Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему *Метод автоматизованого формування семантичного ядра цифрових текстів*

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-20-1*		*В.О. Купрійчук*
	Курс, група виконавця	Підпис	Ініціали, прізвище
Керівник:	*викладач кафедри КН*		*П.М. Радюк*
	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

____ _____ 2021 р.

Хмельницький 2021

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* *2021* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ МАГІСТРА

1. Тема кваліфікаційної роботи магістра: *«Метод автоматизованого формування семантичного ядра цифрових текстів»*

2. Завдання видано студенту *Купрійчуку Владиславу Олександровичу*
(прізвище, ім'я, по батькові)

3. Керівник роботи *викладач кафедри КН Радюк Павло Михайлович*
(прізвище, ім'я, по батькові)

4. Затверджені наказом університету від «*25*» *серпня 2021* р. № *102*

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

Мета роботи полягає у розробці методу автоматизованого формування семантичного ядра цифрових текстів, що призначений для автоматизованого формування множини ключових семантичних одиниць за цифровим текстом та множинами слів і словосполучень цього тексту з показниками їх важливості, одержуючи сформовані семантичні ядра за різними показниками важливості, зокрема із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах. Провести прикладне дослідження розробленого методу автоматизованого формування семантичного ядра цифрових текстів і виконати аналіз результатів використання відповідної інформаційної системи..

Реферат

Кваліфікаційна робота магістра розв’язує науково-технічну задачу автоматизованого формування множини ключових семантичних одиниць за цифровим текстом та множинами слів і словосполучень цього тексту з показниками їх важливості, одержуючи сформовані семантичні ядра за різними показниками важливості, зокрема із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Актуальність теми. Аналізуючи граматичні структури й конструкції речень та відслідковуючи відношення між словами у визначеному контексті, методи семантичного аналізу текстів дозволяють комп’ютеру розпізнавати, розуміти та інтерпретувати речення, абзаци та повноцінні документи – усю цифрову інформацію, з якою людина працює кожного дня.

Однією із найбільших проблем на IT ринку, досі не вирішених, є відсутність або ж недосконалість технологій та сервісів автоматичної смислової обробки неструктурованої текстової інформації. Ця проблема ускладнюється ще й тим, що для автоматизованої змістовної обробки цифрових текстів необхідний комплекс методів, які зможуть забезпечити реалізацію алгоритмів і програмного забезпечення повного комп'ютерного лінгвістичного аналізу текстів природною мовою та автоматичної змістовної обробки інформації.

Для реалізації обчислювального процесу автоматичної смислової обробки інформації в комп'ютерних системах широкого застосування повинні бути розроблені способи організації високоефективного обчислювального процесу, що забезпечують формування результатів пошуку та аналітичної обробки інформації в реальному масштабі часу. Виходячи з наведеного, розробка методів і засобів формування семантичного ядра цифрових текстів у вигляді обмеженої множини ключових слів та словосполучень є актуальною задачею інформаційних технологій.

Мета і задачі роботи. Мета кваліфікаційної роботи магістра полягає у розробці методу автоматизованого формування семантичного ядра цифрових текстів, що призначений для автоматизованого формування множини ключових семантичних одиниць за цифровим текстом та множинами слів і словосполучень цього тексту з показниками їх важливості, одержуючи сформовані семантичні ядра за різними показниками важливості, зокрема із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах. Для досягнення поставленої мети розробки методу автоматизованого формування семантичного ядра цифрових текстів були поставлені й вирішені наступні завдання:

1. Проведено аналіз предметної області семантичного аналізу цифрових текстів та відомих підходів до автоматизації формування семантичного ядра цифрових текстів.

2. Вдосконалено метод автоматизованого формування семантичного ядра цифрових текстів.

3. Розроблено інформаційну технологію автоматизованого формування множини ключових семантичних одиниць.

4. Розроблено інформаційну систему автоматизованого формування множини ключових семантичних одиниць.

5. Проведено прикладне дослідження методу автоматизованого формування семантичного ядра цифрових текстів у складі інформаційної технології автоматизованого формування множини ключових семантичних одиниць і виконано аналіз результатів використання відповідної інформаційної системи.

Об’єкт дослідження – процес формування семантичного ядра цифрових текстів.

Предмет дослідження – інформаційні технології, моделі, методи та алгоритми для автоматизації процесів формування множин ключових слів і словосполучень цифрових текстів.

Методи дослідження, застосовані для вирішення поставлених завдань: для розв'язання поставлених задач використовуються основні положення методів аналізу даних і теорії множин, а для реалізації інформаційної системи автоматизованого формування множини ключових семантичних одиниць – методології проектування інформаційних систем і об'єктно-орієнтований підхід.

