Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

Рівень вищої освіти другий (магістерський)
(рівень вищої освіти)

Інтеграція текстової інформації на основі семантичного аналізу з використанням великих графів
(тема)

Виконала:
студентка 2 курсу, групи СШІм-18-1
Вахрушина Г.В.
(прізвище, ініціали)

Спеціальність 122 – Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма Системи штучного інтелекту (СШІ)
(повна назва освітньої програми)

Керівник к.т.н., доц. Вітько О.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ В.О. Філатов
(підпис) (прізвище, ініціали)

2019 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Штучного інтелекту

Рівень вищої освіти другий (магістерський)

Спеціальність 122 – Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 2019 р.

ЗАВДАННЯ

НА АТЕСТАЦІЙНУ РОБОТУ

студентові Вахрушиній Ганні Вячеславівні
(прізвище, ім'я, по батькові)

1. Тема роботи Інтеграція текстової інформації на основі семантичного аналізу з використанням великих графів
затверджена наказом по університету від 04 листопада 2019 р. № 1623Ст
2. Термін подання студентом роботи до екзаменаційної комісії ____ _____ 2019 р.
3. Вихідні дані до роботи Науково-технічні публікації, дані Інтернет-джерел та відомих наукових проектів щодо дослідження та реалізації інтелектуального аналізу потоків даних природними мовами
4. Перелік питань, що потрібно опрацювати в роботі Аналіз предметної області та постановка задачі, задача інтелектуального аналізу потоків даних, задачі опрацювання природних мов, задача семантичного аналізу джерел природної мови, дослідження методів використання графових структур для інтеграції інформації, проблема великих даних, методи роботи з великими графами

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри)

Рисунок 2.1 – Діаграма кроків методу інтеграції даних, Рисунок 2.2 – Вміст та початковий граф першого тексту, Рисунок 2.3 – Вміст та початковий граф другого тексту, Рисунок 2.4 – Результат обробки текстів за запропонованим методом, Рисунок 2.5 – Загальна мережа відношень першого тексту, Рисунок 3.1 – Схема компонентів застосунку, Рисунок 3.2 – Схема організації даних в графовій БД, Рисунок 3.3 – Діаграма послідовності роботи сценарію обробки, Рисунок 3.4 – Діаграма послідовності сценарію інтеграції даних, Рисунок 3.5 – Діаграма класів компонента опрацювання текстів, Рисунок 3.6 – Діаграма класів компонента інтеграції текстів, Рисунок 3.7 – Діаграма класів компонента аналізу тексту, Рисунок 3.8 – Екран вводу текстової інформації, Рисунок 3.9 – Екран успішної обробки користувацького вводу, Рисунок 3.10 – Екран операції інтеграції текстів, Рисунок 3.11 – Екран семантичного аналізу текстів

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		Підпис	дата
Основна частина	к.т.н., доц. Вітько О.В		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на атестаційну роботу	04.11.19	виконано
2	Аналіз предметної області	04.11.19-10.11.19	виконано
3	Постановка завдання та узгодження з	11.11.19-12.11.19	виконано
4	Дослідження методів використання графових структур для інтеграції інформації	12.11.19-19.11.19	виконано
5	Розробка алгоритму	20.11.19-27.11.19	виконано
6	Розробка програмного забезпечення	22.11.19-03.12.19	виконано
7	Написання пояснювальної записки	03.12.19-10.12.19	виконано
8	Нормоконтроль	11.12.19	виконано
9	Попередній захист	17.12.19	виконано
10	Захист перед ЕК	19.12.19	

Дата видачі завдання 04 листопада 2019 р.

Студент _____
(підпис)

Керівник роботи _____ к.т.н., доц. Вітько О.В.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 95 с., 16 рис., 9 табл., 4 дод., 24 джерела.

ВЕЛИКІ ГРАФИ, ІНТЕГРАЦІЯ ІНФОРМАЦІЇ, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, ОБРОБКА ПРИРОДНОЇ МОВИ, СЕМАНТИЧНА МЕРЕЖА, СЕМАНТИЧНИЙ АНАЛІЗ

Об'єкт дослідження – інтелектуальний аналіз потоків текстових даних.

Предмет дослідження – семантичне представлення текстової інформації, що подана природною мовою.

Мета роботи – розробка та імплементація методу аналізу та інтеграції текстової інформації, використовуючи інструменти роботи з семантичними мережами.

У ході роботи були досліджені існуючі літературні джерела за темою, проаналізовані сучасні підходи до рішення проблеми, спроектовано та втілено метод обробки інформації, що базується на використанні графових структур для синтезу джерел природної мови.

РЕФЕРАТ

Пояснительная записка: 95 с., 16 рис., 9 табл., 4 прил., 24 источника.

БОЛЬШИЕ ГРАФЫ, ИНТЕГРАЦИЯ ИНФОРМАЦИИ, ИНТЕЛЛЕКТУАЛЬНИЙ АНАЛИЗ ДАННЫХ, ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА, СЕМАНТИЧЕСКАЯ СЕТЬ, СЕМАНТИЧЕСКИЙ АНАЛИЗ

Объект исследования – интеллектуальный анализ потоков текстовых данных.

Предмет исследования – семантическое представление текстовой информации, поданной на естественном языке.

Цель работы – разработка и имплементация метода анализа и интеграции текстовой информации, используя инструменты работы с семантическими сетями.

В ходе работы были исследованы существующие литературные источники по теме, проанализированы современные подходы решения проблемы, был спроектирован и реализован метод обработки информации, что базируется на использовании графовых структур для синтеза источников естественного языка.

ABSTRACT

Master thesis: 95 p., 16 fig., 9 tabl., 4 ann., 24 references.

BIG GRAPHS, INFORMATION INTEGRATION, INTELLECTUAL DATA ANALYSIS, NATURAL LANGUAGE PROCESSING, SEMANTIC ANALYSIS, SEMANTIC NETWORK

The object of research is the process of text data intelligent analysis.

The subject of study is the semantic representation of text information in natural language.

The purpose of the work is to develop and implement a method of analyzing and integrating textual information using means for working with semantic networks.

In the scope of the work, the existing literature sources on the topic were studied, the modern approaches of solving the problem were analyzed, an information processing method that is based on the utilization of graph structures for natural language sources synthesis was designed and implemented.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ СКОРОЧЕНЬ І ТЕРМІНІВ

БД – база даних;

СУБД – система управління базами даних;

п\н – порядковий номер;

AI – Artificial Intelligence – штучний інтелект;

API – Application Programming Interface – прикладний програмний інтерфейс;

HPLSA – Hierarchical Probabilistic Latent Semantic Analysis – ієрархічний імовірнісний латентно-семантичний аналіз;

JDBC – Java DataBase Connectivity – підключення до бази даних на Java;

LDA – Latent Dirichlet allocation – латентне розміщення Діріхле;

LSA – Latent Semantic Analysis – латентно-семантичний аналіз;

MASHA – Multinomial ASymmetric Hierarchical Analysis – багаточленний асиметричний ієрархічний аналіз;

NER – Named Entity Recognition – розпізнавання іменованої сутності;

NLA – Natural-Language Analysis – аналіз природної мови;

NLG – Natural-Language Generation – генерація природної мови;

NLP – Natural-Language Processing – обробка природної мови;

OCR – Optical Character Recognition – оптичне розпізнавання символів;

PG – Paraphrase Generation – генерація парафрази;

PI – Paraphrase Identification – ідентифікація парафрази;

pLSA – Probabilistic Latent Semantic Analysis – імовірнісний латентно-семантичний аналіз;

pLSI – Probabilistic Latent Semantic Indexing – імовірнісна латентно-семантична індексація;

UI – User Interface – інтерфейс користувача.

ВСТУП

Інтелектуальний аналіз потоків даних, як область знань, зазнав швидкого просування у розробці методів аналізу і обробки інформації завдяки сучасним тенденціям збільшення її об’єму. Найкраще це ілюструє приклад таких популярних зараз конструктів як соціальні мережі та інформаційні веб-ресурси: їх розробка та швидке наповнення контентом призвело до організації масштабних репозиторіїв даних різноманітного характеру. Зокрема, сьогоденний веб є технологічним тригером, що спонукає користувачів до створення великих об’ємів текстового контенту у формі, що є легкою до обробки і зберігання, проте не якісного машинного аналізу. Зростаюча кількість такої текстової інформації, метою обробки якої є, в тому числі, і використання її множиною різних програмних застосунків, створила необхідність у вдосконалені алгоритмічних підходів у дизайні компонентів, що обробляють дані: ціллю таких сучасних технічних рішень є створення інтелектуальних систем, що є динамічними та можуть легко масштабуватися.

Поточна ситуація стрімкого розвинення сучасних інформаційних технологій спонукає виявленню нових та модернізації існуючих методів обробки ресурсів інформації, зокрема її найбільшої часки в сучасності – джерел природної мови – використання машинних підходів для аналізу, розуміння та генерування людських мов. Однією з головних задач таких систем програмного забезпечення є вирішення проблеми аналізу мови з точки зору розуміння смислових значень її одиниць – семантики.

Станом на сьогодні, існує множина програмних рішень, що працюють з задачею семантичного аналізу даних. Здебільшого, такі системи використовують засоби алгоритмів латентно-семантичного аналізу та його похідних, утилізують інструменти роботи з нейронними мережами. І хоча їх застосування має широку популярність, вони стикаються з певними обмеженнями та мають ряд характерних недоліків.

В якості рішення існуючих проблем опрацювання семантики текстів, в даній роботі буде запропоновано застосовувати принципи роботи з графовими структурами для побудови семантичних мереж, що відтворюють машинне уявлення про структуру та смисловий вміст оброблюваних текстів, на основі цього підходу буде представлено метод інтеграції текстової інформації для утворення нових джерел природної мови.

Таким чином, метою поточної роботи є розробка та імплементація методу інтеграції текстової інформації на основі її семантичного аналізу з використанням інструментів роботи з великими графами.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Семантичний аналіз текстової інформації та його актуальність в сучасності

Сучасні об’єми інформації, що опрацьовуються денно, драматично збільшуються завдяки як програмному, так і апаратному вдосконаленню інформаційних технологій. Так, за даними на листопад 2019, у рамках опитування джерела Netcraft було визначено, що на поточний час зареєстровано 1477803927 веб-сайтів, 229586773 унікальних доменних імен, і 8366753 доступних у всесвітній мережі комп’ютерів [1]. Лавиноподібне зростання маси різноманітної інформації в сучасному суспільстві отримало назву «інформаційного вибуху», спостерігаючи за яким можна відмітити відповідне збільшення попиту до якісного опрацювання ресурсів – розробки ефективних альтернатив для дослідження, аналізу та виявлення знань з даних.

Методи збирання та обробки знань із надмірної кількості електронної інформації існують з 1970-х років, в їх число, крім інших, входять злиття та добування даних (англ. Data mining та Data fusion відповідно), техніка якісного дослідження (англ. Qualitative research) [2], графові імовірнісні моделі, такі як, наприклад, імовірнісні байєсовські мережі [3] та багато інших. Усі вони пропонують певні засоби для інтелектуального аналізу даних, їх організації та використання.

1.1.1 Методи інтелектуального аналізу даних

Вирішення проблеми перенавантаження даними включає такі процеси, як збір, фільтрація, пошук та вилучення інформації, класифікація та кластеризація потоків даних, кластеризація та узагальнення

інформаційних структур різних типів, і т. д.. Мета цих процесів – допомогти користувачам мати кращий доступ до інформаційних ресурсів, що задовольняло б їхнім інформаційнім потребам. Загалом, ці потреби можна описати як необхідність виявлення або отримання певної нової інформації, пошук шаблонів чи паттернів в даних, відокремлення зайвих даних. У контексті текстових джерел інформації, ці обчислювальні процеси є складовими дослідницької галузі видобутку даних з документів (англ. Document Mining).

Протягом новітньої історії існували численні підходи до зберігання, організації, аналізу, пошуку та виявлення інформації з документів. Ці підходи варіюються від простих статистичних підходів до вельми складних рішень аналізу контенту [4]. Перший ґрунтується на простій вичерпній обробці шляхом накопичення статистичної інформації про зміст документа, що може виконуватися повністю механічно: використання методів індексування; конкордансів – типів словників, в яких кожне слово або поняття розташоване в алфавітному порядку з мінімальним контекстом і всіма випадками його вживання у тексті; словників синонімів та інше. Остання група підходів – головним чином залежить від частки інтелектуального аналізу як людей, так і (або) обчислювальних машин, що в тому числі можуть бути обладнані програмами штучного інтелекту. Існують також гібридні способи, що поєднують як статистичну, так і аналітичну обробку вмісту документа.

З точки зору прогресу в дослідженні видобутку знань з документів, відбулося зростання та вдосконалення ймовірнісних методик. Наприклад, в завданнях пошуку інформації це стосується вимірювання частоти слів у відповідних та невідповідних документах, використання термінових частотних заходів для регулювання умовної ваги, наданої різним словам [5]. З боку поприща дослідження штучного інтелекту, можна вказати на відповідні спроби проводити інтелектуальний аналіз інформації автоматично за допомогою нейронних мереж [6].

Однією з головних сфер застосування згаданих технік і методів інтелектуального опрацювання текстової інформації є обробка найбільшої з сучасних підкатегорії циркулюючих джерел даних – обробка природної мови.

1.1.2 Особливості роботи з сучасними формами текстової інформації

Обробка природної мови (англ. Natural-language processing, NLP) – це підполе лінгвістики, інформатики та дослідження штучного інтелекту, що стосується взаємодії між комп'ютерними та людськими, чи природними, мовами, що, зокрема, вирішує таку задачу як програмування комп'ютерів для обробки та аналізу великої кількості даних, що подані природною мовою – заповнення проміжку між людським спілкуванням та комп'ютерним розумінням. Хоча обробка природних мов не є новою наукою, ця технологія швидко розвивається завдяки підвищеному інтересу до комунікацій між людиною та машиною, а також аспекту наявності великих даних та можливостей потужних обчислень.

Для розуміння сучасних тенденцій у сфері роботи з NLP далі подано короткий огляд історії розвитку підходів до опрацювання джерел природних мов.

Британська вчена Карен Спарк Джонс виділяє чотири етапи розвитку обробки природної мови [7].

Перша фаза розвитку обробки природної мови припадала на період кінця 1940-х до кінця 1960-х років. Робота у даний час була сфокусована на машинному перекладі. У 1952 році відбулась перша міжнародна конференція, фокусом якої став машинний переклад. Першим прикладом машинного перекладу, який був представлений на Джорджтаунському експерименті – демонстрація машинного перекладу, 1954 р. – був англо-російський елементарний автоматичний переклад. 1954-й рік став знаковим не лише завдяки першій демонстрації машинного перекладу, а й

завдяки публікації першого випуску журналу «Механічний переклад». Кульмінацією першого етапу стала Теддінгтонська міжнародна конференція з машинного мовного перекладу та прикладного мовного аналізу, що була проведена у 1961 році, на якій було представлено досягнення різних країн світу у галузях морфології, синтаксису, семантики та інтерпретації. Незважаючи на низький технологічний розвиток, що значно ускладнював обробку даних, дослідники активно вивчали предмет та вирішували складні завдання, які поставали перед ними у цей період.

