


Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

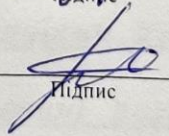
КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему Метод рекомендації текстових документів за запитаннями користувачів засобами NLP

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

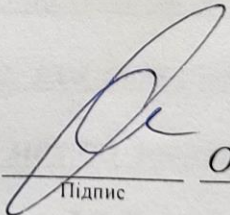
Виконав: студент 2 курсу, група КНМ-22-1  О.В. Козенко
Курс, група виконавця Підпис Ініціали, прізвище

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

12 грудня 2023 р.

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет інформаційних технологій

Кафедра комп'ютерних наук

Освітній ступінь магістр

Галузь знань 12 – Інформаційні технології

Спеціальність 122 – Комп'ютерні науки

ЗАТВЕРДЖУЮ

Завідувач кафедри комп'ютерних наук

(підпис)

д.т.н., професор О.В. Бармак

« 01 » вересня 2023 року

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ МАГІСТРА

1. Тема кваліфікаційної роботи магістра: «Метод рекомендації текстових документів за запитаннями користувачів засобами NLP»

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

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

4. Затверджені наказом університету від « 15 » серпня 2023 р. № 30

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

Мета кваліфікаційної роботи магістра – вирішення задачі автоматизованої рекомендації текстових документів за запитаннями користувачів засобами NLP, для чого слід розробити відповідний метод рекомендації текстових документів за запитаннями користувачів, спроектувати компоненти метода та архітектуру відповідної програмної системи, яка використовує розроблений метод; після чого слід програмно реалізувати метод рекомендації текстових документів за запитаннями користувачів засобами NLP у вигляді спроектованої програмної системи й дослідити його ефективність.

Реферат

Кваліфікаційна робота магістра розв'язує науково-технічну задачу автоматизованої рекомендації текстових документів за запитаннями користувачів. Результатом роботи є метод, призначений для пошуку релевантних рекомендацій з бази текстових порад за текстовим запитом користувача, що перетворює вхідні дані у вигляді користувацького запиту, датасету із набором попередньо оброблених лематизованих асоціативних запитів та векторизованого корпусу у вихідні дані у формі текстової рекомендації релевантної до запиту, і програмна реалізація розробленого методу для предметної області служби психологічної підтримки.

Актуальність теми. У сучасному світі спостерігається зростання попиту на послуги психологічної підтримки. Це пов'язано з низкою факторів, зокрема, з ускладненням життя, збільшенням кількості стресових ситуацій, поширенням психосоматичних захворювань, особливо в умовах воєнного стану. У зв'язку з цим виникає необхідність у розробці ефективних методів надання психологічної допомоги.

Одним із перспективних напрямків розвитку психологічної допомоги є застосування інформаційних технологій. Інтернет-ресурси використовуються для надання консультацій, проведення тренінгів, організації спільнот для підтримки людей, які страждають від психологічних проблем, однак, за даними досліджень, в майбутнього прогнозується зріст людей з психологічними проблемами, і наявні інформаційні ресурси не зможуть повноцінно задовольнити наявний попит. У зв'язку з цим виникає необхідність у розробці методів, що дозволяють рекомендувати користувачам релевантні текстові документи, які можуть допомогти їм у вирішенні їхніх проблем.

Розроблений у кваліфікаційній роботі метод має ряд переваг у порівнянні з існуючими методами. Зокрема, він дозволяє враховувати контекст запиту користувача, особливості його мовлення, а також структуру бази текстових порад. Це дозволяє підвищити точність і ефективність рекомендацій.

Таким чином, кваліфікаційна робота магістра має значну наукову та практичну цінність. Вона може бути використана для розробки нових інформаційних систем і сервісів, призначених, зокрема, для надання психологічної підтримки.

Мета і задачі роботи. *Метою кваліфікаційної роботи магістра є вирішення задачі автоматизованого формування вибірок рекомендованих текстових документів за запитаннями користувачів, а також відповідна програмна реалізація методу для предметної області служби психологічної підтримки. Для досягнення мети необхідно виконати наступне:*

- Провести аналіз предметної області, виділити актуальні напрями та окреслити сферу для автоматизації рекомендації документів за запитами користувачів.

- Створити метод рекомендації текстових документів за запитаннями користувачів засобами NLP та описати його кроки та компоненти.

- Спроекувати інформаційну систему на базі методу рекомендації текстових документів за запитаннями користувачів засобами NLP.

- Виконати вибір засобів розробки інформаційної системи на базі методу рекомендації текстових документів.

- Розробити відповідну програмну реалізацію методу рекомендації текстових документів за запитаннями користувачів засобами NLP для предметної області служби психологічної підтримки.

- Дослідити ефективність програмно реалізованого методу формування вибірок рекомендованих текстових документів за запитаннями користувачів.

Об’єкт дослідження – процес знаходження релевантних текстових документів за запитами.

Предмет дослідження – моделі, методи, алгоритми пошуку ключових слів для знаходження релевантних текстових документів за запитами.

Методи дослідження, що застосовані для вирішення поставлених завдань, наступні: положення методів аналізу даних та теорії множин, методології проектування інформаційних систем, об'єктно-орієнтований підхід.

Наукова новизна одержаних результатів. Результати виконання кваліфікаційної роботи магістра містять *інновації та наукову новизну*, зокрема було удосконалено метод рекомендації текстових документів за запитаннями користувачів засобами NLP, який дозволяє автоматизовано виконувати формування вибірок рекомендованих текстових документів за запитаннями користувачів, виконуючи при цьому аналіз відповідності запиту користувача як за семантичною відповідністю до типового запиту до текстового документу, так і, у випадку її відсутності, знаходячи релевантні текстові документи за методом косинусної подібності, таким чином перетворюючи вхідні дані у вигляді користувацького запиту, вибірки із набором лематизованих асоціативних запитів та векторизованого корпусу текстових документів у вихідні дані у вигляді ранжованої множини релевантних рекомендованих текстових документів.

Практичне значення одержаних результатів. Було створено інформаційну систему автоматизованого визначення рекомендацій текстових документів за користувацькими запитами, яка є прикладною програмною реалізацією метода рекомендації текстових документів за запитаннями користувачів, що призначений для пошуку релевантних рекомендацій з бази текстових порад за текстовим запитом користувача та вхідними даними має набір даних з групами асоційованих запитів та відповідних їм асоційованих відповідей та користувацький запит, що перетворює їх у вихідні дані у вигляді текстової рекомендації релевантної до запиту

Інформаційна структура системи складається із набору даних та трьох підсистем: «Підсистеми роботи з корпусами текстів», «Підсистеми редактора навчальної вибірки документів запитів-відповідей» та головної «Підсистеми автоматизованого підбору документів відповідей до запитів», що дозволяють

знаходити релевантний текстовий документ за користувачьким текстовим запитом.

Апробація результатів кваліфікаційної роботи магістра та публікації.

Одержані наукові і практичні результати кваліфікаційної роботи магістра доповідались в доповіді на тему «Метод автоматизованого підбору відповідей на користувачькі запитання за семантичною подібністю» на Міжнародній науково-практичній конференції «Глушковські читання» (26 жовтня 2023 року) та на тему «Використання метрик косинусної схожості та індексу Жаккара для інтелектуального аналізу семантичної подібності текстових документів» на XV Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2023» (17-18 листопада 2023 року); за темою роботи автором виконано наукову публікацію:

Козенко О.В., Мазурець О.В., Молчанова М.О., Собко О.В. Використання метрик косинусної схожості та індексу Жаккара для інтелектуального аналізу семантичної подібності текстових документів. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2023». Хмельницький, 2023. с. 146-147.

Структура і обсяг роботи. Кваліфікаційна робота магістра складається з наступного: реферату, завдання, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань з 41 найменування й 5-х додатків. Обсяг основного тексту кваліфікаційної роботи магістра становить 85 сторінок. В роботі наведено 47 світлин й 5 таблиць.

Ключові слова: NLP, TF-IDF, косинусна подібність, індекс Жаккара, рекомендації текстових документів за запитом, інформаційна система.

Зміст

Перелік скорочень	4
Вступ.....	5
Розділ 1 Аналіз предметної області пошуку подібності текстових документів засобами NLP	9
1.1 Особливості рекомендації текстових документів за запитаннями користувачів для напрямку психологічної підтримки	9
1.2 Аналіз методів обробки природної мови для задачі рекомендації текстових документів за запитаннями користувачів	10
1.3 Аналіз існуючих програмних рішень та сучасних наукових вишукувань щодо підбору текстових документів за запитаннями.....	14
1.4 Мета, задачі та вимоги до реалізації інформаційної системи	18
Висновки до першого розділу.....	19
Розділ 2 Метод рекомендації текстових документів за запитаннями користувачів засобами NLP	21
2.1 Формування навчальної вибірки даних у вигляді множини запитів та текстів відповідей.....	21
2.2 Метрики для знаходження рекомендації текстових документів за запитаннями користувачів	25
2.3 Підхід до пошуку рекомендації текстових документів за запитаннями користувачів.....	27
2.4 Етапи методу рекомендації текстових документів за запитаннями користувачів.....	28
Висновки до другого розділу	32
Розділ 3 Інформаційна система автоматизованого визначення рекомендацій текстових документів за користувацькими запитаннями	34
3.1 Схема інформаційної системи	34
3.2 Схема та функції підсистеми формування навчальної вибірки документів запитів-відповідей.....	36
3.3 Схема та функції підсистеми автоматизованого підбору документів відповідей до запитів	37
3.4 Формування комбінації засобів розробки інформаційної системи	39
3.5 Вибір спеціалізованих програмних розширень	42

3.6 Програмна архітектура інформаційної системи автоматизованого визначення рекомендацій текстових документів	45
Висновки до третього розділу	49
Розділ 4 Дослідження ефективності методу рекомендації текстових документів за запитаннями користувачів засобами NLP	51
4.1 Особливості розробки прикладних компонентів інформаційної системи рекомендації текстових документів	51
4.2 Прикладне тестування інформаційної системи	58
4.3 Особливості використання інформаційної системи автоматизованого визначення рекомендацій текстових документів	65
4.4 Дослідження ефективності та інтерпретація отриманих результатів.....	75
Висновки до четвертого розділу.....	79
Загальні висновки.....	81
Перелік посилань.....	83
Додатки	

Перелік скорочень

Скорочення, термін, позначення	Пояснення
TF	Term Frequency
TF-IDF	Term Frequency – Inverse Document Frequency
FAQ	Frequently asked questions
GUI	Графічний інтерфейс користувача
NLP	Natural language processing
NLTK	Natural Language Toolkit
ІС	Інформаційна система
ІТ	Інформаційні технології
КРМ	Кваліфікаційна робота магістра
КН	Комп'ютерні науки
ПЗ	Пояснювальна записка
ПП	Програмний продукт
ХНУ	Хмельницький національний університет.
ПТСР	Посттравматичний стресовий розлад
ІІ	Штучний інтелект

Вступ

Кваліфікаційна робота магістра розв'язує науково-технічну задачу автоматизованої рекомендації текстових документів за запитаннями користувачів. Результатом роботи є метод, призначений для пошуку релевантних рекомендацій з бази текстових порад за текстовим запитом користувача, що перетворює вхідні дані у вигляді користувацького запиту, датасету із набором попередньо оброблених лематизованих асоціативних запитів та векторизованого корпусу у вихідні дані у формі текстової рекомендації релевантної до запиту, і програмна реалізація розробленого методу для предметної області служби психологічної підтримки.

Актуальність теми. У сучасному світі спостерігається зростання попиту на послуги психологічної підтримки. Це пов'язано з низкою факторів, зокрема, з ускладненням життя, збільшенням кількості стресових ситуацій, поширенням психосоматичних захворювань, особливо в умовах воєнного стану. У зв'язку з цим виникає необхідність у розробці ефективних методів надання психологічної допомоги.

Одним із перспективних напрямків розвитку психологічної допомоги є застосування інформаційних технологій. Інтернет-ресурси використовуються для надання консультацій, проведення тренінгів, організації спільнот для підтримки людей, які страждають від психологічних проблем, однак, за даними досліджень, в майбутнього прогнозується зріст людей з психологічними проблемами, і наявні інформаційні ресурси не зможуть повноцінно задовольнити наявний попит. У зв'язку з цим виникає необхідність у розробці методів, що дозволяють рекомендувати користувачам релевантні текстові документи, які можуть допомогти їм у вирішенні їхніх проблем.

Розроблений у кваліфікаційній роботі метод має ряд переваг у порівнянні з існуючими методами. Зокрема, він дозволяє враховувати контекст запиту користувача, особливості його мовлення, а також структуру бази текстових порад. Це дозволяє підвищити точність і ефективність рекомендацій.

Таким чином, кваліфікаційна робота магістра має значну наукову та практичну цінність. Вона може бути використана для розробки нових інформаційних систем і сервісів, призначених, зокрема, для надання психологічної підтримки.

Мета і задачі роботи. *Метою кваліфікаційної роботи магістра є вирішення задачі автоматизованого формування вибірок рекомендованих текстових документів за запитаннями користувачів, а також відповідна програмна реалізація методу для предметної області служби психологічної підтримки. Для досягнення мети необхідно виконати наступне:*

- Провести аналіз предметної області, виділити актуальні напрями та окреслити сферу для автоматизації рекомендації документів за запитами користувачів.

- Створити метод рекомендації текстових документів за запитаннями користувачів засобами NLP та описати його кроки та компоненти.

- Спроекувати інформаційну систему на базі методу рекомендації текстових документів за запитаннями користувачів засобами NLP.

- Виконати вибір засобів розробки інформаційної системи на базі методу рекомендації текстових документів.

- Розробити відповідну програмну реалізацію методу рекомендації текстових документів за запитаннями користувачів засобами NLP для предметної області служби психологічної підтримки.

- Дослідити ефективність програмно реалізованого методу формування вибірок рекомендованих текстових документів за запитаннями користувачів.

Об’єкт дослідження – процес знаходження релевантних текстових документів за запитами.

Предмет дослідження – моделі, методи, алгоритми пошуку ключових слів для знаходження релевантних текстових документів за запитами.

Методи дослідження, що застосовані для вирішення поставлених завдань, наступні: положення методів аналізу даних та теорії множин, методології проектування інформаційних систем, об’єктно-орієнтований підхід.

Наукова новизна одержаних результатів. Результати виконання кваліфікаційної роботи магістра містять *інновації та наукову новизну*, зокрема було удосконалено метод рекомендації текстових документів за запитаннями користувачів засобами NLP, який дозволяє автоматизовано виконувати формування вибірок рекомендованих текстових документів за запитаннями користувачів, виконуючи при цьому аналіз відповідності запитання користувача як за семантичною відповідністю до типового запиту до текстового документу, так і, у випадку її відсутності, знаходячи релевантні текстові документи за методом косинусної подібності, таким чином перетворюючи вхідні дані у вигляді користувацького запиту, вибірки із набором лематизованих асоціативних запитів та векторизованого корпусу текстових документів у вихідні дані у вигляді ранжованої множини релевантних рекомендованих текстових документів.

Практичне значення одержаних результатів. Було створено інформаційну систему автоматизованого визначення рекомендацій текстових документів за користувацькими запитом, яка є прикладною програмною реалізацією метода рекомендації текстових документів за запитаннями користувачів, що призначений для пошуку релевантних рекомендацій з бази текстових порад за текстовим запитом користувача та вхідними даними має набір даних з групами асоційованих запитів та відповідних їм асоційованих відповідей та користувацький запит, що перетворює їх у вихідні дані у вигляді текстової рекомендації релевантної до запиту

Інформаційна структура системи складається із набору даних та трьох підсистем: «Підсистеми роботи з корпусами текстів», «Підсистеми редактора навчальної вибірки документів запитів-відповідей» та головної «Підсистеми автоматизованого підбору документів відповідей до запитів», що дозволяють знаходити релевантний текстовий документ за користувацьким текстовим запитом.

Апробація результатів кваліфікаційної роботи магістра та публікації. Одержані наукові і практичні результати кваліфікаційної роботи магістра

доповідались в доповіді на тему «Метод автоматизованого підбору відповідей на користувачькі запитання за семантичною подібністю» на Міжнародній науково-практичній конференції «Глушковські читання» (26 жовтня 2023 року) та на тему «Використання метрик косинусної схожості та індексу Жаккара для інтелектуального аналізу семантичної подібності текстових документів» на XV Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2023» (17-18 листопада 2023 року); за темою роботи автором виконано наукову публікацію:

Козенко О.В., Мазурець О.В., Молчанова М.О., Собко О.В. Використання метрик косинусної схожості та індексу Жаккара для інтелектуального аналізу семантичної подібності текстових документів. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2023». Хмельницький, 2023. с. 146-147.
URL: <https://kn.khmnmu.edu.ua/wp-content/uploads/sites/18/apkn-2023-corporpaper.pdf>

Структура і обсяг роботи. Кваліфікаційна робота магістра складається з наступного: реферату, завдання, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань з 41 найменування й 5-х додатків. Обсяг основного тексту кваліфікаційної роботи магістра становить 85 сторінок. В роботі наведено 47 світлин й 5 таблиць.

Розділ 1 Аналіз предметної області пошуку подібності текстових документів засобами NLP

1.1 Особливості рекомендації текстових документів за запитаннями користувачів для напрямку психологічної підтримки

Психічний стан і психічне здоров'я під час військових конфліктів є одними з ключових факторів здоров'я нації. Багато осіб відчують виснаженість, стрес, тривогу та паніку [1]. Ряд фахівців з психології розробили корисні поради для управління собою в екстремальних умовах, надання підтримки близьким та забезпечення здорового психічного стану в умовах кризи, проте проблема знаходження актуальних порад у поточний момент часу залишається до кінця не розв'язаною.

Війна може сильно вплинути на психічне благополуччя людей, які переживають конфлікт, а також на тих, хто надає допомогу в цих умовах. Першою поширеною проблемою, яка з'являється в умовах війни – травматичні досвіди. Умови війни часто супроводжуються травматичними подіями, які можуть викликати посттравматичний стресовий розлад (ПТСР) та інші психічні проблеми не тільки у ветеранів, цивільних жертв, а також у осіб, які працюють в гуманітарних організаціях та надають психологічну допомогу.

Людина в умовах невизначеності відчуває страх і тривогу. Небезпека, страх і постійна тривога можуть вплинути на психічне здоров'я та спричинити тривожні розлади, а втрата близьких та руйнування житла часто викликають глибокі сум і депресію.

Ще однією проблемою є соціальна ізоляція. В умовах війни люди часто відчувають себе ізольованими від світу навколо них, що може призвести до відчуття самотності та відчуженості.

Через зріст психологічних проблем спостерігається зменшення доступності до психічної допомоги. У військових конфліктах може бути обмежений доступ до психологічної допомоги та медичного обслуговування. За даними експертів, від 40% до 50% населення будуть потребувати психологічної

допомоги. В окремих категоріях населення ця потреба буде така: серед військових та ветеранів - близько 1,8 мільйонів осіб, серед людей похилого віку - близько 7 мільйонів, а серед дітей та підлітків – приблизно 4 мільйони. Прогнозується, що на первинному рівні медичного обслуговування буде потреба в 27 мільйонах консультацій щодо психічного здоров'я. Водночас, приблизно 3-4 мільйони українців матимуть психічні розлади різного ступеня важкості – від помірних до тяжких [2]. Тому автоматизація процесу своєчасного надання необхідних рекомендацій за запитами користувача є актуальною та потребує автоматизації.

1.2 Аналіз методів обробки природної мови для задачі рекомендації текстових документів за запитами користувачів

Для забезпечення користувачів необхідною інформацією та рекомендаціями для поліпшення їхнього психічного стану та емоційного благополуччя необхідно автоматизувати процес надання відповідей за запитами користувачів.

Галузь обробки природної мови (NLP) відноситься до галузі інформатики, а саме до сфери штучного інтелекту (ШІ). Її основна мета – дати комп'ютерам можливість розуміти текст і голос, подібно до того, як це роблять люди. NLP об'єднує обчислювальну лінгвістику (моделювання людської мови на основі правил) зі статистичними моделями, методами машинного навчання та глибокого навчання. Всі ці технології разом дозволяють комп'ютерам обробляти людську мову у вигляді тексту або аудіо-даних та бути зрозумілими, а також розуміти повне значення тексту, враховуючи наміри та настрої спікера або письменника [3]. Приклади застосування NLP включають системи розпізнавання мовлення, узагальнення текстових документів, машинний переклад, виявлення спаму, розпізнавання іменованих сутностей, відповіді на запитання, автоматичне доповнення тексту та багато інших.

На початку історії Інтернету, пошукові системи використовували прості алгоритми, які спиралися на порівняння ключових слів у запитах користувачів і тексту, що містився на вебсторінках. Такий підхід часто призводив до неправильних результатів та не задовольняв потреби користувачів.

Зараз із розвитком технологій штучного інтелекту в пошукових системах почали використовувати алгоритми машинного навчання, які дозволяють аналізувати великі обсяги даних та робити більш точні прогнози щодо того, що саме користувачі шукають.

Одним з важливих кроків у цьому напрямку було впровадження алгоритму Google RankBrain у 2015 році. RankBrain – це система на основі штучного інтелекту, яка допомагає Google ефективніше розуміти та обробляти пошукові запити. Вона використовує машинне навчання для аналізу користувацької поведінки та виявлення шаблонів, що дозволяє зробити більш точні прогнози щодо найбільш релевантних результатів пошуку для конкретного запиту. З моменту впровадження RankBrain він став невідомою частиною пошукового алгоритму Google, і його успіх стимулював інші пошукові системи застосовувати схожі підходи на основі штучного інтелекту [4].

На сьогодні існує багато методів пошуку релевантних документів за запитом користувача. Один з найпоширеніших методів – це використання метрик подібності. Метрики подібності використовуються для вимірювання подібності між запитом та документом. Чим більша подібність між запитом та документом, тим більш релевантний документ [5].

Нижче наведено деякі поширені метрики подібності [5]:

– TF-IDF – це метрика, яка враховує частоту виникнення слів у документі та частоту виникнення слів у наборі документів.

– Cosine similarity – це метрика, яка враховує кут між векторами запиту та документа.

– Jaccard similarity – це метрика, яка враховує кількість спільних слів у запиті та документі.

Також поширеним підходом до вирішення проблеми знаходження релевантних документів за запитом користувача є методи машинного навчання. Машинні моделі можуть бути навчені на наборі документів, позначених як релевантні та нерелевантні до запиту. Після навчання модель може використовуватися для класифікації нових документів як релевантних або нерелевантних.

Практично всі з вищеописаних метрик так чи інакше використовують у роботі підхід ключових слів. У аналізі текстової інформації та в пошукових системах ключове слово представляє собою спеціальний набір слів, які отримуються за допомогою лінгвістичних і математичних методів. Ці слова відображають зміст тексту та є важливими, загальноприйнятними, конкретними та інформативними для певної теми. Вони надають високорівневий опис змісту тексту, що забезпечує коротке та лаконічне уявлення про зміст тексту для читача і полегшує збереження інформації у пам'яті. Можна виділити наступну сукупність ознак (рисунок 1.1) [6].



Рисунок 1.1 – Класифікація методів пошуку ключових слів [6]

Найбільш класичний метод побудови вектора представлення текстів вважають мішок слів [7]. Ключове припущення цього підходу полягає в тому, що текст може бути виражений за допомогою неупорядкованого 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.

TF-IDF є способом векторизації тексту, який відбиває важливість слова у документі, а не лише частоту його появи. Він надає вагу словам, які в кінцевому рахунку є більш значущими у всьому корпусі слів, словам, які не з'являються неймовірно часто, а також словами, які не є надзвичайно рідкісними.

Term Frequency (TF) – частота терміну, що вимірює частоту зустрічання даного терміну в обраному документі. Оскільки в великих документах термін буде зустрічатися більшу кількість раз ніж в маленьких, просто кількість знаходжень цього слова не вистачає. Тому використовується відносна частота – відношення числа входження слова до загальної кількості слів у документі.

Inverse Document Frequency (IDF) – зворотня частотність документа, яка визначає важливість терміна в конкретній колекції документів. Деякі слова, такі як прийменники, часто зустрічаються у всіх документах, але мають мало впливу на сутність тексту. Під час розрахунку частоти терміну ми враховуємо кожен токен як рівноцінний, але IDF дозволяє враховувати це, зменшуючи вагу термінів, що зустрічаються в усіх документах. Значення IDF відображає логарифм від відношення загальної кількості документів у колекції до кількості документів, в яких присутній обраний термін. Цей розрахунок дозволяє надавати вагу рідко зустрічаються словам, відзначаючи їхню важливість у контексті відмінностей між документами порівняно з загальним словником.

Отже, з проведеного аналізу щодо шляхів автоматизації з обробки природної мови для системи рекомендацій документів за запитами користувачів на базі служби психологічної підтримки будуть використані метрики косинусної

схожості та індексу Жаккара, а для векторизації текстового подання буде використано TF-IDF.

1.3 Аналіз існуючих програмних рішень та сучасних наукових вишукувань щодо підбору текстових документів за запитами

Розглядається задача рекомендації текстових документів за запитаннями користувачів за вхідними даними які містять набір даних з групами асоційованих запитів та відповідних їм асоційованих відповідей, причому для апробації має бути використано предметну область служби психологічної підтримки.

Тож, на сьогоднішній день проблема психологічної підтримки є актуальною, і вже працюють велика кількість онлайн ресурсів та гарячі лінії психологічної допомоги. Одним із таких сайтів є «Лайфлайн Україна» (рисунок 1.2). Головна діяльність спрямована на підтримку людей у складних ситуаціях, зокрема тих, хто має суїцидальні думки або переживає емоційні кризи. Фахівці надають допомогу через гарячу лінію, яка доступна 24 години на добу та 7 днів на тиждень, і безкоштовна для користувачів основних операторів мобільного зв'язку в Україні. Номер гарячої лінії – 7333. За час з моменту запуску гарячої лінії в жовтні 2019 року, компанія встигла надати допомогу більше ніж 30 000 разів тим, хто опинився в скрутних життєвих обставинах. Зараз, внаслідок повного вторгнення в Україну, попит на психологічну підтримку зріс приблизно втричі [8].

Для людей, що віддають перевагу письмовим зверненням також надається підтримку через чат. Люди, що потребують допомоги, зможуть зв'язатися з фахівцями психологічної підтримки через функцію чату з веб-сайту, через Facebook Messenger або через Telegram [8].

Також з безкоштовних ресурсів психологічної підтримки є «Teenergizer» - це рух, який об'єднує підлітків та молодь з регіону Східної Європи та Центральної Азії. Він був створений у 2016 році в Україні та зараз має осередки в 10 містах регіону [9].

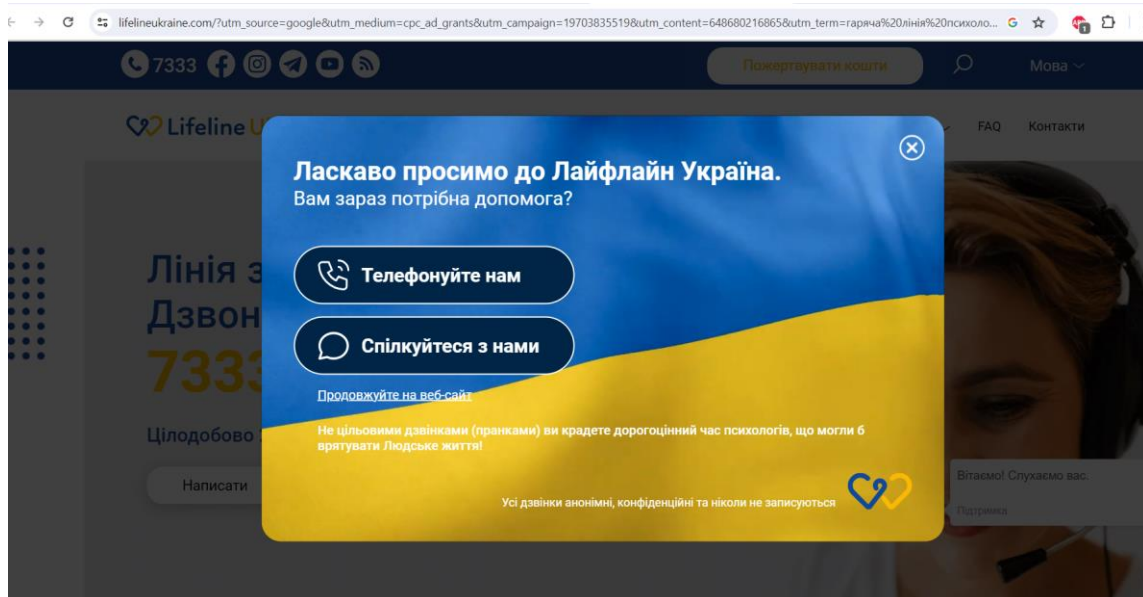


Рисунок 1.2 – Вигляд ресурсу «Lifeline Ukraine»

Мета Teenergizer – створити світ, в якому кожен підліток може реалізувати свій потенціал. Рух працює над досягненням цієї мети, проводячи освітні заходи, спрямовані на підвищення обізнаності про ВІЛ/СНІД, сексуальне здоров'я та права людини. Teenergizer також створює можливості для підлітків та молоді брати участь у громадському житті та робити позитивний вплив на своє суспільство. Також сайт має чат, де можна отримати консультацію (рисунок 1.3).

