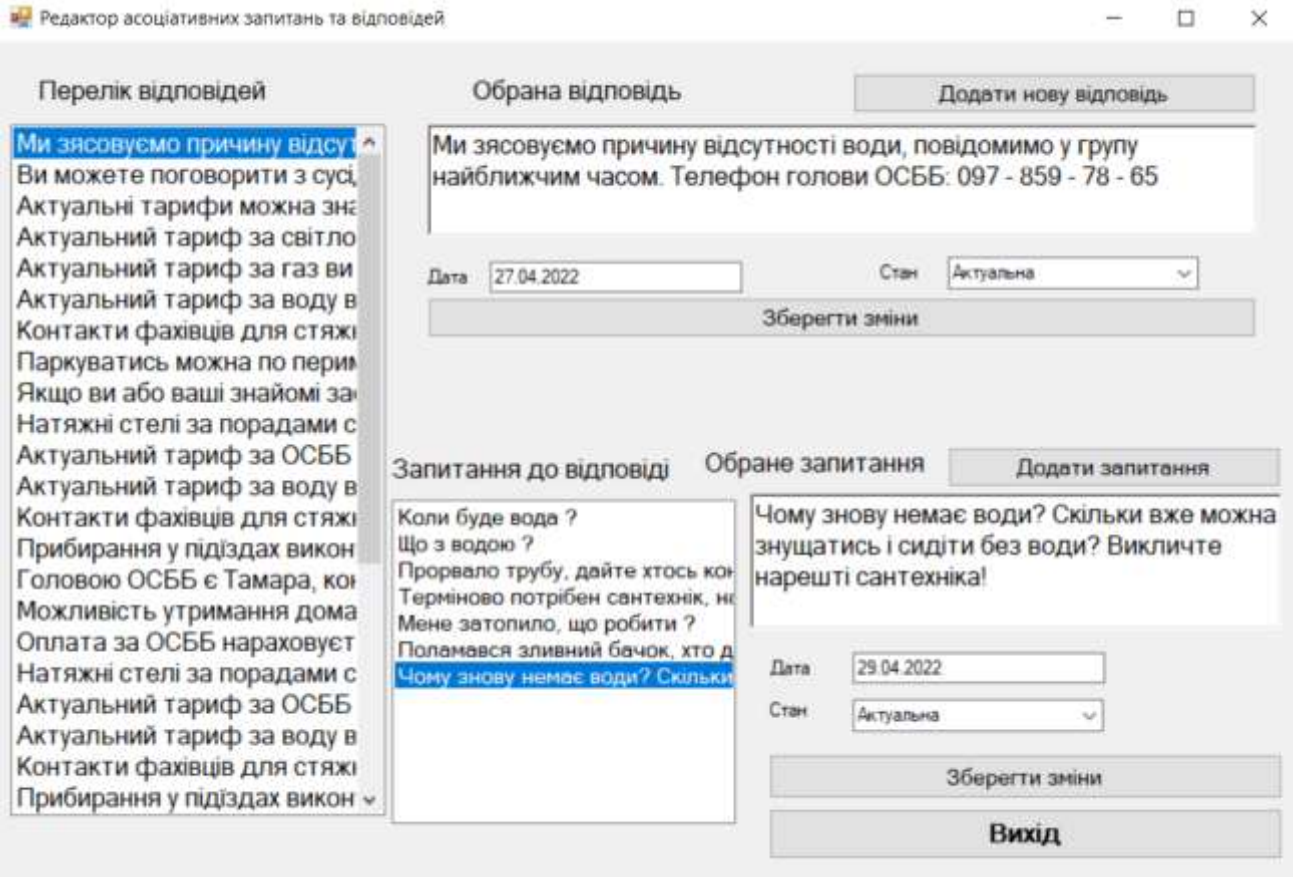
Додаток Б

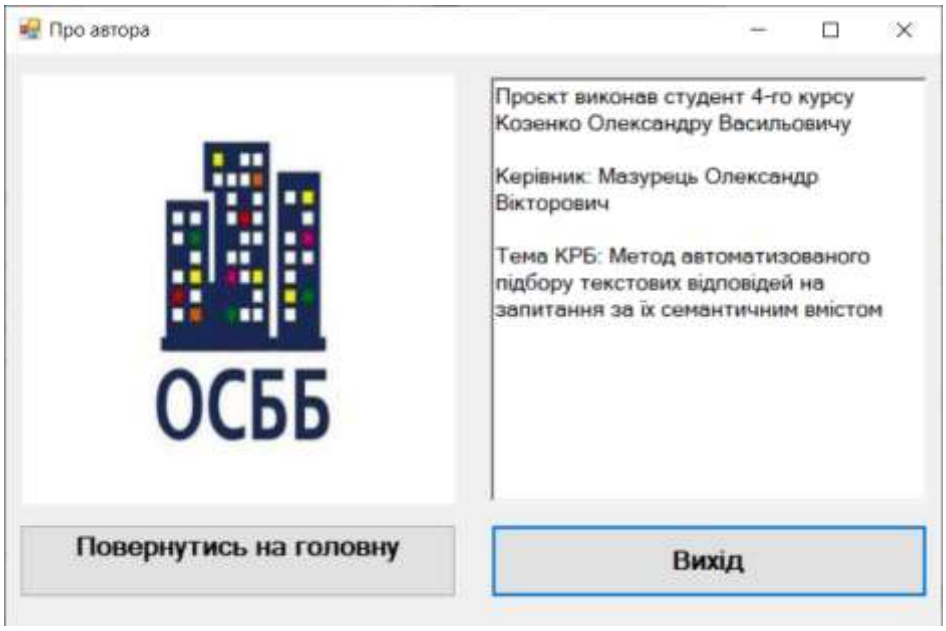
Розгорнута структура класів інформаційної системи автоматизованого підбору відповідей за семантичною подібністю