Друга фаза розвитку обробки природної мови припала на кінець 1960-х до кінця 1970-х років та пов'язана зі штучним інтелектом. Дослідники цього періоду найбільшу увагу приділяють знанню про світ та формуванню певних значень у мовленні. Першою ранньою програмою розуміння природної мови була SHRDLU, розроблена у 1972 році. Дана програма розуміла велику кількість англомовних слів та могла робити певні висновки. Це було великим досягненням у дослідженнях штучного інтелекту, проте програми такого типу стикались із ситуаціями реального світу, з якими їм впоратись не вдавалось.

Третя фаза розвитку обробки природної мови тривала з кінця 1970-х та до кінця 1980-х років. В ній тісно переплелись такі галузі як штучний інтелект та семантика. Застосування штучного інтелекту в обробці природної мови вимагало розробки обчислювальної граматики. В практичному сенсі лінгвісти розробили цілу низку граматичних типів, наприклад, функціональний та категоріальний, які були орієнтовані на обчислюваність. Даний період також характеризується стрімким розвитком логічного програмування з метою навчання програм обробки певних текстів.

Четверта фаза почалась у 1990-х роках, на початку якої здобув популярність лексичний підхід до граматики. Одне з провідних місць в обробці природної мови займав статистичний підхід, який дозволяв не просто аналізувати дані, але й справді застосовувати цей метод для

обробки природної мови. Останнє десятиліття 20-го століття характеризувалось застосуванням методики спрощення текстів для виокремлення важливих одиниць з певного потоку інформації.

Розуміння природної мови іноді вважають AI-повною задачею, тому що розпізнавання живої мови потребує величезних знань системи про навколишнє середовище та можливості взаємодіяти з ним. Саме означення змісту слова «розуміти» – одна з головних задач штучного інтелекту. В наш час значну роль у вирішенні задач з обробки даних природними мовами відіграють онтології, наприклад, WordNet, UWN. У процесі дослідження обробки природної мови було досягнуто значних результатів, серед яких розробка потужних лексикографічних систем, програм для машинного перекладу, електронних словників та ін.. Однак, існує проблема, яка досі не є вирішеною, адже вона коріниться у самій природі людської мови – проблема розуміння людського мовлення полягає у його неоднозначності. Так, можна виділити наступні види неоднозначності [8]:

а) синтаксична неоднозначність: у прислів'ї «Час – не кінь, не підженеш і не зупиниш» для систем обробки природної мови буде абсолютно неясним те, про що саме йдеться у реченні, про коня чи про час;

б) смислова неоднозначність: у питанні «Де знайти ключ до того замку?» слово замок може мати два абсолютно різні значення, зважаючи на поставлений наголос;

в) відмінкова неоднозначність: у фразах «Усі були схвильовані перед концертом» та «Не треба давати перед!» слово перед означає час або місце, що абсолютно змінює сенс фрази;

г) референційна неоднозначність: у фразі «Відкрий поличку та дістань мокру парасольку, я хочу її висушити» займенник її за смисловим значенням матиме відношення до мокрої парасольки, проте для машини, у якої повністю відсутнє розуміння реальності, даний займенник відноситиметься як до полички, так і до парасольки.

Одним із викликів, який виникає у процесі обробки природної мови, можна вважати проблему синонімії, в результаті якої одне поняття може бути вираженим декількома різними словами. Як наслідок, релевантні документи, в яких використано синоніми понять, що було вказано користувачем у запиті, може бути не визначено системою.

Вплив зазначених вище явищ є особливо відчутним при створенні систем машинного перекладу. Проблема полягає у складності встановлення конкретного відображення дійсної семантико-синтаксичної структури речення у його внутрішнє логічне уявлення, яке автоматично генерується системою [9].

Розв'язання таких типів неоднозначності можливе за допомогою введення додаткових значень, які збільшать знання програми про ту чи іншу галузь. Сьогодні програм, які «розуміють» усі типи неоднозначності у великому спектрі галузей, не існує, проте є програми, що можуть коректним чином реагувати на неоднозначності у дуже вузьких сферах.

Проблеми обробки природних мов часто включають розпізнавання мовлення, розуміння природної мови та синтез нових ресурсів природними мовами.

Далі наведено перелік деяких найбільш часто досліджуваних завдань з обробки природних мов. Деякі з них знаходять прямі застосування в реальному світі, а інші частіше є підзадачами, які вирішуються для допомоги розв’язання більш складних проблем. Хоча завдання з обробки природних мов тісно переплетені, вони часто підрозділяються на категорії для зручності, приблизний їх поділ може представлятися так:

а) видобування даних: вивчення даних, пошук зв’язків та закономірностей між ними;

б) синтез мовлення: озвучення (прочитання) тексту (документа, повідомлення і т. д.) голосом, який є наближеним до природного;

в) розпізнавання мови: виведення (розпізнавання) тексту з картинок, відсканованих документів або файлів у PDF форматі. Сюди ж входить розпізнавання мовлення, продуковане людським голосом;

г) генерування природної мови: конвертування комп'ютерних даних у природну мову людини;

ґ) машинний переклад: автоматичний переклад з однієї людської мови на іншу. Дане завдання є надзвичайно складним, адже машина не володіє тими знаннями, якими володіє людина, що робить їх «розуміння» тих чи інших фраз абсолютно різним;

д) питально-відповідальні системи: відповіді на питання, поставлені людською мовою. Зазвичай питання є конкретизованими, наприклад, «Де знаходиться Ейфелева Вежа?», проте існують питання, на які немає конкретної відповіді, наприклад, «Чому всі люди різні?», що робить дане завдання надзвичайно складним для виконання;

е) розпізнавання чи визначення теми: поділ тексту на частини з подальшим визначенням провідної теми для кожної з них;

є) інформаційний пошук: пошук, розпізнавання та видобування інформації;

ж) добування даних: отримання семантичної інформації з тексту;

з) отримання зв’язків: визначення відносин між об'єктами в певній області тексту;

и) спрощення тексту: зміна, розширення або інша обробка інформації для спрощення структури або граматики тексту зі збереженням основної думки;

і) розв’язання лексичної багатоманітності: надання списку можливих значень конкретного багатозначного слова, серед яких можна вибрати найбільш підходяще відповідно до контексту;

ї) розпізнавання абревіатур та заголовків;

й) детектування окремих лінгвістичних одиниць;

к) морфологічна декомпозиція: перетворення окремих термінів (наприклад, медичних або технічних) у зрозумілу форму.

Для максимально ефективного рішення наведених задач процес опрацювання природних мов має виконуватися у шість рівнів [10]:

а) фонологічний аналіз – дослідження організації та інтерпретації звуків мовлення у мові. базовими правилами фонологічного аналізу вважаються фонетичні, фонемні та просодичні;

б) морфологічний аналіз – аспект дослідження, що полягає у ідентифікації, аналізі та описі структури або форм слів у мові;

в) лексичний аналіз – поділ тексту на розділи, абзаци, речення та (або) слова;

г) синтаксичний аналіз – аналіз слів у реченні задля розуміння його граматичної структури. слова перетворюються в структури, що показують, який зв'язок існує між словами. окремі сполучення слів може бути виконано у зв'язку з порушенням граматичних правил або правил комбінування слів у мові;

ґ) семантичний аналіз – визначення значень слів, фраз та речень у мові. він сприяє дослідженню можливих смислів речення у контексті;

д) прагматичний аналіз – аспект дослідження, що дозволяє зрозуміти, як комбінуються речення з різними контекстами для формування абзаців, текстів або діалогів. прагматичний аналіз полягає в інтерпретації окремих речень у відповідних для них контекстах.

Далі у роботі буде розглядатися аспект семантичного аналізу текстової інформації, оскільки він є фокусом даного проекту.

1.1.3 Робота з даними на основі семантичного аналізу

Семантичний аналіз – це процес знаходження зв’язку синтаксичних структур, аналізуючи текст від рівнів слів, словосполучень, речень та абзаців до рівня писемності в цілому, до деяких мовно незалежних

значень. Він також включає усунення особливостей, характерних для конкретного мовленнєвого та культурного контексту, якщо така умовність є можливою. Такі мовні елементі, як ідіоми, фразеологізми та образні висловлювання, будучи культурними, часто також транслюються в відносно інваріантні значення в процесі семантичного аналізу.

Існує декілька загальних методів семантичного аналізу – дослідження семантичних полів слів, що використовуються в роботі мовознавців, психологів та психолінгвістів [11].

Мовознавці досить часто використовують кількісну характеристику лексичного та граматичного складу висловлювань, що передбачає підрахунок частоти поєднання самостійних та службових частин мови при складанні опису чи розповіді про окремий предмет, явище або подію. Досить розповсюдженими є і методи що забезпечують якісну характеристику семантичних зв’язків. Тобто визначають відносини, способи поєднання та характеризують взаємодію компонентів семантичного поля на фонологічному, морфологічному, словотворчому, лексичному та синтаксичному рівнях і представлені зв’язками парадигматичного, синтагматичного та епідигматичного типів. На цьому історичному етапі існує шість основних методів вивчення семантичних полів слів. З їх допомогою розкриваються логічні, суворо послідовні зв'язки між сутністю і явищем, статистичні, сталі та нежорсткі зв'язки між формою вираження і сутністю явища. Досліджуючи семантичне поле науковці на протязі вже майже двох століть вдосконалюють критерії аналізу мовленнєвих явищ з метою встановлення системного характеру парадигматичних, синтагматичних та епідигматичних зв'язків слів.

Описовий метод – один з найбільш широко поширених в наукових практиках, що вивчають факти, предмети і явища мовного оточення людини. До того ж він служить підставою для подальшого застосування в дослідженнях інших методів, бо, перш ніж застосувати їх, треба описати

основні властивості розглянутого предмета. Описовий метод часто використовується також і паралельно з іншими методами дослідження.

Основними компонентами описового методу є спостереження, узагальнення, інтерпретація та класифікація отриманих результатів. Для якомога точнішого опису всіх мовленнєвих явищ і взаємодій при використанні описового методу досить часто застосовують кількісну інтерпретацію висловлювань. Однак, найчастіше описовий метод використовується як основний прийом аналізу семантичних явищ мови в навчальній і науковій літературі. Значення лексичних і фразеологічних одиниць в національних словниках також тлумачяться вербальним шляхом, за допомогою словесних визначень, синонімів, гіпонімів, антонімів. Описовий метод ліг в основу створення тлумачних словників. У них парадигматичний аспект значення слова представлений синонімами, антонімами, гіпоніми, словами тієї ж тематичної групи, до якої входить обумовлена лексема. Синтагматичних аспект значення реалізується у вигляді типових словосполучень, що містяться в словникової статті. Епідигматичний аспект значення розкривається в цитатах із класичних творів найбільш видатних представників даної мови. Досить чітко зазначається різниця між багатозначними і омонімічними словами, хоча і неоднаково, в різних словниках. У лексикографічних джерелах розкриваються і описуються також денотативний і сигніфікативний аспекти значення.

Метод контекстуального аналізу застосовували в своїх роботах В. Порциг, О. Духачек, Е. Оксаар, К. Ройнінг та інші вчені. Цей метод заснований на необхідності вивчення одних слів в їх відносинах з іншими словами в тексті. Семантичні явища вивчаються з допомогою порівняння, як правило, разом з формою вираження. Ступінь зв'язку досліджуваних явищ тим вище, чім повніше збігаються у них план вираження і план змісту. Спільність плану вираження і плану змісту мовних одиниць у

різних мов може пояснюватися походження з одного джерела (прамови), а також впливом однієї мови на іншу в процесі їх контактування.

У лінгвістиці та психолінгвістиці існує також метод дистрибутивного аналізу, який ґрунтується на вивченні сукупності оточуючих факторів, в яких даний елемент може зустрічатися, на відміну від моментів, в яких даний елемент зустрічатися не може. Однакова дистрибуція слів вказує на близькість їх значень. Це дає підставу вважати, що семантичне поле дійсно об'єднує слова з родинними значеннями. Метод дистрибутивного аналізу можна вважати досить надійним, лише при дослідженні великого обсягу матеріалу. Чим нижче частотність слова, тим більша величина вибірки потрібна для того, щоб виявити його дистрибуційні властивості. Звідси громіздкість аналізу і, отже, фактичні труднощі виконання досліджень подібного роду. А. Я. Шайкевич розглядає дистрибутивно-статистичний аналіз в семантиці в цілому і конкретно в лексичній семантиці. Дистрибутивно-статистичний аналіз спрямований на опис мови і спирається лише на розподіл (дистрибуцію) заданих елементів у тексті. Такий аналіз постійно використовує кількісну інформацію, а значить, носить статистичний характер. Кінцевою метою даного аналізу має бути формальний опис мови, а проміжними − опис систем спеціальних мов. У дистрибутивно-статистичному аналізі результати оформлюються у вигляді таблиць або карт. Семантичні карти зберігають індивідуальність відповідних текстів − це специфічна риса цього методу. З допомогою дистрибутивно-статистичного аналізу вдається дослідити перерозподіл семантичних центрів під час перебудови семантичних полів, збільшення та зменшення напруги всередині поля.

Сутність і призначення методу компонентного аналізу зводиться до того, що в сукупності досліджуваних мовних одиниць виділяються ті ознаки, за допомогою яких одні одиниці різняться між собою, а інші, навпаки, об'єднуються в групи або сукупності. Інакше кажучи, опис фактів здійснюється набором ознак, що входять до їх плану змісту. Ознаки, за

допомогою яких значущі одиниці відрізняються одна від іншої, називаються диференційними, а ознаки, які сприяють об'єднанню одиниць − інтегральними. Одна і та ж ознака може бути диференціальною і інтегральною залежно від того, які одиниці зіставляються між собою. Метод компонентного аналізу розроблено Н. С. Трубецьким під час вивчення фонем і в силу своєї ефективності та універсальності поширився на дослідження граматичних, а потім лексичних значень.

Вивчаючи граматичні значення, метод компонентного аналізу вперше був застосованим Р. О. Якобсоном, використовувався для опису категорії відмінювання іменників. Опис шести відмінків досягається за допомогою трьох запропонованих вченим семантичних ознак: спрямованості, об'ємності і периферійності. Такі мовні одиниці, як слова, можуть досліджуватися за допомогою компонентного аналізу і лексичними групами. Опис лексичних груп будується шляхом виділення у змісті компонентів за допомогою яких одні слова в групі різняться між собою, інші − ототожнюються. Опис значень окремих слів ґрунтується на прихованому протиставленні їх іншим словами, в результаті чого виділяється ціла сукупність семантичних компонентів (значень).

Статистичний метод використав В. А. Москович при описі поля кольороутворення. Визначення зв'язку між двома словами (прикметниками) проводилося дослідником шляхом виконання ряду обчислень: підраховувалася ймовірність утворення кожного прикметника кольору, математична ймовірність виникнення в мовленні суб’єкта кожного прикметника кольору з кожним іменником і т. д. і, нарешті, вираховувався коефіцієнт кореляції кожної пари прикметників кольору.