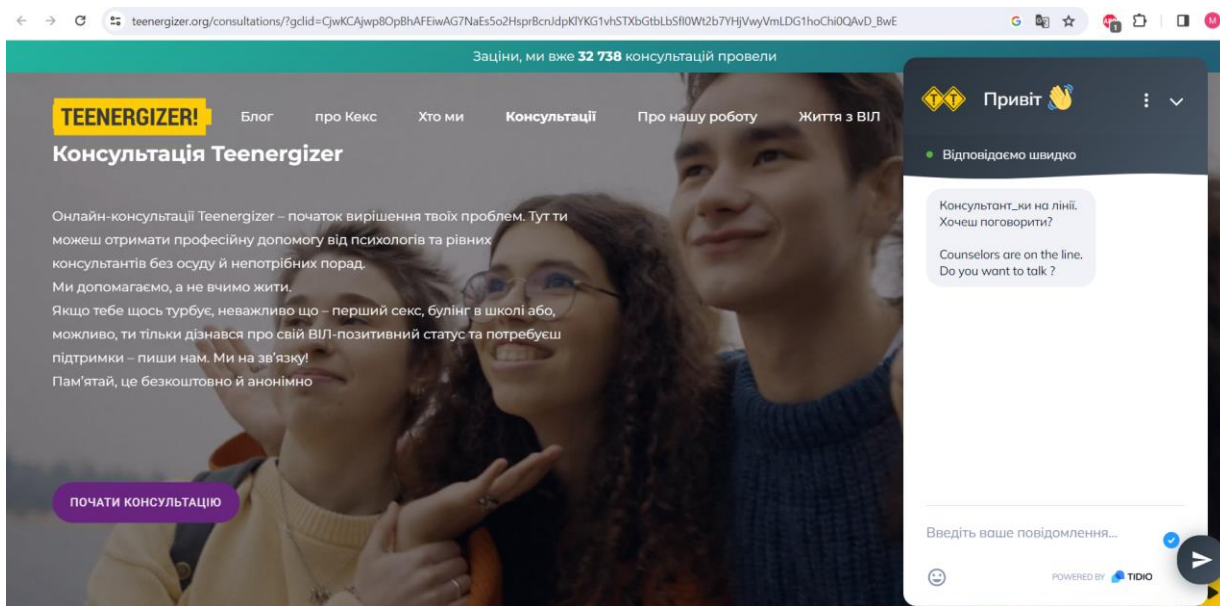


Рисунок 1.3 – Вигляд ресурсу «Teenergizer»

Окрім існуючих сайтів з номерами гарячих ліній та чат-ботів, цією проблемою також опікуються науковці. Зокрема, автором [10] було визначено найбільш уразливі категорії населення та виокремлено загальні загрози та ризики, пов'язані з другорядною травматизацією. Детально описані ситуації, які потребують негайної, подальшої та третинної допомоги в галузі психологічного та психічного здоров'я. Виокремлено необхідність організації соціально-психологічної підтримки та супроводу для громадян у форматі онлайн з метою забезпечення допомоги на другому етапі. На основі цього розроблено чотири онлайн-модулі, які активуються під час військових подій.

Перший модуль, "Залучення", використовується для онлайн і офлайн активностей, щоб продемонструвати доступність та наявність надавачів супроводу, що може бути співвіднесимо з метафоричною "розтягнутою рукою допомоги". Проблемою цього модуля є потреба враховувати онлайн-поведінку громадян та соціодемографічне спрямування.

Другий модуль, "Психологічного просвітництва", спрямований на забезпечення розуміння постраждалими своєї ситуації, стимулювання мотивації до змін, інструктування стосовно дій в кризових ситуаціях для збереження, відновлення та розвитку. Проблема полягає в необхідності враховувати маркетингове позиціонування платформ, з яких надається інформація.

Третій модуль, "Пошуку рішень", включає в себе електронні ресурси для обміну соціальною підтримкою - інформаційною, емоційною, оцінковою, мережевою та матеріалізованою. Проблема полягає в необхідності стимулювання практик самопомоги для запобігання другорядній травматизації.

Четвертий модуль, "Супроводу", передбачає тривалу соціально-психологічну підтримку постраждалих у відеоформаті, що базується на заздалегідь укладених угодах. Проблема полягає в дотриманні психологічної безпеки та етики, а також в компенсації неможливості фізичного контакту.

Зазначається, що ці модулі можуть використовуватися окремо або як частина інтегрованої очної соціально-психологічної підтримки особистості.

Авторами [11] зазначено, що серед семи мільйонів українських дітей 4,3 мільйона були вимушені залишити свої домівки. Ця надзвичайна ситуація призвела до трагічних наслідків, таких як загибель 232 дітей і поранення 427 дітей. В цих обставинах діти стали найбільш вразливою категорією населення в Україні, яка потребує професійної психолого-педагогічної підтримки. Особливу увагу також приділено дорослим, які надають цю підтримку, і зазначено, що інформаційно-технологічне забезпечення цього процесу, а також оцінка і систематизація існуючих ресурсів, є надзвичайно важливими аспектами.

Авторами виявлено, що протягом майже трьох місяців воєнних дій в Україні до надання психолого-педагогічної підтримки приєдналися численні державні та неурядові організації, педагоги, психологи та психотерапевти. Наприклад, при введенні запиту "Психолого-педагогічна підтримка дітей в умовах війни" в Google було знайдено 7 450 000 результатів за 0,55 секунди. Такий обсяг інформації потребує аналізу, систематизації, оцінки та узагальнення.

Список використаних авторами [12] основних джерел інформаційно-технологічного забезпечення включає в себе веб-сайт Міністерства освіти та науки України, веб-сайт Інституту модернізації змісту освіти, веб-сайт освітнього омбудсмена України, відеоінструкції та інфографіку, розміщені на YouTube, телеграм-канали (такі як "Підтримай дитину", "EdEra", "ПТСП і всі інші", "Психологічна Допомога в Умовах Війни", "Психологічна підтримка та інші"), чат-боти (наприклад, "Dovidka Info", "Друг" і інші), інформаційні комікси, курси на онлайн-платформах "Prometheus" і "EdEra", сторінки психотерапевтів і психологів у Facebook, блоги (як "Діти в місті", "Mental Code" і інші). Визначено переваги, недоліки та ризики цих ресурсів, а також обговорено перспективи їх подальшого використання.

Авторами [12] досліджено, що у зв'язку з бойовими діями та значними внутрішніми міграціями, українське населення стикається з численними

перешкодами у доступі до послуг охорони здоров'я. Пошкоджена інфраструктура робить ситуацію більш складною і, в деяких випадках, ускладнює навіть логістику. Вимушено переміщені особи потребують більше адаптації, психологічної підтримки та інформаційного супроводження для отримання медичних послуг.

Результати використання цифрових інструментів для покращення доступності медичних послуг показали, що українське суспільство успішно адаптувало наявні цифрові інструменти в умовах війни і навіть розробило нові, щоб забезпечити їх сталість. Ця ініціатива стосується як державного сектору, так і громадських ініціатив.

Оцінка продуктивності поточних цифрових інструментів показала, що використання чат-ботів для поліпшення доступності медичних послуг є важливим для лікування людей, які живуть з ВІЛ. Аналіз продемонстрував, що 32% користувачів належать до цільової аудиторії, і понад 65% з них успішно використовують функцію пошуку, щоб знайти найближчий медичний заклад за допомогою чат-бота.

Отже, відповідно до опрацьованих джерел та програмних продуктів у сфері охорони психологічного здоров'я населення, можна зробити висновок, що розробка програмного забезпечення для вирішення задачі автоматизованого підбору рекомендацій документів за запитами користувачів на базі служби психологічної підтримки є актуальною та потребує подальшої автоматизації.

1.4 Мета, задачі та вимоги до реалізації інформаційної системи

Метою роботи є розробка методу рекомендації текстових документів за запитами користувачів засобами NLP, а також відповідна програмна реалізація методу для предметної області служби психологічної підтримки.

У рамках досягнення мети необхідно виконати наступні задачі:

– Провести аналіз предметної області, виділити актуальні напрями та окреслити сферу для автоматизації рекомендації документів за запитами користувачів.

– Створити метод рекомендації текстових документів за запитаннями користувачів засобами NLP та описати його кроки.

– Виконати проєктування інформаційної системи на базі методу рекомендації текстових документів за запитаннями користувачів засобами NLP для предметної області служби психологічної підтримки.

– Виконати вибір засобів розробки інформаційної системи та розробити відповідну програмну реалізацію методу.

– Виконати дослідження ефективності програмно реалізованого методу рекомендації текстових документів за запитаннями користувачів засобами NLP.

Висновки до першого розділу

У процесі написання першого розділу було виконано дослідження предметної області пошуку подібності текстових документів засобами NLP:

1. Досліджено область рекомендації документів за запитами користувачів на базі служби психологічної підтримки та виділено процеси, які потрібно автоматизувати. Оскільки через значний зріст психологічних проблем, спостерігається зменшення доступності до психічної допомоги. Особливо гостро проблема стосується військових, у яких через збройний конфлікт обмежений доступ до психологічної допомоги та медичного обслуговування. За даними експертів, від 40% до 50% населення будуть потребувати психологічної допомоги. Вченими прогнозується, що на первинному рівні медичного обслуговування буде потреба в 27 мільйонах консультацій щодо психічного здоров'я. Водночас, приблизно 3-4 мільйони українців матимуть психічні розлади різного ступеня важкості – від помірних до тяжких. Саме тому автоматизація процесів рекомендації документів за запитами користувачів на

базі служби психологічної підтримки є актуальною задачею інформаційних технологій.

2. Здійснено пошук шляхів автоматизації процесу рекомендації документів за запитами користувачів на базі служби психологічної підтримки, у рамках якого визначено доцільним використання метрик косинусної схожості та індексу Жаккара для визначення подібності запиту користувача та наявних відповідей, а для векторизації текстового подання бази відповідей буде використано метод TF-IDF.

3. Проведено аналіз існуючих програмних рішень та наукових вишукувань за темою дослідження, де у відповідності до опрацьованих джерел та програмних продуктів зокрема у сфері охорони психологічного здоров'я населення, було зроблено висновок, що розробка програмного забезпечення для вирішення задачі автоматизованого підбору рекомендацій документів за запитами користувачів на базі служби психологічної підтримки є актуальною та потребує подальшої автоматизації.

4. Виділені нерозв'язані задачі та поставлена мета відповідно до предметної області дослідження, зокрема розробки методу рекомендації текстових документів за запитаннями користувачів засобами NLP, а також відповідної програмної реалізації методу рекомендації текстових документів для предметної області служби психологічної підтримки.

Розділ 2 Метод рекомендації текстових документів за запитаннями користувачів засобами NLP

2.1 Формування навчальної вибірки даних у вигляді множини запитів та текстів відповідей

Формування навчальної вибірки даних передбачає формування навчального датасету, який містить множину типових запитів та зіставлених текстів відповідей до них.

Передбачається прикладне застосування розробленого методу для предметної області служби психологічної підтримки. Психічне здоров'я включає наше емоційне, психологічне та соціальне благополуччя. Психічне здоров'я є невід'ємною частиною здорового, збалансованого життя. Це впливає на те, як людина думає, відчуває та діє. Це також допомагає визначити, як людина справляється зі стресом, ставимося до інших і робить свій вибір. Емоційне та психічне здоров'я є важливим, оскільки воно є важливою частиною життя людини та впливає на думки, поведінку, емоційні прояви. Емоційне здоров'я може сприяти продуктивності й ефективності таких видів діяльності, як робота, навчання чи догляд.

У роботі буде використано датасет «Mental Health Conversational Data» та «Mental Health FAQ for Chatbot». «Mental Health Conversational Data» містить понад 1000 діалогів між людьми, які шукають психологічну підтримку, та фахівцями з психічного здоров'я. Датасет охоплює широкий спектр тем, включаючи тривогу, депресію, стрес, самогубство та інші проблеми з психічним здоров'ям. Містить основні розмови, поширені запитання про психічне здоров'я, розмови про класичну терапію та загальні поради, надані людям, які страждають від тривоги та депресії [13].

Цей набір даних можна використовувати для навчання моделі чат-бота, який може поводитися як терапевт, щоб надавати емоційну підтримку людям із тривогою та депресією. Вигляд даних наведено на рисунку 2.1.

```

{
  "tag": "about",
  "patterns": ["Who are you?", "What are you?", "Who you are?", "Tell me more about yourself.", "What is your name?", "What should I call you?", "What's your name?", "Tell me about y",
  "responses": ["I'm Pandora, your Personal Therapeutic AI Assistant. How are you feeling today", "I'm Pandora, a Therapeutic AI Assitant designed to assist you. Tell me about yourse
},
{
  "tag": "skill",
  "patterns": ["What can you do?"],
  "responses": ["I can provide general advice regarding anxiety and depression, answer questions related to mental health and make daily conversations. Do not consider me as a sub
},
{
  "tag": "creation",
  "patterns": ["Who created you?", "How were you made?", "How were you created?"],
  "responses": ["I was created by >.", "I was trained on a text dataset using Deep Learning & Natural Language Processing techniques", "The real question is: Who created you?"]
},
{
  "tag": "name",
  "patterns": ["My name is ", "I am name.", "I go by "],
  "responses": ["Oh nice to meet you. Tell me how was your week?", "Nice to meet you. So tell me. How do you feel today?", "That's a great name. Tell me more about yourself."]
},
{
  "tag": "help",
  "patterns": ["Could you help me?", "give me a hand please", "Can you help?", "What can you do for me?", "I need support", "I need help", "Support me please"],
  "responses": ["Sure. Tell me how can i assist you", "Tell me your problem so that i can assist you", "Yes, sure. How can I help you?"]
},
{
  "tag": "sad",
  "patterns": ["I am feeling lonely", "I am so lonely", "I feel down", "I feel sad", "I am sad", "I feel so lonely", "I feel empty", "I don't have anyone"],
  "responses": ["I'm sorry to hear that. I'm here for you. Talking about it might help. So, tell me why do you think you're feeling this way?", "I'm here for you. Could you tell me wh
},
{
  "tag": "stressed",
  "patterns": ["I am so stressed out", "I am so stressed", "I feel stuck", "I still feel stressed", "I am so burned out"],
  "responses": ["What do you think is causing this?", "Take a deep breath and gather your thoughts. Go take a walk if possible. Stay hydrated", "Give yourself a break. Go easy on
},
{
  "tag": "worthless",
  "patterns": ["I feel so worthless.", "No one likes me.", "I can't do anything.", "I am so useless", "Nothing makes sense anymore"],
  "responses": ["It's only natural to feel this way. Tell me more. What else is on your mind?", "Let's discuss further why you're feeling this way.", "I first want to let you know
},
{
  "tag": "depressed",
  "patterns": ["I can't take it anymore", "I am so depressed", "I think i'm depressed.", "I have depression"],
  "responses": ["It helps to talk about what's happening. You're going to be okay", "Talk to me. Tell me more. It helps if you open up yourself to someone else.", "Sometimes when
},
}

```

Рисунок 2.1 – Набір даних для дослідження

Набір даних містить так звані наміри, прогнозовані майбутні дії, які можуть стати реальними. Наприклад, якщо сказати «я сумний», намір у цьому випадку був би «сумний». Залежно від наміру існує набір шаблонів та відповідей, які відповідають наміру. Під шаблонами розуміється набір прикладів повідомлень користувача, які відповідають намірам, тоді як відповіді – це відповіді, які програма надає відповідно до заданих намірів. Визначаються різні наміри, а їхні шаблони та відповіді використовуються як навчальні дані моделі для визначення конкретного наміру.

Набір даних «Mental Health FAQ for Chatbot» складається з поширених запитань про психічне здоров'я, а саме:

- Види психічних захворювань.
- Симптоми психічних захворювань.
- Причини психічних захворювань.
- Лікування психічних захворювань.
- Ресурси для людей з психічними захворюваннями.

Набір даних містить 98 запитань та відповідей на них. Запитання стосуються різних аспектів психічного здоров'я, таких як депресія, тривога,

біполярний розлад, шизофренія, посттравматичний стресовий розлад (ПТСР) та інші.

Набір даних був зібраний з різних джерел, включаючи вебсайти, блоги та статті про психічне здоров'я. Дані були очищені та відредаговані, щоб забезпечити їхню точність та актуальність.

Приклад даних з набору зображено на рисунку 2.2.

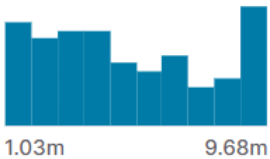
Question_ID	Questions	Answers
unique id of each question	question related to mental health	corresponding answer
	98 unique values	98 unique values
1590140	What does it mean to have a mental illness?	Mental illnesses are health conditions that disrupt a person's thoughts, emotions, relationships, ...
2110618	Who does mental illness affect?	It is estimated that mental illness affects 1 in 5 adults in America, and that 1 in 24 adults have a...
6361820	What causes mental illness?	It is estimated that mental illness affects 1 in 5 adults in America,

Рисунок 2.2 – Дані датасету «Mental Health FAQ for Chatbot» [14]

Оскільки запропонований метод повинен працювати з україномовним текстом, дані датасети для подальшого використання буде об'єднано та перекладено на українську мову.

Так як структура датасетів мала різні теги, дані було об'єднано у формат .json з розміткою, наведеною на рисунку 2.3

```

{
  "tag": "допомога",
  "patterns": [
    "Не могли б ви мені допомогти?",
    "дайте мені руку, будь ласка",
    "Ви можете допомогти?",
    "Що ти можеш зробити для мене?",
    "Мені потрібна підтримка",
    "Мені потрібна допомога",
    "Підтримай мене, будь ласка"
  ],
  "responses": [
    "Звичайно. Розкажіть, чим я можу вам допомогти",
    "Розкажіть про свою проблему, щоб я міг вам допомогти",
    "Так, звичайно. Чим я можу вам допомогти?"
  ]
},
{
  "tag": "смуток",
  "patterns": [
    "Я відчуваю себе самотнім",
    "Я така самотня",
    "Я почуваюся пригніченим",
    "Я почуваюсь засмученим",
    "Мені сумно",

```

Рисунок 2.3 – Приклад об'єднаних даних

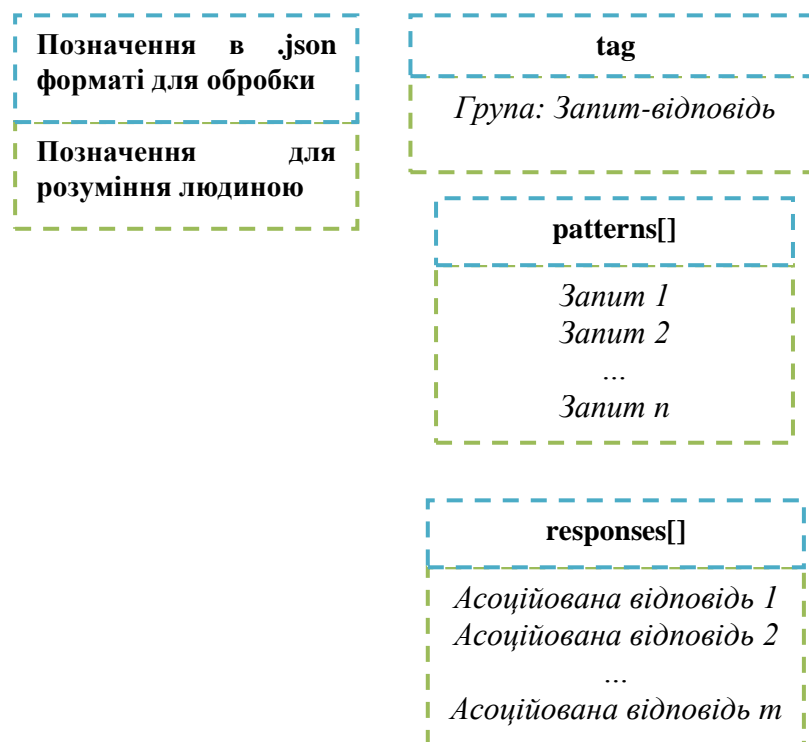


Рисунок 2.4 – Типова конструкція елемента «запит-відповідь»

Відповідно, дані представляють собою певні набори, де підтеми об'єднано в групи типу «Запит-відповідь». На рисунку 2.4 наведена типова конструкція елемента «запит-відповідь» для подальшого використання програмними засобами.

Отже, наведеним чином було підготовлено набір даних на основі датасетів «Mental Health Conversational Data» та «Mental Health FAQ for Chatbot». «Mental Health Conversational Data» містить понад 1000 діалогів між людьми, які шукають психологічну підтримку, та фахівцями з психічного здоров'я та охоплює широкий спектр тем, включаючи тривогу, депресію, стрес, самогубство та інші проблеми з психічним здоров'ям, а «Mental Health FAQ for Chatbot» даних містить 98 запитань та відповідей на них. Запитання, як в попередньому датасеті, стосуються різних аспектів психічного здоров'я, таких як депресія, тривога, біполярний розлад, шизофренія, посттравматичний стресовий розлад (ПТСР) та інші. Описані дані були перекладені на українську мову та будуть використовуватись в рамках створення методу рекомендації текстових документів за запитаннями користувачів засобами NLP.

2.2 Метрики для знаходження рекомендації текстових документів за запитаннями користувачів

Для знаходження рекомендації текстових документів за запитаннями користувачів можуть бути використані різноманітні метрики. За результатами аналізу, виконаному у першому розділі, було обрано для використання метрику косинусної схожості та індексу Жаккара.

Основна ідея косинусної схожості полягає в тому, що якщо два вектори спрямовані в одному напрямку (кут між ними дорівнює 0 градусів), то їхні вектори мають найвищу косинусну схожість та вважаються дуже схожими. Навпаки, якщо кут між векторами дорівнює 90 градусів, то косинусна схожість дорівнює 0, що вказує на максимальну різницю між ними.

Для обчислення косинусної схожості між двома векторами тексту (наприклад, словниковими представленнями документів), використовується наступна формула [15]:

$$\text{Cosine Similarity } (A, B) = (A \cdot B) / (||A|| * ||B||),$$

де A та B – вектори тексту або документів, $A \cdot B$ – скалярний добуток векторів A і B , $||A||$ та $||B||$ - норми (довжини) векторів A і B .

Значення косинусної схожості зазвичай лежать в діапазоні від -1 (повна протилежність) до 1 (повна ідентичність), де 0 означає відсутність схожості.

Цей метод широко використовується в пошукових системах, рекомендаційних системах, аналізі тексту та класифікації документів для визначення ступеня схожості між текстами та відбору найбільш релевантних результатів. У роботі він буде застосований як одна із метрик подібності запиту користувача до наявних шаблонів запитів з метою отримання релевантної відповіді.

Косинусна подібність є мірою, яка використовується для обчислення схожості між двома векторами у багатовимірному просторі, вимірюючи косинус кута між ними [16, 17]. Ця метрика є особливо корисною у сферах, де вектори представляють текстові документи або фрази, перетворені в числовий формат, наприклад, через процеси, такі як TF-IDF векторизація. Косинусна подібність дозволяє визначити, наскільки два документи або текстові вирази є схожими з точки зору їх змісту, що є фундаментальним у системах рекомендацій, кластеризації текстів та інших застосунках NLP.

Індекс Жаккара (Jaccard index), також відомий як Жаккардова схожість чи коефіцієнт Жаккара, є метрикою схожості, яка використовується для порівняння множин елементів. Ця метрика визначає ступінь схожості двох множин шляхом вимірювання кількості спільних елементів в обох множинах відносно загальної кількості унікальних елементів.

Індекс Жаккара обчислюється за наступною формулою:

$$J(A, B) = |A \cap B| / |A \cup B|,$$

де $J(A, B)$ - індекс Жаккара між множинами A і B , $|A \cap B|$ - кількість спільних елементів між множинами A і B , $|A \cup B|$ - кількість унікальних елементів у множинах A і B разом.

Значення Індексу Жаккара можуть лежати в діапазоні від 0 до 1, де 0 означає відсутність спільних елементів, а 1 означає повну ідентичність множин. Значення, близькі до 1, вказують на високу схожість множин, тоді як значення, близькі до 0, свідчать про низьку схожість.

Індекс Жаккара широко використовується в різних галузях, включаючи аналіз тексту, рекомендаційні системи, пошукові системи та біологічні дослідження [18]. В роботі буде використовуватися для визначення схожості між ключовими словами користувачького запиту та наявних рекомендацій.

Обидві ці метрики надають цінну інформацію щодо структурної та змістовної схожості в даних, і їх використання дозволяє підвищити точність та ефективність алгоритмів обробки даних та систем штучного інтелекту.

Отже, для пошуку рекомендацій текстових документів будуть використовуватись розглянуті метрики косинусної схожості та індекс Жаккара, які вказують на схожість між текстовими представленнями документів.

2.3 Підхід до пошуку рекомендації текстових документів за запитаннями користувачів

Загальна схема підходу до пошуку рекомендації текстових документів за запитаннями користувачів наведена на Рисунку 2.5. Користувач пише свій запит, що проходить попередню обробку тексту, і який порівнюється із наявними запитами у групах асоціативних запитів Q_1, \dots, Q_n , де $Q_n = \{q_1, q_2, \dots, q_n\}$, а q_1, q_2, \dots, q_n є відповідними запитами. За ключовими словами користувачького та найближчого за схожістю асоційованого запиту з групи запитів підбирається асоційована рекомендація з групи асоційованих рекомендації R_1, \dots, R_n , де $R_n = \{r_1, r_2, \dots, r_n\}$, а r_1, r_2, \dots, r_n є відповідними асоційованими відповідями.

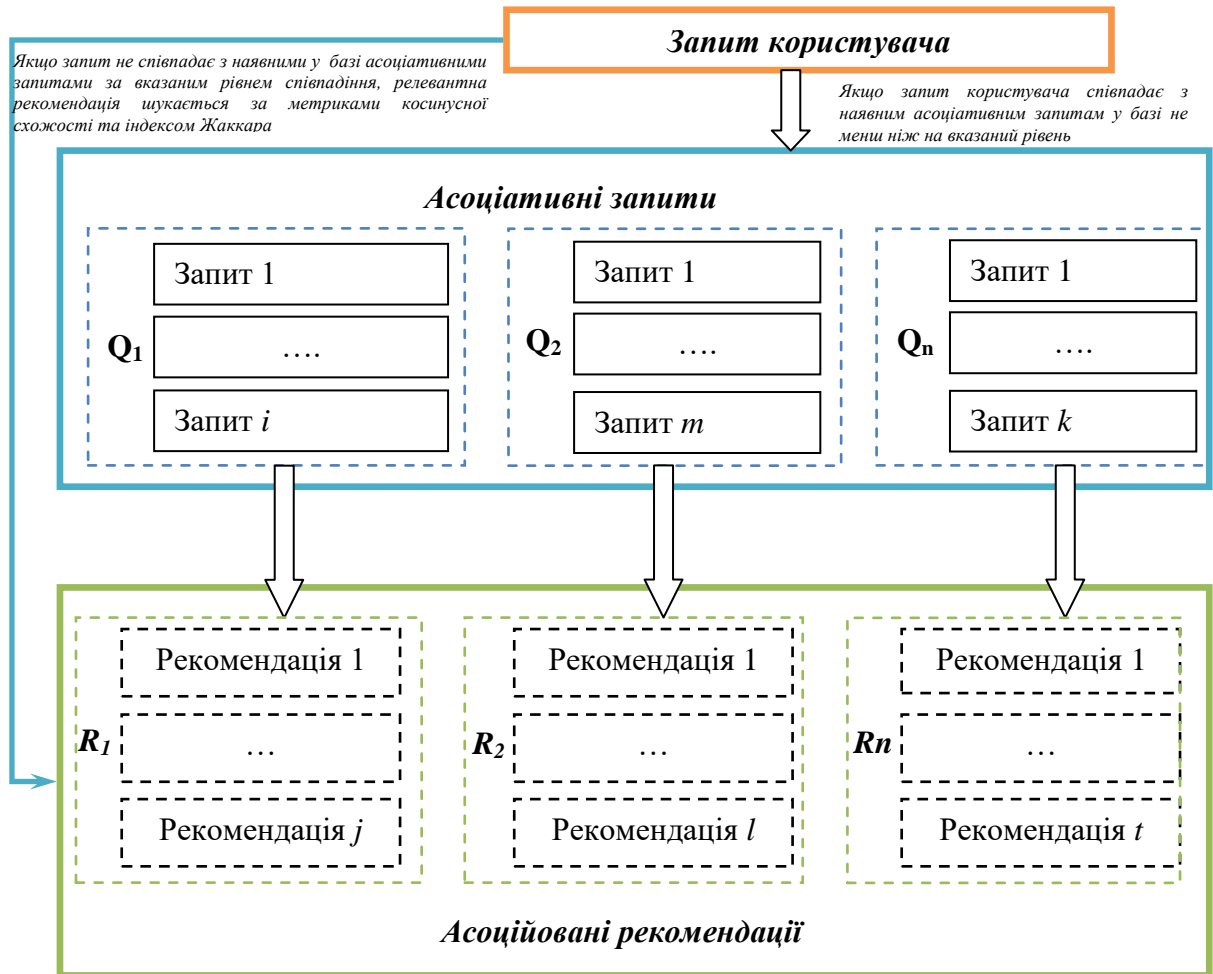


Рисунок 2.5 – Схема пошуку рекомендації текстових документів за запитаннями користувачів

У випадку відсутності співпадінь на зазначеному рівні за ключовими словами, шукається релевантна рекомендація за метриками косинусної схожості та індексом Жаккара.

Таким чином, було схематично проілюстровано пошук рекомендації текстових документів за запитаннями користувачів.

2.4 Етапи методу рекомендації текстових документів за запитаннями користувачів

Метод рекомендації текстових документів за запитаннями користувачів призначений для пошуку релевантних текстових документів з навчального

датасету (множини текстів запитів та відповідей до них) до користувацького запиту з використанням засобів обробки природної мови.