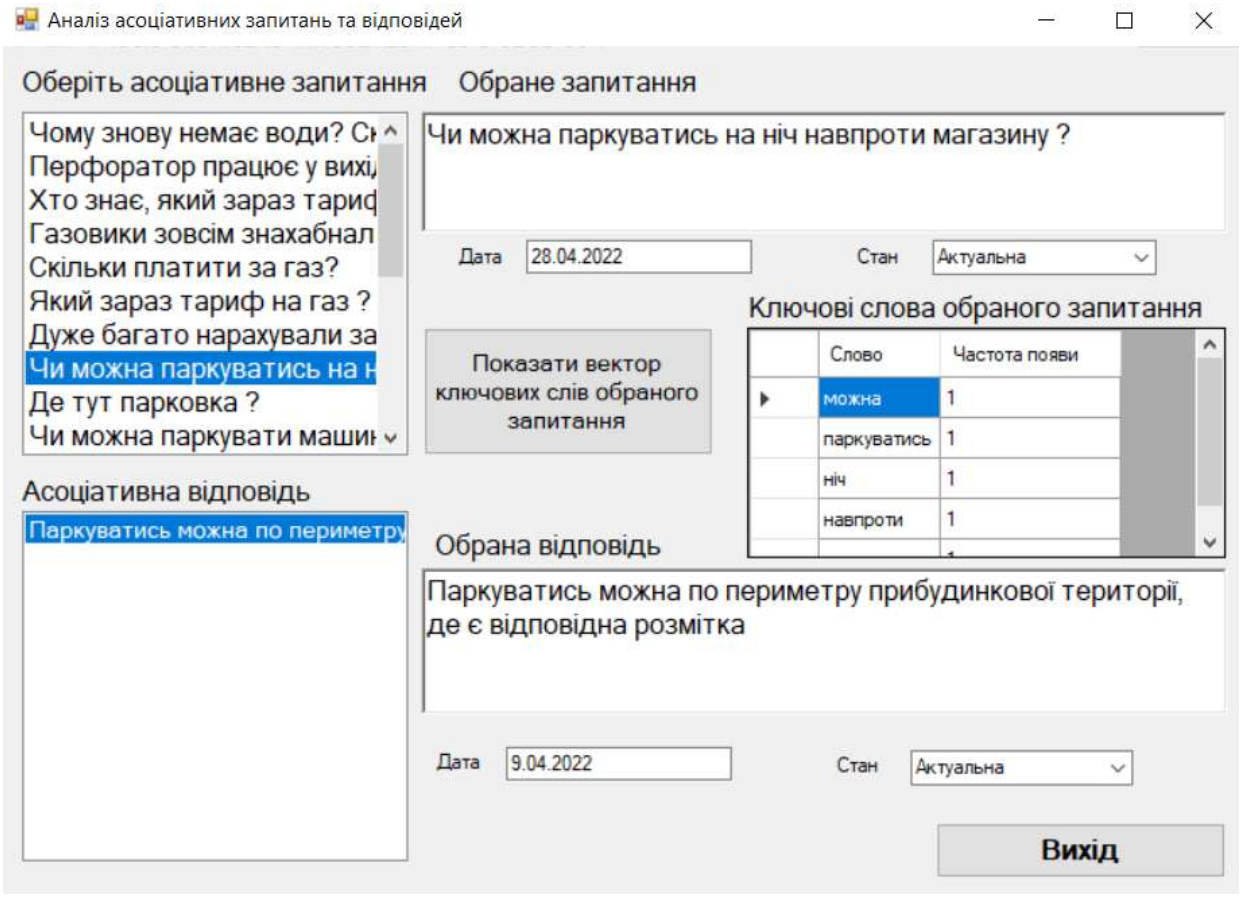


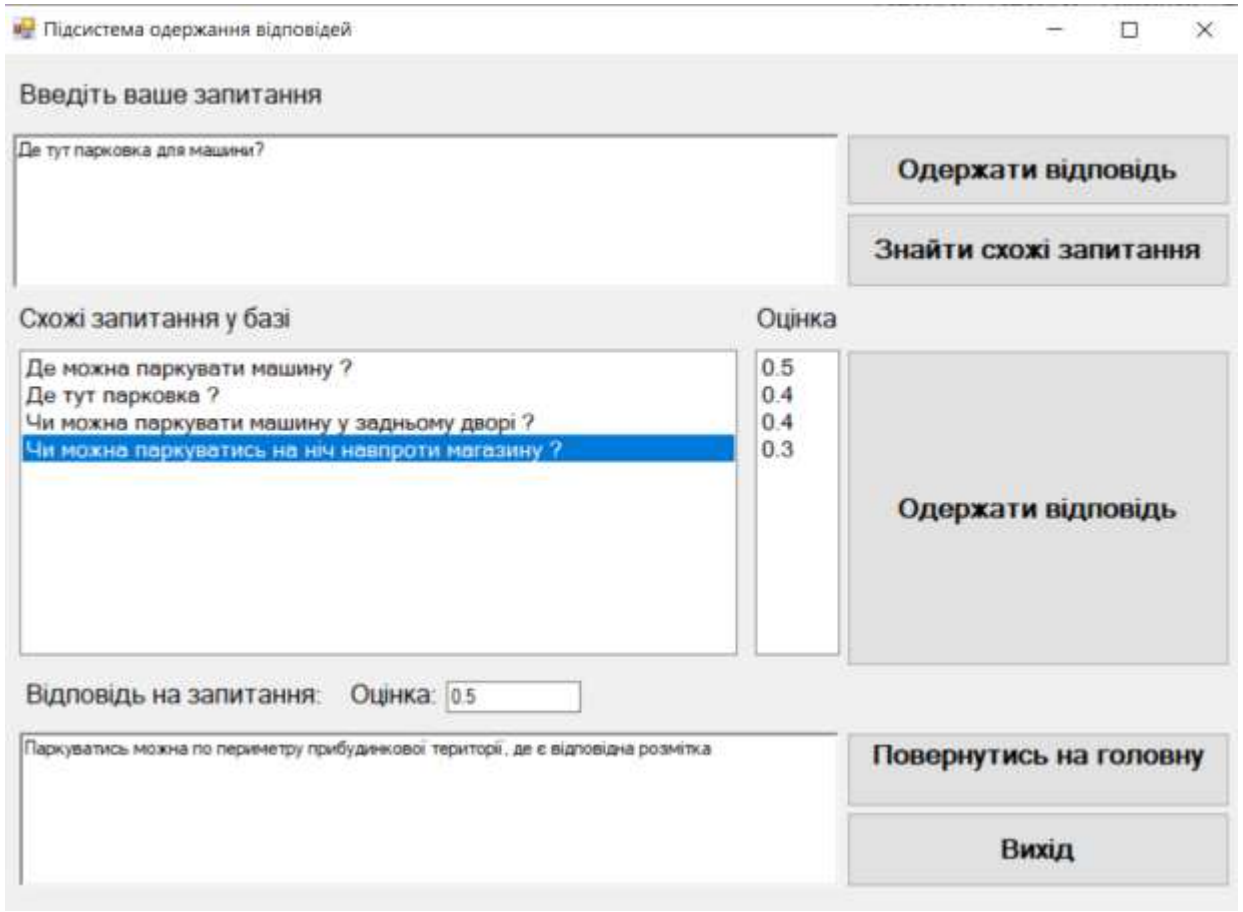
Додаток В

Світлини екрану з результатами роботи інформаційної системи автоматизованого підбору відповідей за семантичною подібністю






Аналіз асоціативних запитань та відповідей
Оберіть асоціативне запитання
Обране запитання
Чому знову немає води? Сн
Перфоратор працює у вихі
Хто знає, який зараз тариф
Газовики зовсім знахабнал
Скільки платити за газ?
Який зараз тариф на газ ?
Дуже багато нарахували за
Чи можна паркуватись на н
Де тут парковка ?
Чи можна паркувати машин
Чи можна паркуватись на ніч навпроти магазину ?
Дата 28.04.2022
Стан Актуальна
Ключові слова обраного запитання
Показати вектор ключових слів обраного запитання
Слово
Частота появи
можна 1
паркуватись 1
ніч 1
навпроти 1
Асоціативна відповідь
Паркуватись можна по периметру
Обрана відповідь
Паркуватись можна по периметру прибудинкової території, де є відповідна розмітка
Дата 9.04.2022
Стан Актуальна
Вихід


Підсистема одержання відповідей
Введіть ваше запитання
Де тут парковка для машини?
Одержати відповідь
Знайти схожі запитання
Схожі запитання у базі
Оцінка
Де можна паркувати машину ?
Де тут парковка ?
Чи можна паркувати машину у задньому дворі ?
Чи можна паркуватись на ніч навпроти магазину ?
0.5
0.4
0.4
0.3
Одержати відповідь
Відповідь на запитання:
Оцінка: 0.5
Паркуватись можна по периметру прибудинкової території, де є відповідна розмітка
Повернутись на головну
Вихід

Додаток Г

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

МЕТОД АВТОМАТИЗОВАНОГО ПІДБОРУ ТЕКСТОВИХ ВІДПОВІДЕЙ НА ЗАПИТАННЯ ЗА ЇХ СЕМАНТИЧНИМ ВМІСТОМ

Виконав:
студент 4 курсу, групи КН-18-1
Козенко Олександр Васильович

Керівник:
к.т.н., доцент кафедри КН
Мазурець Олександр Вікторович

Актуальність

Спілкування між людьми є невід'ємною частиною буття людини. З розвитком технологій спілкування перейшло з листування у конвертах на рівень електронних листів, а далі і до месенджерів та форумів.