Статистичний метод дав можливість кількісно вимірювати відстані між словами одного і того ж семантичного поля. Звідси випливає, що даний метод може дати результати більш об'єктивні, надійні, ніж результати, одержувані інтуїтивно. Незважаючи на всі переваги

статистичного методу, вивчення семантичного поля з його допомогою представляє собою трудомісткий процес, і це утрудняє більш широке впровадження даного методу в дослідження семантичних полів.

Останнім часом великий інтерес викликають застосування психолінгвістичних методів. Їх суть полягає в тому, що з їх допомогою передбачається обробка й аналіз тих мовних фактів, які можна отримати від носіїв мови, в результаті спеціально організованих експериментів. Психолінгвістичні методи вивчення семантики мови неоднорідні. Всі їх види можна узагальнити в два типи: дослідження фізіологічних реакцій організму людини в процесі його мовленнєвої діяльності (відтворенні або сприйманні мови) і аналіз мовних реакцій і оцінок мовних явищ досліджуваними при впливі на них довільних або цілеспрямованих мовних стимулів.

Сутність першого типу методів зводиться до того, що фахівці розкривають смислові зв'язки між мовними одиницями за допомогою реєстрації змін фізіологічних реакцій людського організму на ті чи інші мовні стимули, пропоновані піддослідним (зміна частоти пульсу, розширення зіниць очей, судинні реакції шкіри і ін.) В результаті об'єктивного спостереження за реакціями організму у відповідь на запропоновані людині слова-стимули розкривається системний характер лексики з семантичної точки зору, принцип її організації в людській свідомості та ін.. Одна з таких методик розроблена А. Р. Лурія. У випробуваного виробляється фізіологічна реакція на певне слово, семантичні зв'язки якого з іншими словами потрібно встановити: спочатку людині повідомляється слово, а його наступне звучання супроводжується подразненням шкіри випробуваного слабким електричним струмом. У результаті повторення цієї процедури у досліджуваного при подальшому сприйманні даного слова реєструється за допомогою спеціальних приладів розширення шкірних судин і без супроводу слова електричним струмом. Потім випробуваному пред'являються інші слова, що тематично пов'язані з

розглянутим, або слова, семантичні зв'язки яких із словом-стимулом потрібно виявити. Виявляється при цьому, що семантично пов'язані з заданим словом лексичні одиниці також викликають розширення шкірних судин випробуваного, причому, чим сильніший семантичний зв'язок вихідного слова з іншим словом, тим виразніша реакція. Ця обставина дозволяє визначити ступінь смислового зв'язку слів, що допомагає, у свою чергу, розкривати семантичну структуру лексичної групи. Слова, не пов'язані за змістом із словом-стимулом, подібної фізіологічної реакції у випробуваного не викликають. Організація експериментів, аналогічних описуваного, являє собою складну процедуру, що вимагає спеціальної підготовки як з боку дослідників, так і з боку випробовуваних, використання технічних засобів, клінічних умов проведення дослідів і т. д., що, по-перше, не сприяє широкому поширенню подібних прийомів і, по-друге, ці умови є штучно створеними, а значить не відображають реальної картини взаємодії слів. Тому мовознавці вважають за краще використовувати психолінгвістичні методи іншого типу, зокрема аналіз асоціацій у випробуваного зі словами-стимулами. Психолінгвістичні методи, засновані на аналізі словесних асоціацій випробовуваних, відносяться до асоціативних експериментів, серед яких розрізняють два види: вільні і спрямовані.

При вільному асоціативному експерименті досліджувані дають відповіді залежно від установки і мети дослідження на слово-стимул або одним першим словом, що прийшло на розум, чи цілим рядом слів, що спливають в їхній свідомості протягом певного відрізку часу (часто, однієї хвилини). Відповіді піддослідних реєструються, обробляються і зводяться до переліку, де, як правило, розташовуються по частоті, тобто чим частіше реакції певних слів, тим ближче вони поміщаються до слова-стимулу (є складовими його семантичного ядра). Всі лексичні одиниці, пов'язані зі словом-стимулом у свідомості досліджуваних, що виявлені в результаті експериментів, утворюють асоціативне поле, потужність якого

визначається кількістю елементів, що входять до його складу. Семантичний характер зв'язку цих слів зі словом-стимулом може бути найрізноманітнішим (синтагматичним, парадигматичним чи епідигматичним).

Спрямований асоціативний експеримент застосовується для отримання на слово-стимул певних семантичних зв'язків (синонімічних, антонімічних, фразеологічних і т. д.). Результати експерименту обробляються, і залежно від його мети дослідник визначає ланцюжок асоціацій до заданого слова. Сила семантичних зв'язків визначається в цих випадках частотою поєднання слів − реакцій зі словами-стимулами: чим частіше зустрічаються слова, тим сильніше між ними семантичний зв'язок.

Семантичний аналіз тексту перетинається з безліччю інших областей досліджень. Наприклад, з лексикологією, прагматикою, синтаксисом, етимологією та іншими. Відповідно, в кожній з цих областей поняття семантики по-різному сприймається і носить різні функції.

У лінгвістиці процес семантичного аналізу деякого тексту, починаючи з визначення зв’язку між окремими словами, закінчуючи більш комплексним оглядом того чи іншого висловлювання, вимагає розуміння лексичної ієрархії мови, включаючи гіпонімію – явище існування слів з вужчим значенням, яке називає предмет (властивість, ознаку) як елемент класу (множини); гіпернімію – слова з ширшим значенням, яке виражає загальне, родове поняття, назву класу (множини) предметів (властивостей, ознак); мерономію – явище вичленення певних значущих частин, що створюють структуру об’єкта; полісемію – наявність у мовній одиниці (слові, фраземі, граматичній формі, синтаксичній конструкції) кількох значень; синоніми – слова однієї частини мови, проте різних за звучанням і написанням, які мають дуже близьке або тотожне лексичне значення; антоніми – слова, протилежні за значенням; омоніми – слова, які однаково звучать та пишуться, але мають різне значення [12]. Процес такого аналізу також охоплює опрацювання таких понять, як конотація (чи

семіотика) – сумарне чи тотальне значення слова, як описове, так і емоційне, та колокація – комбінація двох або більше слів, які формують стійке словосполучення, що відіграє важливе значення для природного звучання мови.

В інформатиці процес семантичного аналізу полягає у виділенні семантичних відносин між складовими тексту, формуванні семантичного представлення текстів.

На базовому рівні автономне семантичне опрацювання природної мови включає такі задачі як дрібнення мови на більш прості, елементарні одиниці мовлення та організація їх у певні структури даних, що мають на меті представлення вихідного тексту взаємозв’язками його компонентів, поєднання та аналіз яких можуть описати картину шуканого значення.

Варто зауважити, що формальне уявлення про значення, отримане за допомогою машинної обробки природних мов, відрізняється від того, що природньо можна вважати значенням тексту. Протягом комп’ютерного опрацювання, основна увага приділяється отриманню відповідних, однозначних та оперативних представлень опрацьовуваного джерела мовлення.

1.2 Аналіз теоретичного базису існуючих рішень

Умовно семантичні операції в NLP можна поділити на дві основні галузі – аналіз природних мов (англ. Natural-language analysis, NLA) та породження природних мов (англ. Natural-language generation, NLG). Лексичний, синтаксичний, семантичний, прагматичний та морфологічний аналіз тексту вивчається в NLA, у той час як генерування красномовних багатозначних або багатоабзаційних текстів – в NLG [13]. Обидва ці напрямки NLP певним чином стикаються у вирішенні проблеми визначення семантичної подібності мовленнєвих структур, до рішення якої існує два підходи – парафразування та двонаправлене залучення.

Парафразування можна описати як перетворення значення уривку за допомогою використання інших слів, що відповідають вихідному смислу. Відповідно до задачі, що розглядається в даній роботі – NLG – парафразування використовується як підхід до збільшення різноманітності тексту, що утворюється в процесі обробки [14], використовуватися парафрази можуть на рівнях слова, фрази, речення або дискурсу цілком. Задача перефразування має щонайменше три підкатегорії: генерація парафрази (англ. Paraphrase Generation, PG), отримання парафрази та ідентифікація парафрази (англ. Paraphrase Identification, PI).

PG вважається проблемою NLG – завданням генерації альтернативного тексту вхідної лексеми [15], отримання чи вилучення якої передбачає існування номінальних парафраз або можливості визначення їх із поточного тексту чи деякого загального джерела [16].

PI – це завдання розпізнавання співвідношень, що можуть бути перефразованими, на вхідних текстах, де пошук таких зв'язків у текстах – це завдання визначити з-поміж пари текстових фрагментів, чи є значення одного пов'язаним з іншим [17]. Так, парафраза може розглядатися як двонаправлене відношення: текст А є парафразою тексту В тоді і лише тоді, коли А означає В, а В означає А [18]. Для визначення таких залежностей, використовуються техніки непідконтрольного та контрольованого навчання. Під першим розуміється спроба знайти приховану структуру в невизначених наперед чи безструктурних даних. Контрольоване ж навчання – задача машинного навчання виявити функціональні залежності, опираючись на заготовлені навчальні дані.

У контексті проблеми, що розглядається в рамках даної роботи, далі буде йти мова про семантичний аналіз текстових джерел без нагляду, який у цьому розділі буде розглядатися на прикладі визначення семантичної подібності текстів.

Схожість між двома текстами зазвичай вимірюється за допомогою простого підходу лексичного узгодження та отримання оцінки схожості на

основі кількості лексичних одиниць, що мають місце в обох вхідних сегментах. Для вдосконалення цього методу розглядалися такі допоміжні підходи як видалення стоп-слів, тегування частин мови, визначення найдовших відповідних послідовностей мовних конструкцій, використання різних коефіцієнтів зважування та нормалізації [19], [20]. Ці методи, хоча певною мірою є успішними, не можуть визначати подібність між реченнями, які використовують різні, але синонімічні слова та вирази, що мають однакове значення.

Проблему відсутності можливості проаналізувати зв'язок між близькими за змістом текстами, в яких використовується різна лексика – у таких випадках об'єднання має відбуватися не тільки на основі подібності, а ще й на основі семантичної суміжності або асоціативності, можна вирішити за допомогою складання тезаурусів – тлумачних або тематичний словників, що прагнуть максимально охопити лексику конкретної мови, що можуть також містити приклади використання слів в тексті; таких словників, в яких слова, що належать до якої-небудь галузі знань, розташовано за тематичним принципом, де показано семантичні зв'язки між лексичними одиницями.

На сьогоднішній час, для визначення семантичної подібності тексту, найбільш широко використовуються метод латентного семантичного аналізу (англ. Latent semantic analysis, LSA) [21] та його модифікації.

Латентно-семантичний аналіз – метод обробки інформації природною мовою, зокрема, дистрибутивної семантики, що дозволяє аналізувати взаємозв'язок між набором документів і термінами, які в них зустрічаються, шляхом створення набору понять. LSA припускає, що слова, близькі за значенням, зустрічатимуться в подібних фрагментах тексту – дистрибутивна гіпотеза. З певної великої частини тексту створюється матриця, що вміщує інформацію про параграфи: рядки містять унікальні слова, а стовпці – текст кожного параграфа. За допомогою математичного методу, що називається сингулярним

розкладом матриці, кількість рядків матриці зменшують, зберігаючи при цьому структуру подібності у стовпцях. Потім слова порівнюють за допомогою обчислення косинуса кута між двома векторами, що утворено будь-якими двома рядками – скалярний добуток векторів, поділений на добуток їх модулів. Значення, близькі до 1, вважаються дуже схожими словами, тоді як значення, близькі до 0, представляють дуже різнорідні слова [22].

Імовірнісний латентно-семантичний аналіз (англ. Probabilistic latent semantic analysis, pLSA), також відомий як імовірнісне латентно-семантичне індексування (англ. Probabilistic latent semantic indexing, pLSI) – це статистичний метод аналізу кореляції двох типів даних.

Існують такі види pLSA:

а) ієрархічні розширення:

1) асиметричне: MASHA (англ. Multinomial ASymmetric Hierarchical Analysis, MASHA – поліноміальний асиметричний ієрархічний аналіз) [23],

2) симетричне: HPLSA (англ. Hierarchical Probabilistic Latent Semantic Analysis, HPLSA – ієрархічний імовірнісний латентно-семантичний аналіз) [24];

б) прихований розподіл Діріхле – додає розподіл Діріхле в якості апріорного розподілу тематик за документами;

в) дані вищого порядку: pLSA можна застосувати і для даних більш високого порядку (трирівневих і вище), тобто він може моделювати спільну поведінку трьох і більше змінних. У симетричному формулюванні це робиться простим додаванням умовного розподілу ймовірностей для додаткових змінних. Це імовірнісний аналог невід’ємної тензорної факторизації.

У порівнянні зі звичайним латентно-семантичним аналізом, що заснований на лінійній алгебрі і є способом зниження розмірності матриці (як правило, за допомогою розкладання діагональної матриці за

сингулярними значеннями), імовірнісний латентно-семантичний аналіз заснований на змішаному розкладанні, що, в свою чергу, бере свій початок з моделі прихованих класів. Даний підхід є більш принциповим, оскільки має основу в області статистики.

Латентне розміщення Діріхле (англ. Latent Dirichlet allocation, LDA) – породжуюча модель, що дозволяє пояснювати результати спостережень за допомогою неявних груп, завдяки чому є можливим виявлення причин подібності деяких частин даних. Наприклад, якщо спостереженнями є слова, зібрані в документи, стверджується, що кожен документ являє собою суміш невеликої кількості тем і що поява кожного слова пов'язана з однією з тем документа. LDA є одним з методів тематичного моделювання і вперше був представлений в якості графічної моделі для виявлення тематик [23]. Цей підхід є схожим з pLSA з тією різницею, що в LDA передбачається, що розподіл тематик має в якості апріорі розподіл Діріхле. На практиці результатом є більш коректний набір тематик.

Аналізуючи можливі шляхи рішення розглядуваної задачі, можна зазначити, що перший шлях – отримання тезаурусів – є досить трудомістким і повністю визначається поставленим завданням: створення універсального тезаурусу є потенційно неможливим, якщо мається на увазі його використання за рамками певної предметної області. Останній зі згаданих алгоритмічний шлях також має свої недоліки. Найбільшою проблемою є певна непрактичність розглянутих методів і неочевидність їх застосування для текстових даних. Наприклад, LDA вимагає умови нормальності розподілу, що при вирішенні лінгвістичних завдань не завжди задовольняється. Більш того, як правило, всі ці алгоритми мають велику кількість параметрів, визначення яких є емпіричним і може істотно вплинути на якість рішення: наприклад, скорочення сингулярних значень діагональної матриці в LSA має нелінійний вплив на результат.