Вхідними даними методу пошуку рекомендації текстових документів за запитаннями користувачів засобами NLP є набір даних з групами асоційованих запитів та відповідних їм асоційованих відповідей, а також користувацький запит.

Схема попередньої обробки набору даних наведена на рисунку 2.6. У процесі попередньої обробки датасету відбувається робота з множиною асоціативних запитів, що включає видалення стоп-символів та стоп слів, переведення у нижній регістр, лематизація.



Рисунок 2.6 – Схема підготовки вхідних даних для методу пошуку рекомендації текстових документів за запитаннями користувачів

Після попередньої обробки множини запитів, здійснюється формування та векторизація корпусу із множини асоційованих відповідей. Спершу

відбувається об'єднання множини асоційованих відповідей у один корпус, який у подальшому векторизується методом TF-IDF.

Вихідними даними етапу підготовки вхідних даних методу пошуку рекомендації текстових документів за запитаннями користувачів є датасет із набором очищених лематизованих асоціативних запитів та векторизоване представлення корпусу у вигляді матриці TF-IDF.

Загальна схема методу рекомендації текстових документів за запитаннями користувачів на базі служби психологічної підтримки призначений для пошуку релевантних рекомендацій з бази текстових порад за текстовим запитом користувача наведена на рисунку 2.7.

Початковим етапом методу пошуку рекомендації текстових документів є попередня обробка користувацького запитання, що включає видалення стоп-символів, стоп слів, переведення слів у нижній регістр, лематизація користувацького запиту.

Наступним етапом є пошук запитання користувача у асоціативних запитах. Здійснюється він шляхом обрахунку оцінки семантичної подібності лематизованого користувацького запиту із кожним лематизованим асоціативним запитом із бази запитань.

Оцінка семантичної подібності P_i i -го користувацького запиту A , що є лематизованою очищеною від стоп-слів та стоп-символів множиною слів користувацького запиту до наявного асоціативного запиту B_i , що є i -ою лематизованою очищеною від стоп-слів та стоп-символів множиною слів, що належить множині асоційованих запитів B до деякої асоційованої групи рекомендацій визначається наступним чином:

$$P_i = \frac{\{A\} \cap \{B_i\}}{\text{len}(\{A\})}, i = \overline{1, n} \quad (2.1)$$

де n – кількість запитів у групі асоційованих запитів, $\text{len}(\{A\})$ - довжина лематизованої очищеної від стоп-слів та стоп-символів множини слів.



Рисунок 2.7 – Схема методу пошуку рекомендації текстових документів за запитаннями користувачів

Результуюча оцінка для вибору рекомендації визначається як максимальна з обрахованих оцінок P_i :

$$P = (\max P_i) \cdot 100\%, i = \overline{1, n} \quad (2.2)$$

Запит із найбільшою кількістю співпадінь (проте не менше 60%) вважається релевантний і здійснюється пошук відповідної рекомендації із множини асоційованих відповідей, що відповідають найбільшому співпадінню з запитом.

Якщо ж оцінка співпадіння із асоціативними запитами менше 60 %, здійснюється векторизація користувацького запиту за допомогою TF-IDF та

відбувається знаходження найближчої релевантної рекомендації за зваженою оцінкою значень метрик косинусної схожості та індексом Жаккара. Зважена оцінка визначається так:

$$WS = \alpha \cdot \text{CosSimilarity} + \beta \cdot \text{JaccardIndex} \quad (2.3)$$

де α та β - вагові коефіцієнти для значень косинусної подібності та індексу Жаккара. По замовчуванню рівні 0,5, проте у рамках роботи їх вплив буде досліджуватись.

Вихідними даними методу пошуку рекомендації текстових документів за запитаннями користувачів є рекомендація текстового документа, що релевантна до користувацького запитання.

Отже, таким чином був розроблений метод рекомендації текстових документів за запитаннями користувачів, що призначений для швидкого знаходження релевантних рекомендацій із бази асоційованих відповідей за текстовим запитом користувача та вхідними даними має датасет із набором асоціативних запитів та відповідей до них та користувацький запит, що перетворює їх у вихідні дані у форматі текстової рекомендації семантично подібної до запиту.

Висновки до другого розділу

У рамках виконання другого розділу кваліфікаційної роботи магістра, що присвячений розробці методу рекомендації текстових документів за запитаннями користувачів засобами NLP, було одержано результати:

1. Підготовлено набір даних на основі англomовних датасетів «Mental Health Conversational Data» та «Mental Health FAQ for Chatbot», які було перекладено на українську мову та об'єднано в структуру, зручну для подальшої обробки. Датасети присвячені тематиці психічного здоров'я та охоплюють широкий спектр тем, включаючи тривогу, депресію, стрес, самогубство, ПТСР та інші проблеми з психічним здоров'ям.

2. Були обрані метрики для пошуку рекомендацій текстових документів, а саме: будуть використовуватись метрики косинусної схожості та індекс Жаккара, які вказують на схожість між текстовими представленнями документів. Вказані відповідні шляхи їх пошуку.

3. Було наведено загальну схему підходу до пошуку рекомендації текстових документів за запитаннями користувачів, що базується на пошуку семантичної схожості запиту користувача і асоційованого запиту з набору даних. У випадку відсутності співпадінь на зазначеному рівні за ключовими словами, шукається релевантна рекомендація за зваженою оцінкою, що базується на використанні метрик косинусної схожості та індексом Жаккара з заданими користувачем ваговими коефіцієнтами.

4. Було розроблено метод рекомендації текстових документів за запитаннями користувачів, що призначений для пошуку релевантних рекомендацій з бази текстових порад за текстовим запитом користувача та вхідними даними має набір даних з групами асоційованих запитів та відповідних їм асоційованих відповідей та користувацький запит, що перетворює їх у вихідні дані у вигляді текстової рекомендації релевантної до запиту.

Розділ 3 Інформаційна система автоматизованого визначення рекомендацій текстових документів за користувацькими запитами

3.1 Схема інформаційної системи

Інформаційна система автоматизованого визначення рекомендацій текстових документів за користувацькими запитами є прикладною програмною реалізацією методу пошуку рекомендації текстових документів за запитом користувачів, що призначений для пошуку релевантних рекомендацій з бази текстових порад за текстовим запитом користувача та вхідними даними має набір даних з групами асоційованих запитів та відповідних їм асоційованих відповідей та користувацький запит, що перетворює їх у вихідні дані у вигляді текстової рекомендації релевантної до запиту.

Загальна схема інформаційної структури системи автоматизованого визначення рекомендацій текстових документів наведена на рисунку 3.1. Інформаційна структура системи складається із набору даних та трьох підсистем: «Підсистеми роботи з корпусами текстів», «Підсистеми редактора навчальної вибірки документів запитів-відповідей» та головної «Підсистеми автоматизованого підбору документів відповідей до запитів».

Підсистеми роботи з корпусами текстів призначена для роботи з корпусом текстів. Має ряд функцій, зокрема видалення стоп-символів з корпусу, видалення стоп-слів з корпусу, лематизація текстових даних, векторизація текстових даних методом TF-IDF, а також збереження результатів обробки корпусу.

Набір даних .JSON включає в себе дані на основі англійських датасетів «Mental Health Conversational Data» та «Mental Health FAQ for Chatbot», які було перекладено на українську мову та об'єднано в структуру, зручну для подальшої обробки. Предметна область наборів даних присвячена тематиці психічного здоров'я та охоплює широкий спектр тем, включаючи тривогу, депресію, стрес, самогубство, ПТСР та інші проблеми з психічним здоров'ям.

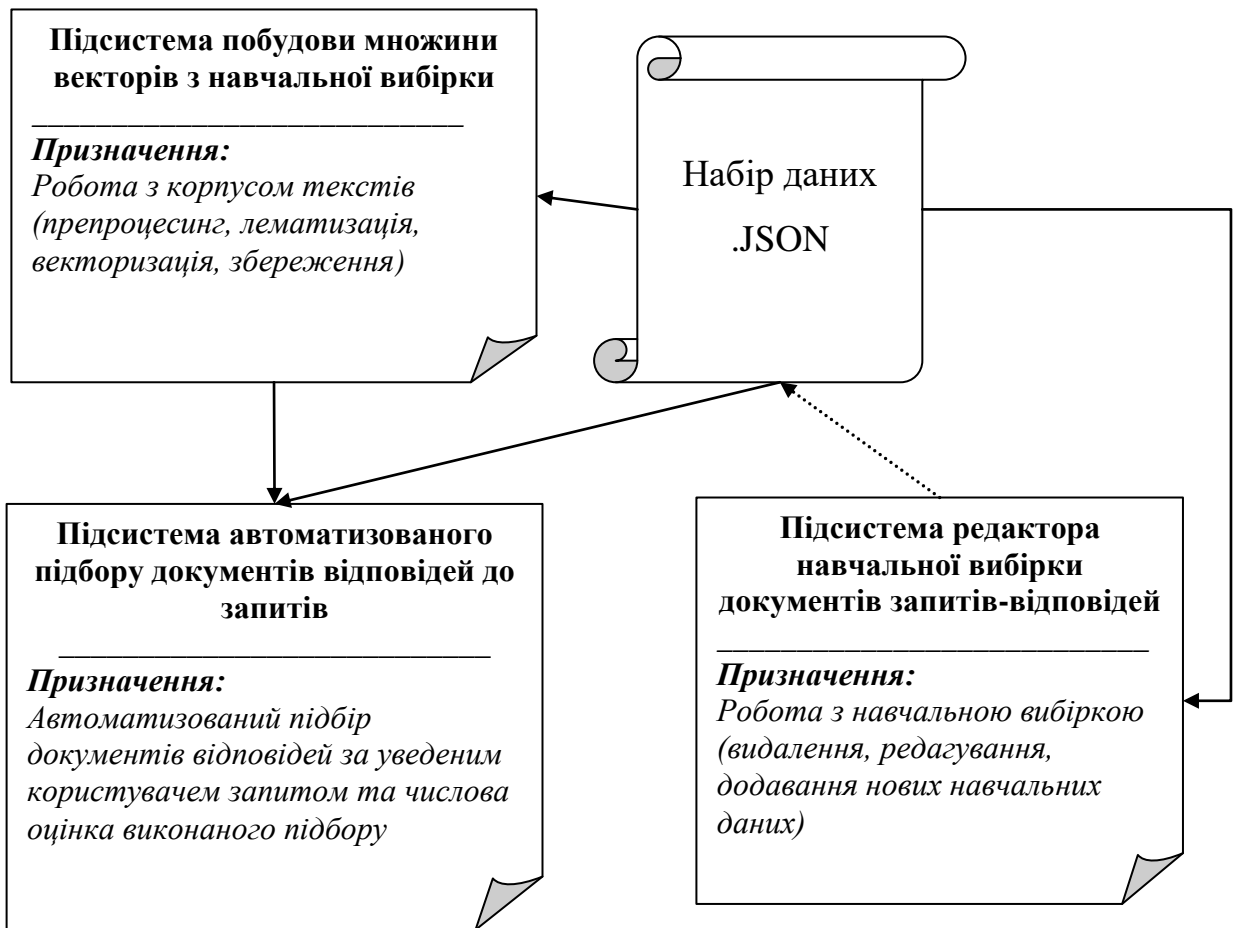


Рисунок 3.1 – Інформаційна структура системи автоматизованого визначення рекомендацій текстових документів

Підсистема редактора навчальної вибірки документів запитів-відповідей призначена для роботи з навчальною вибіркою, а саме для роботи з групами користувацьких запитів-відповідей (видалення, редагування, додавання нового) користувацькими запитом (видалення, редагування, додавання нового) та асоціативними відповідями навчальної вибірки (видалення, редагування, додавання нових навчальних асоціативних відповідей). Змінені дані записуються у набір даних .JSON.

Підсистема автоматизованого підбору документів відповідей до запитів є головною підсистемою інформаційної системи, та має призначення виконувати автоматизований підбір документів відповідей за уведеним користувачем запитом та показувати числову оцінку виконаного підбору.

Отже, було спроектовано інформаційну структуру системи автоматизованого визначення рекомендацій текстових документів на основі методу пошуку рекомендації текстових документів за запитам користувачів, що за вхідними даними датасету з групами асоційованих запитів та відповідних їм асоційованих відповідей та користувацького запиту перетворює їх у вихідні дані у вигляді текстової рекомендації релевантної до запиту.

3.2 Схема та функції підсистеми формування навчальної вибірки документів запитів-відповідей

Підсистема формування навчальної вибірки документів запитів-відповідей, що основним призначенням має роботу з навчальною вибіркою, наведена на рисунку 3.2.

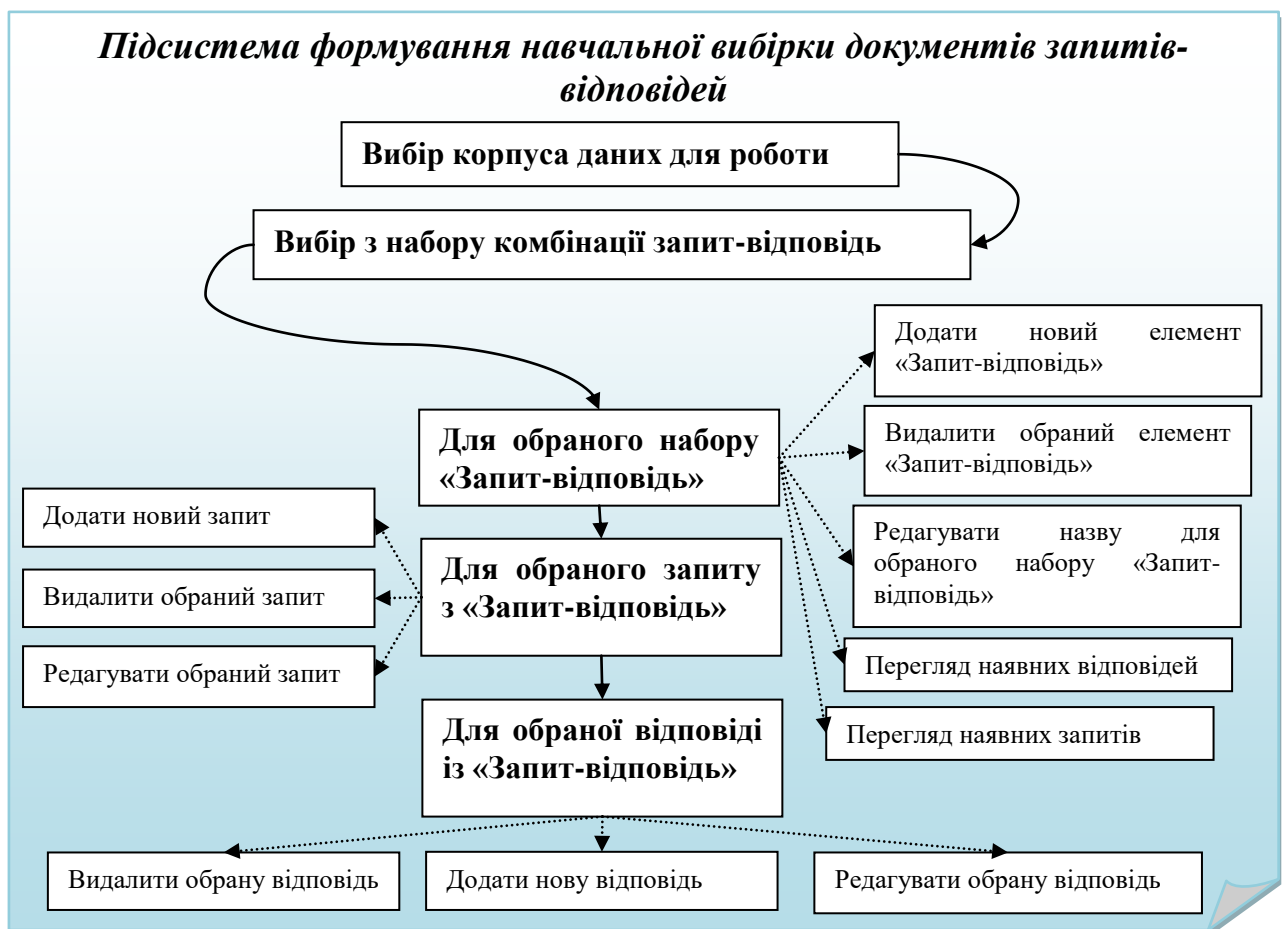


Рисунок 3.2 – Схема та функції підсистеми формування навчальної вибірки документів

Першою функцією у даній підсистемі є вибір корпусу даних для роботи. В залежності від обраного корпусу даних надалі можна обирати комбінацію «Запит-відповідь», яка має групу функцій, таких як можливість відредагувати, видалити або ж можна додати до навчальної вибірки нову комбінацію «Запит-відповідь», можливість перегляду наявних запитів щодо обраної комбінації «Запит-відповідь» та можливість перегляду наявних асоціативних відповідей щодо обраної комбінації «Запит-відповідь».

Наступною групою функцій є функції роботи з обраним з користувачьких запитів можна здійснити редагування або ж видалити обраний запит. Також є можливість додати новий запит до комбінації «Запит-відповідь».

Для обраної відповіді із комбінації «Запит-відповідь» є група функцій з можливості редагування обраної відповіді, можливості видалення обраної відповіді або ж можливості додавання нової відповіді в множину асоційованих відповідей.

Отже, спроектовано функціональні особливості підсистеми редактора навчальної вибірки документів, що є складовою інформаційної системи автоматизованого визначення рекомендацій текстових документів на основі методу пошуку рекомендації текстових документів за запитами користувачів.

3.3 Схеми та функції підсистеми автоматизованого підбору документів відповідей до запитів

Підсистема автоматизованого підбору документів відповідей до запитів є основною підсистемою інформаційної системи автоматизованого визначення рекомендацій текстових документів за користувачькими запитами, та має основне призначення виконувати автоматизований підбір документів відповідей за уведеним користувачем запитом, а також показувати числову оцінку виконаного підбору. Схеми та основні функції підсистеми наведено на рисунку 3.3.



Рисунок 3.3 – Схема та функції підсистеми автоматизованого підбору документів відповідей до запитів

Підбір релевантних документів щодо запиту користувача здійснюється на основі обраного корпусу документів для роботи.

Наступною функцією підсистеми після вибору корпусу є функція введення користувацького запиту, який можна ввести як вручну, так і обрати з переліку існуючих.

Наступною функцією є функція підбору текстової рекомендації на основі введеного запиту та обраного корпусу. Дана функція автоматично обирає метод, яким відповідь буде більш точна з двох підходів – пошуку семантичної подібності за ключовими словами, а якщо даний метод показує результат нижче

60%, пошук семантичної подібності здійснюється шляхом зваженої оцінки подібності з використанням індексу Жаккара і косинусної схожості.

Після здійсненого підбору релевантної відповіді вона буде відображена користувачу. Також можна подивитись схожі відповіді, що підібрані за спадом максимальної оцінки подібності.

Наступною функцією є можливість подивитись оцінки та метод, за яким були знайдені релевантні відповіді до запиту користувача.

Отже, було побудовано схему та розглянуто основні функції підсистеми автоматизованого підбору документів відповідей до запитів, яка є головною підсистемою проектованої інформаційної системи.

3.4 Формування комбінації засобів розробки інформаційної системи

Для майбутньої реалізації прикладного програмного забезпечення на базі методу рекомендації текстових документів за запитаннями користувачів засобами NLP для предметної області служби психологічної підтримки, необхідно обрати мову програмування, додаткові засоби та середовище розробки програмного коду.

Python, завдяки своїй універсальності та високому рівню абстракції, вважається однією з найпривабливіших мов програмування для розробки програмного забезпечення, що використовує штучний інтелект. Основними причинами цього є легкість у вивченні, чистота синтаксису та здатність до глибокої інтеграції з іншими мовами програмування. Ці фактори сприяють не лише ефективному навчанню нових спеціалістів у галузі ШІ, але й забезпечують гнучкість у розробці складних систем.

Завдяки своїм властивостям, Python підтримує швидку розробку і прототипування, що є критично важливим у світі штучного інтелекту, де швидкість інновацій є ключовим фактором. Його гнучкість у підтримці різних парадигм програмування дозволяє розробникам вибирати найбільш підходящі

підходи для конкретних задач, сприяючи таким чином оптимізації процесу розробки.

Читабельний та інтуїтивно зрозумілий синтаксис Python сприяє підвищенню ефективності командної розробки та зменшує витрати часу на звикання до коду, що особливо важливо при роботі над великими проектами. Активна спільнота розробників, що постійно розширюється, забезпечує підтримку та постійне оновлення широкого спектру бібліотек і фреймворків, що охоплюють практично всі аспекти штучного інтелекту, від машинного навчання до обробки природної мови.

Python також відомий своєю спроможністю ефективно інтегруватися з іншими мовами програмування, такими як C або C++, що дозволяє досягти високої продуктивності та оптимізувати використання системних ресурсів. Така інтеграція є особливо важливою для розробки високопродуктивних програмних рішень в сфері ШІ, де часто виникає необхідність обробки великих обсягів даних.

Python ефективно використовується для роботи з великими наборами даних та масштабованими системами, що є ключовим аспектом у багатьох додатках ШІ. Завдяки бібліотекам для обробки даних та паралельного виконання, Python може ефективно обробляти великі обсяги інформації.

Python має велику кількість бібліотек і фреймворків, які спрощують реалізацію складних алгоритмів ШІ та машинного навчання. Бібліотеки як TensorFlow, PyTorch, Keras, Scikit-learn надають готові до використання, оптимізовані алгоритми, які можуть значно скоротити час розробки. Також Python підтримує різні парадигми програмування, які дозволяють розробникам вибрати найбільш ефективний підхід для кожного конкретного проекту. Ця гнучкість є важливою для швидкого прототипування та ітеративного розвитку проектів ШІ [20].

Одним із найкращих середовищ написання програмного коду є PyCharm [21]. PyCharm є інтегрованим середовищем розробки (IDE – Integrated Development Environment), створеним компанією JetBrains, що спеціалізується на

створенні інструментів для програмістів. Як IDE, PyCharm забезпечує комплексний підхід до розробки, особливо для мови програмування Python, включаючи розробку додатків, що використовують штучний інтелект та машинне навчання [22].

PyCharm пропонує розширені можливості для редагування коду Python, включаючи підсвічування синтаксису, автоматичне завершення коду, перевірку коду на наявність помилок та попередження в реальному часі. PyCharm включає інструменти для розробки, такі як інтегрований дебагер, інструменти для профілювання, а також підтримку різних систем управління версіями коду, включаючи Git. Середовище може інтегруватися з великою кількістю інших інструментів та сервісів, що робить його вигідним у складних і багатошарових проектах.

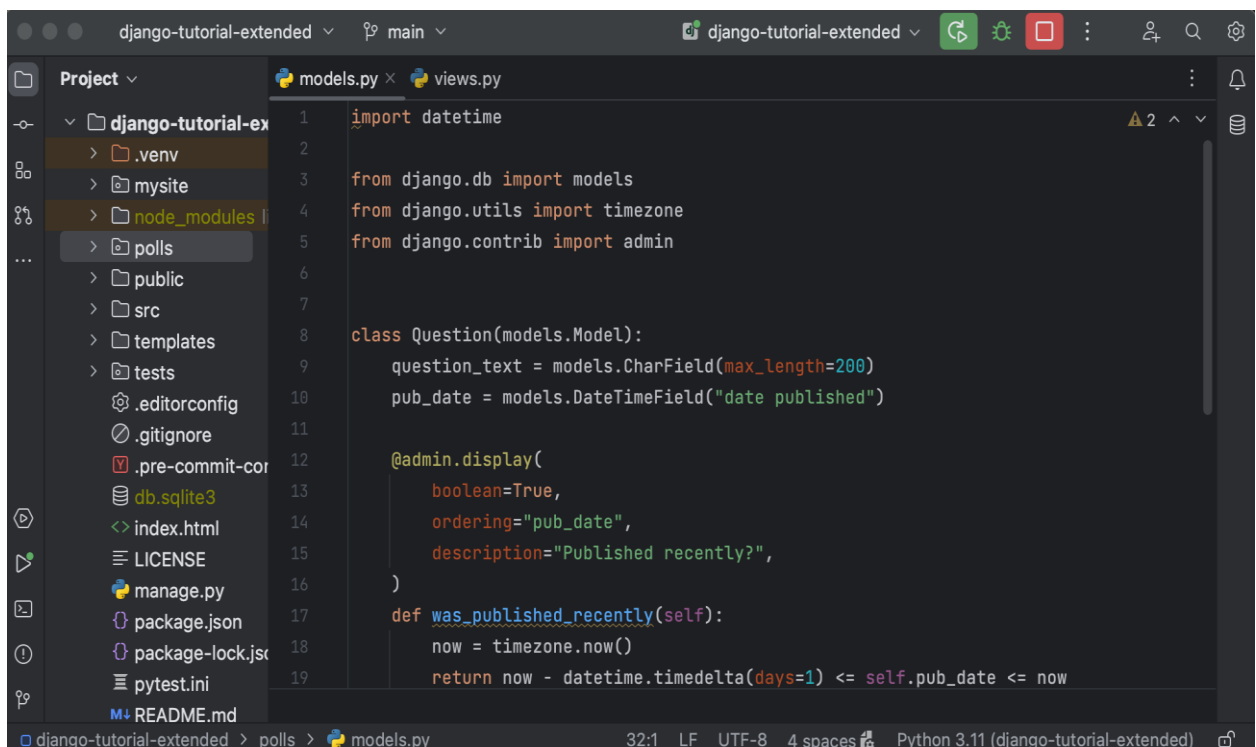


Рисунок 3.5 – Вигляд інтерфейсу середовища розробки програмного коду PyCharm [23]

PyCharm ефективно спрощує та автоматизує багато аспектів розробки на Python, дозволяючи розробникам зосередитись на логіці та архітектурі програмного продукту, замість рутинних завдань. Підтримка наукової роботи та

великого набору інструментів роблять PyCharm особливо привабливим для дослідників, інженерів-програмістів та розробників в області штучного інтелекту та машинного навчання. Вигляд інтерфейсу середовища розробки наведено на рисунку 3.5.

Для розробки програмного застосунку на базі методу рекомендації текстових документів за запитаннями користувачів засобами NLP, було обрано мову програмування Python та інтегроване середовище розробки PyCharm. Такий вибір зумовлений низкою переваг цих інструментів, які оптимально відповідають потребам проекту. Python, завдяки своїй гнучкості, багатій екосистемі бібліотек для NLP та машинного навчання, є ідеальним вибором для створення складних систем обробки даних. PyCharm, в свою чергу, надає розширені можливості для ефективної розробки та налагодження коду, що значно підвищує продуктивність робочого процесу. Використання цих інструментів дозволяє створити надійний та функціональний застосунок, який ефективно задовольнить запити користувачів за допомогою методів NLP.

3.5 Вибір спеціалізованих програмних розширень

Для реалізації програмного застосунку, заснованого на методах рекомендації текстових документів у відповідності до запитань користувачів за допомогою технологій обробки природної мови (NLP), важливим є вибір відповідного інструменту для створення графічного інтерфейсу користувача (GUI).

PyQt5 [24,25], як інтегрований фреймворк для створення GUI в Python, має ряд переваг, що відповідають вимогам сучасних програмних продуктів, особливо тих, що залучають обробку природної мови. Він надає широкий спектр віджетів та інструментів для створення інтуїтивно зрозумілих, візуально привабливих інтерфейсів, що є ключовим для забезпечення зручного користувацького досвіду. Функціональність PyQt5, що включає підтримку подій, налаштовуваних віджетів, інтеграції з базами даних та інших важливих

компонентів, забезпечує гнучкість та широкі можливості для реалізації складних функціональних вимог. Приклад використання PyOt5 наведено на рисунку 3.6.

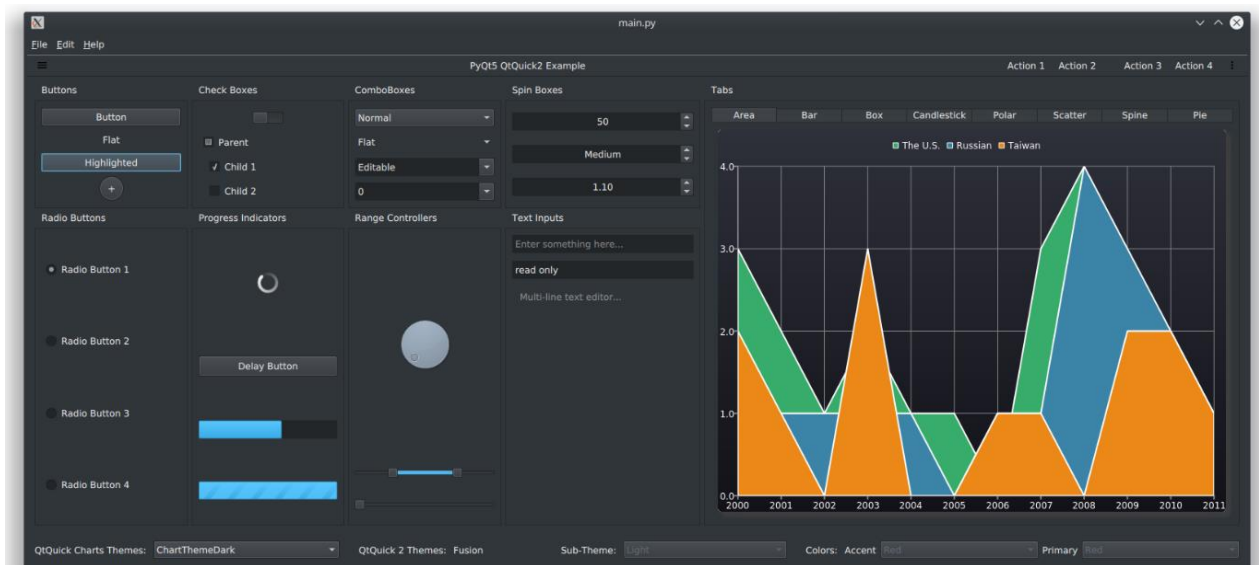


Рисунок 3.6 – Приклад використання бібліотеки PyOt5 для реалізації GUI [26]

Додатково, PyQt5 підтримує крос-платформність, що дозволяє розробникам створювати застосунки, які можуть працювати на різних операційних системах без значних змін у коді. Ця особливість є важливою для розширення аудиторії користувачів та забезпечення універсальності програмного продукту.