Обробка природньої мови є одним із ключових напрямів ШІ, який працює з аналізом, розумінням та генерацією живих мов для взаємодії з комп'ютером і шляхом усного спілкування, і шляхом письмового замість звичного для комп'ютера способу машинних кодів. Відповідно, текстові дані пошуку в мережі аналізуються з метою надання таргетованого рекламного контенту. Також аналізуються тексти у листах, та навіть просто набрані у певних текстових редакторах із задачею пошуку та виправлення орфографічних помилок, тощо. Обробка природньої мови також і присутня у новітніх гаджетах, таких як розумний будинок, сірі, гул-асистент. І навіть підбірка користувацьких новин також виконується завдяки аналізу пошукових запитів користувача.

Зазвичай, спілкування у чатах нагадує балаган, де хтось запитує щось по суті, а хтось має бажання просто поспілкуватись. І важливі повідомлення можуть значно просісти у стрічці загальних повідомлень і залишитись без відповіді. Оскільки більшість запитань схожі одні на одні або періодично повторюються, виникає потреба згрупувати певні асоціативні запитання та сформувати для них актуальний перелік відповідей. Відповідно, автоматизація процесів підбору текстових відповідей на запитання за їх семантичним вмістом є актуальною задачею.

Мета і задачі роботи

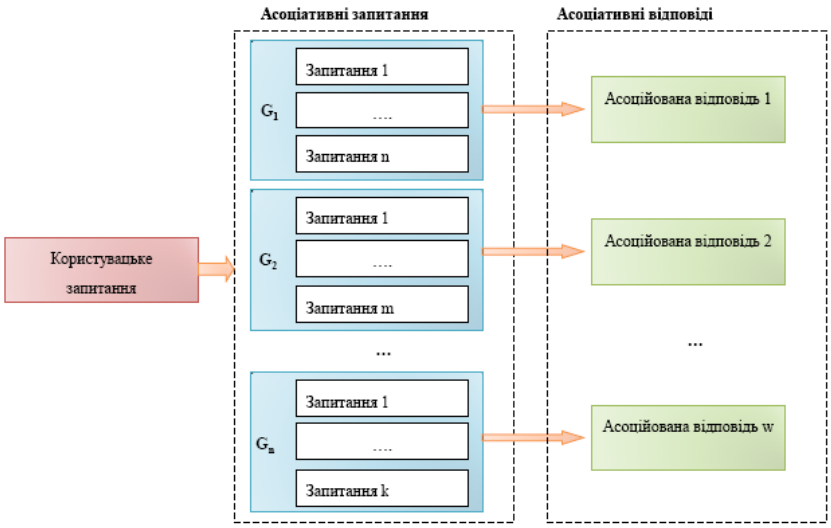
Метою кваліфікаційної роботи бакалавра є розробка й апробація методу автоматизованого підбору відповідей на запитання за семантичною подібністю, для чого слід вирішити задачі:

1. Провести аналіз предметної області.
2. Розробити метод автоматизованого підбору відповідей на запитання за семантичною подібністю.
3. Виконати проєктування інформаційної системи на базі методу автоматизованого підбору відповідей на запитання за семантичною подібністю.
4. Зробити вибір засобів розробки інформаційної системи.
5. Розробити програмну реалізацію методу автоматизованого підбору відповідей на запитання за семантичною подібністю, провести її тестування.

Розроблена програмна реалізація методу автоматизованого підбору відповідей на запитання за семантичною подібністю в вигляді інформаційної системи на платформі.NET має виконувати наступні основні групи функцій:

- семантичний аналіз наявних запитань і тестового користувацького запитання;
- обрахунок оцінок семантичної подібності користувацького запитання до кожного з наявних у базі запитань з асоційованими відповідями;
- знаходження відповіді у базі на користувацьке запитання.