1.3 Постановка задачі

Розглянувши алгоритми аналізу текстів у попередньому пункті, було виявлено ряд недоліків їх прикладного використання у програмних застосунках для користувачів: надлишковість, велика кількість емпіричних параметрів роботи та певна неочевидність їх застосування щодо різнорідних текстових даних. Резюмуючи, можна зробити висновок, що зазначені особливості негативно впливають на продуктивність роботи систем, що їх використовують: наприклад, застосовуючи LSA для рішення нового завдання, такі параметри як алгоритми зважування, кількість збережених вимірів та методи попередньої обробки тексту повинні бути оптимізовані для цього конкретного завдання, оскільки вони впливають на ефективність LSA. Говорячи про застосування таких алгоритмів, використовуючи їх як складові застосунків повного семантичного аналізу множин текстів, можна сказати, що вони є неоптимальними.

Для рішення описаних недоліків пропонується використовувати метод семантичного аналізу та обробки текстів на основі роботи з семантичними графами.

Пропонований метод опрацювання текстів базується на відтворенні машинного розуміння семантичної структури вхідної інформації шляхом визначення зв’язків між складовими вхідних даних: словами, фразами і параграфами, що дозволяє більш точно аналізувати зміст документів. Наявність таких структур, що можуть відтворювати сенс оброблюваного тексту, таких структур, що є природньо властивими до масштабування і збагачення семантичною інформацією, структур, що не передбачають попереднє настроювання параметрів їх обробки, потенційно може якісно покращити процес розуміння змісту та спростити шляхи до його утилізації: пропонований підхід до представлення інформації може демонструвати значну перспективу для смислового аналізу та для покращення існуючих процесів обробки даних.

Для реалізації зазначеного необхідно вирішити такі основні завдання:

а) вивчити особливості предметної області, абстрагувати проблеми, які необхідно вирішити, оглянути теоретичну базу використовуваних концептів;

б) сформувати принципи подання інформації у вигляді семантичних мереж, розробити методи роботи з отриманими структурами: їх формування та обробку;

в) спроектувати архітектуру системи, продумати робочий процес її функціонування;

г) реалізувати відповідне програмне забезпечення для демонстрації роботи алгоритму;

ґ) окреслити перспективи розвитку застосунка для підвищення ефективності його використання.

2 РОЗРОБКА МЕТОДУ ІНТЕГРАЦІЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ НА ОСНОВІ ЇЇ СЕМАНТИЧНОГО АНАЛІЗУ

2.1 Граф як структура представлення інформації

Пропоноване вище семантичне уявлення текстової інформації можна представити графом, чи семантичною мережею, що відображає відносини між складовими її частинами – смисловими одиницями тексту. Глибина подання такого семантичного опису може бути різною, у той час як в реалізації поточного підходу буде будуватися багатостороннє уявлення про текст і його окремі фрагменти.

У загальному розумінні, граф – це абстрактний математичний об'єкт, який представляє собою множину вершин графа і певний набір ребер, тобто з'єднань – зв’язків між парами вершин. Математично його можна описати наступним чином:

$$G := (V, E), \tag{2.1}$$

де G – граф;

V – непорожня множина вершин, чи вузлів;

E – множина пар вершин, в разі неорієнтованого графа – невпорядкованих, що називаються ребрами.

За винятком поданого вище формального виду, графи можна описувати за допомогою:

а) матриць суміжності – таблиць, де як стовпці, так і рядки відповідають вершинам графа. В кожному осередку таких матриць записується число, що визначає наявність зв'язку від вершини–рядка до вершини–колонки;

б) матриць інцидентності – таблиць, де рядки відповідають вершинам графа, а стовпці відповідають ребрам графа. У осередок матриці на перетині рядка *i* зі стовпцем *j* записується певне математичне значення,

що означає наявність зв’язку між об’єктами. Для неорієнтованих графів прийнято вживати пару *(1; 0)*, а у випадку орієнтованого – значення чарунки визначається так :

1 – в разі, якщо зв'язок *j* «виходить» з вершини *i*;

-1 – якщо зв'язок «входить» в вершину;

0 – у всіх інших випадках: тобто якщо зв'язок є петлею або зв'язок не інцидентний вершині.

В контексті вирішуваної задачі, застосування графів для описання інформаційних структур було зумовлено їх властивостями: за допомогою таких мереж можна представляти виявлені зв’язки оброблюваних компонент, що дає змогу, відстежуючи їх, аналізувати та опрацьовувати інформацію щодо семантичного представлення текстів – шукати нетривіальні смислові залежності, визначати шаблони та взаємозв’язки, аналізувати структуру та вміст вхідних даних.

У ході роботи з текстом пропонується використовувати такі графи, що утворюються з відносин між словами, реченнями та параграфами, що об’єднані у мережеву структуру опису інформації про оброблюваний текст. Складові структури, у ході їх опрацювання, нараджується збагачувати додатковою допоміжною інформацією, детальніше – далі, розширюючи коло можливостей використання складнішої логіки під час аналізу.

2.2 Пропонований метод аналізу і інтеграції текстової інформації

Для реалізації можливості аналізу та інтеграції джерел даних пропонується використовувати метод, що базується на застосуванні мережевих структур для визначення, зберігання та обробки семантичних зв’язків складових оброблюваних текстів – ребр графів, що описують бінарні взаємовідносини типу «належить» та «має зв’язок з» між різними частинами тексту, що аналізуються: такими як слова, речення та

параграфи, надаючи можливість комплексно розуміти текст, як складну структуру взаємозв’язків. Перетворюючи таким чином потоки даних у збагачені мережі, можна взаємно їх вивчати та модифікувати, генеруючи на основі їх структур, що, умовно, перетинаються, якісно нові результати.

Так, оперуючи побудованими семантичними мережами, що являють графи відношень слів, речень та параграфів у тексті, використовуючи отриману інформацію про їх зв’язки та взаємні відношення, реалізацію задачі інтеграції двох текстових джерел можна описати наступним чином.

Маючи деякий вхідний текст, та виконуючи описані нижче операції для другого, пропонується використовувати такий метод його обробки, що представлено на рисунку 2.1.

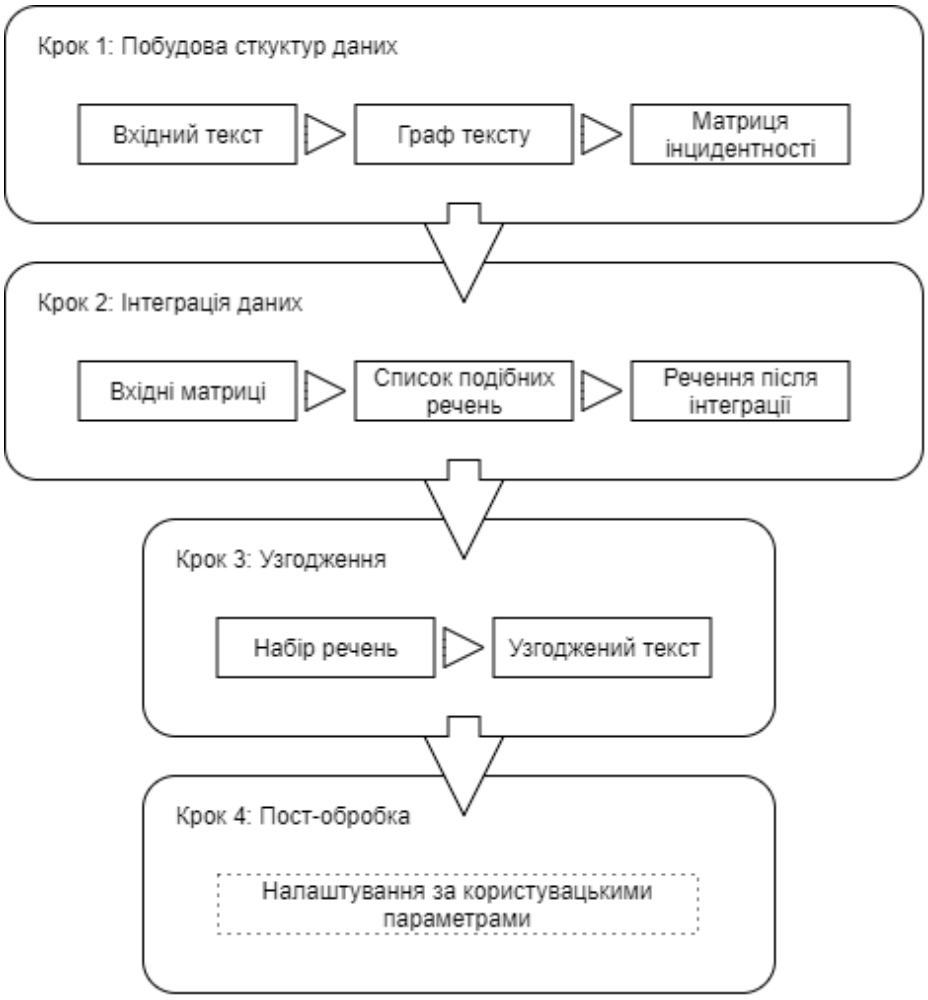

Рисунок 2.1 – Діаграма кроків методу інтеграції даних

Детальніше кожен з кроків, що наведені на рисунку 2.1, можна описати наступним чином.

Крок 1: побудова структур даних. В рамках цього кроку для деякого вхідного тексту пропонується сформувати його семантичний граф, опис якого надано у наступних підрозділах, та побудувати матрицю інцидентності слів цього графа. Під словами далі передбачаються отримані у ході попередньої обробки мовленнєві леми, що є канонічними формами лексем. Така матриця моделює умовну карту використовуваності слів в тексті, що у поєднанні з інформацією про їх зв'язки у реченні, дає змогу визначати умовну подібність конструкцій речень між множиною текстів – так пропонується визначати найбільш подібні речення текстів, що оброблюються.

Крок 2: інтеграція даних. Під час нього, аналізуючи отриману множину подібних речень, та беручи за базу будь-який з доступних текстів, визначаються далі опрацьовувані параграфи та речення, послідовно оброблюючи які шляхом довільної заміни слів на слова протилежних речень чи на синонімічні їм словосполучення, отримується певний набір частин тексту, що є результатом механічної інтеграції інформації.

Крок 3: узгодження. Оскільки отримана структура складається зі слів, що не мають мовленнєвих характеристик, притаманних природній мові, для вихідного тексту необхідно провести ряд перетворень, що узгоджували б час, рід та число модифікованих компонентів. Так, за поточної реалізації пропонується використовувати таку послідовність дій, маючи на увазі, що під час попередніх кроків зберігалася мета-інформація щодо проведених перетворень: послідовно для кожного речення тексту, послідовно для кожного модифікованого слова в контексті речення, відповідно до раніше застосованих перетворень, визначається набір із мовленнєвих характеристик речення, якщо воно просте, та складових речення у протилежному випадку. Отримані дані щодо сполучень

відсортовуються за рангом та, за результатами, обирається найбільш вживане, відносно якого опрацьовувані конструкції узгоджуються. Таким чином отримується деякий набір атомарно-узгоджених речень, що потребує виконання аналогічної фінальної модифікації відносно часу, проте на рівні тексту у цілому.

Крок 4: Пост-обробка. За потреби, можна розглядати додаткове проведення передуючого етапу – а отже і моделювання відповідного алгоритму – відносно наперед заданих користувацьких побажань щодо результуючого тексту.

Для ілюстрації описуваної ідеї використовується приклад, що подано на рисунках 2.2, 2.3 та 2.4, таблицях 2.1, 2.2, 2.3, 2.4, 2.5 та 2.6.

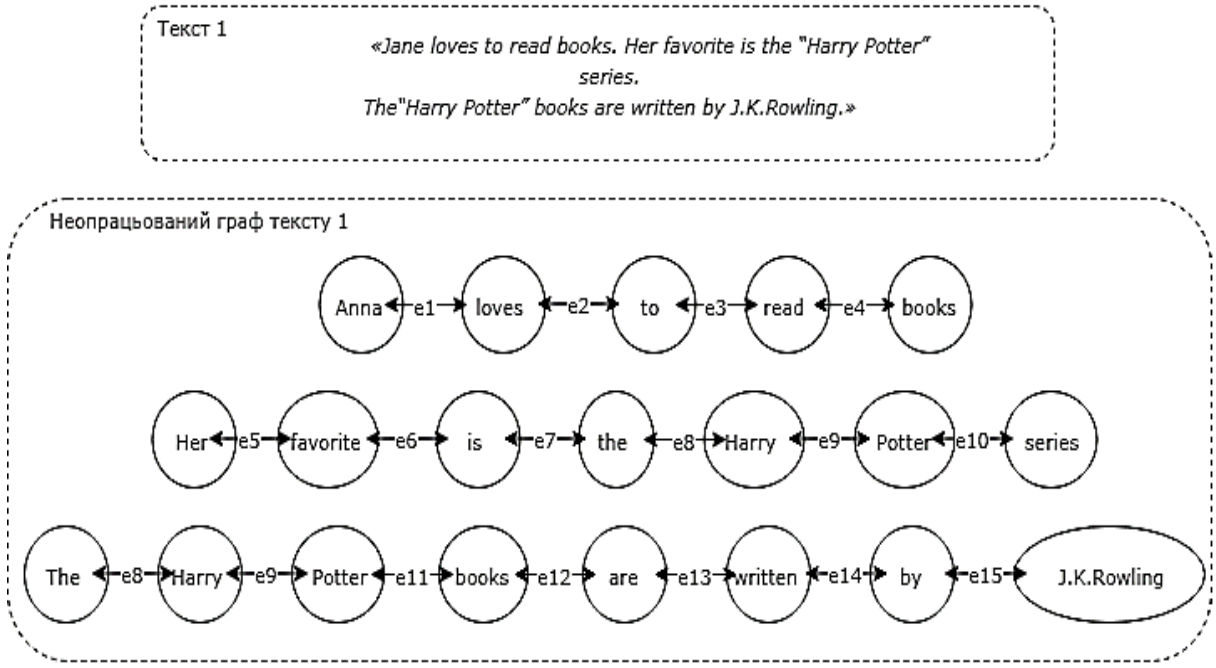

Рисунок 2.2 – Вміст та початковий граф першого тексту

Таблиця 2.1 – Матриця інцидентності графа першого тексту

п.н	слово \ ребро	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10	e11	e12	e13	e14	e15
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	Anna	1														

Продовження таблиці 2.1

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
2	love (loves) *синоніми: adore*	1	1													
3	To		1	1												
4	read			1	1											
5	book (books) *синоніми: series, novel*				1							1	1			
6	she (her)					1										
7	favorite					1	1									
8	to be (is) (are)						1	1					1	1		
9	the							1	2							
10	harry								2	2						
11	potter									2	1	1				
12	series										1					
13	write (written)													1	1	
14	By														1	1
15	J.K.Rowling															1