Підсумовуючи, вибір PyQt5 для розробки GUI в контексті програмного застосунку на базі методів NLP для рекомендації текстових документів є обґрунтованим та ефективним, оскільки він сприяє створенню високоякісного інтерфейсу, оптимізує процес розробки та водночас забезпечує високий рівень задоволення кінцевих користувачів.

Також, одним з кроків методу є лематизація та препроцесинг, тож необхідно обрати відповідні бібліотеки та засоби для вирішення цієї задачі. Natural Language Toolkit (NLTK) є провідною бібліотекою Python, яка використовується для роботи з природною мовою. Ця бібліотека широко використовується в області обробки природної мови (NLP) для різноманітних завдань, включаючи лематизацію та препроцесинг тексту [27, 28].

Лематизація в NLTK включає процес перетворення слова до його базової форми (леми) [29, 30]. На відміну від стемінгу [31, 32], який просто відкидає закінчення слова, лематизація звертає увагу на контекст та використовує морфологічний аналіз для зведення слова до його канонічної форми. Наприклад, слова "running", "runs" будуть лематизовані до "run" [33].

NLTK надає клас WordNetLemmatizer [34], який використовує WordNet – велику лінгвістичну базу даних англійської мови – для лематизації. Цей клас дозволяє точно ідентифікувати канонічну форму слова, враховуючи його частину мови.

Препроцесинг тексту за допомогою бібліотеки NLTK є фундаментальним кроком у процесі обробки природної мови, де виконується ряд операцій для оптимізації тексту перед його аналізом. До цього процесу належать токенізація, видалення стоп-слів та визначення частини мови [35].

Токенізація включає розділення тексту на основні компоненти - слова або фрази. NLTK надає різноманітні токенізатори, які дозволяють ефективно розділяти текст на індивідуальні токени, будь то слова чи речення, забезпечуючи таким чином базу для подальшого аналізу.

Стоп-слова – це слова, які часто вживаються в мові, але не несуть важливої інформації для розуміння контексту (наприклад, "і", "але", "на"). Видалення таких слів з тексту допомагає сконцентруватися на більш значущих словах та зменшує рівень шуму в даних [36].

Визначення частини мови (POS tagging) полягає у визначенні частини мови для кожного слова в тексті (наприклад, іменник, дієслово, прикметник). POS tagging є важливим для розуміння структури речень та може бути використаний для подальшого синтаксичного та семантичного аналізу [37].

NLTK забезпечує інструменти для кожного з цих етапів препроцесингу, роблячи її надзвичайно цінною бібліотекою для комплексної підготовки тексту. Ці інструменти дозволяють розробникам та дослідникам ефективно готувати текстові дані до більш глибокого аналізу та обробки, що є ключовим аспектом у багатьох застосунках NLP [38].

Для векторизації було обрано `scikit-learn`, що є однією з найпопулярніших бібліотек машинного навчання для мови програмування Python. Вона надає широкий спектр інструментів для статистичного моделювання та машинного навчання, включаючи різні алгоритми класифікації, регресії, кластеризації, а також засоби для препроцесингу даних і вимірювання якості моделей [39].

Однією з ключових особливостей `scikit-learn` є її здатність виконувати векторизацію тексту, зокрема за допомогою TF-IDF (Term Frequency-Inverse Document Frequency). Векторизація TF-IDF є методом перетворення тексту в числову форму, яка може бути використана машинними алгоритмами.

В TF-IDF, "Term Frequency" відноситься до частоти слова у документі, а "Inverse Document Frequency" зменшує вагу тих слів, які часто зустрічаються в наборі документів. Це дозволяє зменшити вплив часто вживаних, але менш інформативних слів на аналіз. `scikit-learn` надає клас `TfidfVectorizer`, який автоматично перетворює колекцію сирих документів у матрицю TF-IDF ознак.

В процесі розробки та оцінювання моделей машинного навчання та алгоритмів обробки природної мови (NLP), використання об'єктивних метрик є важливим для точного вимірювання ефективності та точності цих моделей. Серед різноманітних метрик, що застосовуються у цих цілях, особливо важливими є косинусна подібність, що входить в пакет `sklearn.metrics.pairwise` та індекс Жаккара.

Таким чином, було визначено допоміжні бібліотеки для реалізації прикладного застосунку на базі методу рекомендації текстових документів за запитаннями користувачів засобами NLP .

3.6 Програмна архітектура інформаційної системи автоматизованого визначення рекомендацій текстових документів

Після проєктування схеми інформаційної системи та опису основних її складових та функцій, а також після вибору засобів розробки доцільно переходити до створення програмної архітектури інформаційної системи.

Схема запропонованої програмної архітектури для інформаційної системи автоматизованого визначення рекомендацій текстових документів для предметної області служби психологічної підтримки наведена на рисунку 3.7.

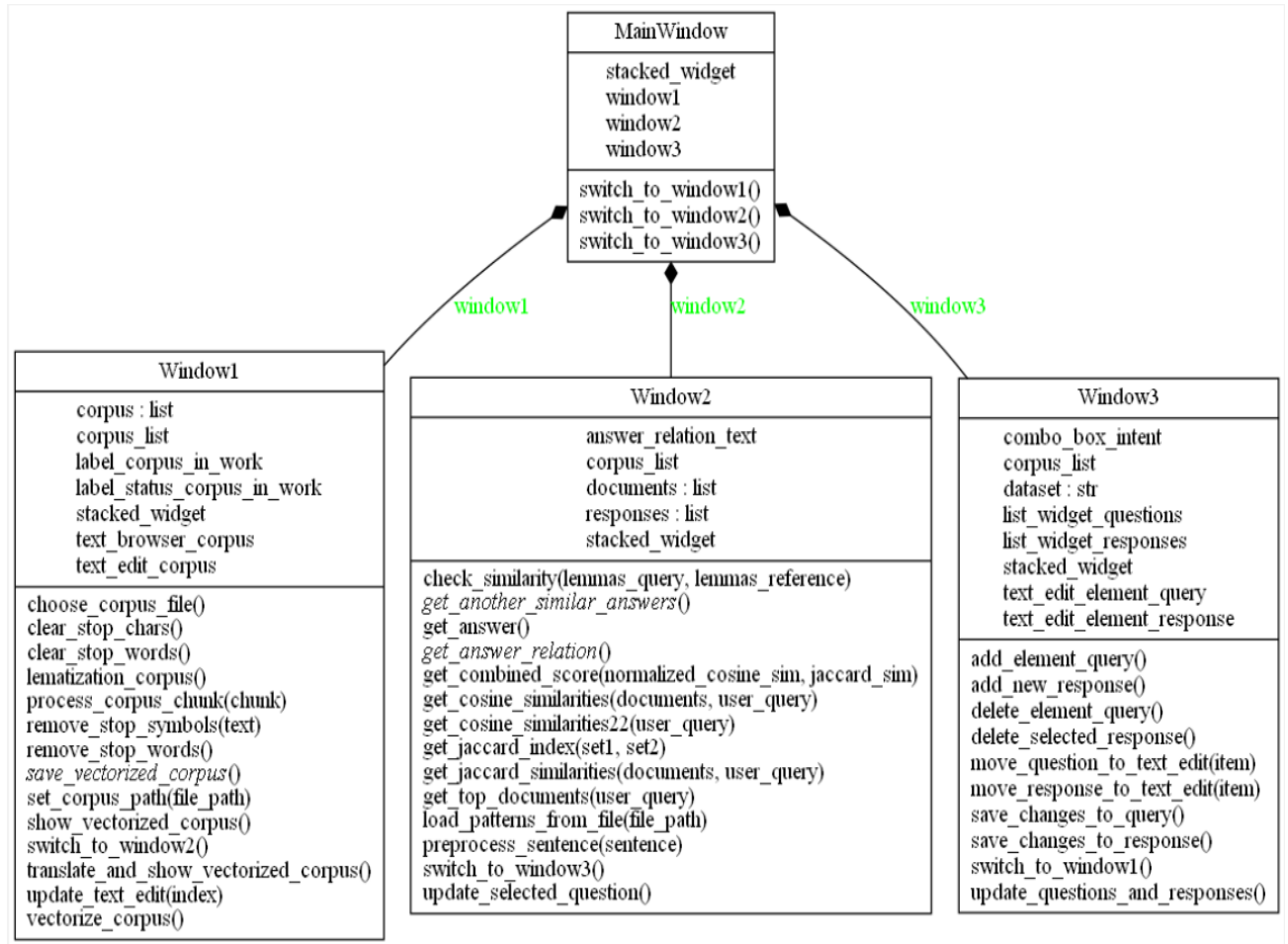


Рисунок 3.7 – Програмна архітектура системи

Програмна архітектура системи є об'єктно-орієнтованою та складається з 4-х класів: «MainWindow», та класів «Window1», що реалізує роботу з корпусами текстів, «Window2», що реалізує пошук релевантних відповідей, та «Window3», що реалізує можливість взаємодії з навчальними текстовими вибірками.

Клас «Window1», що основним призначенням має побудову множини векторів відповідей з навчальної вибірки має у своїй реалізації ряд методів. Метод `choose_corpus_file()` призначений для відкриття вікно вибору файлу за допомогою `QFileDialog`, де користувач може вибрати файл корпусу. Обрана назва файлу передається методу `set_corpus_path()`. Головним методом даного

класу є метод `vectorize_corpus()`, що призначений для векторизації корпусу тексту з використанням TF-IDF, основна ідея якого полягає в тому, щоб визначити важливість кожного слова у конкретному документі в межах колекції документів.

Метод `lematization_corpus()` призначений для лематизації даних, зчитаних з JSON-файлу. Метод `set_corpus_path()` встановлює шлях корпусу, та виводить назву корпусу у віджеті тексту. Метод `remove_stop_words()` використовує бібліотеку NLTK для завантаження стоп-слів української мови та видаляє їх з тексту. Метод `clear_stop_words()` використовує метод `remove_stop_words` для обробки відповідей у JSON-файлі та виведення результату в віджет тексту.

Метод `remove_stop_symbols()` та `clear_stop_chars()` Видаляють зайві символи з тексту (символи, які не є літерами, цифрами або пробілами).

Клас «Window2» використовує різні методи для порівняння користувацьких запитань з еталонними патернами і визначення відповідей на основі схожості. Крім того, він виводить результати в інтерфейсі за допомогою різних віджетів, таких як текстові поля та комбобокси. Головним методом є метод `get_answer()`, що отримує користувацький запит, завантажує патерни з файлу, порівнює їх з користувацьким запитом та визначає відповіді. Якщо є збіг патерну, виводить вибрану відповідь. В іншому випадку викликає метод `get_top_documents()` для пошуку найбільш схожих документів за метриками індексу Жаккара та косинусної схожості. Метод `preprocess_sentence(self, sentence)` виконує попередню обробку користувацького запиту. Він використовує токенизацію слів, лематизацію та фільтрацію за довжиною та типом символу. Результатом є список лем, які відображають токени, підготовлені для порівняння.

Метод `check_similarity(self, lemmas_query, lemmas_reference)` порівнює два списки лем за допомогою косинусної подібності. Він визначає відсоток схожості між двома списками лем та повертає його. Метод `get_cosine_similarities(self, documents, user_query)` обчислює косинусну

подібність між користувацьким запитом та іншими документами в корпусі за допомогою TF-IDF векторизації.

Метод `get_combined_score(self, normalized_cosine_sim, jaccard_sim)` обчислює комбінований індекс на основі нормалізованої косинусної подібності та індексу Жаккара. Метод `get_top_documents(self, user_query)` отримує топ-3 документів на основі комбінованого індексу, виводить їх та повертає список цих документів.

Клас «Window3» здійснює роботу з елементами інтерфейсу та набору навчальних даних для додавання, видалення та редагування асоційованих запитів і відповідей. Метод `add_element_query(self)` призначений для додавання нового запиту до списку запитів, якщо воно не є порожнім та для збереження доданого запиту в документі навчальних даних. Метод `delete_element_query(self)` видаляє обране запитання зі списку запитань та з документу навчальних даних. Метод `save_changes_to_query(self)` зберігає внесені зміни до тексту обраного запитання.

Метод `move_response_to_text_edit(self, item)` переміщає обрану відповідь до текстового редактора для редагування. Метод `save_changes_to_response(self)` призначений для збереження внесених зміни до тексту обраної відповіді та документу навчальних даних. Метод `delete_selected_response(self)` призначений для видалення обраної відповіді зі списку відповідей. Метод `add_new_response(self)` призначений для додавання нової відповіді до списку відповідей, якщо вона не є порожньою, а також до документу навчальних даних.

Метод `update_questions_and_responses(self)` призначений для оновлення списків запитань і відповідей відповідно до вибраного тегу наміру з датасету. Використовується для взаємодії з віджетами типу `QListWidget` та `QComboBox`.

Отже, таким чином було спроектовано програмну архітектуру інформаційної системи для предметної області служби психологічної підтримки, що є прикладною програмною реалізацією розробленого методу рекомендації текстових документів за запитаннями користувачів засобами NLP.

Висновки до третього розділу

В ході виконання третього розділу було виконано формування інформаційної системи автоматизованого визначення рекомендацій текстових документів за користувацькими запитамі для предметної області служби психологічної підтримки, в рамках якого:

1. Виконано проектування інформаційної структури системи автоматизованого визначення рекомендацій текстових документів на основі методу пошуку рекомендації текстових документів за запитамі користувачів, що за вхідними даними датасету з групами асоційованих запитів та відповідних їм асоційованих відповідей та користувацького запиту перетворює їх у вихідні дані у вигляді текстової рекомендації релевантної до запиту.

2. Наведено основні схеми та функції основних підсистем інформаційної системи, а саме редактора навчальної вибірки документів запитів-відповідей та підсистеми автоматизованого підбору документів відповідей до запитів, яка є головною підсистемою спроектованої інформаційної системи.

3. Виконано формування комбінації засобів розробки інформаційної системи для предметної області служби психологічної підтримки, які складаються з мови програмування Python та інтегрованого середовища розробки PyCharm. Вибір зумовлений низкою переваг цих інструментів, які оптимально відповідають потребам проекту. Python, завдяки своїй гнучкості, багатій екосистемі бібліотек для NLP та машинного навчання, є ідеальним вибором для створення складних систем обробки даних. PyCharm, в свою чергу, надає розширені можливості для ефективного розробки та налагодження коду, що значно підвищує продуктивність робочого процесу.

4. Виконано вибір спеціалізованих програмних розширень для реалізації прикладного застосунку на базі методу рекомендації текстових документів за запитаннями користувачів засобами NLP для предметної області служби психологічної підтримки, якими стали PyQt5 для розробки інтерфейсу

користувача, бібліотека `scikit-learn`, що використовується для векторизації та пошуку косинусної відстані, бібліотеку `NLTK` для лематизації та препроцесингу.

5. Виконано проектування програмної архітектури інформаційної системи, що є прикладною програмною реалізацією для предметної області служби психологічної підтримки на основі розробленого методу рекомендації текстових документів за запитаннями користувачів засобами `NLP`. Програмна архітектура налічує 4 класи, функціональність яких детально описано.

Розділ 4 Дослідження ефективності методу рекомендації текстових документів за запитаннями користувачів засобами NLP

4.1 Особливості розробки прикладних компонентів інформаційної системи рекомендації текстових документів

За створеною програмною архітектурою інформаційної системи для предметної області служби психологічної підтримки, що є об'єктно-орієнтованою та складається з 4-х класів («MainWindow», «Window1», «Window2» та «Window3»), що реалізують заявлений функціонал інформаційної системи, надалі створюються окремі її складові.

Клас Window1 своїм основним призначенням має побудову множини векторів відповідей з навчальної вибірки. Головним методом класу є метод `vectorize_corpus()`, що призначений для векторизації корпусу текстових даних за допомогою TF-IDF. Умовно метод проходить декілька етапів, схематично наведених на рисунку 4.1.

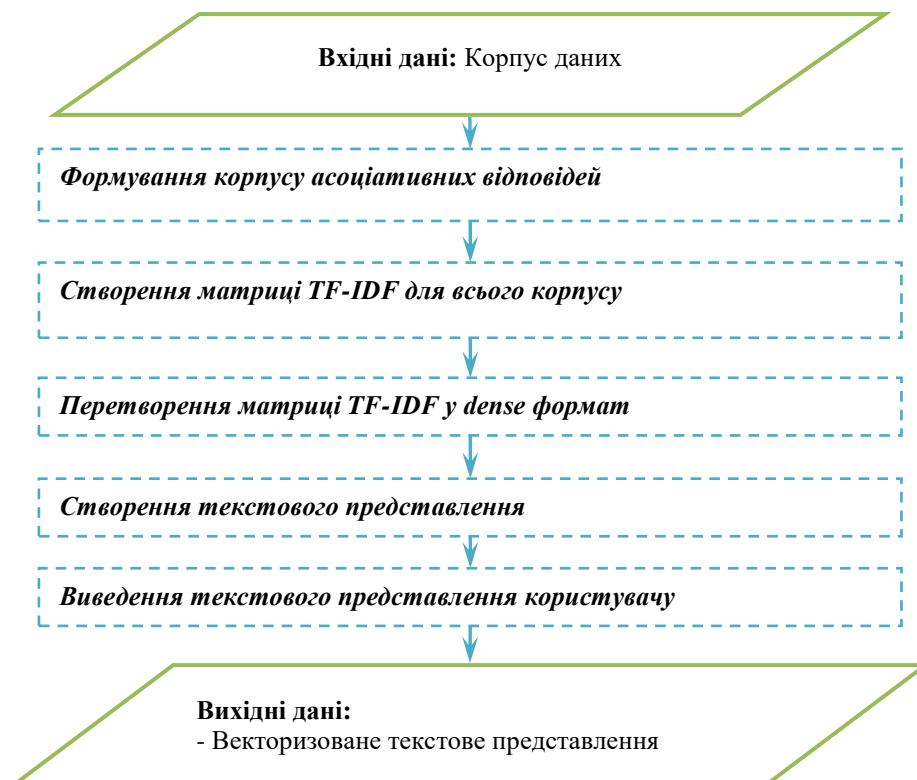


Рисунок 4.1 – Схема роботи методу `vectorize_corpus()`

Вхідними даними є зчитаний з json-формату корпус асоціативних запитань та відповідей. Оскільки векторизувати потрібно не весь корпус, а лише відповіді, є потреба у формуванні корпусу асоціативних відповідей.

Кожен намір (елемент «Запит-відповідь») містить список відповідей. Метод витягує ці відповіді з кожного наміру та об'єднує їх в загальний корпус `corpus`.

Наступним кроком є створення матриці TF-IDF для всього корпусу. Використовуючи бібліотеку `scikit-learn` та клас `TfidfVectorizer`, метод створює матрицю TF-IDF для всього корпусу. Кожен рядок у матриці представляє одну відповідь, а кожний стовпчик представляє термін. Матриця `tfidf_matrix` містить числові значення для кожного терміну у відповіді, враховуючи його важливість в контексті всього корпусу.

Наступним кроком є перетворення матриці TF-IDF у `dense` формат. Оскільки матриця TF-IDF розріджена, її потрібно перетворити у `dense` формат за допомогою методу `toarray()`. Це полегшує подальшу обробку та виведення.

Наступним кроком методу є створення текстового представлення. Метод перебирає кожен рядок у `dense` матриці та створює текстове представлення для кожної відповіді. Кожен термін у відповіді із матриці TF-IDF має своє числове значення, яке додається до рядка. Створені рядки додаються до списку `individual_text_representations`. Відповідно, створюється результуючий рядок, де кожна відповідь між `Next Text` розділеною рядками відокремлюється від іншої.

Останнім кроком методу є виведення текстового представлення користувачу в інтерфейс. Результат вставляється в `QTextEdit` (віджет для текстового введення/виведення) з ідентифікатором `text_edit_corpus`.

Вихідними даними методу є векторизоване текстове представлення. Приклад роботи даного методу наведено на рисунку 4.2.

Рисунок 4.3 – Схема роботи методу `get_answer()`

Отримання тексту користувацького запиту. Метод використовує `self.findChild` для отримання вмісту `QTextEdit` із ідентифікатором `'input_question'`. Введений текст отримується за допомогою `input_text.toPlainText()`. Якщо

користувач обирає питання з переліку, воно також відображається у QTextEdit, і обробляється вищеописаним способом.

Для формування корпусу асоціативних запитів з необробленого корпусу даних викликається метод `load_patterns_from_file()`, що завантажує дані з файлу та повертає список об'єктів, де кожен об'єкт містить патерни (асоціативні запити).

На етапі формування лематизованого представлення корпусу асоціативних запитів з необробленого корпусу даних метод перебирає кожен об'єкт у списку отриманих шаблонів (`pattern_set`) та їх патерни. Кожен патерн лематизується за допомогою `preprocess_sentence()` для подальшого порівняння із введеним лематизованим текстом користувача.

На етапі пошуку співставлень користувацького запиту та асоціативних запитів якщо знаходиться патерн, схожий на введений текст за допомогою виклику методу `check_similarity()`, і відповідність більше ніж 60%, він вважається знайденим. Обирається ТОП-3 відповіді, за результатом порівняння з асоційованими запитами, що мають найвищу подібність, а також результат знаходження релевантних документів у вигляді відсоткового значення і методу.

Якщо відповіді не знайдено, викликається метод `get_top_documents()` для отримання топових документів за допомогою зваженої оцінки на основі косинусної схожості та відстані Жаккара. Пошук косинусної відстані між вектором користувацького запиту та кожною асоційованою відповіддю з корпусу векторів асоційованих відповідей здійснюється методом `get_cosine_similarities()`, який повертає масив косинусних відстаней для кожного з документів корпусу.

Пошук відстані Жаккара між вектором користувацького запиту та кожною асоційованою відповіддю з корпусу векторів асоційованих відповідей здійснюється викликом методу `get_jaccard_similarities()`, який повертає масив відстаней Жаккара для кожного з документів корпусу.

Обрахунок зваженої оцінки та формування рейтингу документів за схожістю обраховується методом `get_combined_score()`, який на основі двох попередніх кроків обрахує середню зважену оцінку.

Далі відбувається формування ТОП-3 документів з найвищою точністю співставлення та відповідних оцінок і методу отримання, користувачу виводиться релевантна відповідь.

Вихідними даними є збережений список ТОП-3 текстових рекомендацій та їх оцінки. Приклад роботи підсистеми на основі описаного методу на рисунку 4.4.

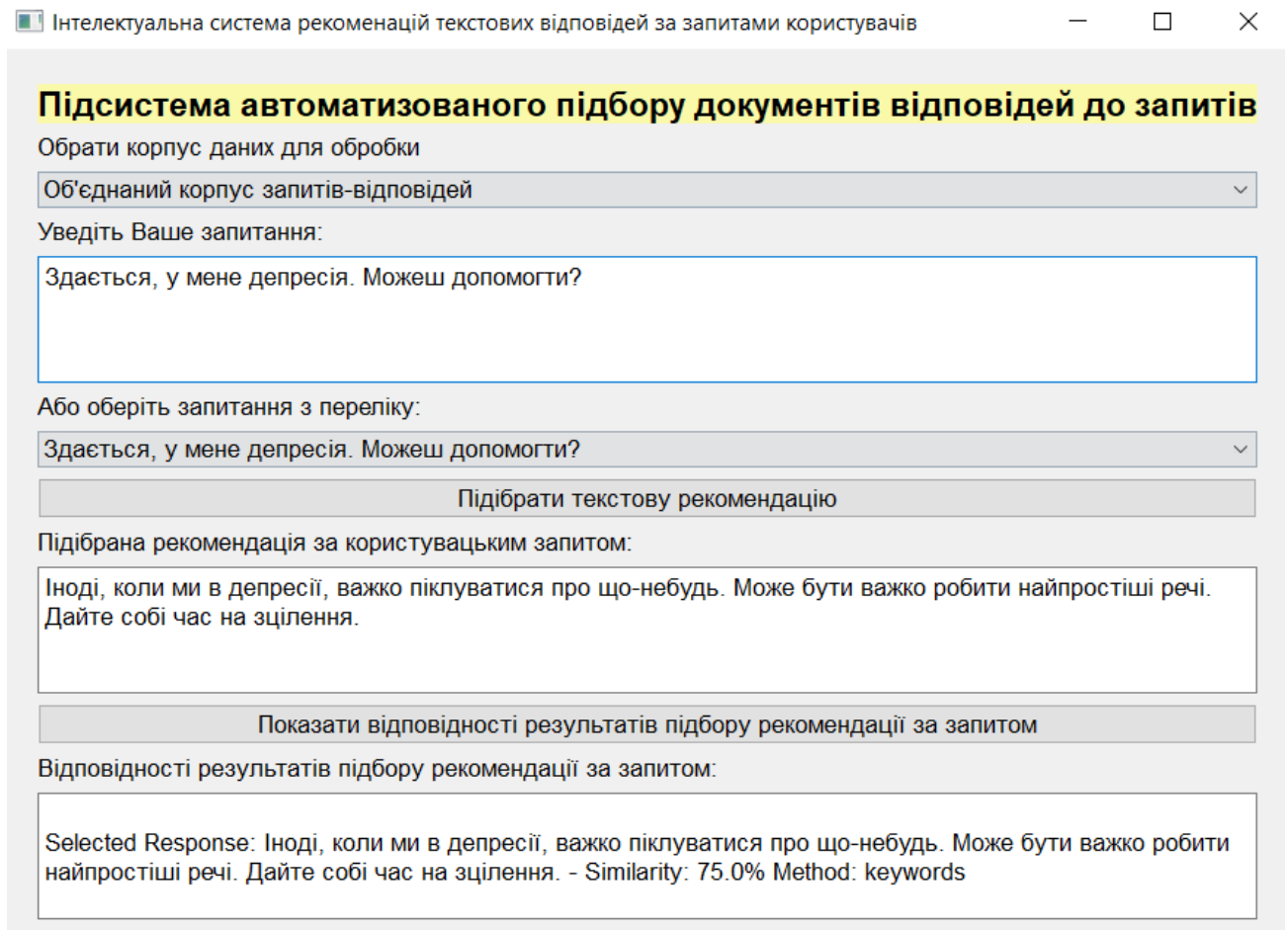


Рисунок 4.4 – Підбір документу за запитом користувача

Клас «Window3» призначений для організації роботи через елементи інтерфейсу з набором навчальних даних для додавання, видалення та редагування асоційованих запитів і відповідей.

Клас має набір методів, умовно які можна розбити по грапам: робота з запитами, робота з асоційованими відповідями, робота з елементами «Запит-відповідь». Приклад користувацького інтерфейсу для реалізації даного класу наведено на рисунку 4.5

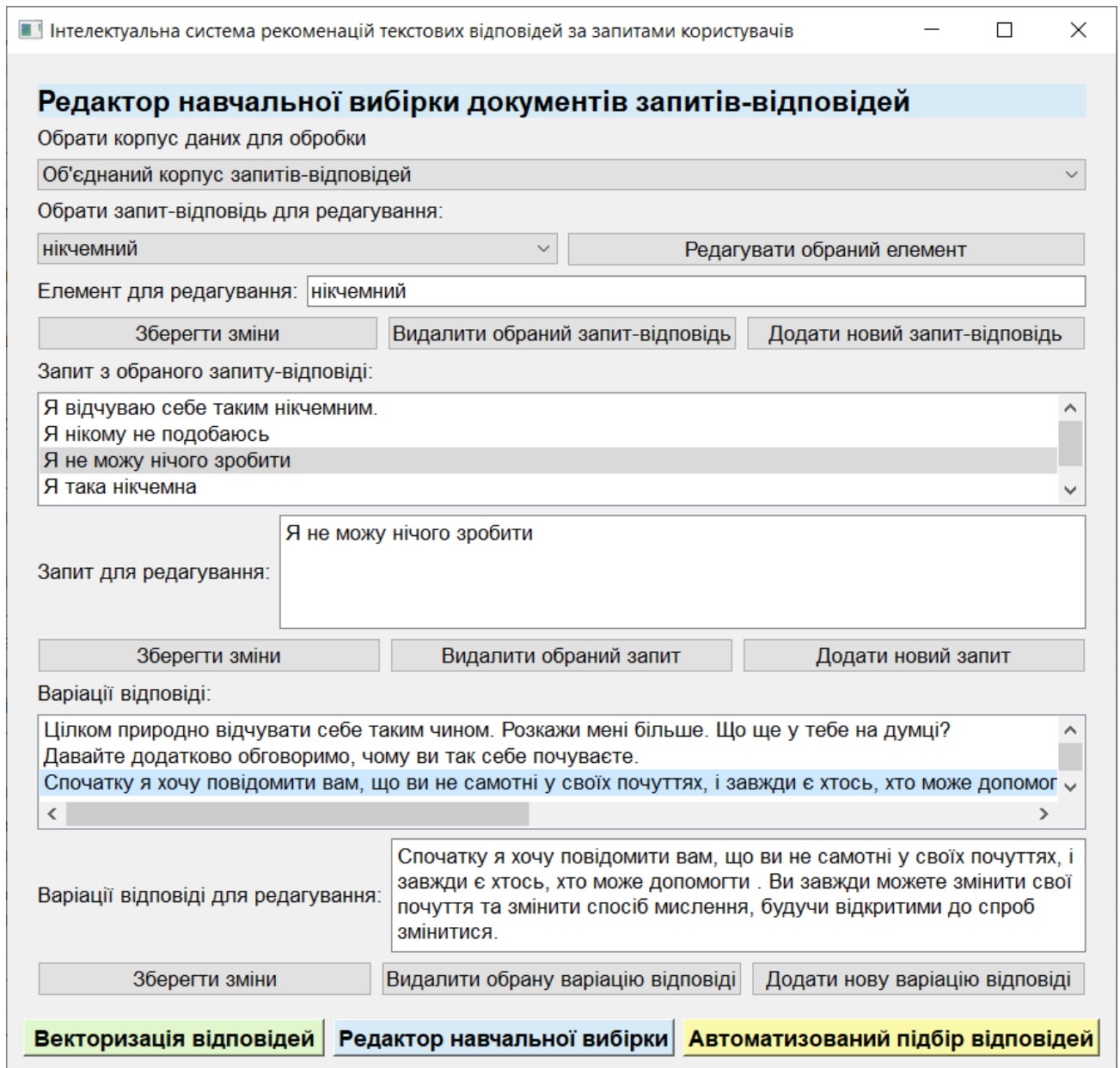


Рисунок 4.5 – Функціонал класу «Window3»

Отже, таким чином було описано прикладні особливості розробки компонентів інформаційної системи, що є об'єктно-орієнтованою та складається з 4-х класів які реалізують заявлений функціонал інформаційної системи

автоматизованого визначення рекомендацій текстових документів за користувацькими запитам.