Підхід до автоматизованого підбору відповіді на користувацькі запитання



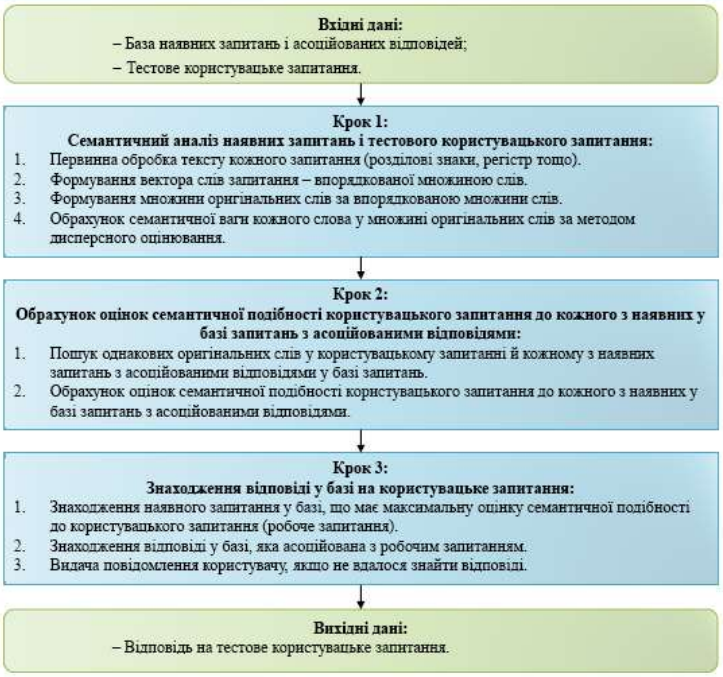


Схема методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом

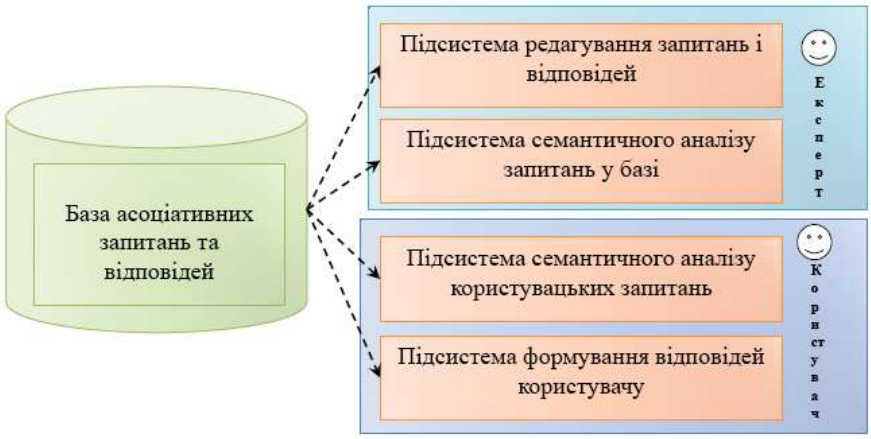
Математична модель оцінки семантичної подібності користувацького запитання до наявного запитання

Оцінка семантичної подібності $P_{a,b}$ користувацького запитання *a* до наявного запитання *b* до деякої асоційованої відповіді визначається наступним чином:

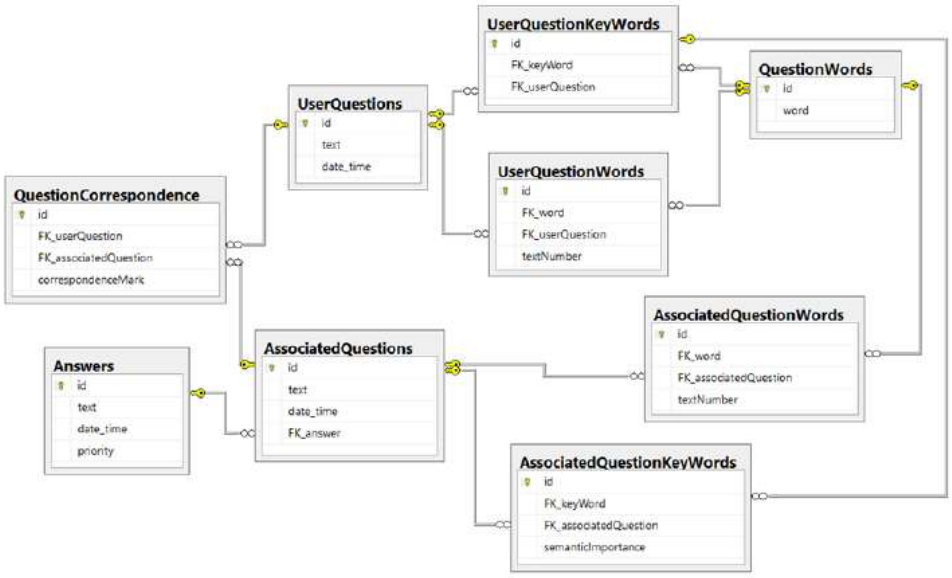
$$P_{a,b} = \sum_{i=1}^{n} D_{a,i} D_{b,i},$$

де *n* – кількість однакових оригінальних слів у користувацькому запитанні *a* й наявному запитанні *b* до деякої асоційованої відповіді, $D_{a,i}$ – оцінка семантичної важливості слова *i* у користувацькому запитанні *a*, $D_{b,i}$ – оцінка семантичної важливості слова *i* у наявному запитанні *b* до деякої асоційованої відповіді.