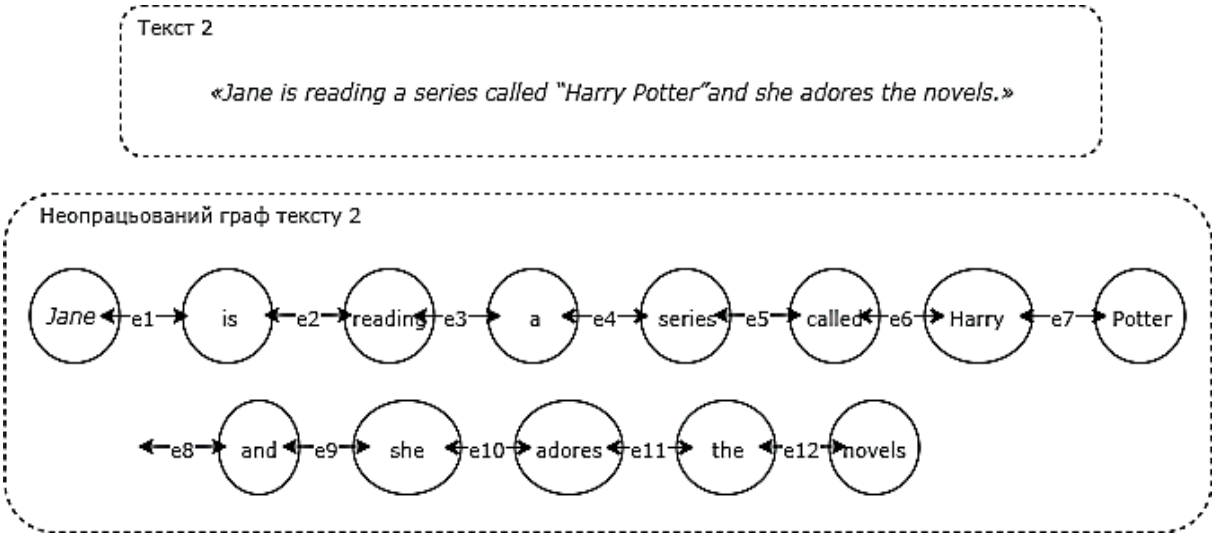

Рисунок 2.3 – Вміст та початковий граф другого тексту

Таблиця 2.2 – Матриця інцидентності графа другого тексту

п.н	слово \ ребро	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10	e11	e12
1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	jane	1											

Продовження таблиці 2.2

1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	to be (is)	1	1										
3	read (reading)		1	1									
4	a			1	1								
5	series				1	1							
6	called					1	1						
7	harry						1	1					
8	potter							1	1				
9	and								1	1			
10	she									1	1		
11	adore (adores) *синоніми: love*										1	1	
12	the											1	1
13	novel (novels) *синоніми:* series, book												1

Таблиця 2.3 – Частка використовування слів в тексті

Слово	Підрахунок	Результат
anna	1	1
love	1 * 2	2
read	1 * 2 + 1 * 2	4
book	1 * 3	3
…		
love – adore	2 + 2	4
book – novel – series	3 + 1 + 2	5

Таблиця 2.4 – Зведена таблиця вживаних слів за відносним рангом

Слово	Результат	Посилання на речення	Посилання на параграф	Посилання на текст
book	5	1, 3	1,2	*1*
novel	5	1	1	*2*
series	5	2, 1	1, 1	**1,2**
love	4	1	1	*1*
adore	4	1	1	*2*
read	4	1,1	1,1	**1,2**
harry	4	2,3,1	1,2,1	**1,1,2**

В наведеній таблиці 2.4 стовпці «Посилання на речення», «Посилання на параграф» та «Посилання на текст» мають взаємно відповідні значення, які потрібно розуміти таким чином: для слова book

посилання на речення 1 має відповідне посилання на параграф – 1 і текст 1; для посилання на речення 3 – посилання на параграф – 2 і текст 1.

Жирним накресленням відзначено входження вживаних слів в різних текстах, курсивним – посилання на різні тексти слів-синонімів.

Таблиця 2.5 – Зведена таблиця подібності речень

Текст 1	Текст 2	Ранг
речення 1	речення 1	4
речення 2	речення 1	2
речення 3	речення 1	2

За результатами, що приведені в таблиці 2.5 можна зробити висновок, що найбільш подібними було визначено пару перших речень з обох текстів.

Для обраних подібних речень, шляхом заміни слів відносно таблиці 2.6 та узгодження їх у вихідному реченні, отримується результат, що подано на рисунку 2.4. Відповідні корективи на ньому позначені виділенням певних областей тексту: результат заміни позначено жирним накресленням, частина тексту, що не підлягала модифікації – сірим кольором .

Таблиця 2.6 – Таблиця використаних парафраз

Слово тексту 1	Слово тексту 2
book	novel, series
love	adore
read	read (може бути замінено на слово-синонім)

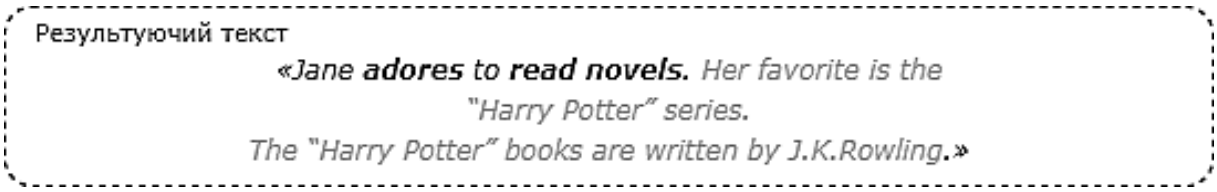

Рисунок 2.4 – Результат обробки текстів за запропонованим методом

2.3 Детальний опис методу інтеграції даних

Для організації роботи описаного у попередньому пункті алгоритму, набори вхідних даних необхідно попередньо обробляти та формувати в відповідні структури даних таким чином, щоб охопити зберігання необхідної кількості мета-інформації щодо текстів, що опрацьовуються. Наведені далі форми організації даних зумовлюють більш глибоке та цілісне розуміння про семантичну значущість компонентів вхідних даних.

2.3.1 Подання тексту як семантичної мережі

Семантична мережа довільного тексту, у рамках рішення проблеми, шо розглядається, містить сукупну інформацію щодо складових цього тексту з точки зору їх синтаксису. У загальному вигляді, така мережа відношень у тексті описує взаємні залежності її складових частин та представляє певну деревовидну структуру, що охоплює опис слів, речень, параграфів та їх умовних відносин – залежності між «сусідніми елементами» – та залежності «відношення». Підмножина слів такої мережі утворює повний граф слів тексту, що розглядається.

Відносно описуваного прикладу з попереднього пункту підрозділу, ілюстрація зазначеного подана на рисунку 2.5, який зображує схематичну мережеву організацію для першого тексту.

Поданий підхід до семантичного представлення аналізує інформацію шляхом реєстрації та роботи над смисловими зв'язками, що існують у тексті. Описана семантична обробка ґрунтується на повному синтаксичному аналізі та обширному семантичному моделюванні, які разом утворюють цілісне відображення текстових складових відносно представлення смислу.

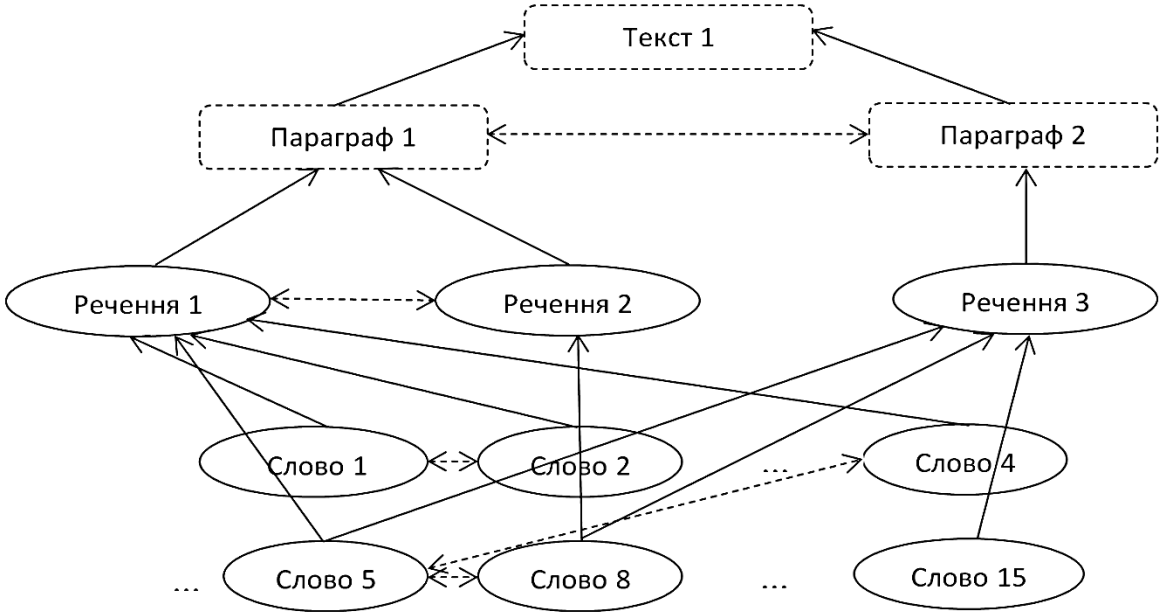

Рисунок 2.5 – Загальна мережа відношень першого тексту

2.3.2 Структури зберігання інформації

Детальніше пояснюючи рівні абстракції, далі наведено умовний опис структур зберігання інформації щодо оброблюваних компонентів вхідного тексту.

Для більш якісного семантичного аналізу даних, пропонується використовувати такі структури організації інформації про параграфи, речення та слова, приклад яких подано в таблицях 2.7, 2.8 і 2.9 відповідно.

Таблиця 2.7 – Приклад представлення доповняльної інформації про параграф

Характеристика	Значення
тип об’єкту	параграф
п\н	1
відношення між параграфами	[2,3]
посилання на текст	[1]

Таблиця 2.8 – Приклад представлення доповняльної інформації про речення

Характеристика	Значення
тип об’єкту	речення
п\н	1
посилання на слова	[1, 2, 3, 4, 5]
1	2
відношення між реченнями	[2]
посилання на параграф	[1]

Таблиця 2.9 – Приклад представлення доповняльної інформації про слово

Характеристика	Значення
тип об’єкту	слово
п\н	5
Терм	book
Синоніми	[volume, novel, series, tome, work, publication]
використання в тексті	[book, books]
відношення між словами	[4, 8, 11]
посилання на речення	[1, 3]

Для кожного з приведених нижче описів представлена інформація щодо значущих параметрів відносно типу об’єкта, взаємних відношень між сусідніми об’єктами одного типу та посилання на відношення до узагальнюючого типу. Така структура потенційно надає можливість проводити пошук за елементами за більш короткий термін виконання.

Варто зазначити, що для створення таких структур необхідно виконати попередню обробку вхідних даних, що включає:

а) переклад всіх букв в тексті в нижній або верхній регістри;

б) видалення цифр або заміна їх на текстовий еквівалент;

в) видалення зайвої пунктуації;

г) видалення символів зайвих пробілів;

ґ) токенізацію – розбиття потоку даних на атомарні частини;

д) лематизацію – отримання слова в його початковому поданні.

2.3.3 Робота з великими графами

Важливим аспектом пропонованого методу обробки текстової інформації є розуміння її об’єму. Так, за опрацюванням текстів відносно великого розміру, отримувані масиви даних, їх матричне і графове представлення та зберігання доповняльної інформації, за рахунок її складності та багаторівневості, може спонукати до використання засобів роботи з великими даними, та, відносно уживаних форм організації даних, засобів роботи з великими графами.

Станом на сьогодні, існує ряд популярних технологічних рішень, що надають можливості ефективно оперувати великою кількістю структурованих у мережу даних:

а) деякі системи графових баз даних підтримують обробку транзакцій в реальному часі, як, наприклад, СУБД Neo4j. Однак варто зазначити, що використовувані методи доступу до графів в деяких графових БД не враховують локальність даних, і при обробці графів, використовується практично випадковий доступ до даних. Для великих графів, які неможливо зберігати в пам'яті, випадковий доступ до диска може стати «вузьким місцем» опрацювання, що потребує додаткової обробки;

б) інфраструктура MapReduce, що розроблена компанією Google, дозволяє створювати кластери зі звичайних комп'ютерів і програмувати їх для обробки великих обсягів даних за один прохід. На відміну від Neo4j, інфраструктура MapReduce не призначена для обробки запитів в реальному часі. MapReduce оптимізована для аналізу великих обсягів даних, розподілених на сотнях комп'ютерів. Завдяки своїй простоті і масштабованості, в промисловій і науковій областях широко популярна

open-source (англ. Open-Source Software – відкрите програмне забезпечення) інфраструктура Apache Hadoop, що призначена для виконання розподілених обчислень при роботі з великими обсягами даних і побудована на основі принципів MapReduce. Однак, Hadoop і пов'язані з нею технології, наприклад, Pig і Hive, в цілому не призначені для масштабованої обробки графових даних;

в) Giraph – розподілена відмовостійка система, що заснована на моделі паралельних обчислень (англ. Bulk Synchronous Parallel) для обробки великих обсягів графових даних із застосуванням алгоритмів паралелізації;

г) Graphlab – заснована на графах високопродуктивна інфраструктура розподілених обчислень, написана на C ++.

Для реалізації задач даної роботи було обрано СУБД Neo4j, як головний інструмент для інтерактивного опрацювання великих графів. Її функціоналу цілком достатньо для демонстрації роботи описаного методу обробки даних, проте розглядаючи перспективу росту обчислювальних потреб проектованої системи, має сенс розглянути можливий перехід на інші, згадані вище, технології чи реалізацію деяких модифікацій відносно поточної.

3 ПРОГРАМНА РЕАЛІЗАЦІЯ МЕТОДУ

Відповідно до поставлених задач та спроектованого методу інтеграції текстової інформації, у рамках поточної роботи, було відтворено програмний застосунок, опис якого надано далі.

3.1 Загальна архітектура системи

Реалізована система у цілому є монолітним веб-орієнтованим ресурсом, що містить поєднані у єдиний компонент шари користувацького інтерфейсу та власне компонента обробки і стороннє NoSQL-сховище даних, що є інструментом для зберігання інформації у формі графів, схематичне зображення внутрішнього устрою застосунка подано на рисунку 3.1.