4.2 Прикладне тестування інформаційної системи

Для перевірки коректності інформаційної системи автоматизованого визначення рекомендацій текстових документів за користувацькими запитам для предметної області служби психологічної підтримки було створено ряд тест-кейсів.

Першим тестовим випадком буде перевірка коректності завантаження обраного корпусу текстів та відображення його змісту на екрані користувача з підсистеми побудови множини векторів відповідей. Кроки тест-кейса наведено у таблиці 4.1.

Таблиця 4.1 – Тест-кейс K001

Тест-кейс ID: K001	Пріоритет: 1	Створено: 29.11.2023, Козенко О.
Назва: Перевірка коректності завантаження обраного корпусу текстів та відображення його змісту на екрані користувача		
Кроки		Очікуваний результат
1. Запустити програму; 2. Перейти на підсистему «Підсистема побудови множини векторів відповідей з навчальної вибірки» 3. Обрати корпус даних для обробки		Відкрився застосунок Відкрилась підсистема побудови множини векторів відповідей з навчальної вибірки Назва корпусу відображена в полі «Корпус в роботі», вміст множини відповідей з навчальної вибірки показано на екрані користувача
Результат виконання тест-кейсу: пройдено успішно		

Результат успішного виконання тест-кейсу K001 наведено на рисунку 4.6.

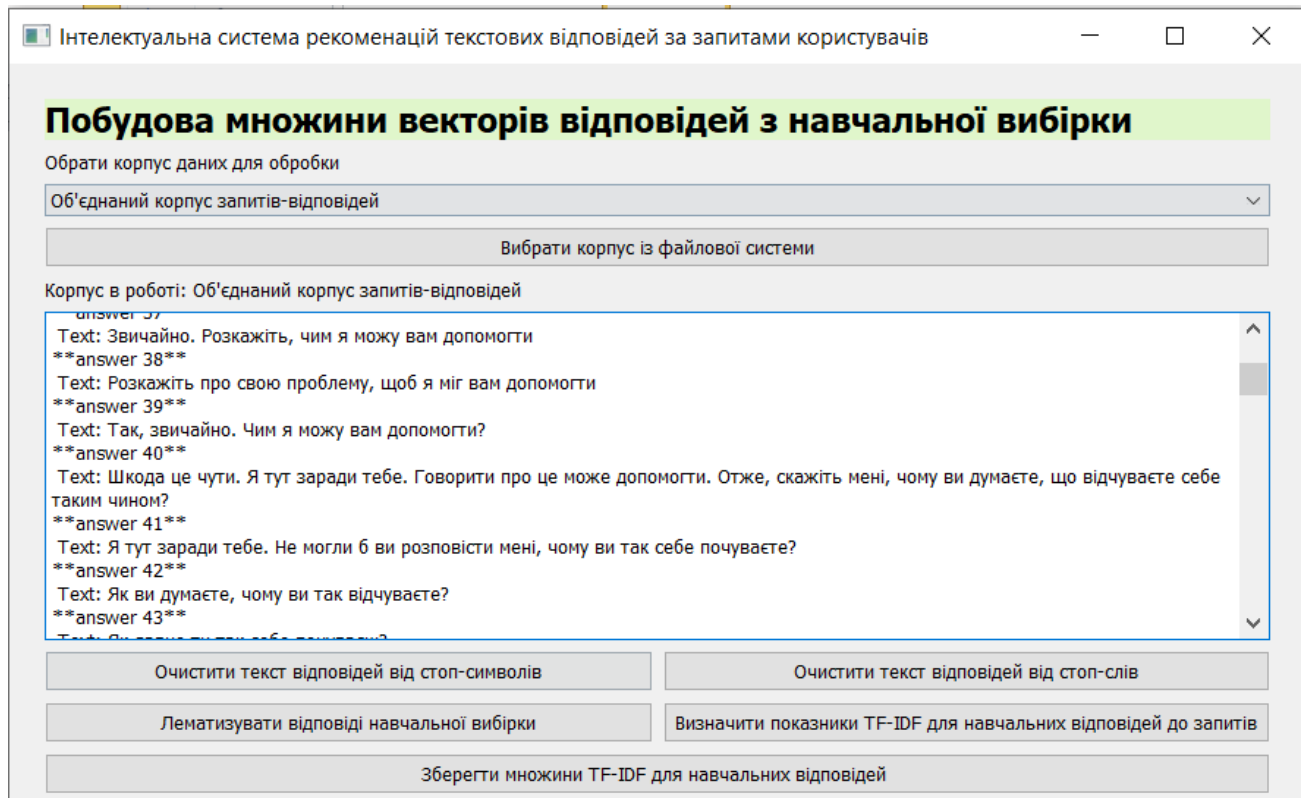


Рисунок 4.6 – Результат виконання тест-кейсу K001

Наступним тестовим випадком буде перевірка коректності перетворення завантаженого обраного корпусу текстів у векторне представлення. Кроки тест-кейса наведено у таблиці 4.2.

Таблиця 4.2 – Тест-кейс K002

Тест-кейс ID: K002	Пріоритет: 1	Створено: 30.11.2023, Козенко О.
Назва: Перевірка коректності перетворення завантаженого обраного корпусу текстів у векторне представлення		
Кроки	Очікуваний результат	
<ol style="list-style-type: none"> Запустити програму; Перейти на підсистему «Підсистема побудови множини векторів відповідей з навчальної вибірки» Обрати корпус даних для обробки Натиснути кнопку «Визначити показники TF-IDF для навчальних відповідей до запитів» 	<p>Відкрився застосунок</p> <p>Відкрилась підсистема побудови множини векторів відповідей з навчальної вибірки</p> <p>Назва корпусу відображена в полі «Корпус в роботі», вміст множини відповідей з навчальної вибірки показано на екрані користувача</p> <p>Виведено векторне представлення у поле «Результат обробки корпусу»</p>	
Результат виконання тест-кейсу: пройдено успішно		

Таблиця 4.3 – Тест-кейс K003

Тест-кейс ID: K003	Пріоритет: 1	Створено: 1.12.2023, Козенко О.
Назва: Перевірка коректності редагування обраного запиту з множини асоціативних запитів		
Кроки	Очікуваний результат	
<ol style="list-style-type: none"> Запустити програму. Перейти на підсистему «Редактор навчальної вибірки документів запитів-відповідей». Обрати корпус даних для обробки «Об'єднаний корпус запитів-відповідей». Обрати шаблон «Допомога» з множини шаблонів «Запит-відповідь» для редагування. Обрати для редагування другий напис, двічі по ньому натиснувши. Змінити текст запиту. Натиснути кнопку «Зберегти зміни». 	<p>Відкрився застосунок Відкрилась підсистема редактору навчальної вибірки документів запитів-відповідей</p> <p>Заповнились поля з асоційованими запитами та відповідями</p> <p>Текст другого запиту відображено на формі. Текст змінено Змінений текст відображено в користувацькому застосунку та збережено до файлу .json</p>	
Результат виконання тест-кейсу: пройдено успішно		

Результат успішного виконання тест-кейсу K003 наведено на рисунку 4.8.

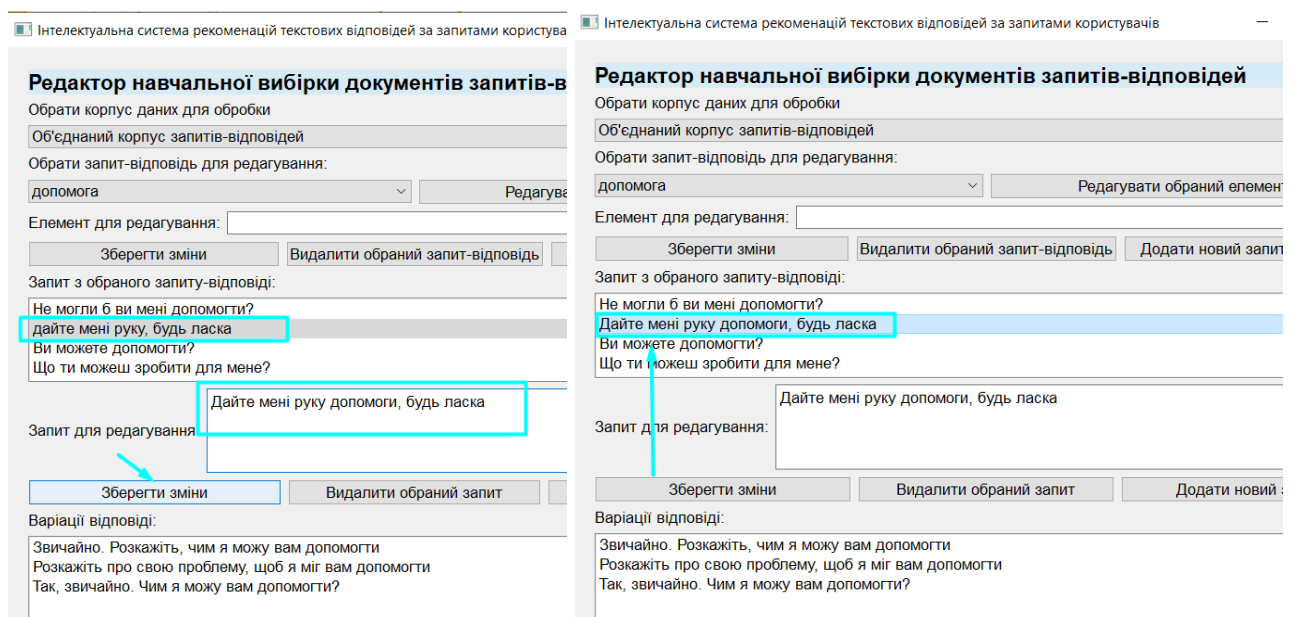


Рисунок 4.8 – Результат виконання тест-кейсу K003

Також необхідно переконатись, що дані успішно збереглись не лише в інтерфейсі користувача, а й в .json документі. При пошуку відповідного набору шаблонів «Запит-відповідь» було помічено, що дані записались коректно (рисунок 4.9).

```
{
  "tag": "допомога", |
  "patterns": [
    "Не могли б ви мені допомогти?",
    "Дайте мені руку допомоги, будь ласк",
    "Ви можете допомогти?",
    "Що ти можеш зробити для мене?",
    "Мені потрібна підтримка",
    "Мені потрібна допомога",
    "Підтримай мене, будь ласка"
  ],
}
```

Рисунок 4.9 – Збереження змінених даних

Наступним тестовим випадком буде перевірка коректності додавання нового запиту до множини асоціативних запитів для підсистеми «Редактор навчальної вибірки документів запитів-відповідей». Кроки тест-кейса наведено у таблиці 4.4.

Результат успішного виконання тест-кейсу K003 наведено на рисунках 4.10 – 4.11.

Рисунок 4.10 – Уведення нового запиту

Таблиця 4.4 – Тест-кейс К004

Тест-кейс ID: К004	Пріоритет: 1	Створено: 1.12.2023, Козенко О.
Назва: Перевірка коректності додавання нового запиту до множини асоціативних запитів		
Кроки		Очікуваний результат
<ol style="list-style-type: none"> 1. Запустити програму. 2. Перейти на підсистему «Редактор навчальної вибірки документів запитів-відповідей». 3. Обрати корпус даних для обробки «Об'єднаний корпус запитів-відповідей». 4. Обрати шаблон «стрес» з множини шаблонів «Запит-відповідь». 5. В полі «Запит для редагування» написати текст «Я перебуваю в дикому стресовому стані». 6. Натиснути кнопку «Додати новий запит». 		<p>Відкрився застосунок Відкрилась підсистема редактору навчальної вибірки документів запитів-відповідей</p> <p>Заповнились поля з асоційованими запитами та відповідями</p> <p>Текст додано до списку існуючих запитів та збережено до файлу .json</p>
Результат виконання тест-кейсу: пройдено успішно		

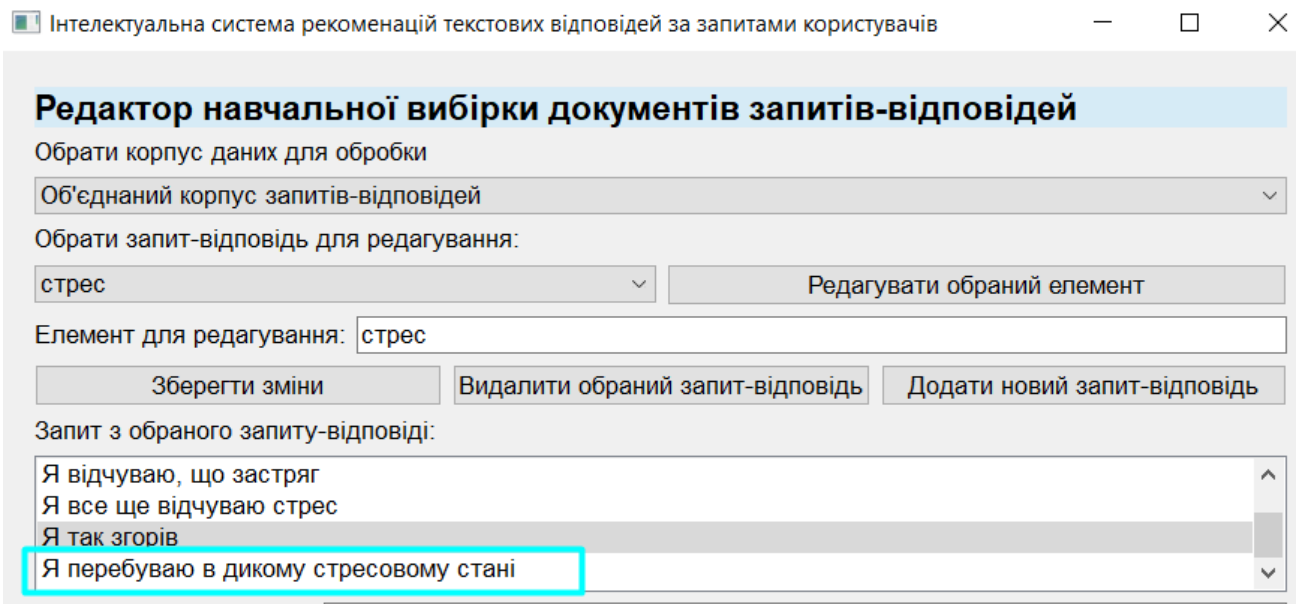


Рисунок 4.11 – Додавання нового запиту

Наступним тестовим випадком буде перевірка коректності підбору рекомендації за користувацьким запитом підсистеми «Автоматизованого підбору документів відповідей до запитів». Кроки тест-кейса наведено у таблиці 4.5.

Таблиця 4.5 – Тест-кейс K005

Тест-кейс ID: K005	Пріоритет: 1	Створено: 3.12.2023, Козенко О.
Назва: Перевірка коректності підбору рекомендації за користувацьким запитом		
Кроки		Очікуваний результат
<ol style="list-style-type: none"> 1. Запустити програму. 2. Перейти на підсистему «Автоматизованого підбору документів відповідей до запитів». 3. Обрати корпус даних для обробки «Об'єднаний корпус запитів-відповідей». 4. У поле «Уведіть Ваше запитання» ввести «Розкажи про obsesivno-kompulsivnii rozlad osobistosti». 5. Натиснути кнопку «Підібрати текстову рекомендацію». 		<p>Відкрився застосунок</p> <p>Відкрилась підсистема автоматизованого підбору документів відповідей до запитів</p> <p>Підібрана релевантна відповідь</p>
Результат виконання тест-кейсу: пройдено успішно		

Результат успішного виконання тест-кейсу K003 наведено на рисунку 4.12. Як видно з рисунку, підібраний текст відповідає запиту.

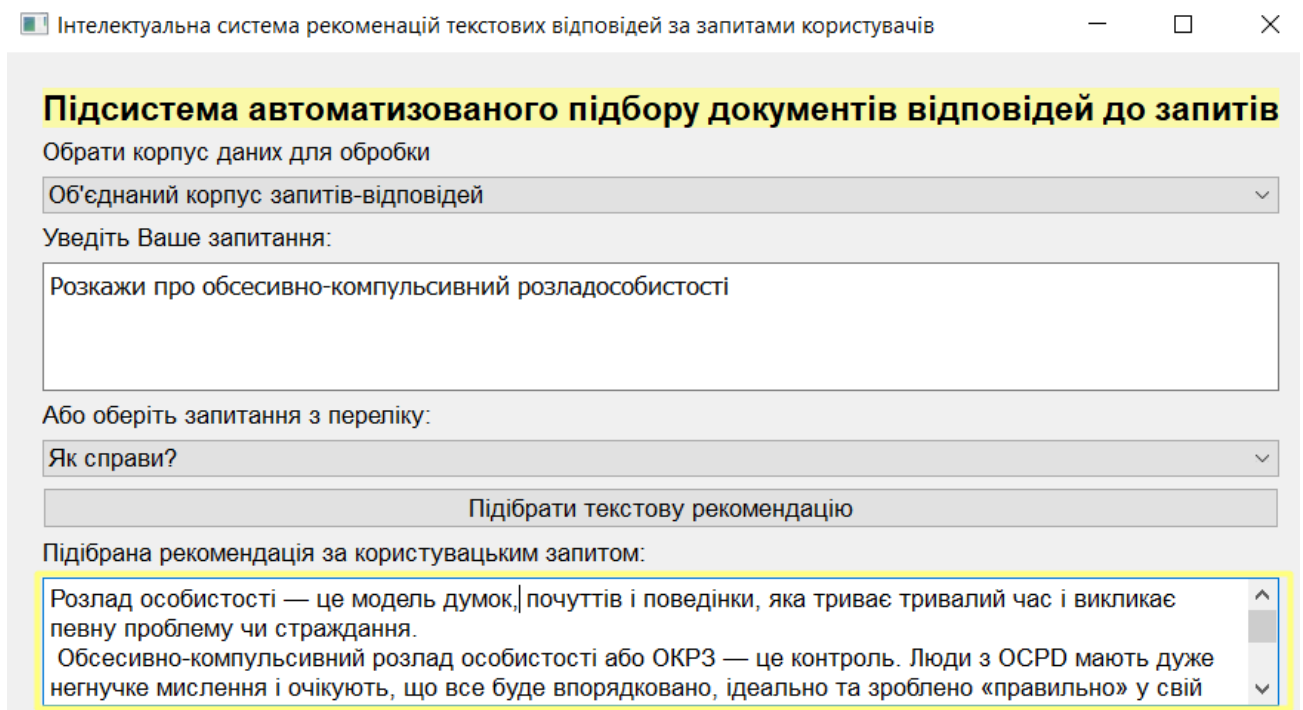


Рисунок 4.12 – Результат виконання тест-кейсу K005

Отже, таким чином виконано тестування розробленого програмного застосунку, що реалізовує метод пошуку рекомендації текстових документів за запитаннями користувачів для предметної області служби психологічної підтримки, що призначений для знаходження релевантних рекомендацій з бази текстових порад за текстовим запитом користувача та вхідними даними має набір даних з групами асоційованих запитів та відповідних їм асоційованих відповідей та користувацький запит, що перетворює їх у вихідні дані у вигляді текстової рекомендації релевантної до запиту. Під час проведення тестування некоректної роботи функцій не спостерігалось. Весь функціонал працює згідно до поставленої задачі.

4.3 Особливості використання інформаційної системи автоматизованого визначення рекомендацій текстових документів

Інформаційна система автоматизованого визначення рекомендацій текстових документів складається з трьох підсистем та головного вікна, з якого можна перейти на потрібну підсистему. По замовчуванню завантажується підсистема побудови множини векторів відповідей з навчальної вибірки.

Схема можливих переходів між підсистемами наведена на рисунку 4.13. Як видно з рисунку 4.13, внизу вікна розташована панель з кнопками для переходів на інші підсистеми застосунку.

Відповідно, для коректної роботи методу необхідно спершу здійснити попередню обробку корпусу даних, з яким буде працювати метод. Для цього користувачу необхідно обрати на підсистемі побудови множини векторів відповідей з навчальної вибірки корпус для роботи (рисунок 4.14).

Як видно з рисунку 4.14, завантажений корпус виводиться на екран. Далі з завантаженим корпусом можна виконувати очистку, векторизацію та збереження результатів.

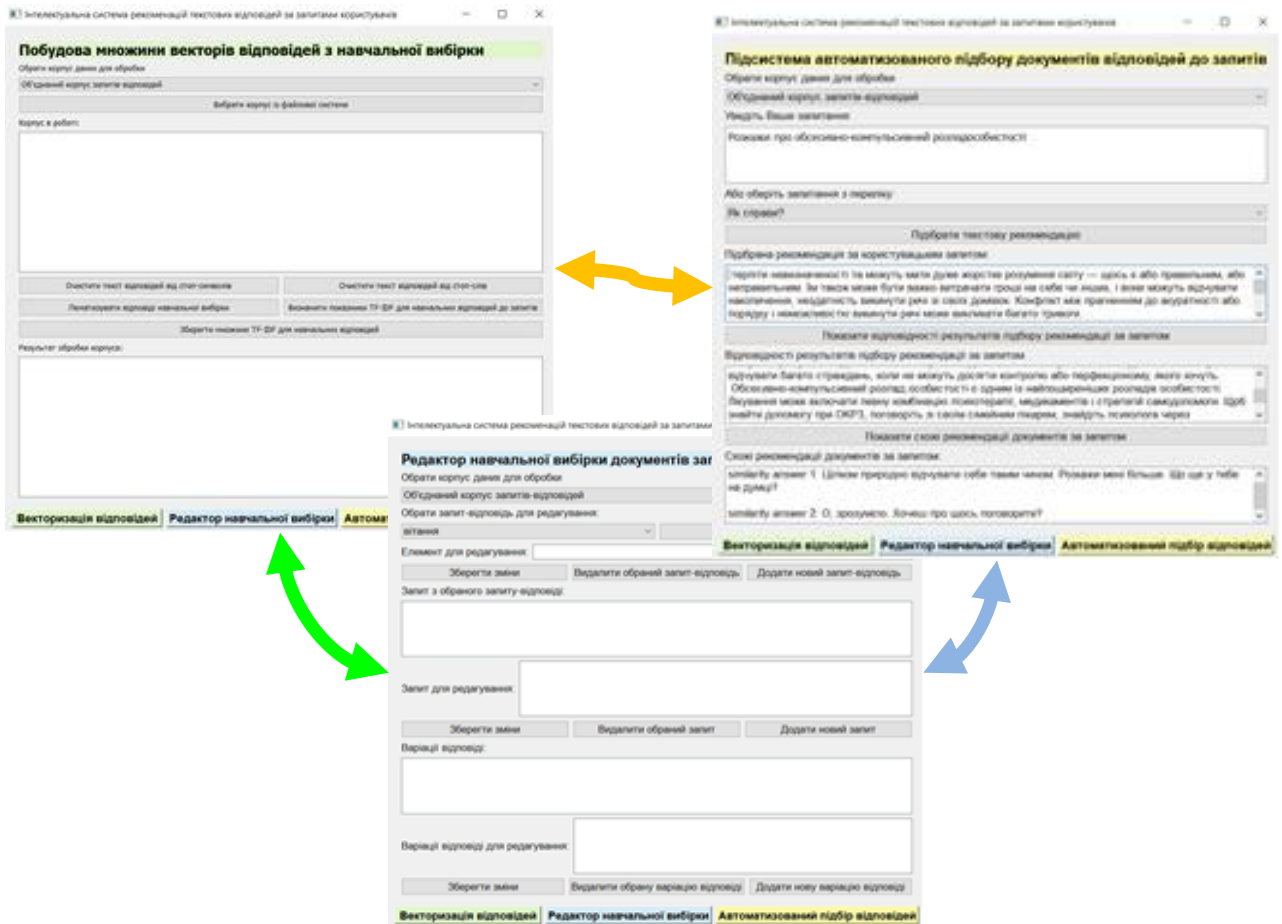


Рисунок 4.13 – Реалізація переходів між підсистемами

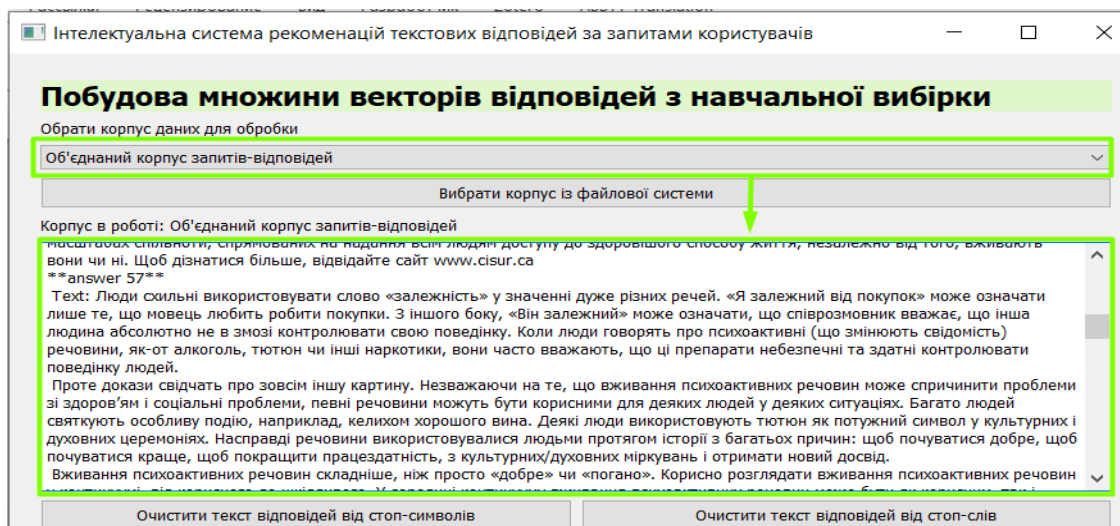


Рисунок 4.14 – Завантаження обраного корпусу в роботу

Для того, щоб очистити корпус від стоп-символів, необхідно натиснути кнопку «Очистити текст відповідей від стоп-символів». Результат очищення тексту від стоп-символів наведено на рисунку 4.15.