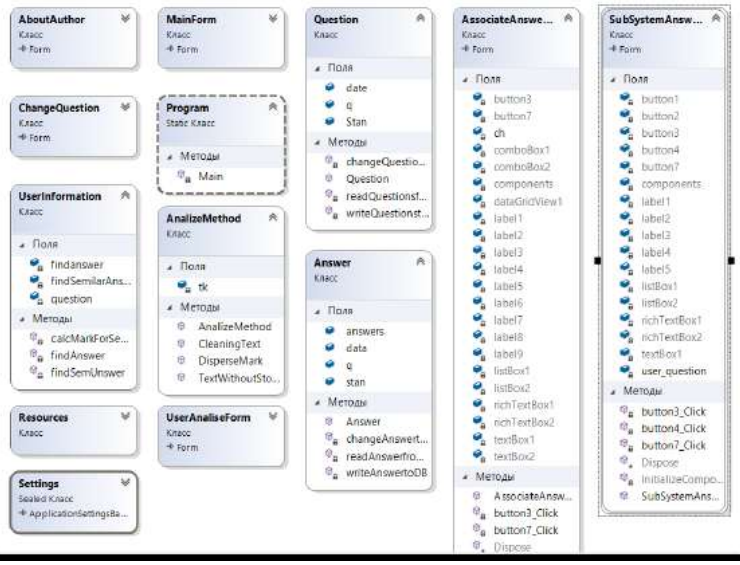
Підхід до автоматизованого підбору відповіді на користувацькі запитання



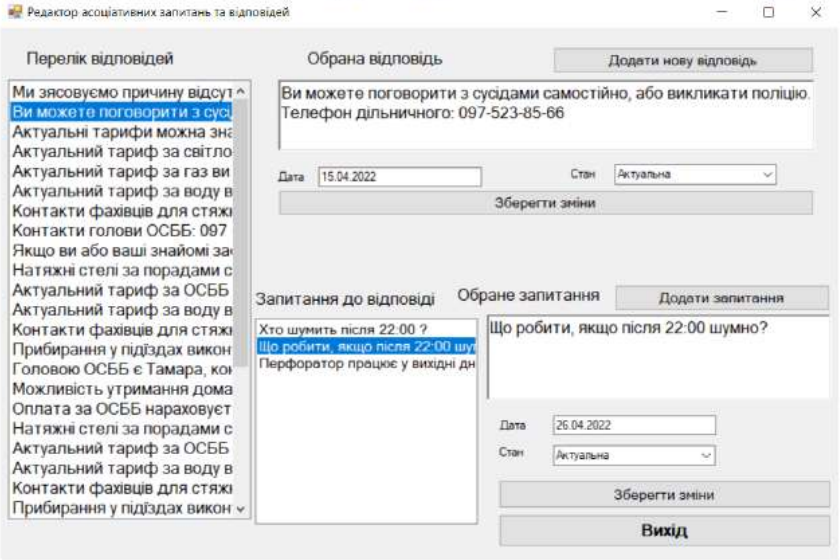
Даталогічна модель бази даних



Діаграма класів програмної реалізації методу автоматизованого підбору асоціативної відповіді за семантичною подібністю

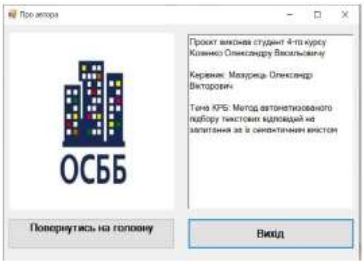


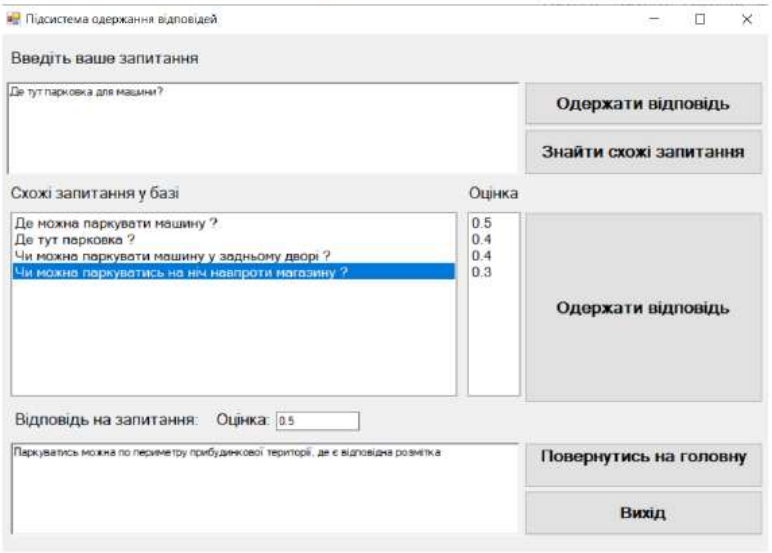
Програмна реалізація методу автоматизованого підбору асоціативної відповіді за семантичною подібністю



Форма редагування компонентів асоціативного запитання

Програмна реалізація методу автоматизованого підбору асоціативної відповіді за семантичною подібністю





Автоматизований підбір відповідей за семантичною подібністю

Висновки

У рамках виконання кваліфікаційної роботи бакалавра було виконано **розробку й апробацію методу** автоматизованого підбору відповідей на запитання за семантичною подібністю. Зокрема, було проведено аналіз предметної області й досліджено сучасні підходи до автоматизованого пошуку ключових слів, розглянуто існуючі програмні реалізації за цим напрямком.