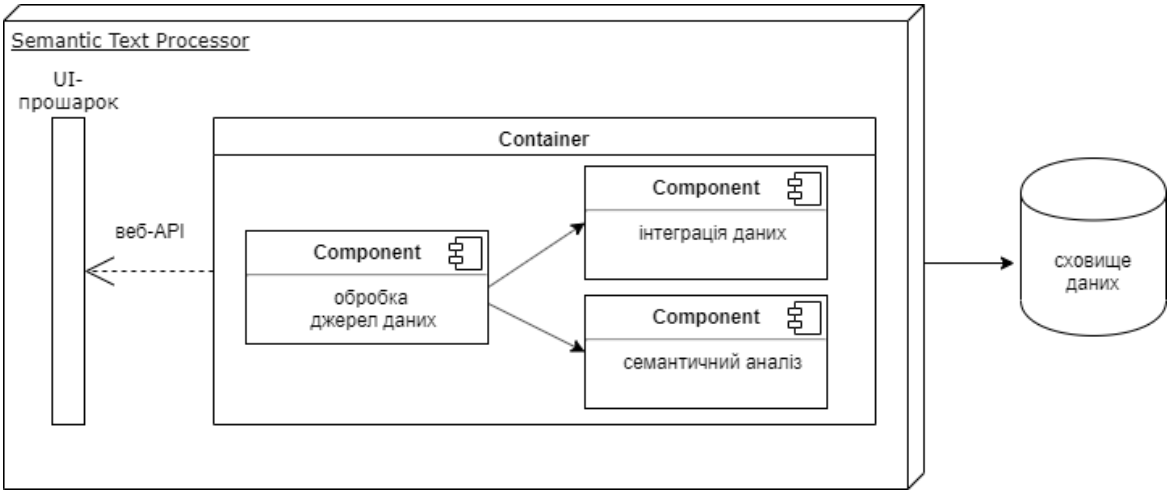

Рисунок 3.1 – Схема компонентів застосунка

Зазначений на схемі прошарок користувацького інтерфейсу (англ. User Interface, UI) реалізовано за допомогою використання мови розмітки, написання скриптів та стилів HTML версії 5, JavaScript версії 1.6 та CSS версії 3 відповідно, з допоміжною бібліотекою Bootstrap версії 4, бібліотеки Thymeleaf версії 3.0.11; інтеграція між UI і веб-API (англ.

Application Programming Interface, API) застосунка відтворена за допомогою можливостей модуля фреймворка Spring – Spring MVC; технологіями написання власне функціонального модуля є мова програмування Java версії 1.8, фреймворк Spring Boot версії 2.0, бібліотека Hibernate версії 5.4.4, бібліотека Apache Commons Math версії 3.6.1 для роботи з графами і матрицями; в якості СУБД використовувалася Neo4j версії 3.5.11 та відповідний JDBC-драйвер.

Оскільки в даній роботі використовується NoSQL СУБД, дані, що в ній зберігаються, не мають стійкої структури. Внутрішню організацію структур даних для спроектованої програми можна описати використовуючи поняття зі сфери роботи з графами: описати множину типових вершин і множину їх зв’язків. Так, поточна реалізація організації даних має:

а) Вузли даних типу «Слово», «Речення», «Параграф» і «Текст». Через те що детальна інформація щодо допоміжних даних може варіюватися від екземпляра до екземпляра одного типу вершин, у схемі вона опускається;

б) Ребра вузлів типу «Зв’язок» – неорієнтоване ребро, та «Посилання» – орієнтоване.

Спрощена ілюстрація такого підходу подана на рисунку 3.2.

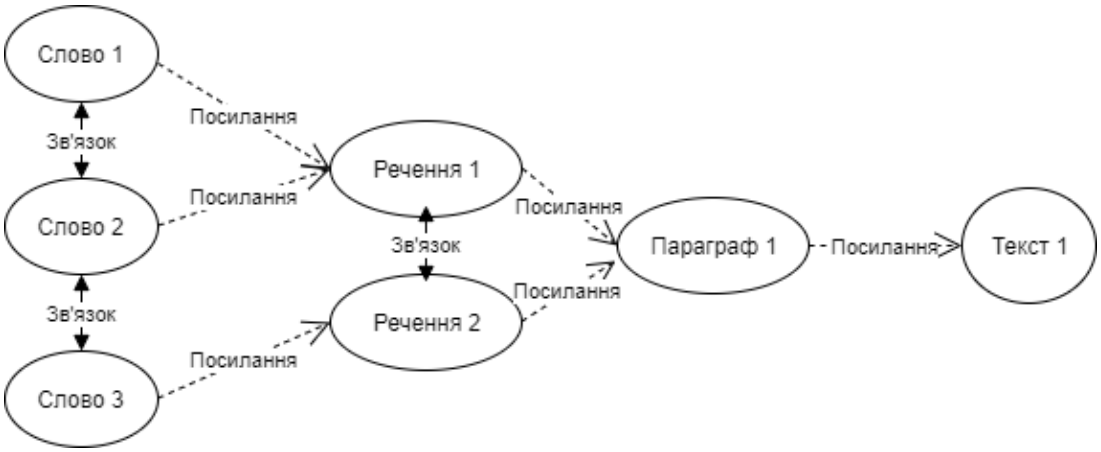

Рисунок 3.2 – Схема організації даних в графовій БД

Описуючи систему з точки зору взаємодії з її користувачами, на рисунках 3.3 і 3.4 надано UML-діаграми послідовності головних сценаріїв користування побудованим програмним застосунком – обробка та інтеграція наданих текстових джерел.

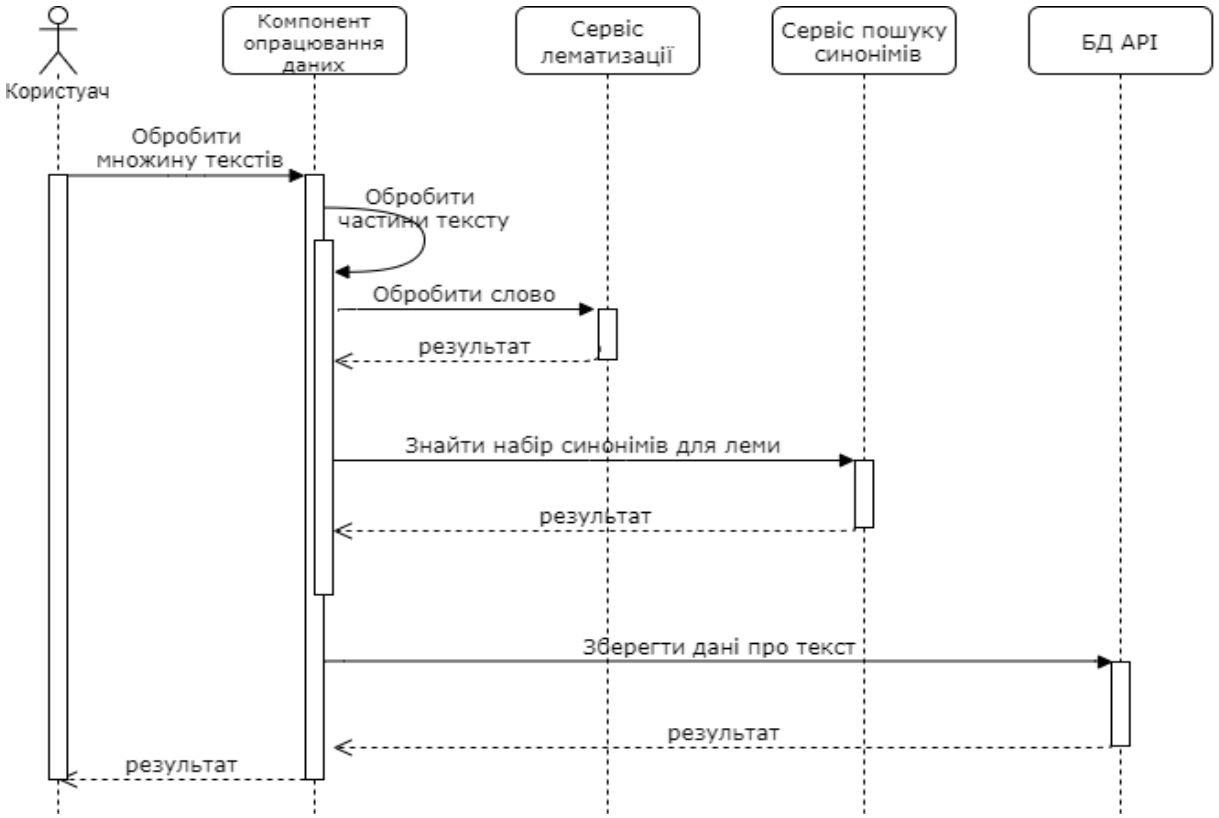

Рисунок 3.3 – Діаграма послідовності роботи сценарію обробки даних

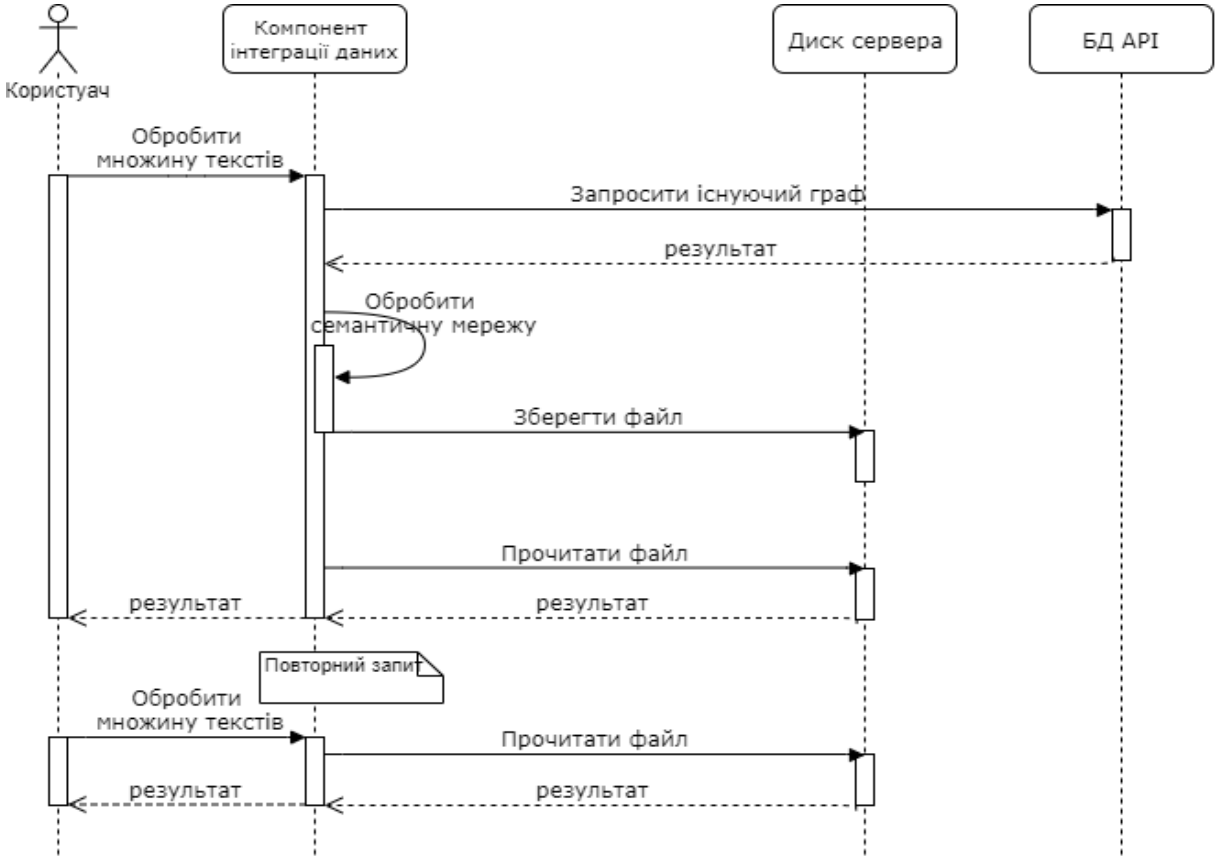

Рисунок 3.4 – Діаграма послідовності сценарію інтеграції даних

З точки зору детальнішого опису складових частин, серверний компонент створеного програмного забезпечення представляє собою систему з трьох функціональних модулів, що відповідно послідовності їх роботи, обробляють вхідні дані природною мовою, перетворюють та аналізують результати опрацювання.

3.1.1 Компонент обробки тексту

Головною задачею компоненту обробки є перетворення вхідного набору текстів природною мовою на відповідну кожному з них семантичну мережу, що містила б необхідну інформацію для виконання подальших операцій з користувацькими даними.

Так, реалізований алгоритм опрацювання текстів можна описати наступним чином: маючи у якості вхідних даних деякий набір текстів,

обробка кожного з яких відбувається одночасно – у паралельних потоках виконання, для кожного тексту, після передуючої нормалізації даних, шляхом дроблення за спеціальними символами, визначаються параграфи, речення та слова у їх первинний формі. Обробка кожного слова включає пошук його канонічного виду, словника синонімів та мовленнєвої інформації щодо роду, числа та відмінка. Останнім етапом опрацювання текстів є їх взаємне узгодження – злиття вузлів даних відносно слів, що зустрічаються багаторазово. Інформація щодо кожного зі структурних елементів зберігається як вузол певного типу в БД з відповідними їх зв'язками.

Внутрішній устрій модуля подано у вигляді спрощеної UML-діаграми класів на рисунку 3.5, уривки відповідного програмного коду – в Додатку А.

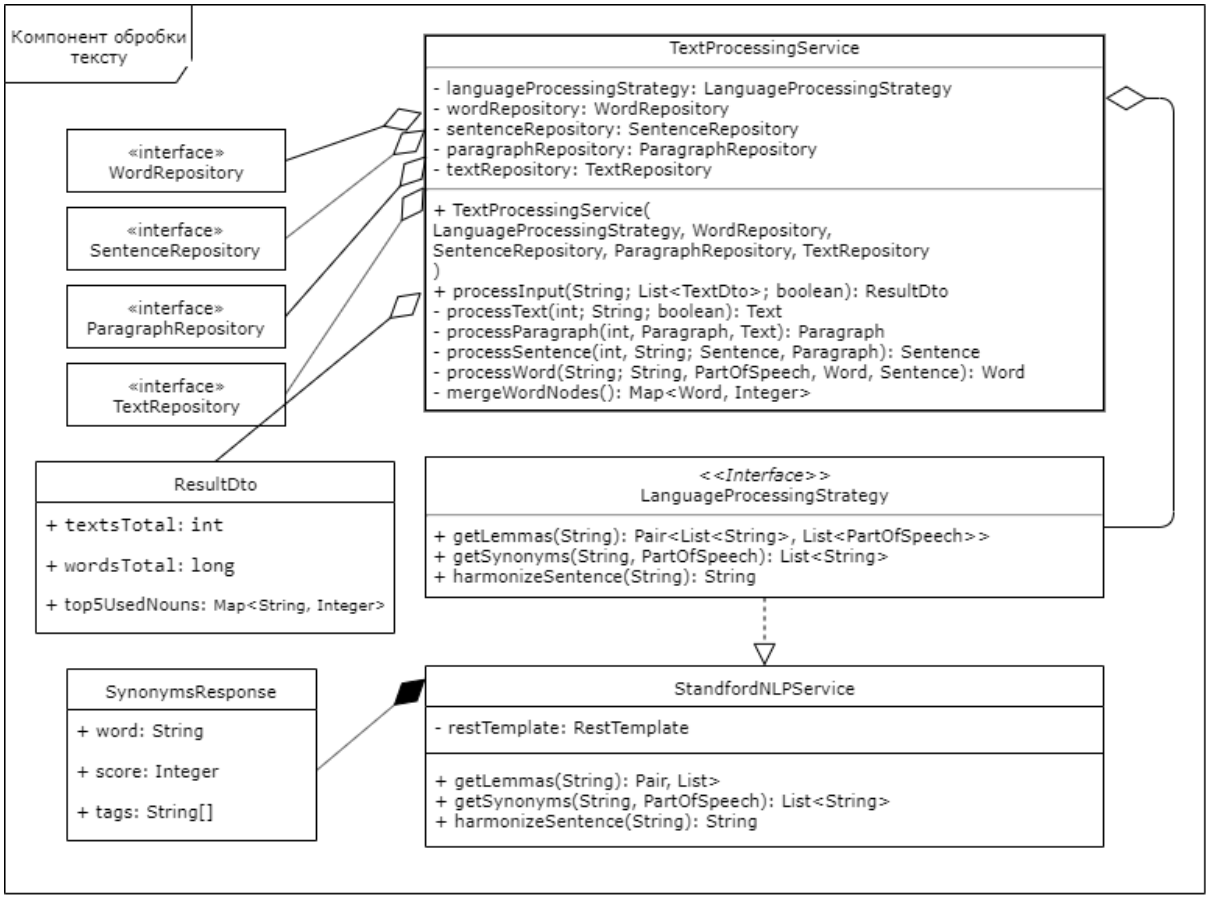

Рисунок 3.5 – Діаграма класів компонента опрацювання текстів

3.1.2 Компонент інтеграції текстів

Після виконаної обробки поданих текстів, система опрацювання оперує результуючим семантичним графом їх множини. Відповідно до описаного у розділі 2 методу інтеграції даних, відтворений алгоритм синтезу нового тексту з існуючих використовує побудовану матрицю інцидентності для вузлів типу «Слово», відфільтрувавши їх за умовою приналежності слова до значущих частин мови, визначає набір подібних з-поміж текстів речень, та, шляхом описаної техніки парафразування, утворює нових текст, узгоджує його та зберігає на диск сервера в якості результату.