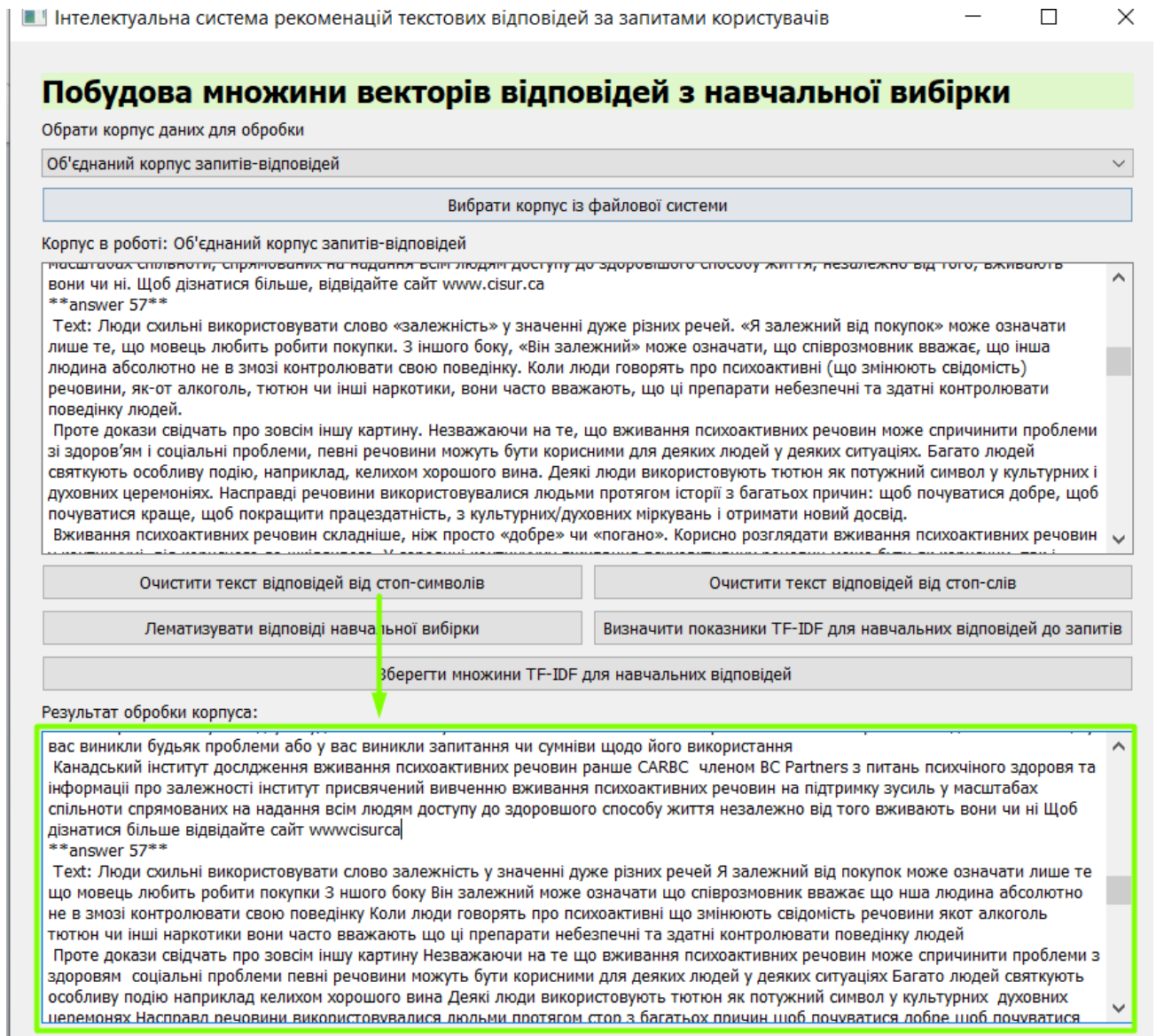


Рисунок 4.15 – Очищення тексту від стоп-символів

Після очищення від стоп-символів можна очистити вибірку від стоп-слів. Для цього необхідно натиснути на кнопку «Очистити текст від стоп-слів». Для Лематизації відповідей необхідно натиснути на кнопку «Лематизувати відповіді навчальної вибірки», після чого текст буде лематизовано. Для векторизації множини навчальної вибірки необхідно натиснути на кнопку «Визначити показники TF-IDF для навчальних відповідей до запитів». Результат векторизації попередньо-обробленої навчальної множини наведено на рисунку 4.16.

Інтелектуальна система рекомендацій текстових відповідей за запитами користувачів

Побудова множини векторів відповідей з навчальної вибірки

Обрати корпус даних для обробки

Об'єднаний корпус запитів-відповідей

Вибрати корпус із файлової системи

Корпус в роботі: Об'єднаний корпус запитів-відповідей

масштабах співжиття, сприятливих на подання всім людям доступу до здорового способу життя, незалежно від того, вживають вони чи ні. Щоб дізнатися більше, відвідайте сайт www.cisur.ca

****answer 57****

Text: Люди схильні використовувати слово «залежність» у значенні дуже різних речей. «Я залежний від покупок» може означати лише те, що мовець любить робити покупки. З іншого боку, «Він залежний» може означати, що співрозмовник вважає, що інша людина абсолютно не в змозі контролювати свою поведінку. Коли люди говорять про психоактивні (що змінюють свідомість) речовини, як-от алкоголь, тютюн чи інші наркотики, вони часто вважають, що ці препарати небезпечні та здатні контролювати поведінку людей.

Проте докази свідчать про зовсім іншу картину. Незважаючи на те, що вживання психоактивних речовин може спричинити проблеми зі здоров'ям і соціальні проблеми, певні речовини можуть бути корисними для деяких людей у деяких ситуаціях. Багато людей святкують особливу подію, наприклад, келихом хорошого вина. Деякі люди використовують тютюн як потужний символ у культурних і духовних церемоніях. Насправді речовини використовувалися людьми протягом історії з багатьох причин: щоб почуватися добре, щоб почуватися краще, щоб покращити працездатність, з культурних/духовних міркувань і отримати новий досвід.

Вживання психоактивних речовин складніше, ніж просто «добре» чи «погано». Корисно розглядати вживання психоактивних речовин

Очистити текст відповідей від стоп-символів

Очистити текст відповідей від стоп-слів

Лематизувати відповіді навчальної вибірки

Визначити показники TF-IDF для навчальних відповідей до запитів

Зберегти множини TF-IDF для навчальних відповідей

Результат обробки корпусу:

```
0.0822, 0.0000, 0.0341, 0.0428, 0.0390, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0362, 0.0000, 0.0000,
0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0428, 0.0000, 0.0000, 0.0323, 0.1912, 0.0308, 0.0000, 0.0341, 0.0000, 0.0000, 0.0000,
0.0428, 0.0390, 0.0428, 0.0779, 0.0000, 0.0582, 0.0000, 0.0000, 0.0341, 0.0428, 0.0390, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000,
0.0274, 0.0616, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0925, 0.0000, 0.0000, 0.0428, 0.0000, 0.0000, 0.0779, 0.0000, 0.0000,
0.0000, 0.0362, 0.0000, 0.0390, 0.0362, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0308, 0.0390, 0.0000, 0.0000, 0.0000, 0.0000,
0.0000, 0.0000
```

Next Text

****answer 57****

TF-IDF Vector: 0.0002, 0.0000, 0.0021, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000,
0.0000, 0.0000, 0.0000, 0.0040, 0.0000, 0.0000, 0.0000, 0.0000, 0.0005, 0.0000, 0.0577, 0.0451, 0.0000, 0.0000, 0.0000, 0.0680, 0.0410,
0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0001, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0649, 0.0000, 0.0000, 0.0381,
0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0451, 0.0000, 0.0000, 0.0000, 0.0000, 0.0451, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000,
0.0000, 0.0381, 0.0000, 0.0451, 0.0000, 0.0000, 0.0000, 0.0000, 0.0598, 0.0000, 0.0451, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000

Векторизація відповідей

Редактор навчальної вибірки

Автоматизований підбір відповідей

Рисунок 4.16 – Перетворення навчальних відповідей у векторний формат

Після перетворення множини навчальних відповідей у векторне представлення необхідно зберегти результат роботи, натиснувши на кнопку «Зберегти множини TF-IDF для навчальних відповідей».

Після цього перший етап підготовки даних закінчено, можна переходити до підсистеми «Редагування навчальної множини», або ж до «Підсистеми Автоматизованого підбору відповіді». Наступним буде розглянуто функціонал та особливості використання підсистеми «Редактора навчальної вибірки документів запитів-відповідей». Першим що необхідно зробити, це обрати корпус даних для

роботи, а після тег множини «Запит-відповідь», з якими користувач буде взаємодіяти з метою редагування (рисунок 4.17).

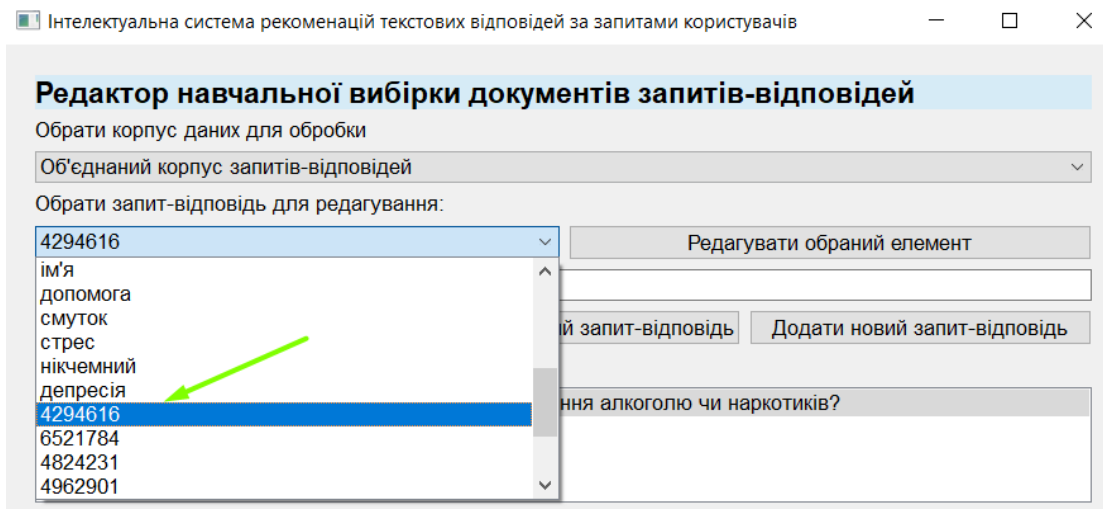


Рисунок 4.17 – Вибір даних з обраної множини «Запит-відповідь»

Після вибору комбінації «Запит-відповідь», користувач побачить наявні варіанти асоціативних запитів та асоціативних відповідей, що є в обраному корпусі (рисунок 4.18).

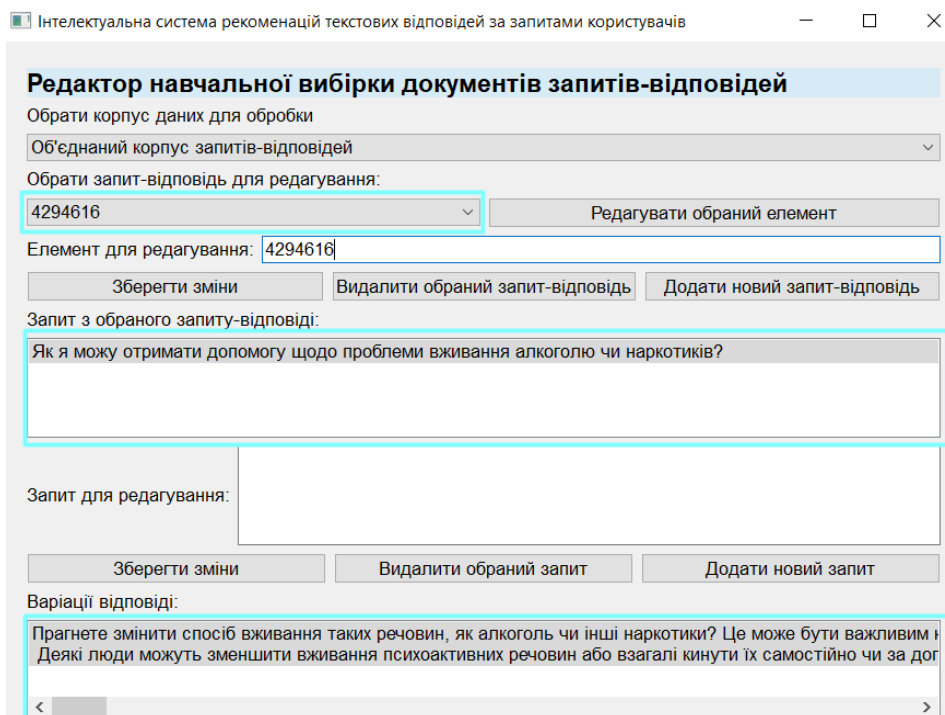


Рисунок 4.18 – Перегляд комбінації «Запит-відповідь» 4294616

Для редагування обраного запиту необхідно двічі натиснути на нього, і текст запиту буде поміщено в поле «Запит для редагування» (рисунок 4.19).

Рисунок 4.19 – Редагування запиту

Текст запиту можна відредагувати, та після завершення редагування натиснути кнопку «Зберегти зміни». Окрім редагування є можливість додати в базу запитів ще варіанти асоційованих запитів. Для цього в поле «Запит для редагування» необхідно помістити текст запиту, який потрібно додати до бази запитів, та натиснути кнопку «Додати новий запит». Запит буде додано до датасету, а також відображено у списку «Запит з обраного запиту-відповіді» (рисунок 4.20).

Рисунок 4.20 – Додавання нового запиту

Також доданий запит повинен додатись у набір даних (рисунок 4.21).

```
{
  "tag": "429/4616",
  "patterns": [
    "Як я можу отримати допомогу щодо проблеми вживання алкоголю чи наркотиків?",
    "Я маю проблеми з вживанням наркотичних речовин."
  ],
  "responses": [
    "Прагнете змінити спосіб вживання таких речовин, як алкоголь чи інші наркотики?"
  ]
}
```

Рисунок 4.21 – Додавання нового запиту до бази запитів

Також аналогічний функціонал є для редагування, додавання нової чи видалення асоційованих відповідей. Для редагування потрібно двічі натиснути на відповідь, яку потрібно відредагувати. Її зміст буде відображено в полі «Варіації відповіді для редагування» (рисунок 4.22).

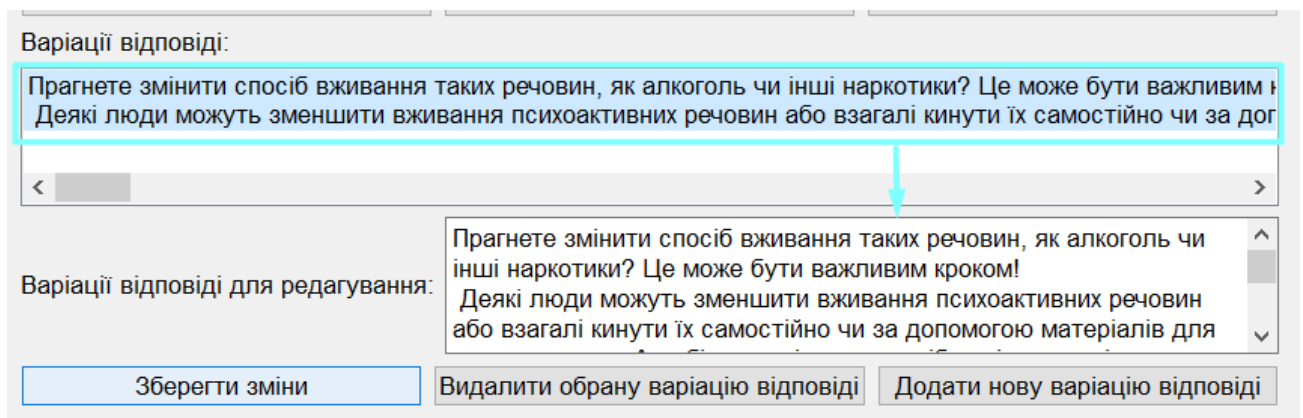


Рисунок 4.22 – Редагування асоційованої відповіді

Для збереження внесених змін потрібно натиснути на кнопку «Зберегти зміни». Для додавання нової відповіді необхідно ввести текстову відповідь і натиснути кнопку «Додати нову варіацію відповіді». Для видалення відповіді необхідно курсор миші поставити на відповідь в полі «Варіації відповіді» на відповідь, яку потрібно видалити, після чого натиснути кнопку «Видалити обрану варіацію відповіді».

Останньою підсистемою є головна підсистема, що призначена безпосередньо для пошуку асоційованих відповідей у відповідь на користувацький запит. Вікно підсистеми зображено на рисунку 4.23.

Інтелектуальна система рекомендацій текстових відповідей за запитами користувачів

Підсистема автоматизованого підбору документів відповідей до запитів

Обрати корпус даних для обробки

Об'єднаний корпус запитів-відповідей

Уведіть Ваше запитання:

Або оберіть запитання з переліку:

Як справи?

Підібрати текстову рекомендацію

Підібрана рекомендація за користувацьким запитом:

Показати відповідності результатів підбору рекомендації за запитом

Відповідності результатів підбору рекомендації за запитом:

Показати схожі рекомендації документів за запитом

Схожі рекомендації документів за запитом:

Векторизація відповідей | Редактор навчальної вибірки | Автоматизований підбір відповідей

Рисунок 4.22 – Підсистема автоматизованого підбору документів відповідей до запитів

Перш за все необхідно обрати з випадаючого списку корпус, з якого будуть підбиратись асоціативні відповіді (рисунок 4.23).

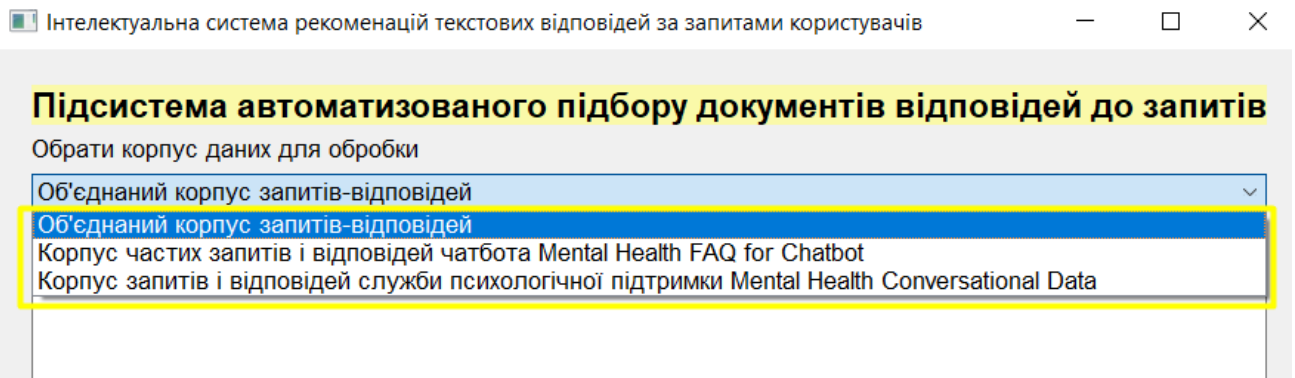


Рисунок 4.23 – Вибір корпусу для роботи

Після вибору корпусу користувач може або ввести запит самостійно, або обрати з переліку стандартних запитів (рисунок 4.24).

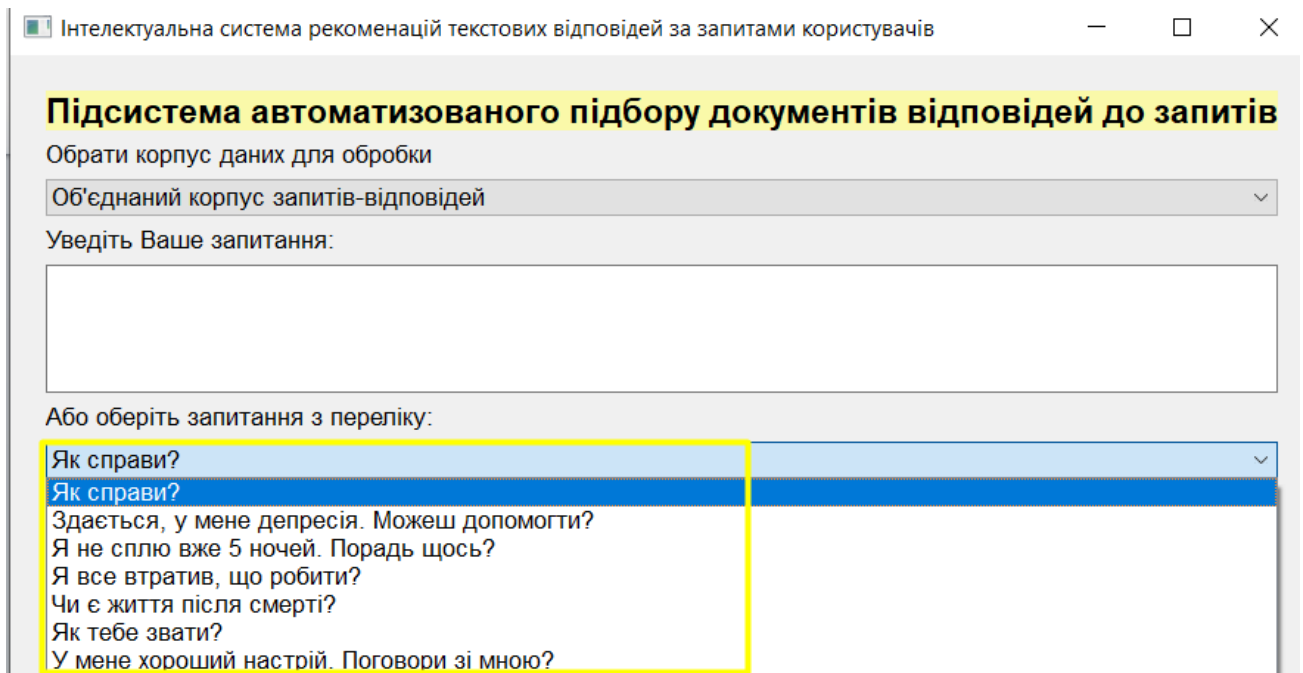


Рисунок 4.24 – Вибір запитання зі списку

При виборі запитання зі списку воно відобразиться у полі «Уведіть Ваше запитання». Після того, як запит введено, для пошуку релевантної відповіді до запиту необхідно натиснути кнопку «Підібрати текстову рекомендацію» (рисунок 4.25).

Інтелектуальна система рекомендацій текстових відповідей за запитами користувачів

Підсистема автоматизованого підбору документів відповідей до запитів

Обрати корпус даних для обробки
Об'єднаний корпус запитів-відповідей

Уведіть Ваше запитання:
Почуваюсь нікчемним і самотнім. Мені здається це через стрес

Або оберіть запитання з переліку:
Як справи?

Підібрати текстову рекомендацію

Підібрана рекомендація за користувацьким запитом:
Шкода це чути. Я тут заради тебе. Говорити про це може допомогти. Отже, скажіть мені, чому ви думаєте, що відчуваєте себе таким чином?

Показати відповідності результатів підбору рекомендації за запитом

Рисунок 4.25 – Підбір рекомендації

Після підбору текстової рекомендації можна переглянути оцінку та метод, яким було знайдено дану рекомендацію (рисунок 4.26). Для цього потрібно натиснути на кнопку «Показати відповідності результатів підбору рекомендації за запитом».

Підібрана рекомендація за користувацьким запитом:
Шкода це чути. Я тут заради тебе. Говорити про це може допомогти. Отже, скажіть мені, чому ви думаєте, що відчуваєте себе таким чином?

Показати відповідності результатів підбору рекомендації за запитом

Відповідності результатів підбору рекомендації за запитом:
Шкода це чути. Я тут заради тебе. Говорити про це може допомогти. Отже, скажіть мені, чому ви думаєте, що відчуваєте себе таким чином?
(39, 0.18598122)
Method: combined

Рисунок 4.25 – Підбір рекомендації

Також, можна переглянути і інші схожі відповіді, для цього необхідно натиснути на кнопку «Показати схожі рекомендації документів за запитом» (рисунок 4.26).

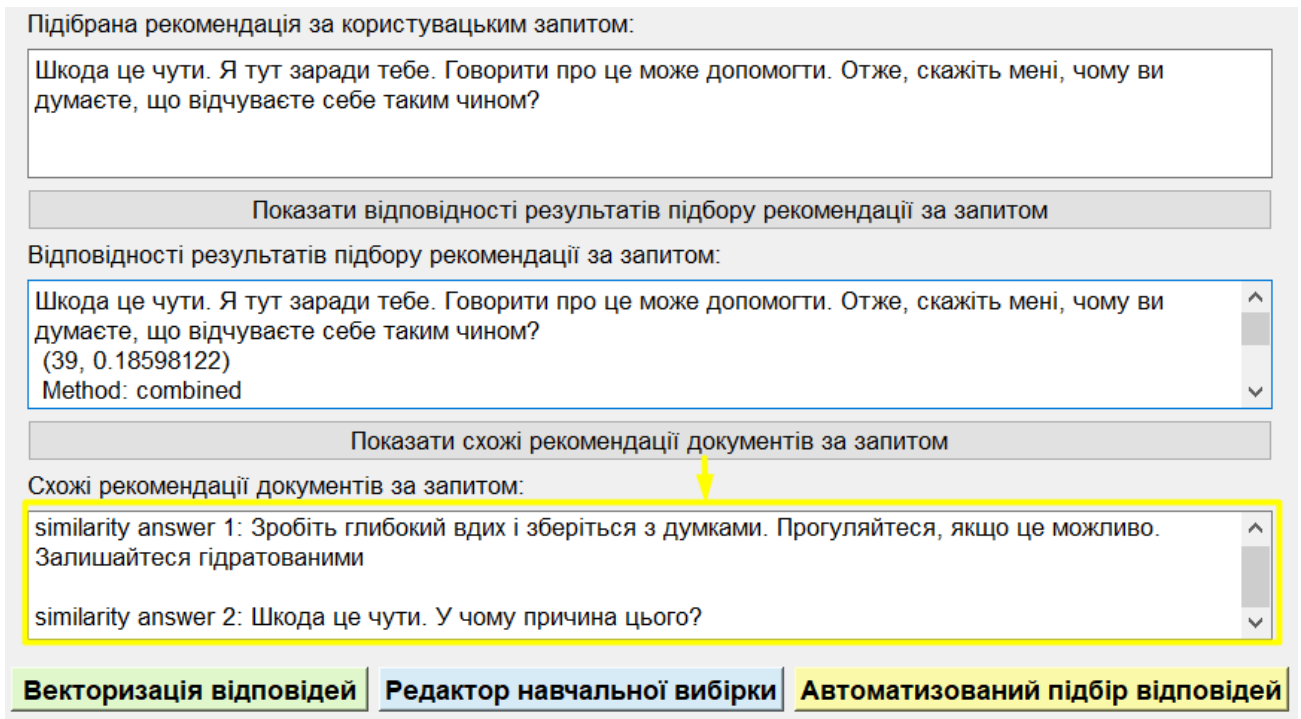


Рисунок 4.26 – Підбір схожих рекомендації

Отже, таким чином було описано особливості використання інформаційної системи автоматизованого визначення рекомендацій текстових документів з огляду функціонування трьох підсистем.

4.4 Дослідження ефективності та інтерпретація отриманих результатів

Для дослідження ефективності інформаційна система автоматизованого визначення рекомендацій текстових документів за користувацькими запитами було сформовано набір експериментальних даних, які мали позначку співвіднесення до відповідної комбінації «Запит-відповідь» і налічував 50 тестових запитів користувачів та відповідну мітку «Запит-відповідь».

Виконаний аналіз вибірки тестових даних по відношенню до вибірки навчальних даних має такий розподіл, як на рисунку 4.27.



Рисунок 4.27 – Розподіл запитів у навчальній та тестовій вибірках

При програмному тестуванні ефективності розробленого програмного забезпечення, воно показало високу ефективність, що склала 92 % коректного підбору (46 з 50 зразків коректно ідентифіковано). Також було проведено детальний аналіз некоректно ідентифікованих зразків, що виявив, що в запитах містилась інформація суперечлива, яку дійсно можна віднести до декількох категорій «Запит-відповідь». До прикладу, користувацький запит: *«Відчуваю себе нікчемно і самотньо. Життя для мене втратило сенс.»*. Перша відповідь видана програмою це *«Шкода це чути. Я тут заради тебе. Говорити про це може допомогти. Отже, скажіть мені, чому ви думаєте, що відчуваєте себе таким чином?»*, що належить до групи «Запит-відповідь» - смуток. Однак, в тестовому наборі даних воно було під тегом «нікчемний» (рисунок 4.28).

```

{
  "tag": "смуток",
  "patterns": [
    "Я відчуваю себе самотнім",
    "Я така самотня",
    "Я відчуваюся пригніченим",
    "Я відчуваюся засмученим",
    "Мені сумно",
    "Мені так самотньо",
    "Я відчуваюся порожньо",
    "У мене нікого немає"
  ],
  "responses": [
    "Шкода це чути. Я тут заради тебе.
  ],
  "tag": "нікчемний",
  "patterns": [
    "Я відчуваю себе таким нікчемним.",
    "Я нікому не подобаюсь",
    "Я не можу нічого зробити",
    "Я така нікчемна",
    "Ніщо більше не має сенсу"
  ],
  "responses": [
    "Цілком природно відчувати себе таким чином. Рі
    "Давайте додатково обговоримо, чому ви так себе
    "Спочатку я хочу повідомити вам, що ви не само

```

Рисунок 4.28 – Аналіз причин некоректної ідентифікації

Поряд з цим, варто зауважити що відповіді з групи «Запит-відповідь» - «нікчемний» були знайдені по семантичній схожості на 2-й та 3-й позиції (рисунок 4.29).

Відповідно, можна зробити висновок що в даному випадку користувач отримав коректний висновок, а недоліком є недосконалість навчальної множини, що була перекладена автоматизованим шляхом. Решта некоректних ідентифікацій також має схожий зміст, тому можна стверджувати що метод рекомендації текстових документів за запитаннями користувачів призначений для пошуку релевантних рекомендацій з бази текстових порад за текстовим запитом користувача та вхідними даними має набір даних з групами асоційованих запитів та відповідних їм асоційованих відповідей та користувацький запит, що перетворює їх у вихідні дані у вигляді текстової рекомендації релевантної до запиту є релевантним та може бути масштабованим.