Розроблена **програмна реалізація** методу автоматизованого підбору відповідей на запитання за семантичною подібністю на платформі .NET виконує наступні основні функції:

- Семантичний аналіз наявних запитань і тестового користувацького запитання.
- Обрахунок оцінок семантичної подібності користувацького запитання до кожного з наявних у базі запитань з асоційованими відповідями.
- Знаходження відповіді у базі на користувацьке запитання.

У якості засобів розробки було обрано фреймворк .NET Framework, мову програмування C#, редактор програмного коду VisualStudio та СКБД MS SQL Server.

UNICHECK

Ім'я користувача:
Кафедра КН

ID перевірки:
1011548994

Дата перевірки:
12.06.2022 10:10:02 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
12.06.2022 10:16:40 EEST

ID користувача:
100005671

Назва документа: Козенко_ЗАПИСКА_short

Кількість сторінок: 69 Кількість слів: 8308 Кількість символів: 64005 Розмір файлу: 3.79 MB ID файлу: 1011421027

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

8.26%
Схожість

Найбільша схожість: 4.42% з джерелом з Бібліотеки (ID файлу: 1011420943)

3.55% Джерела з Інтернету	60	Сторінка 71
6.14% Джерела з Бібліотеки	89	Сторінка 71

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи	7
Підозріле форматування	19 сторінок

Sun Jun 12 09:19:01 EEST 2022, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальное совпадение с одним документом 13.0%

Словари проверки: en_US, ru_RU, ua_UA. **Ошибок в документах: 10%**

	Документ		Суммарное совпадение по Базе Данных	
	Символы	Лексемы	Символы	Лексемы
ID: 105049 Название: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА Добавлено в БД: 2022-06-12 Авторы: на тему Метод автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом Руководители: О.В. Козенко Консультанты: О.В. Мазурець Опоненты:	48701	757	8140 (17%)	128 (17%)

Источник плагиата

ID	Описание	Наличие плагиата в документе	
		Символы	Лексемы
104163	Название: ЗВІТ з професійної практики Добавлено в БД: 2022-05-30 Авторы: О.В. Козенко Руководители: Т.К. Скрипник Консультанты: Опоненты:	6392 (13.0%)	104 (14.0%)

РІШЕННЯ ЕКСПЕРНОЇ КОМІСІЇ КАФЕДРИ КОМП'ЮТЕРНИХ НАУК
ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: *Метод автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом*

Автор: *студент групи КН-18-1 Козенко Олександр Васильович*

Спеціальність: *122 – Комп'ютерні науки*

Освітня програма: *освітньо-професійна*

Науковий керівник: *доцент кафедри КН Мазурець О.В.*

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	*відповідає*
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	

Підтвердження: запозичення, виявлені в роботі О.В. Козенка, є законними і не є плагіатом, оскільки:

1) за програмою Anti-Plagiarism виявлені 13% запозичень вказують на документ автора роботи та містять його ж Звіт з практики.

2) За програмою UNICHECK виявлені 8,26%, які є фрагментарними, не більше 4,42% на джерело – містять поширені конструкції, загальновідомі терміни та визначення.

3) запозичення розміщені в розділах аналізу існуючих аналогів та прототипів, які не описують безпосередньо авторське дослідження і не стосуються результатів роботи.

Керівник роботи _____ *Олександр МАЗУРЕЦЬ*

Гарант ОП _____ *Олександр МАЗУРЕЦЬ*

Завідувач кафедри КН _____ *Олександр БАРМАК*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу бакалавра

студента *гр. КН-18-1 Козенко Олександра Васильовича*

за темою *Метод автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом*

1. Актуальність теми

Незважаючи на досить широкі можливості автоматизації, обробка природньої мови все ще є дуже актуальним напрямом, зокрема для створення чат-ботів та автоматизованих довідників. Оскільки більшість запитань схожі або періодично повторюються, виникає потреба згрупувати певні асоціативні запитання та сформувати для них актуальний перелік відповідей. Відповідно, автоматизація процесів отримання відповідей користувачами є актуальною задачею комп'ютерних наук.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки

Поставлена у кваліфікаційній роботі бакалавра мета стосується розробки методів і технологій отримання, зберігання, обробки, передачі та використання інформації, інтелектуального аналізу даних і прийняття рішень, а саме розробки методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом. При цьому при вирішенні поставлених задач використовуються математичні моделі, методи та алгоритми розв'язання теоретичних і прикладних задач, що виникають при розробці інформаційних технологій. Тому результати виконання кваліфікаційної роботи бакалавра відповідають стандарту бакалавра спеціальності 122 – Комп'ютерні науки.