Програмний код описаного надано в Додатку Б та проілюстровано діаграмою класів на рисунку 3.6.

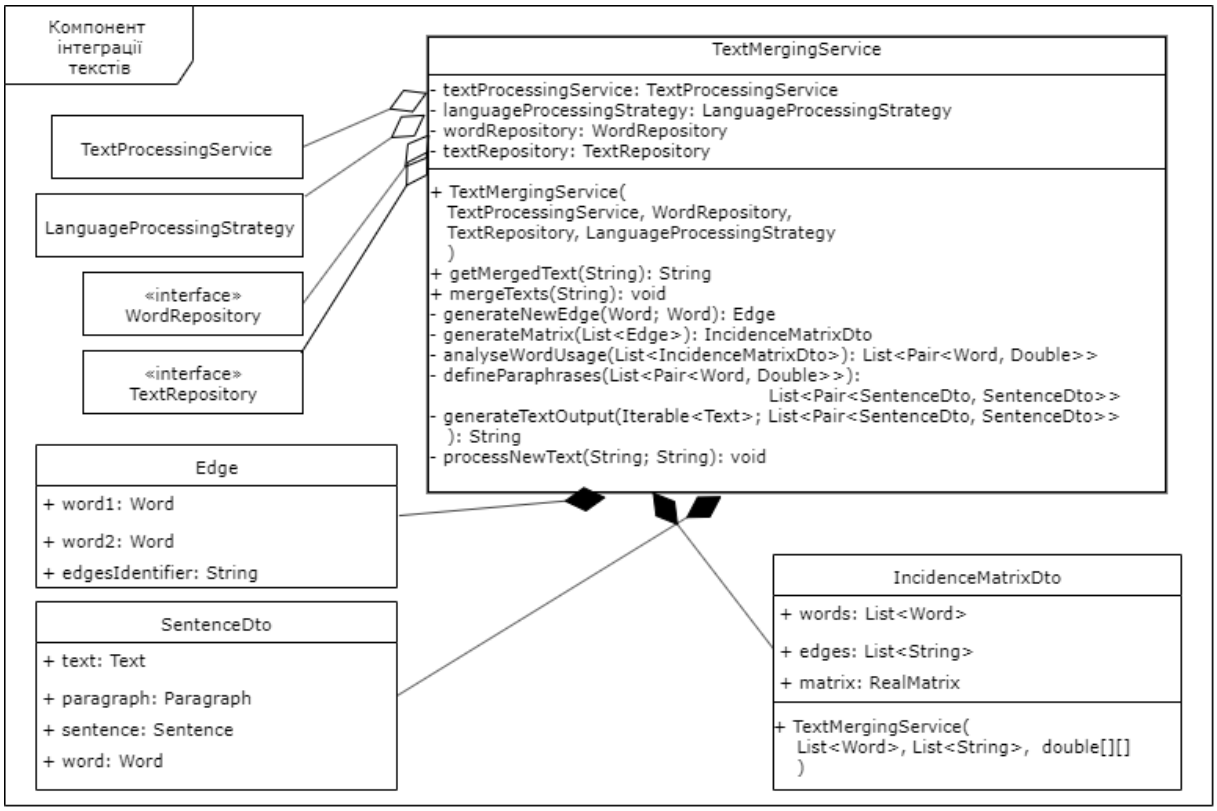

Рисунок 3.6 – Діаграма класів компонента інтеграції текстів

3.1.3 Компонент семантичного аналізу

За бажанням користувача, інформація щодо опрацьованих вхідних та утвореного нового тексту може бути проаналізована. Реалізація такого семантичного аналізу відтворена за допомогою засобів роботи з графовими БД, що використовують властивості цих структур для надання можливості комплексного вивчення даних, що зберігаються.

Побудований програмний інтерфейс для збору семантичної статистики зображено на рисунку 3.7, відповідний код надано у Додатку В.

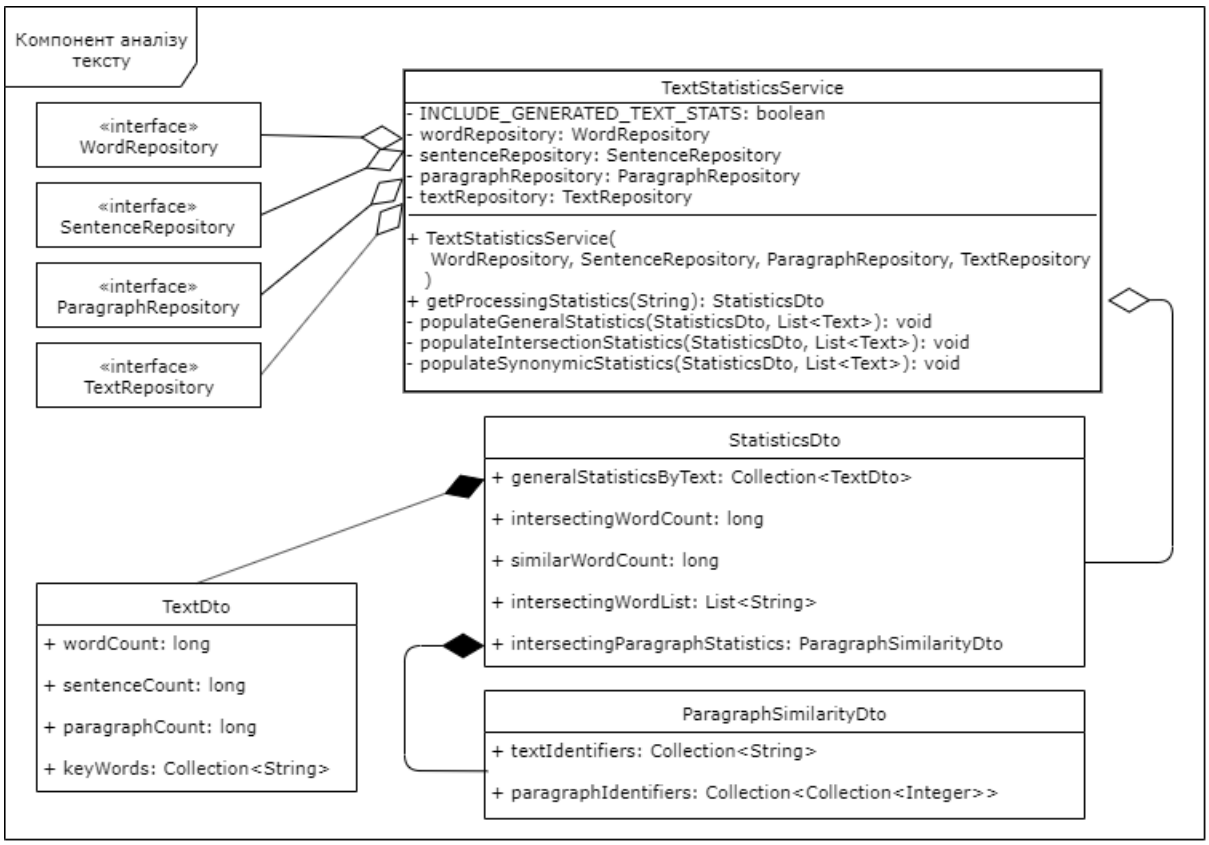

Рисунок 3.7 – Діаграма класів компонента аналізу тексту

3.2 Приклад роботи застосунка

Робота користувача застосунка починається з головного екрана системи (рисунок 3.8), що дозволяє вводити вхідну інформацію шляхом написання її у відповідне вікно чи загрузки текстового файлу.

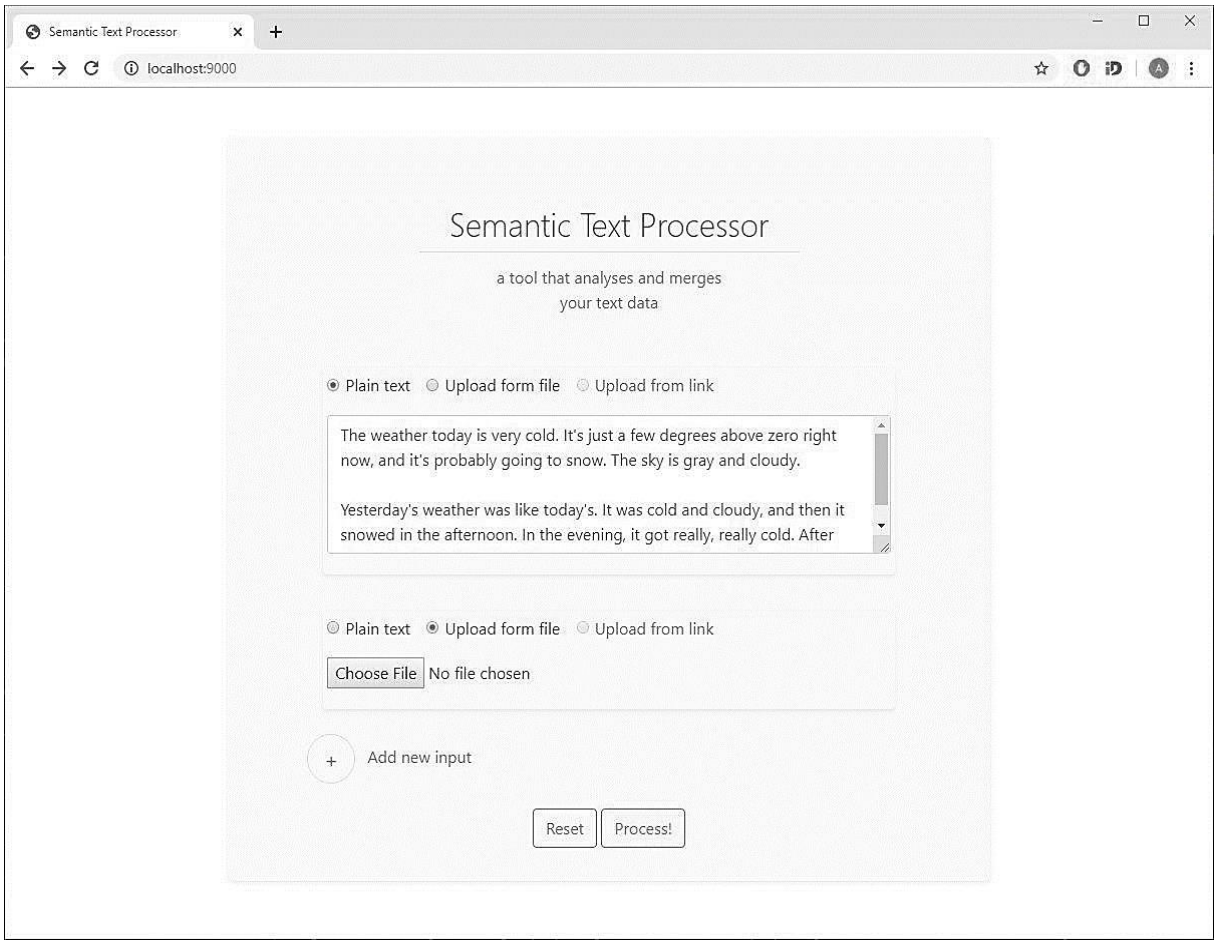

Рисунок 3.8 – Екран вводу текстової інформації

Після успішного обробки наданої інформації, система перенаправляє користувача на екран (рисунок 3.9), що відображає статус опрацювання, надає короткі статистичні відомості щодо роботи системи та пропонує обрати подальшу операцію з обробленими документами.