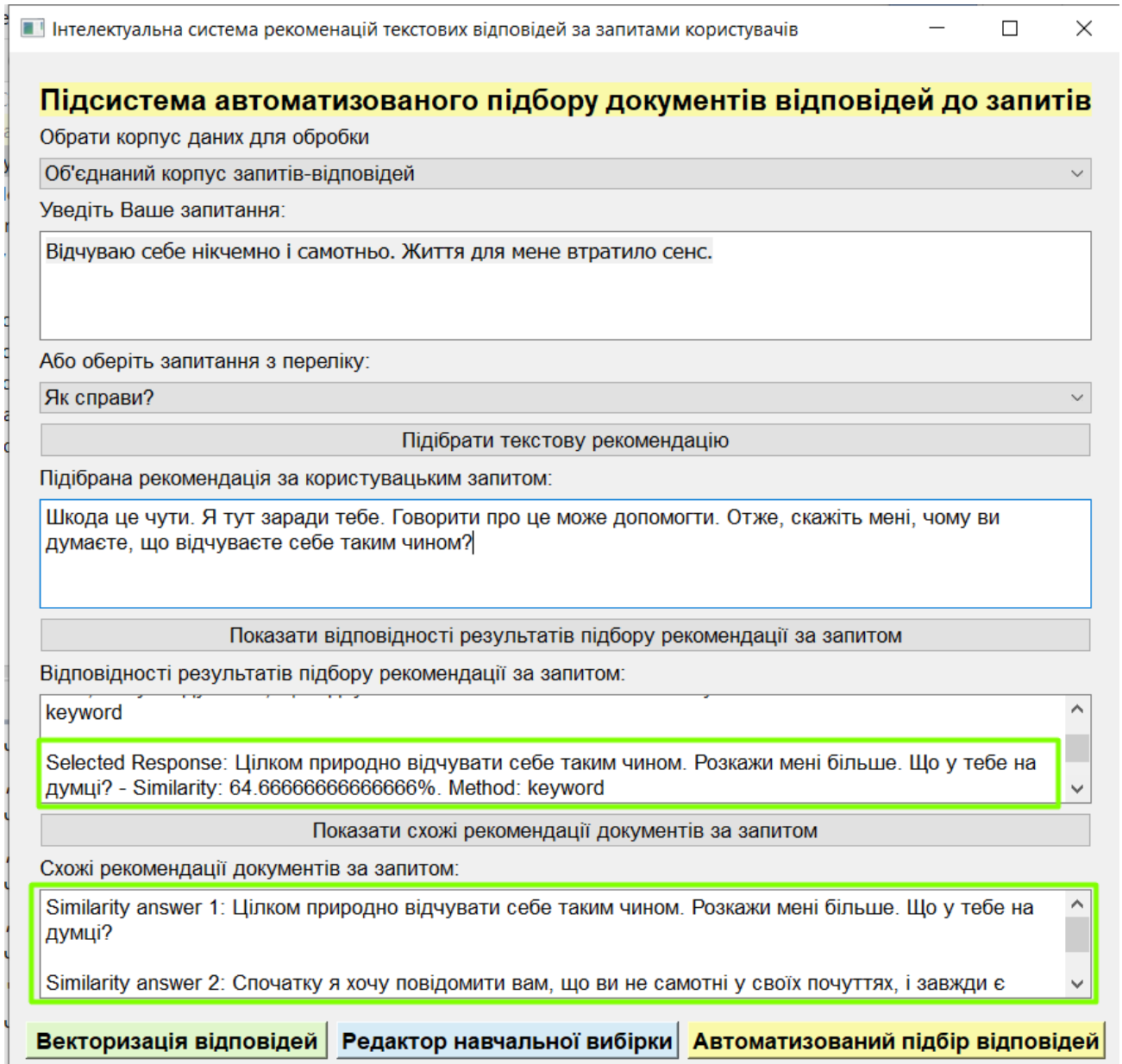


Рисунок 4.29 – Програмний аналіз некоректно ідентифікованого зразку

Також було проведено аналіз, яким методом переважно були знайдені користувацькі рекомендації. Графік розподілу користувацьких рекомендацій до методу пошуку наведено на рисунку 4.30

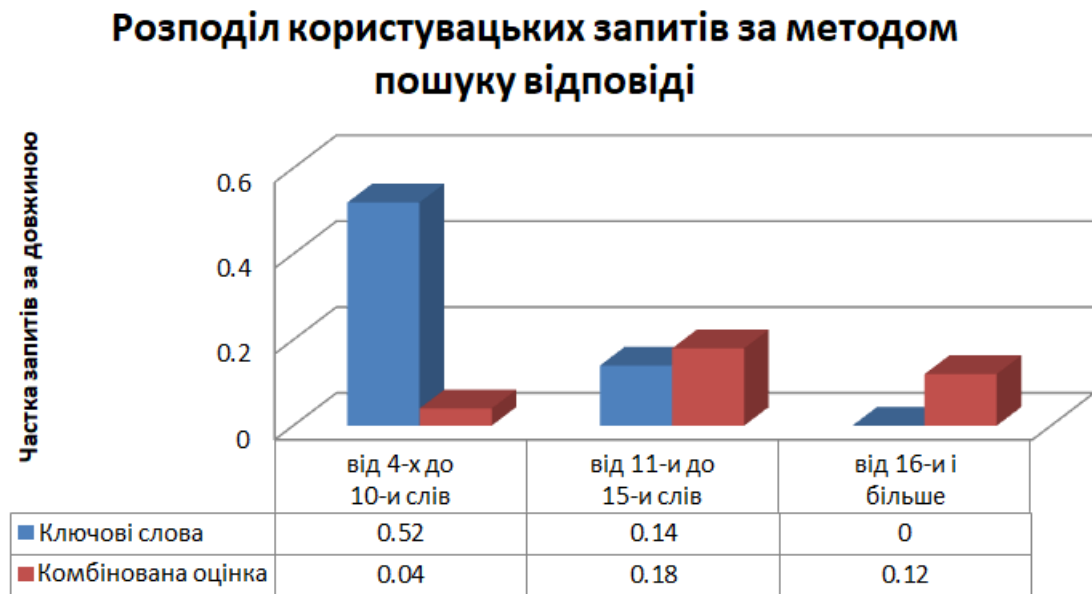


Рисунок 4.30 – Аналіз використовуваного методу пошуку рекомендацій

Як видно з діаграми 4.30 – на коротких запитах перевага надається методу, що працює з ключовими словами, а на довгих – перевага у комбінованого методу пошуку семантичної подібності. Що з стосується середньої довжини, тут результати неоднозначні, з незначною перевагою методу комбінованої оцінки.

Отже, таким чином було виконано дослідження ефективності та інтерпретація отриманих результатів, в ході якого було продемонстровано, що застосунок, оснований на методі рекомендації текстових документів за запитаннями користувачів призначений для пошуку релевантних рекомендацій з бази текстових порад за текстовим запитом користувача є релевантним. У ході досліджень було досягнуто точності ідентифікації 92 % на експериментальній вибірці у 50 зразків.

Висновки до четвертого розділу

В ході виконання четвертого розділу було виконано дослідження ефективності інформаційної системи автоматизованого визначення

рекомендацій текстових документів за користувацькими запитами на прикладі діалогів служби психологічної підтримки, в рамках якого:

1. Було описано прикладні особливості розробки компонентів інформаційної системи, яка є об'єктно-орієнтованою та складається з 3-х підсистем і 4-х класів, які реалізують заявлений функціонал інформаційної системи автоматизованого визначення рекомендацій текстових документів за користувацькими запитами.

2. Проведено прикладне тестування інформаційної системи засобами тест-кейсів, яке показало що весь заявлений функціонал працює коректно. Некоректно працюючих функцій у ході дослідження не виявлено.

3. Описано особливості використання інформаційної системи автоматизованого визначення рекомендацій текстових документів на прикладі діалогів служби психологічної підтримки.

4. Виконано дослідження ефективності, яке показало, що на коротких запитах перевага надається методу, що працює з ключовими словами, а на довгих – перевага у комбінованого методу пошуку семантичної подібності (на зваженій оцінці індексу Жаккара та косинусної схожості). Що ж стосується середньої довжини, тут результати неоднозначні, з незначною перевагою методу комбінованої оцінки. Метод показав високу точність, а саме було досягнуто точності ідентифікації 92 % на експериментальній вибірці у 50 зразків.

Загальні висновки

Кваліфікаційна робота магістра розв'язує задачу автоматизованої рекомендації текстових документів за запитаннями користувачів. Результатом роботи є метод, призначений для пошуку релевантних рекомендацій з бази текстових порад за текстовим запитом користувача, що перетворює вхідні дані у вигляді користувацького запиту, датасету із набором попередньо оброблених лематизованих асоціативних запитів та векторизованого корпусу у вихідні дані у формі текстової рекомендації релевантної до запиту, і програмна реалізація розробленого методу для предметної області служби психологічної підтримки.

За досягнення мети дослідження було вирішено наступне:

- Проведено аналіз предметної області, виділено актуальні напрями та окреслено сферу для автоматизації рекомендації текстових документів за запитаннями користувачів.

- Розроблено метод рекомендації текстових документів за запитаннями користувачів засобами NLP та описано його кроки та компоненти.

- Спроектовано інформаційну систему на базі методу рекомендації текстових документів за запитаннями користувачів засобами NLP.

- Виконано вибір засобів розробки інформаційної системи на базі методу рекомендації текстових документів для формування вибірок рекомендованих текстових документів за запитаннями користувачів.

- Розроблено програмну реалізацію методу рекомендації текстових документів за запитаннями користувачів засобами NLP для предметної області служби психологічної підтримки.

- Досліджено ефективність програмно реалізованого методу рекомендації текстових документів за запитаннями користувачів засобами NLP.

Результати виконання кваліфікаційної роботи магістра містять інновації та наукову новизну, зокрема було удосконалено метод рекомендації текстових документів за запитаннями користувачів засобами NLP, який дозволяє автоматизовано виконувати формування вибірок рекомендованих текстових

документів за запитаннями користувачів, виконуючи при цьому аналіз відповідності запиту користувача як за семантичною відповідністю до типового запиту до текстового документу, так і, у випадку її відсутності, знаходячи релевантні текстові документи за методом косинусної подібності, таким чином перетворюючи вхідні дані у вигляді користувацького запиту, вибірки із набором лематизованих асоціативних запитів та векторизованого корпусу текстових документів у вихідні дані у вигляді ранжованої множини релевантних рекомендованих текстових документів.

Розроблений у кваліфікаційній роботі метод має ряд переваг у порівнянні з існуючими методами. Зокрема, він дозволяє враховувати контекст запиту користувача, особливості його мовлення, а також структуру бази текстових порад. Це дозволяє підвищити точність і ефективність рекомендацій.

За темою роботи автором виконано наукову публікацію [41]. Одержані наукові і практичні результати кваліфікаційної роботи магістра доповідались:

- в доповіді на тему «Метод автоматизованого підбору відповідей на користувацькі запитання за семантичною подібністю» на Міжнародній науково-практичній конференції «Глушковські читання» (Київ, 26 жовтня 2023 року);

- в доповіді на тему «Використання метрик косинусної схожості та індексу Жаккара для інтелектуального аналізу семантичної подібності текстових документів» на XV Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2023» (Хмельницький, 17-18 листопада 2023 року).

Перелік посилань

1. Психологічна підтримка в умовах війни. URL: <https://dovidka.info/psychologichna-pidtrymka-v-umovah-vijny/>
2. Підтримка ментального здоров'я в часи війни. URL: <https://niss.gov.ua/news/komentari-ekspertiv/pidtrymka-mentalnoho-zdorovya-v-chasy-viyny>
3. Natural Language Processing (NLP). URL: <https://www.ibm.com/cloud/learn/natural-language-processing>
4. Роль ІІІ в еволюції інтелектуального пошуку. URL: <https://ts2.space/uk/роль-ііі-в-еволюції-інтелектуального-п/>
5. Cosine Similarity. URL: <https://www.geeksforgeeks.org/cosine-similarity/>
6. Методи та алгоритми вилучення ключових слів. URL: <https://cyberleninka.ru/article/n/metody-i-algoritmy-izvlecheniya-klyuchevyh-slov>
7. Text Document Clustering: Wordnet vs. TF-IDF vs. Word Embeddings. URL: <https://aclanthology.org/2021.gwc-1.24.pdf>
8. Lifeline Ukraine. URL: <https://lifelineukraine.com/activities>
9. Teenergizer. URL: <https://teenergizer.org/articles/>
10. Онлайн-модулі психологічної підтримки громадян України у воєнний час. URL: <https://visnyk.naps.gov.ua/index.php/journal/article/view/307>
11. Питання психолого-педагогічної підтримки дітей під час війни: інформаційно-технологічне забезпечення. URL: <https://ojs.itup.com.ua/index.php/iiu/article/view/475>
12. Диджиталізація для підвищення доступності послуг охорони здоров'я для ВПО в умовах воєнного стану. URL: <https://repo.knmu.edu.ua/bitstream/123456789/31089/1/Збірка%20конф%20ОЗД%20ХНМУ%2028.10.2022.pdf#page=101>
13. Mental Health Conversational Data. URL: <https://www.kaggle.com/datasets/elvis23/mental-health-conversational-data/>

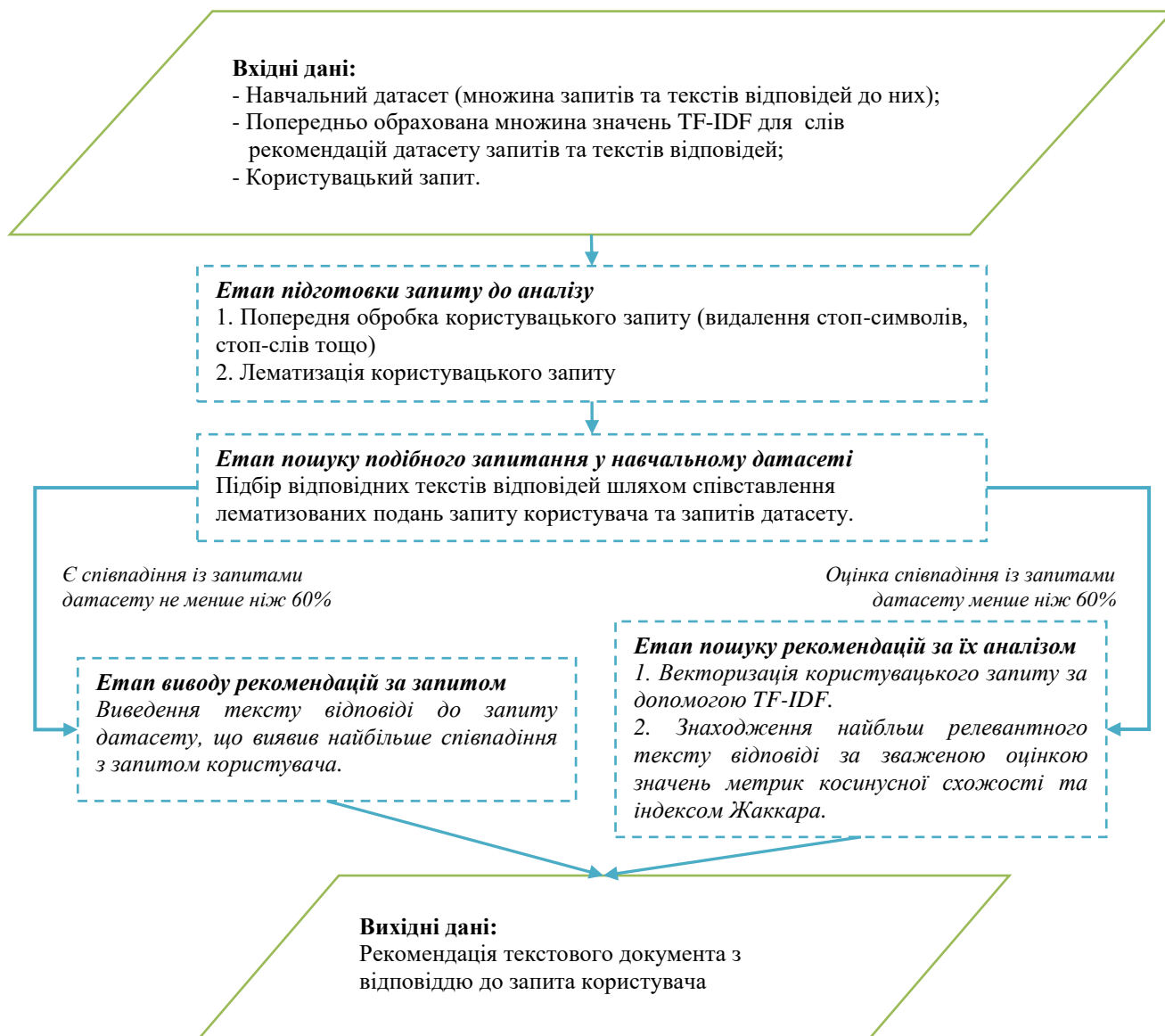
14. Mental Health FAQ for Chatbot. URL: <https://www.kaggle.com/datasets/narendrageek/mental-health-faq-for-chatbot>
15. Cosine Similarity. URL: <https://www.learndatasci.com/glossary/cosine-similarity>
16. Петренко, Владислав Вікторович. Рекомендаційна система на базі алгоритму колаборативної фільтрації. BS thesis. КПІ ім. Ігоря Сікорського, 2023
17. Косинус подібності. URL: https://uk.wikipedia.org/wiki/Косинус_подібності
18. Коефіцієнт Жаккара. URL: https://uk.wikipedia.org/wiki/Коефіцієнт_Жаккара
19. Welcome to Python. URL: <https://www.python.org/>
20. J. Prateek. Artificial intelligence with python. Packt Publishing Ltd, 2017
21. PyCharm Download. URL: <https://www.jetbrains.com/pycharm/download/>
22. Lopez-Carreño, Joseph, et al. "Artificial-Intelligence-Based School Assistant for Detecting the Behavior of University Students." 2022 IEEE 2nd International Conference on Advanced Learning Technologies on Education & Research (ICALTER). IEEE, 2022
23. Вибір середовища розробки для програмування. URL: <https://uk.soringprepair.com/software-for-writing-code/>
24. PyQt5 Documentation. URL: <https://pyqt.org/project/PyQt5/>
25. PyQt. URL: <https://uk.wikipedia.org/wiki/PyQt>
26. PyQt5 Qtquick2 Example. URL: <https://gitplanet.com/project/pyqt5-qtquick2-example>
27. Natural Language Toolkit. URL: <https://www.nltk.org/>
28. Перші кроки в NLP: розглядаємо Python-бібліотеку NLTK в реальному завданні. URL: <https://dou.ua/lenta/articles/first-steps-in-nlp-nltk/>
29. Що таке лематизація? URL: <https://uk.theastrologypage.com/lemmatization>
30. Лематизація. URL: <https://uk.wikipedia.org/wiki/Лематизація>
31. Стемінг. URL: <https://uk.wikipedia.org/wiki/Стемінг>

32. NLTP Python. URL: <https://learn.ztu.edu.ua/mod/resource/view.php?id=170111>
33. What is the difference between lemmatization vs stemming?. URL: <https://stackoverflow.com/questions/1787110/what-is-the-difference-between-lemmatization-vs-stemming>
34. Python | Lemmatization with NLTK. URL: <https://www.geeksforgeeks.org/python-lemmatization-with-nltk/>
35. Обробка природної мови (NLP) у Python з кодом. URL: <https://oleg-dubetsky.medium.com/обробка-природної-мови-nlp-у-python-з-кодом-частина-1-83d588b3ad71>
36. Застосування методів машинного навчання для виявлення психічних захворювань. URL: https://ai.kpi.ua/ua/masters/thesis/28521smat-tsuprun_magistr.pdf
37. Part-of-speech tagging. URL: https://en.wikipedia.org/wiki/Part-of-speech_tagging
38. Mishra, Vikash Kumar, et al. "Implementation of Tokenization in Natural Language Processing Using NLTK Module of Python." Mathematics and Computer Science Volume 1 (2023): 493-504.
39. Raschka, Sebastian, et al. Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python. Packt Publishing Ltd, 2022.
40. Методи векторизації текстів. URL: <https://medium.com/@bigdataschool/4-методи-векторизації-текстів>
41. Козенко О.В., Мазурець О.В., Молчанова М.О., Собко О.В. Використання метрик косинусної схожості та індексу Жаккара для інтелектуального аналізу семантичної подібності текстових документів. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2023». Хмельницький, 2023. с. 146-147. URL: <https://kn.khmnu.edu.ua/wp-content/uploads/sites/18/apkn-2023-corpuspaper.pdf>

ДОДАТКИ

Додаток А

Схема методу пошуку рекомендації текстових документів за запитаннями користувачів



Додаток Б

**Інформаційна структура системи автоматизованого визначення
рекомендацій текстових документів**

Додаток В

**Схема роботи методу обробки запитань користувача `get_answer()`
інформаційної системи автоматизованого визначення рекомендацій
текстових документів**



Додаток Г

Світлини наукових публікацій, виконаних при роботі над кваліфікаційною роботою магістра

(ксерокопії титульної сторінки, сторінки змісту та всіх сторінок із публікацією)

Наукова публікація:

Козенко О.В., Мазурець О.В., Молчанова М.О., Собко О.В. Використання метрик косинусної схожості та індексу Жаккара для інтелектуального аналізу семантичної подібності текстових документів. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2023». Хмельницький, 2023. с. 146-147.

Міністерство освіти і науки України
Хмельницький національний університет



ЗБІРНИК НАУКОВИХ ПРАЦЬ
за матеріалами XV Всеукраїнської науково-практичної конференції
«Актуальні проблеми комп'ютерних наук АПКН-2023»

17-18 листопада 2023

Хмельницький 2023

Козенко О.В., Мазурець О.В., Молчанова М.О., Собко О.В. Використання метрик косинусної схожості та індексу Жаккара для інтелектуального аналізу семантичної подібності текстових документів	146
Комін А.С., Бойко О.В. Архітектурне рішення для підсистеми підтримки управління гібридною енергосистемою з використанням машинного навчання на мобільних пристроях	148
Кузьмін А.А. Концепція інформаційної системи для автоматизованої генерації цифрового контенту на основі штучного інтелекту	153
Кучменко К.Ю., Праворська Н.І. Ігровий застосунок у жанрі «платформер» з інтерфейсом управління на основі голосової взаємодії з використанням технологій Unity	157
Лантєв М.П., Лисий А.М., Сергєєв Є.В., Віжєвський П.В. Метод криптографічного замісту протоколів в засобах комунікації інтернету речей	161
Левандовський А.О., Муляр І.В. Метод аналізу трафіку з метою виявлення атак на комплексні системи захисту інформації	163
Лизун О.О. Методи та засоби виявлення зловмисних дроперів в комп'ютерних системах	166
Мазур К.Р., Пасічник О.А., Скрипник Т.К. Метод виявлення боєприпасів, що не розірвались, за зображенням з тепловізора засобами глибокого навчання	168
Малицький Т.Б., Чешун О.В., Чешун В.М. Математична інтерпретація концепції захисту інформаційних ресурсів корпоративної мережі із застосуванням імовірнісних критеріїв довіри	172
Мандрик А.І., Лисенко С.М. Метод оптимізації планування проектів та формування команд з використанням генетичного алгоритму	177
Манзюк Е.А. Застосування розпаралелювання для криптографії з використанням губчастої структури	181

УДК 004.4

Козенко О.В., Мазурець О.В., Молчанова М.О., Собко О.В.

Хмельницький національний університет

ВИКОРИСТАННЯ МЕТРИК КОСИНУСНОЇ СХОЖОСТІ ТА ІНДЕКСУ ЖАККАРА ДЛЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ СЕМАНТИЧНОЇ ПОДІБНОСТІ ТЕКСТОВИХ ДОКУМЕНТІВ

Робота присвячена дослідженню та застосуванню метрик косинусної схожості та індексу Жаккара у контексті рекомендації текстових документів відповідно до запитань користувачів. Згідно до проведеного аналізу, обрано використання зазначених метрик у контексті порівняння векторів, що представляють текстові документи. Отримані результати аналізу можуть бути використані для подальшого вдосконалення систем рекомендації та інформаційного пошуку, зокрема в областях, де ключовим є семантичне порівняння текстової інформації.

The work is devoted to the study and application of cosine similarity metrics and the Jaccard index in the context of recommending text documents according to user queries. According to the conducted analysis, it was chosen to use the specified metrics in the context of comparing vectors representing text documents. The results of the analysis can be used for further improvement of recommendation systems and information retrieval, in particular in areas where semantic comparison of textual information is key.

Наразі існує багато методів пошуку релевантних документів за запитом користувача. Один з найпоширеніших методів – це використання метрик подібності. Метрики подібності використовуються для вимірювання подібності між запитом та документом. Чим більша подібність між запитом та документом, тим більш релевантний документ [1].

Нижче наведено деякі поширені метрики подібності [1]:

– TF-IDF – це метрика, яка враховує частоту виникнення слів у документі та частоту виникнення слів у наборі документів.

– Косинусна схожість (Cosine similarity) – це метрика, яка враховує кут між векторами запиту та документа.

– Індекс Жаккара (Jaccard similarity) – це метрика, яка враховує кількість спільних слів у запиті та документі.

Метрики косинусної схожості та індексу Жаккара пропонуються використовувати у рамках методу рекомендації текстових документів за запитаннями користувачів на базі служби психологічної підтримки призначений для пошуку релевантних рекомендацій з бази текстових порад за текстовим запитом користувача та вхідними даними має датасет із набором очищених лематизованих асоціативних запитів, векторизований корпус та користувацький запит, що

перетворює їх у вихідні дані у вигляді текстової рекомендації релевантної до запиту.

Основна ідея косинусної схожості полягає в тому, що якщо два вектори спрямовані в одному напрямку (кут між ними дорівнює 0 градусів), то їхні вектори мають найвищу косинусну схожість та вважаються дуже схожими. Навпаки, якщо кут між векторами дорівнює 90 градусів, то косинусна схожість дорівнює 0, що вказує на максимальну різницю між ними [2].

Для обчислення косинусної схожості між двома векторами тексту (наприклад, словниковими представленнями документів), використовується наступна формула:

$$\text{Cosine Similarity}(A, B) = (A \cdot B) / (|A| * |B|),$$

де: A та B – вектори тексту або документів. $A \cdot B$ – скалярний добуток векторів A і B . $|A|$ та $|B|$ – норми (довжини) векторів A і B .

Значення косинусної схожості зазвичай лежать в діапазоні від -1 (повна протилежність) до 1 (повна ідентичність), де 0 означає відсутність схожості.

Цей метод широко використовується в пошукових системах, рекомендаційних системах, аналізі тексту та класифікації документів для визначення ступеня схожості між текстами та відбору найбільш релевантних результатів. У роботі буде застосований як одна із метрик подібності запиту користувача до наявних шаблонів запитів з метою отримання релевантної відповіді.

Індекс Жаккара (Jaccard index), також відомий як Жаккардова схожість чи коефіцієнт Жаккара, є метрикою схожості, яка використовується для порівняння множин елементів. Ця метрика визначає ступінь схожості двох множин шляхом вимірювання кількості спільних елементів в обох множинах відносно загальної кількості унікальних елементів [3].

Індекс Жаккара обчислюється за наступною формулою:

$$J(A, B) = |A \cap B| / |A \cup B|,$$

де: $J(A, B)$ – індекс Жаккара між множинами A і B . $|A \cap B|$ – кількість спільних елементів між множинами A і B . $|A \cup B|$ – кількість унікальних елементів у множинах A і B разом.

Значення Індeksu Жаккара можуть лежати в діапазоні від 0 до 1, де 0 означає відсутність спільних елементів, а 1 означає повну ідентичність множин. Значення, близькі до 1, вказують на високу схожість множин, тоді як значення, близькі до 0, свідчать про низьку схожість.

Індекс Жаккара широко використовується в різних галузях, включаючи аналіз тексту, рекомендаційні системи, пошукові системи та біологічні дослідження. В роботі буде використовуватися для визначення схожості між ключовими словами користувацького запиту та наявних рекомендацій.

Перелік посилань

1. Cosine Similarity. URL: <https://www.geeksforgeeks.org/cosine-similarity/>
2. Cosine Similarity. URL: <https://www.learnatasci.com/glossary/cosine-similarity/>
3. Jaccard Similarity Made Simple: A Beginner's Guide to Data Comparison. URL: <https://medium.com/@mayurdhvajsinhjadeja/jaccard-similarity-34e2c15fb524>

Додаток Д

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

МЕТОД РЕКОМЕНДАЦІЇ ТЕКСТОВИХ ДОКУМЕНТІВ ЗА ЗАПИТАННЯМИ КОРИСТУВАЧІВ ЗАСОБАМИ NLP



Виконав:
студент 2 курсу, групи КНм-22-1
Козенко Олександр Васильович



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

Актуальність

У сучасному світі спостерігається зростання попиту на послуги психологічної підтримки. Це пов'язано з низкою факторів, зокрема, з ускладненням життя, збільшенням кількості стресових ситуацій, поширенням психосоматичних захворювань, особливо в умовах воєнного стану. У зв'язку з цим виникає необхідність у розробці ефективних методів надання психологічної допомоги.

Одним із перспективних напрямків розвитку психологічної допомоги є застосування інформаційних технологій. Інтернет-ресурси використовуються для надання консультацій, проведення тренінгів, організації спільнот для підтримки людей, які страждають від психологічних проблем, однак, за даними досліджень, в майбутнього прогнозується зріст людей з психологічними проблемами, і наявні інформаційні ресурси не зможуть повноцінно задовольнити наявний попит. У зв'язку з цим виникає необхідність у розробці методів, що дозволяють рекомендувати користувачам релевантні текстові документи, які можуть допомогти їм у вирішенні їхніх проблем.

Розроблений у кваліфікаційній роботі метод має ряд переваг у порівнянні з існуючими методами. Зокрема, він дозволяє враховувати контекст запиту користувача, особливості його мовлення, а також структуру бази текстових порад. Це дозволяє підвищити точність і ефективність рекомендацій.

Мета і задачі роботи

Метою кваліфікаційної роботи магістра є розробка методу рекомендації текстових документів за запитаннями користувачів засобами NLP, а також відповідна програмна реалізація для предметної області служби психологічної підтримки.