3. Професійні та особистісні якості бакалавра

При роботі над кваліфікаційною роботою бакалавра Козенко Олександр Васильович проявив себе кваліфікованим фахівцем та дисциплінованим студентом, вчасно виконуючи поставлені етапи дослідження. Як в процесі написання пояснювальної записки, так і при розробці прикладного програмного забезпечення проявив достатні для одержання успішного результату компетентності.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Одержані в роботі результати є наслідком особистої діяльності студента, який самостійно виконував всі поставлені задачі.

5. Ступінь оволодіння методами дослідження

В роботі при розробці та прикладній реалізації методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом виявлено достатній ступінь оволодіння студентом необхідними інструментами та обладнанням, методами, методиками та технологіями предметної області комп'ютерних наук.

6. Повнота та якість розкриття теми роботи

Тема роботи в повній мірі обґрунтована й розкрита, проведено аналіз актуальності та відомих досліджень в межах обраної теми, поставлені завдання, які у роботі виконані, та розроблено програмне забезпечення для автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом.

7. Логічність, послідовність, аргументованість, літературна грамотність викладу матеріалу

Структура роботи та послідовність викладення логічні та відповідні поставленій меті. Викладення матеріалу грамотне та виявляє високий ступінь відповідності стилю.

8. Можливість практичного застосування кваліфікаційної роботи бакалавра, окремих її частин

Запропонований метод автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом може мати практичне використання, зокрема програмно реалізовано наступні функції: первинна обробка тексту кожного запитання, формування вектора слів запитання, формування множини оригінальних слів за впорядкованою множини слів, обрахунок семантичної ваги кожного слова у множині оригінальних слів за методом дисперсного оцінювання, пошук однакових оригінальних слів у користувацькому запитанні й кожному з наявних запитань з асоційованими відповідями у базі запитань, обрахунок оцінок семантичної подібності користувацького запитання до кожного з наявних у базі запитань з асоційованими відповідями, знаходження наявного запитання у базі що має максимальну оцінку семантичної подібності до користувацького запитання, знаходження відповіді у базі яка асоційована з робочим запитанням, видача повідомлення користувачу, якщо не вдалося знайти відповіді.

9. Висновок про можливість допуску кваліфікаційної роботи бакалавра до захисту, на яку оцінку заслуговує робота

*Враховуючи високий рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка «**відмінно**».*

Керівник _____ *к.т.н., доц. каф. КН Олександр МАЗУРЕЦЬ*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студента *гр. КН-18-1 Козенко Олександра Васильовича*
за темою: *Метод автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом*

1. Актуальність обраної теми

Обробка природньої мови все ще є дуже актуальним напрямом, зокрема для створення чат-ботів та автоматизованих довідників. Оскільки більшість запитань схожі або періодично повторюються, виникає потреба згрупувати певні асоціативні запитання та сформувати для них актуальний перелік відповідей. Відповідно, автоматизація процесів отримання відповідей користувачами є актуальною задачею комп'ютерних наук.

2. Повнота розкриття мети та завдань дослідження

Кваліфікаційна робота бакалавра студента Козенка О.В. виконана в повному обсязі, було проаналізовано предметну область, створено методу автоматизованого підбору текстових відповідей на запитання за їх семантичним вмістом та реалізовано програмний застосунок на базу вищезазначеного методу.

3. Зміст кожного розділу роботи

Перший розділ присвячений проведенню аналізу предметної області та визначенню основних параметрів для розв'язку поставленої задачі. Другий розділ присвячений проєктуванню функціональної структури інформаційної системи. Третій розділ присвячений програмній реалізації спроєктованої функціональної структури інформаційної системи. Також кожен розділ підкріплений відповідними висновками та сформовано основний висновок роботи.

4. Оцінка розробленої інформаційної системи, її практична цінність

Створений метод та програмний застосунок на його основі дозволяють користувачеві отримати швидкі відповіді на запитання. Реалізований метод та програмний продукт на його основі можна застосувати чи не в кожній сфері обслуговування клієнтів, зокрема житлово-експлуатаційних організацій, тощо.

5. Якість оформлення кваліфікаційної роботи бакалавра

Робота виконана на належному науково-методичному рівні та відповідає встановленим вимогам щодо оформлення такого роду праць.

6. Недоліки кваліфікаційної роботи бакалавра

Кваліфікаційна робота бакалавра виглядала б привабливіше, якби у ній було розглянуто більше математичних підходів щодо взаємодії документів між собою.

8. Загальний висновок (допускається чи не допускається до захисту), та оцінка якої оцінки заслуговує кваліфікаційна робота.

Враховуючи рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка ***«відмінно»***.

Рецензент Бердатюк Л.П. д. ф-м. н