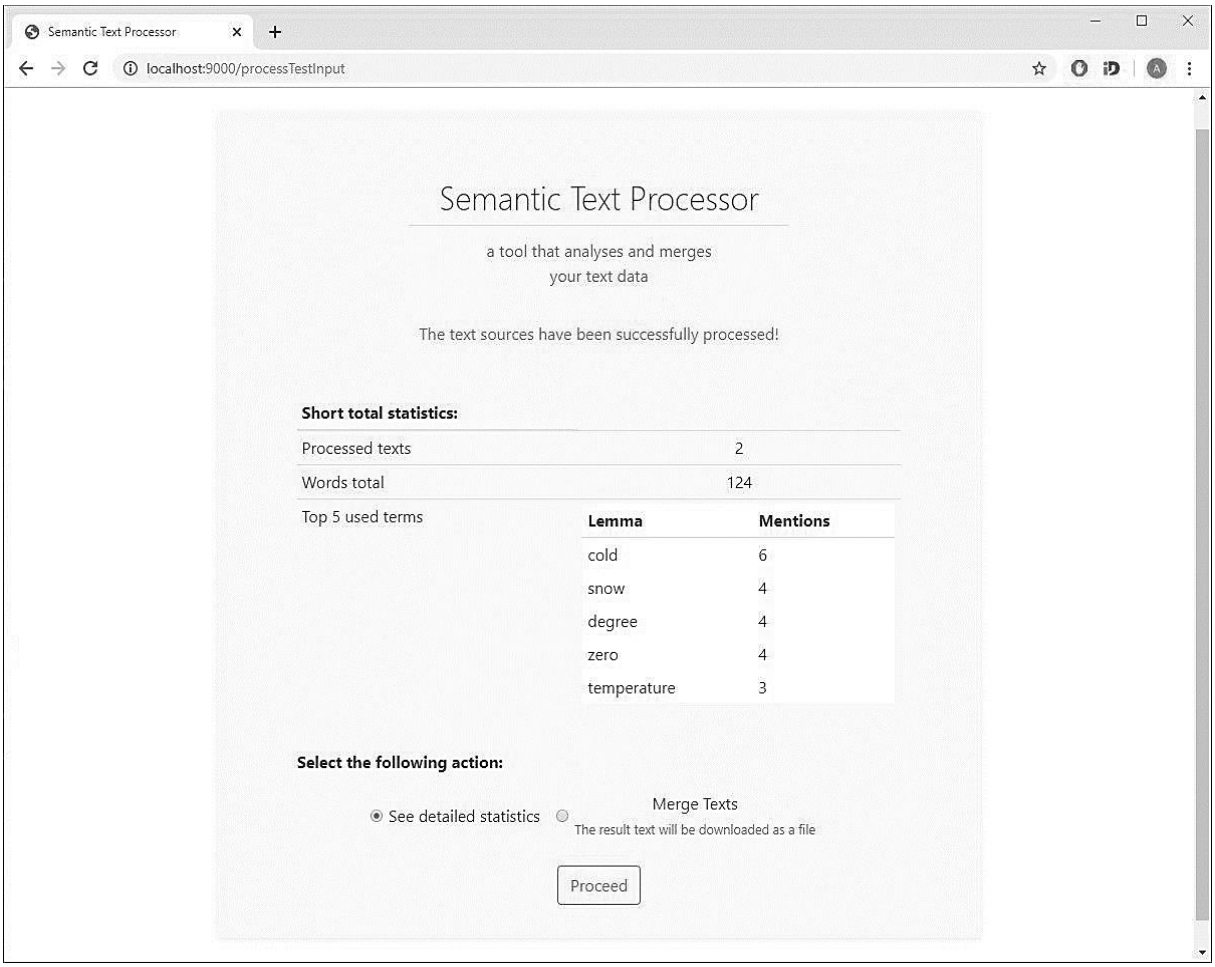

Рисунок 3.9 – Екран успішної обробки користувацького вводу

Обираючи опцію «Злити тексти» (на рисунку 3.9 – «Merge texts»), система виконує операцію інтеграції текстової інформації за описаним алгоритмом, зберігає її результати та відправляє користувачеві текстовий файл з вихідним текстом. Відповідний екран подано на рисунку 3.10.

Опція «Переглянути статистику» (на рисунку 3.9 – «See detailed statistics») відкриває екран (рисунок 3.11), що відображає детальну інформацію щодо оброблених текстових джерел. Зведена статистика включає загальну інформацію про набір текстів, результати семантичного аналізу власне подібних та подібних за синонімічним аналізом слів.

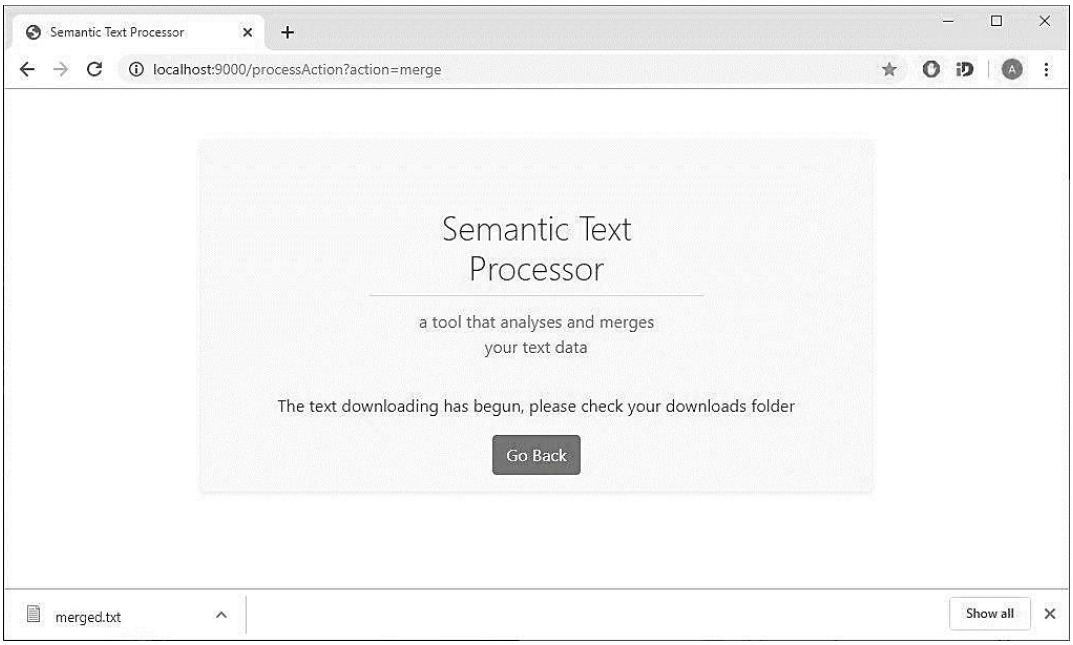

Рисунок 3.10 – Екран операції інтеграції текстів

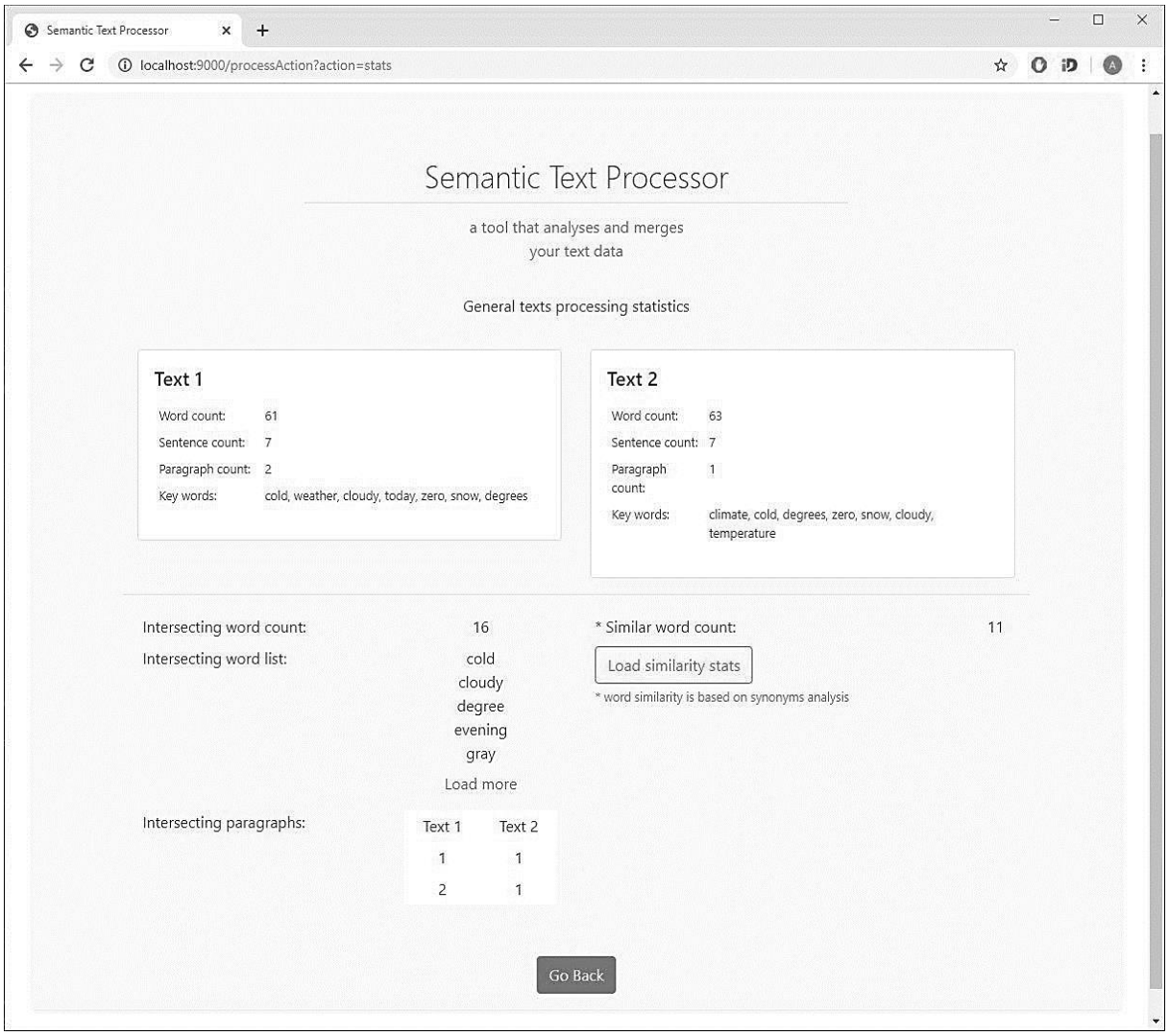

Рисунок 3.11 – Екран семантичного аналізу текстів

3.3 Перспективи вдосконалення розробленої системи

Розроблена архітектура описаного програмного застосунка представляє собою прототип системи, що потенційно здатна оброблювати великі масиви даних. Втім, виконуючи реалізовані операції розподіллено та паралельно, є перспектива досягти покращень в ефективності її роботи.

Так, запроваджуючи засоби роботи з неперервними потоками даних, розділяючи виконання незалежних операцій та утилізуючи кластерні підходи до організації обробки даних, що спонукало б поділ єдиного сховища даних на розподілені компоненти, можна досягти зниження обчислювального навантаження системи, що суттєво покращить показники роботи застосунка.

ВИСНОВКИ

Поточна пояснювальна записка містить результати дослідження проблеми інтелектуального аналізу текстової інформації та відповідні результати роботи розробки та імплементації методу її опрацювання.

У ході роботи, на прикладі проблеми аналізу семантичної подібності текстів, розглядалася задача семантичного аналізу текстів, що є одним зі шляхів застосування методів інтелектуального аналізу даних: були проаналізовані особливості такого типу роботи з мовленнєвими джерелами, були оглянуті існуючі методи та алгоритми опрацювання задачі та був запропонований новий підхід її рішення – використання семантичних графів для збереження та опрацювання смислових зв’язків елементів текстів.

Результатом роботи є відтворений програмний продукт, що реалізує процеси обробки, аналізу та синтезу джерел природної мови, що подається у вигляді текстової інформації. Розроблений метод інтеграції текстів базується на утилізації можливостей опрацювання мережевого представлення даних, використовуючи властивості роботи над ними.

Розроблений програмний застосунок представляє комплекс веб-орієнтованого програмного забезпечення, що складається з власне компонента веб-серверного додатка зі вбудованим користувацьким інтерфейсом, та відповідного сховища даних, що є інструментом роботи з великими графами.

В рамках подальшої роботи за темою, було запропоновано ряд можливих концепційних покращень та окреслено вектор потенційного функціонального розширення реалізованого програмного забезпечення.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. February 2019 Web Server Survey. URL : https://news.netcraft.com/archives/2019/02/28/february-2019-web-server-survey.html (дата звернення: 14.11.2019).

2. Major C. H., Savin-Baden M. An introduction to qualitative research synthesis. London : Routledge, 2010. 200 p.

3. Поляруш О.Н., Витько А.В. Применение байесовских вероятносных сетей для анализа социальных сетей // Системи обробки інформації. 2010. № 6. С. 208-211.

4. Text categorisation: A survey. URL : https://www.cis.uni-muenchen.de/kurse/pmaier/ML_04/material/aas99text.pdf (дата звернення: 14.11.2019).

5. Van Rijsbergen, C. J. Information Retrieval. Wobur : Butterworths, 1979. 152 p.

6. Peat F. D. Artificial Intelligence: How Machines Think. North Carolina : Baen Enterprises, 1985. 356 p.

7. Jones K.S. Natural language processing: a historical review. Computational Linguistics: In Honour of Don Walker : in vol 9. Luxembourg: Springer Dordrecht, 2001. Vol. 9. 598 p.

8. Обработка естественного языка. URL : http://chernykh.net/content/view/1105/1189/ (дата звернення: 1.12.2019).

9. Анисимов А. В., Марченко А. А. Система обработки текстов на естественном языке // Штучний інтелект. 2002. № 4. С. 157.

10. Iroju O. G., Olaleke J. O. A Systematic Review in Natural Language Processing in Healthcare // Information Technology and Computer Science. 2015. № 8. P. 45.

11. Ткач О.М. Методи дослідження семантичних полів слів // Збірник наукових праць Кам'янець-Подільського національного

університету імені Івана Огієнка. Серія : Соціально-педагогічна. 2013. № 23. С. 350-357.

12. Автоматическая Обработка Текста. Первичный семантический анализ. URL : http://aot.ru/docs/seman.html (дата звернення: 14.11.2019).

13. Indurkhya N., Damerau F. Handbook of Natural Language Processing. Cambridge : CRC Press, 2010. 676 p.

14. Paraphrase Identification with Lexico-Syntactic Graph Subsumption / V. Rus, P. McCarthy, M. Lintean, D. McNamara, A. Graesser // Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference. Florida, USA. 2008. P. 201-206.

15. Wubben S., Van den A., and Krahmer E., Paraphrase Generation as Monolingual Translation: Data and Evaluation. URL : http://ilk.uvt.nl/~swubben/publications/INLG2010.pdf (дата звернення: 14.11.2019)

16. Bhagat R., Hovy E., Patwardhan S. Acquiring Paraphrases From Text Corpora. Proceedings of the 5th International Conference on Knowledge Capture. New York, USA. 2009. P. 161-168.

17. Dagan I., Glickman O., Magnini B. The Pascal Recognising Textual Entailment Challenge. Proceedings of the 1st PASCAL Machine Learning Challenges Workshop. Southampton. 2006. P. 177-190.

18. Paraphrase Identification with Lexico-Syntactic Graph Subsumption / V. Rus, P. McCarthy, M. Lintean, D. McNamara, A. Graesser // Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference. Florida, USA. 2008. P. 201-206.

19. Elberrichi Z., Abidi K. Arabic Text Categorization: A Comparative Study of Different Representation Modes // The International Arab Journal of Information Technology. 2012. Vol. 9, No. 5. P. 465-470.

20. Salton G., Buckley C. Term Weighting Approaches in Automatic Text Retrieval // Information Processing and Management. 1988. Vol. 24, No. 5. P. 513-523.

21. Landauer K., Foltz W., Laham D. An Introduction to Latent Semantic Analysis // Discourse Processes. 1998. Vol. 25, No. 2. P. 259-284.

22. Dumais S. T. Latent Semantic Analysis // Annual Review of Information Science and Technology. 2005. №38. P. 188–230.

23. Blei D. M., Andrew Y. Ng, Michael I. J. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. № 3. P. 993-1022.

24. Girolami A., Vinokourov A. Probabilistic Framework for the Hierarchic Organisation and Classification of Document Collections // Journal of Intelligent Information System. 2002. № 18. P. 153–172.