У рамках досягнення мети необхідно виконати наступні задачі:

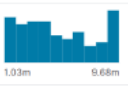
- Провести аналіз предметної області, виділити актуальні напрями та окреслити сферу для автоматизації рекомендації документів за запитаннями користувачів.
- Створити метод рекомендації текстових документів за запитаннями користувачів засобами NLP та описати його кроки.
- Виконати проектування інформаційної системи на базі методу рекомендації текстових документів за запитаннями користувачів засобами NLP для предметної області служби психологічної підтримки.
- Виконати вибір засобів розробки інформаційної системи та розробити відповідну програмну реалізацію методу.
- Виконати дослідження ефективності програмно реалізованого методу рекомендації текстових документів за запитаннями користувачів засобами NLP.

Набори даних для реалізації методу

Датасети: **«Mental Health Conversational Data»** та

«Mental Health FAQ for Chatbot»

```
[{"tag": "about",
  "botturns": ["Who are you?", "What are you?", "Who are you?", "Tell me more about yourself.", "What is your name?", "What should I call you?", "What's your name?", "Tell me about y",
  "responses": ["I'm Randee, your Personal Therapeutic AI Assistant. How are you feeling today?", "I'm Randee, a Therapeutic AI Assistant designed to assist you. Tell me about yourse",
  },
  {"tag": "hello",
  "botturns": ["What are you doing?",
  "responses": ["I can provide general advice regarding anxiety and depression, answer questions related to mental health and make daily conversations. Do not consider me as a sub",
  },
  {"tag": "creation",
  "botturns": ["Who created you?", "How were you made?", "How were you created?",
  "responses": ["I was created by >.", "I was trained on a text dataset using Deep Learning & Natural Language Processing techniques", "The real question is: Who created you?"]
  },
  {"tag": "name",
  "botturns": ["My name is ", "I am name.", "I go by ",
  "responses": ["Oh nice to meet you. Tell me how was your week?", "Nice to meet you. To tell me. How do you feel today?", "That's a great name. Tell me more about yourself."]
  },
  {"tag": "help",
  "botturns": ["Could you help me?", "give me a hand please", "Can you help?", "What can you do for me?", "I need support", "I need help", "Support me please",
  "responses": ["Sure. Tell me how can I assist you?", "Tell me your problem so that I can assist you", "Yes, sure. How can I help you?"]
  },
  {"tag": "sad",
  "botturns": ["I am feeling lonely", "I am so lonely", "I feel down", "I feel sad", "I am sad", "I feel so lonely", "I feel empty", "I don't have anyone",
  "responses": ["It's sorry to hear that. I'm here for you. Talking about it might help. So, tell me why do you think you're feeling this way?", "I'm here for you. Could you tell me ab",
  },
  {"tag": "stressed",
  "botturns": ["I am so stressed out", "I am so stressed", "I feel stuck", "I still feel stressed", "I am so burned out",
  "responses": ["What do you think is causing that?", "Take a deep breath and gather your thoughts. Go take a walk if possible. Stay hydrated", "Give yourself a break. Go easy on",
  },
  {"tag": "useless",
  "botturns": ["I feel so useless.", "No one likes me.", "I can't do anything.", "I am so useless", "Nothing makes sense anymore",
  "responses": ["It's only natural to feel this way. Tell me more. What else is on your mind?", "Let's discuss further why you're feeling this way.", "I first want to let you know",
  },
  {"tag": "depressed",
  "botturns": ["I can't take it anymore", "I am so depressed", "I think I'm depressed", "I have depression",
  "responses": ["It helps to talk about what's happening. You're going to be okay", "Talk to me. Tell me more. It helps if you open up yourself to someone else.", "Sometimes when"]
  },
```

Question_ID	Questions	Answers
unique id of each question	question related to mental health	corresponding answer
	98 unique values	98 unique values
1598148	What does it mean to have a mental illness?	Mental illnesses are health conditions that disrupt a person's thoughts, emotions, relationships, ...
2118618	Who does mental illness affect?	It is estimated that mental illness affects 1 in 5 adults in America, and that 1 in 24 adults have a...
6361820	What causes mental illness?	It is estimated that mental illness affects 1 in 5 adults in America,

Метрики для оцінки семантичної схожості

Косинусна схожість:

$$\text{Cosine Similarity } (A, B) = (A \cdot B) / (||A|| * ||B||),$$

де: A та B – вектори тексту або документів. $A \cdot B$ – скалярний добуток векторів A і B . $||A||$ та $||B||$ – норми (довжини) векторів A і B .

Індекс Жаккара :

$$J(A, B) = |A \cap B| / |A \cup B|,$$

де: $J(A, B)$ - індекс Жаккара між множинами A і B . $|A \cap B|$ - кількість спільних елементів між множинами A і B . $|A \cup B|$ - кількість унікальних елементів у множинах A і B разом.

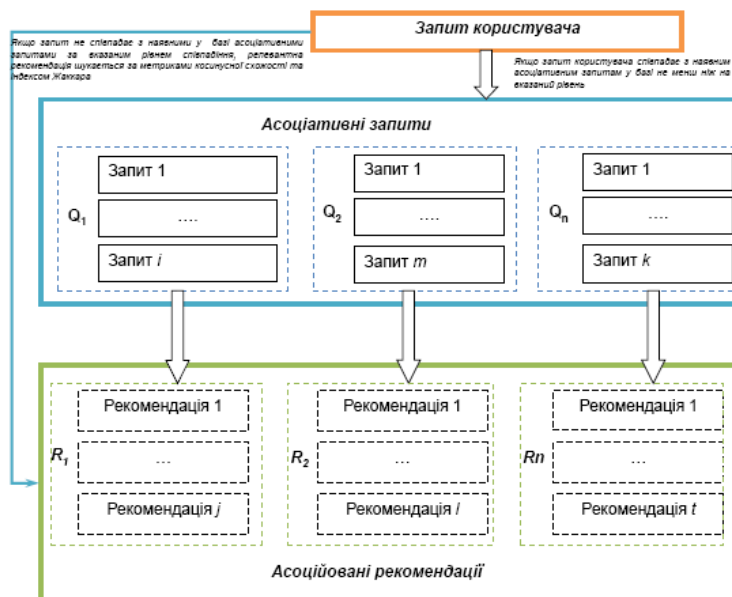


Схема пошуку
рекомендації текстових
документів за
запитаннями
користувачів

```
{
  "tag": "допомога",
  "patterns": [
    "Не могли б ви мені допомогти?",
    "дайте мені руку, будь ласка",
    "Ви можете допомогти?",
    "Що ти можеш зробити для мене?",
    "Мені потрібна підтримка",
    "Мені потрібна допомога",
    "Підтримай мене, будь ласка"
  ],
  "responses": [
    "Звичайно. Розкажіть, чим я можу вам допомогти",
    "Розкажіть про свою проблему, щоб я міг вам допомогти",
    "Так, звичайно. Чим я можу вам допомогти?"
  ]
},
{
  "tag": "смуток",
  "patterns": [
    "Я відчуваю себе самотнім",
    "Я така самотня",
    "Я почувалася пригніченою",
    "Я почувалася засмученою",
    "Мені сумно",
  ]
}
```

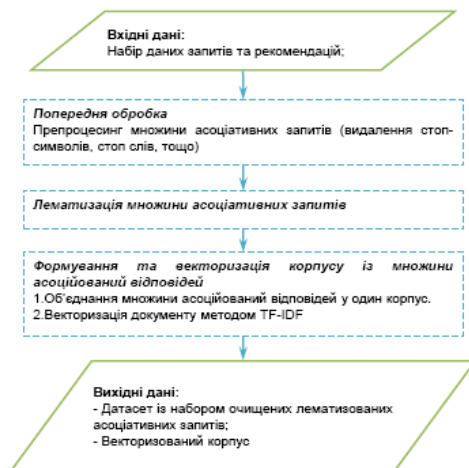


Схема підготовки
вхідних даних методу
пошуку рекомендації
текстових документів
за запитаннями
користувачів

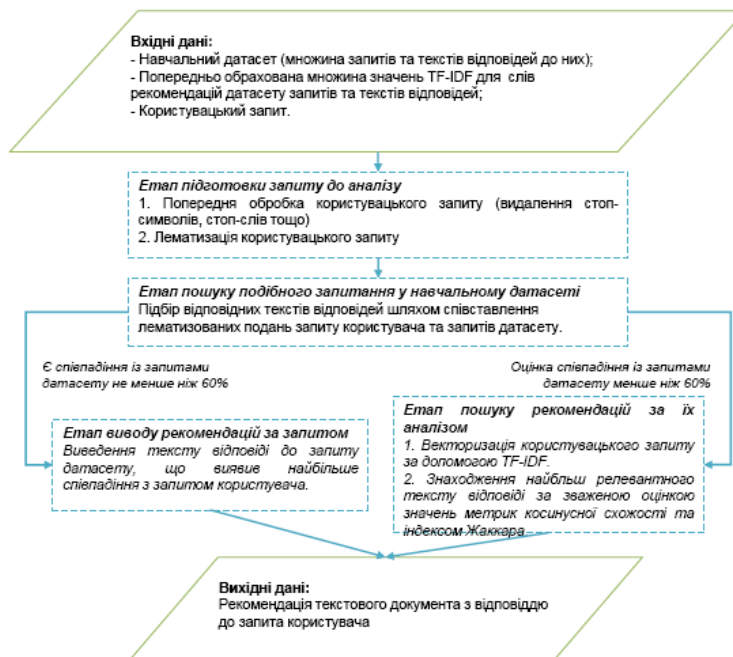


Схема методу пошуку
рекомендації текстових
документів за
запитаннями
користувачів



Інформаційна
структура системи
автоматизованого
визначення
рекомендацій
текстових документів



Схема та функції
головної підсистеми
інформаційної системи

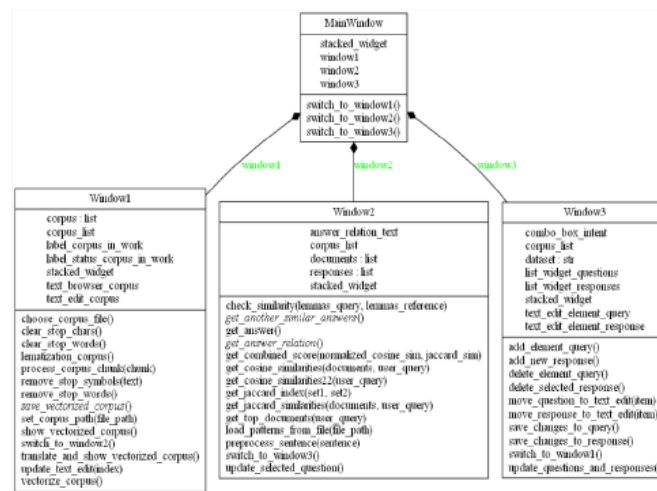
Засоби розробки:

- Мова програмування: *Python*
- Середовище програмування: *PyCharm* (інтегроване середовище розробки)

Спеціалізовані програмні розширення:

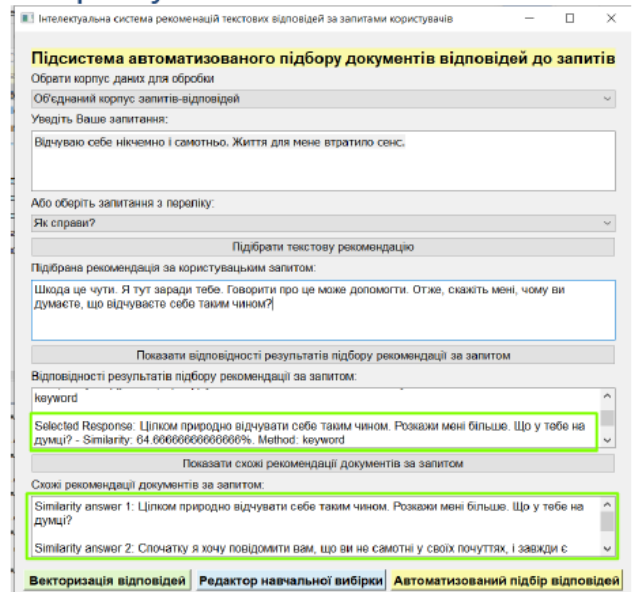
- *PyQt5* – для розробки інтерфейсу
- *NLTK* – лематизація та препроцесинг
- *Scikit-learn* – векторизація та косинусна подібність

Програмна архітектура системи



Програмна реалізація методу автоматизованої рекомендації текстових документів за запитаннями користувачів засобами NLP

Підсистема автоматизованого підбору документів відповідей до запитів

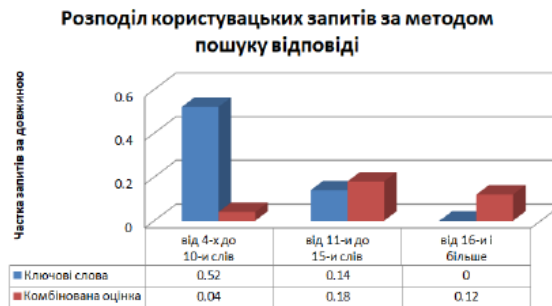


Дослідження ефективності методу

Співвідношення навчальної та тестової вибірок



Дослідження ефективності методу



При програмному тестуванні ефективності розробленого програмного забезпечення, воно показало високу ефективність, що склала 92 % коректного підбору (46 з 50 зразків коректно ідентифіковано).

Дослідження ефективності.

Аналіз некоректно ідентифікованих зразків

Виявлено, що в запитах містилась інформація суперечлива, яку дійсно можна віднести до декількох категорій «Запит-відповідь».

Приклад користувацького запиту: *«Відчуваю себе нікчемно і самотньо. Життя для мене втратило сенс»*.

Перша відповідь видана програмою це *«Шкода це чути. Я тут заради тебе. Поговорити про це може допомогти. Отже, скажіть мені, чому ви думаєте, що відчуваєте себе таким чином?»*, що належить до групи «Запит-відповідь» - «смуток». Однак, в тестовому наборі даних воно було під тегом «нікчемний»

Дослідження ефективності. Аналіз некоректно ідентифікованих зразків

```

{
  "tag": "смуток",
  "patterns": [
    "Я відчуваю себе самотнім",
    "Я така самотня",
    "Я відчуваю пригніченим",
    "Я відчуваю засмученим",
    "Мені сумно",
    "Мені так самотньо",
    "Я відчуваю порожньо",
    "У мене нікого немає"
  ],
  "responses": [
    "Шкода це чути. Я тут заради тебе."
  ]
}

{
  "tag": "нікчемний",
  "patterns": [
    "Я відчуваю себе таким нікчемним.",
    "Я нікому не подобаюсь",
    "Я не можу нічого зробити",
    "Я така нікчемна",
    "Ніщо більше не має сенсу"
  ],
  "responses": [
    "Цілком природно відчувати себе таким чином. Рі",
    "Давайте додатково обговоримо, чому ви так себ",
    "Спочатку я хочу повідомити вам, що ви не само"
  ]
}

```

Порівняння запитів-відповідей «смуток» та «нікчемний»

Дослідження ефективності. Аналіз некоректно ідентифікованих зразків

Поряд з цим, варто зауважити що відповіді з групи «Запит-відповідь» - «нікчемний» були знайдені по семантичній схожості на 2-й та 3-й позиції

Інтелектуальна система рекомендацій текстових відповідей за запитами користувачів

Підсистема автоматизованого підбору документів відповідей до запитів

Обрати корпус даних для обробки
Об'єднаний корпус запитів-відповідей

Уведіть Ваше запитання:
Відчуваю себе нікчемно і самотньо. Життя для мене втратило сенс.

Або оберіть запитання з переліку:
Як справи?

Підбрати текстову рекомендацію

Підбрана рекомендація за користувацьким запитом:
Шкода це чути. Я тут заради тебе. Говорити про це може допомогти. Отже, скажіть мені, чому ви думаєте, що відчуваєте себе таким чином?

Показати відповідності результатів підбору рекомендації за запитом

Відповідності результатів підбору рекомендації за запитом:
keyword

Selected Response: Цілком природно відчувати себе таким чином. Розкажи мені більше. Що у тебе на думці? - Similarity: 64.66666666666666%. Method: keyword

Показати схожі рекомендації документів за запитом

Схожі рекомендації документів за запитом:
Similarity answer 1: Цілком природно відчувати себе таким чином. Розкажи мені більше. Що у тебе на думці?
Similarity answer 2: Спочатку я хочу повідомити вам, що ви не самотні у своїх почуттях, і завжди є

Векторизація відповідей | Редактор навчальної вибірки | Автоматизований підбір відповідей

Висновки

За досягнення мети розробки методу рекомендації текстових документів за запитаннями користувачів було вирішено задачі:

- ❖ Проведено аналіз предметної області, виділено актуальні напрями та окреслено сферу для автоматизації рекомендації текстових документів за запитами користувачів.
- ❖ Розроблено метод рекомендації текстових документів за запитаннями користувачів засобами NLP та описано його кроки та компоненти.
- ❖ Спроектовано інформаційну систему на базі методу рекомендації текстових документів за запитаннями користувачів засобами NLP.
- ❖ Виконано вибір засобів розробки інформаційної системи на базі методу рекомендації текстових документів для формування вибірок рекомендованих текстових документів за запитаннями користувачів.
- ❖ Розроблено програмну реалізацію методу рекомендації текстових документів за запитаннями користувачів засобами NLP для предметної області служби психологічної підтримки.
- ❖ Досліджено ефективність програмно реалізованого методу рекомендації текстових документів за запитаннями користувачів засобами NLP.

Висновки

Кваліфікаційна робота магістра розв'язує задачу автоматизованої рекомендації текстових документів за запитаннями користувачів. Результатом роботи є метод рекомендації текстових документів за запитаннями користувачів призначений для пошуку релевантних текстових документів з навчального датасету (множини текстів запитів та відповідей до них) до користувацького запиту з використанням засобів обробки природної мови.

За темою роботи автором виконано наукову публікацію (тези конференції).

Одержані наукові і практичні результати кваліфікаційної роботи магістра доповідались:

- в доповіді на тему «Метод автоматизованого підбору відповідей на користувацькі запитання за семантичною подібністю» на Міжнародній науково-практичній конференції «Глушковські читання» (Київ, 26 жовтня 2023 року);
- в доповіді на тему «Використання метрик косинусної схожості та індексу Жаккара для інтелектуального аналізу семантичної подібності текстових документів» на XV Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2023» (Хмельницький, 17-18 листопада 2023 року).

Ім'я користувача:
Кафедра КН

Дата перевірки:
12.12.2023 09:10:21 EET

Дата звіту:
12.12.2023 09:28:17 EET

ID перевірки:
1015995975

Тип перевірки:
Doc vs Internet + Library

ID користувача:
100005671

Назва документа: КНМ-22-1 Козенко

Кількість сторінок: 83 Кількість слів: 13740 Кількість символів: 107626 Розмір файлу: 2.91 MB ID файлу: 1015678871

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

7.47%

Схожість

Найбільша схожість: 2.1% з джерелом з Бібліотеки (ID файлу: 1011421027)

6.56% Джерела з Інтернету

593

Сторінка 85

4.99% Джерела з Бібліотеки

106

Сторінка 89

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

2

Підозріле форматування

31
сторінка

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 2.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилки в документах: 11%**

ID: 122631 Назва: КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА на тему Метод рекомендації текстових документів за запитаннями користувачів засобами NLP Додано в БД: 2023-12-12 Автора: О.В. Козенко Керівники: О.В. Мазурець Консультанти: Опоненти:	Документ		Сумарний збіг по Базі Даних	
	Символи	Лексеми	Символи	Лексеми
	92122	1330	4208 (5%)	71 (5%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

РІШЕННЯ ЕКСПЕРНОЇ КОМІСІЇ
КАФЕДРИ КОМП'ЮТЕРНИХ НАУК
ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ МАГІСТРА ДО ЗАХИСТУ
ЗА РЕЗУЛЬТАТАМИ АНАЛІЗУ ЗВІТУ ПОДІБНОСТІ

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

Назва: Метод рекомендації текстових документів за запитаннями користувачів засобами NLP

Автор: Козенко Олександр Васильович

Спеціальність: 122 – Комп'ютерні науки

Освітня програма: освітньо-професійна

Науковий керівник: доцент кафедри КН Мазурець Олександр Вікторович

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.	—

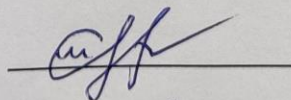
Підтвердження:

Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

- 1) За програмою Anti-Plagiarism виявлені 5%, які є фрагментарними, не більше 2% на джерело – містять поширені конструкції, загальновідомі терміни та визначення.
- 2) За програмою UNICHECK виявлені 7,47%, які є фрагментарними, не більше 2,1% на джерело – містять поширені конструкції, загальновідомі терміни та визначення.

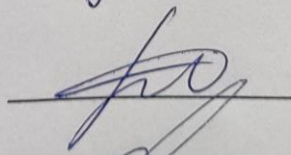
Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ідентичності/схожості, складає 5% і 7,47% відповідно, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи



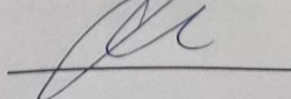
Олександр МАЗУРЕЦЬ

Гарант ОП



Руслан БАГРІЙ

Завідувач кафедри КН



Олександр БАРМАК



ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу магістра

гр. КНм-22-1 Козенка Олександра Васильовича за темою: Метод рекомендації текстових документів за запитаннями користувачів засобами NLP

1. Актуальність обраної теми

Засоби автоматизації підбору текстових документів за запитаннями користувачів відкривають нові перспективи, зокрема, у сфері психологічної підтримки. Так, інтернет-платформи вже активно використовуються для надання онлайн-консультацій, організації тренінгів та створення спільнот для осіб із психологічними труднощами. Проте, згідно з дослідженнями, очікується зростання кількості людей, які потребують психологічної допомоги, і сучасні інформаційні ресурси можуть не встигнути задовольнити цей зростаючий попит. Тому наростає потреба в розробці нових підходів, які б дозволили рекомендувати користувачам відповідні текстові матеріали, котрі можуть допомогти у вирішенні їхніх психологічних проблем. У контексті стрімкого росту обсягів текстових даних у сучасному інформаційному світі, тема "Метод рекомендації текстових документів за запитаннями користувачів засобами NLP" є досить актуальною.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Поставлена у кваліфікаційній роботі магістра мета, пов'язана з створенням методу рекомендації текстових документів за запитаннями користувачів засобами NLP, повною мірою відповідає предметній області спеціальності 122 «Комп'ютерні науки» та вимогам до кваліфікаційної роботи.

3. Професійні та особистісні якості магістранта

У процесі виконання кваліфікаційної роботи магістра, Олександр Васильович Козенко продемонстрував себе як дисциплінований та кваліфікований студент. Всі поставлені завдання були виконані якісно, своєчасно та з старанністю. Студент проявив достатні знання та вміння, які забезпечили успішне досягнення результатів у сфері компетенцій спеціальності 122 «Комп'ютерні науки».

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Результати, отримані в результаті виконання кваліфікаційної роботи бакалавра, є результатом самостійної діяльності студента. Отримані положення наукової новизни та інновації, описані в роботі, дозволили покращити існуючі методи в галузі рекомендації текстових документів за запитаннями користувачів засобами NLP.

5. Наукова новизна та оригінальність запропонованих підходів

Результати виконання кваліфікаційної роботи магістра містять інновації та наукову новизну, зокрема було удосконалено метод рекомендації текстових документів за запитаннями користувачів засобами NLP, який дозволяє автоматизовано виконувати формування вибірок рекомендованих текстових документів за запитаннями користувачів, виконуючи при цьому аналіз відповідності запитання користувача як за семантичною відповідністю до типового запиту до текстового документу, так і, у випадку її відсутності, знаходячи релевантні текстові документи за методом косинусної подібності.

6. Ступінь оволодіння методами дослідження

Магістрант виявив високий ступінь оволодіння необхідними методами дослідження.

7. 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 у роботі виконані, а також проведено аналіз результатів прикладного застосування запропонованих засобів галузі рекомендації текстових документів за запитаннями користувачів засобами NLP

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

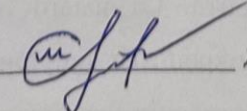
Структура роботи й послідовність викладення логічні та відповідні поставленій меті. Викладення матеріалу грамотне та виявляє високий ступінь відповідності стилю.

9. Можливість практичного застосування кваліфікаційної роботи, окремих її частин

Було створено інформаційну систему автоматизованого визначення рекомендацій текстових документів за користувацькими запитами для предметної області служби психологічної підтримки, яка є прикладною програмною реалізацією метода рекомендації текстових документів за запитаннями користувачів, що призначений для пошуку релевантних рекомендацій з бази текстових порад за текстовим запитом користувача та вхідними даними має набір даних з групами асоційованих запитів та відповідних їм асоційованих відповідей та користувацький запит, що перетворює їх у вихідні дані у вигляді текстової рекомендації релевантної до запиту.

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

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

Науковий керівник  доцент каф. КН Мазурець Олександр Вікторович



ВІДГУК ОПОНЕНТА

на кваліфікаційну роботу магістра

гр. КНм-22-1 Козенка Олександра Васильовича за темою: Метод рекомендації текстових документів за запитаннями користувачів засобами NLP.

1. Актуальність обраної теми

На сучасному етапі спостерігається потреба в психологічній підтримці, що є результатом декількох факторів. Зокрема, це пов'язано з ускладненням життєвих обставин, зростанням стресових ситуацій та поширенням психосоматичних розладів, особливо у контексті воєнних умов. Відповідно до цих викликів, існує актуальна потреба в розробці та застосуванні ефективних методів психологічної допомоги. При застосуванні інформаційних технологій, задача зводиться до автоматизованої рекомендації текстових документів за запитаннями користувачів. Відтак, тема "Метод рекомендації текстових документів за запитаннями користувачів засобами NLP" набуває особливої актуальності у сучасному інформаційному світі, з безперервним зростанням обсягів текстових даних, методи рекомендації текстових документів на основі запитань користувачів з використанням засобів NLP (обробки природної мови) набувають особливої важливості. Таким чином, кваліфікаційна робота магістра має значну наукову та практичну цінність. Зокрема, вона може бути використана для розробки нових інформаційних систем і сервісів, призначених для надання психологічної підтримки.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Обрана тема «Метод рекомендації текстових документів за запитаннями користувачів засобами NLP», в межах якої виконані поставлені задачі, повною мірою відповідає предметній області спеціальності 122 «Комп'ютерні науки» та вимогам до кваліфікаційної роботи магістра.

3. Повнота розкриття мети та завдань дослідження

В роботі автор повністю розкриває мету дослідження та поставленні в межах теми завдання.

4. Наявність наукової новизни

В кваліфікаційній роботі представлена наукова новизна та інновації, відповідна спеціальності 122 «Комп'ютерні науки» в межах обраної області дослідження. Зокрема, було удосконалено метод рекомендації текстових документів за запитаннями користувачів засобами NLP, який дозволяє автоматизовано виконувати формування вибірок

рекомендованих текстових документів за запитаннями користувачів, виконуючи при цьому аналіз відповідності запитання користувача як за семантичною відповідністю до типового запиту до текстового документу, так і, у випадку її відсутності, знаходячи релевантні текстові документи за методом косинусної подібності, таким чином перетворюючи вхідні дані у вигляді користувацького запиту, вибірки із набором лематизованих асоціативних запитів та векторизованого корпусу текстових документів у вихідні дані у вигляді ранжованої множини релевантних рекомендованих текстових документів.

5. Зміст кожного розділу роботи

Робота містить чотири розділи. У першому розділі виконано аналіз сучасного стану області рекомендації текстових документів за запитаннями користувачів засобами NLP. Другий розділ присвячено розробці метода рекомендації текстових документів за запитаннями користувачів засобами NLP. У третьому розділі виконано розробку прикладного програмного застосунку на базі описаного методу. У четвертому розділі виконано дослідження ефективності методу рекомендації текстових документів за запитаннями користувачів засобами NLP.

6. Ступінь розкриття теми роботи

Тема кваліфікаційної роботи достатньою мірою розкрита та обґрунтована, проведено аналіз актуальності та відомих досліджень в межах обраної теми, поставлені завдання, які у роботі виконані, та проведено аналіз результатів прикладного застосування запропонованих методу і засобів.

7. Якість оформлення кваліфікаційної роботи

Оформлення роботи відповідає необхідним нормам та вимогам, які ставляться до оформлення кваліфікаційних робіт.

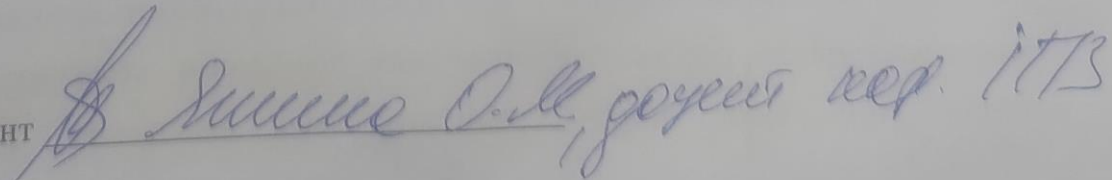
8. Недоліки кваліфікаційної роботи

Доцільно було б навести приклади тестування запропонованого методу за умов наявності в тексті граматичних помилок та сленгу. Надто звужена область прикладного застосування розробленого методу, зокрема розглядається його застосування тільки для предметної області служби психологічної підтримки, хоча можливості розробленого методу суттєво ширші. Також присутні незначні посилки в оформленні роботи, зокрема в ряді випадків рисунки розташовані у тексті раніше, ніж розміщені посилання на них.

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

Враховуючи високий рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка добре.

Опонент


Джуріна О.М. доцент кат. ІТБ