Наукова новизна одержаних результатів. В результаті роботи були отримані інновації і положення наукової новизни:

1. Вдосконалено метод автоматизованого формування семантичного ядра цифрових текстів, що дозволяє за множиною слів і словосполучень цифрового тексту з зіставленими значеннями показників їх семантичної важливості, а також необхідним відсотком щільності ключових слів у тесті автоматизовано формувати відповідні множини ключових слів та словосполучень тексту за різними показниками важливості.

2. Розроблено нову інформаційну технологію автоматизованого формування множини ключових семантичних одиниць, що дозволяє з використанням створеного методу автоматизованого формування семантичного ядра цифрових текстів за вхідними даними у вигляді цифрового тексту та множина слів і словосполучень цього тексту з показниками їх важливості одержувати вихідні дані у вигляді сформованих семантичних ядер із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Практичне значення одержаних результатів. В роботі виконано розробку експериментальної інформаційної системи автоматизованого формування множини ключових семантичних одиниць, реалізовано модулі та складові для інформаційної системи. Інформаційна система використовує створений метод автоматизованого формування семантичного ядра цифрових текстів й забезпечує можливість одержувати дані у вигляді сформованих семантичних ядер із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при

обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Було проведено ряд досліджень, коли для різних цільових відсотків щільності ключових слів у тесті обраховувались компоненти семантичного ядра цифрових текстів за різними способами обрахунку. Середня точність формування семантичного ядра тексту із слів при обрахунку порогу щільності у символах склала 82,93%, середня точність формування семантичного ядра тексту із словосполучень при обрахунку порогу щільності у символах склала 86,19%, середня точність формування семантичного ядра тексту із слів при обрахунку порогу щільності у словах склала 79,63%, а середня точність формування семантичного ядра тексту із словосполучень при обрахунку порогу щільності у словах склала 83,55%.

Загалом формування семантичного ядра тексту із словосполучень (при цьому одне слово тут розглядається як різновид словосполучення) виявило більш точний результат ніж формування семантичного ядра тексту із слів, оскільки дозволило включити в актуальну множину ті слова, які не були визначені важливими але мали семантичну важливість у комбінації з іншими словами. Формування семантичного ядра тексту при обрахунку порогу щільності у символах виявилось більш ефективним за формування семантичного ядра при обрахунку порогу щільності у словах, оскільки дозволяє більш точно зафіксувати значення порогу щільності ключових слів у тексті.

Одержані результати свідчать, що розроблений метод автоматизованого формування семантичного ядра цифрових текстів, який використовує обмеження множин ключових слів і словосполучень за значенням порогу щільності, дозволяє одержати достатньо репрезентативні складові семантичного ядра тексту. Це дає привід стверджувати, що прикладне використання розробленого методу може бути ефективно застосоване при вирішення задач семантичного аналізу текстів відповідно до призначення.

Апробація результатів кваліфікаційної роботи магістра та публікації. Основні наукові та практичні результати кваліфікаційної роботи магістра

доповідались на XIII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН–2021» (15–16 жовтня 2021 року) у доповіді на тему «Інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів»; за темою роботи автором виконано наукову публікацію:

Войчишин О. О., Залуцька О. О., Попов Ю. М., Купрійчук В. О. Інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів. Збірник наукових праць за матеріалами XIII Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2021». Хмельницький, 2021. с. 298-305.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається із реферату, завдання, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань із 29 найменувань та 4 додатків. Загальний обсяг кваліфікаційної роботи магістра становить 102 сторінки, з них 86 сторінок основного тексту та 16 сторінок додатків. У роботі наведено 28 рисунків та 13 таблиць.

Ключові слова: семантичне ядро, семантичний аналіз, ключові слова, ключові словосполучення, цифрові тексти, дисперсійне оцінювання, семантичні одиниці, інформаційна система, інформаційна технологія.

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
АІС	Автоматизована інформаційна система
ІС	Інформаційна система
ІТ	Інформаційні технології
КРМ	Кваліфікаційна робота магістра
КН	Комп’ютерні науки
НТІ	Науково-технічна інформація
ХНУ	Хмельницький національний університет
JVM	Java Virtual Machine
CLR	Common Language Runtime
FCL	Framework Class Library

Вступ

Кваліфікаційна робота магістра розв’язує науково-технічну задачу автоматизованого формування множини ключових семантичних одиниць за цифровим текстом та множинами слів і словосполучень цього тексту з показниками їх важливості, одержуючи сформовані семантичні ядра за різними показниками важливості, зокрема із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Актуальність теми. Аналізуючи граматичні структури й конструкції речень та відслідковуючи відношення між словами у визначеному контексті, методи семантичного аналізу текстів дозволяють комп’ютеру розпізнавати, розуміти та інтерпретувати речення, абзаци та повноцінні документи – усю цифрову інформацію, з якою людина працює кожного дня [1].

Однією із найбільших проблем на IT ринку, досі не вирішених, є відсутність або ж недосконаліть технологій та сервісів автоматичної смислової обробки неструктурованої текстової інформації. Ця проблема ускладнюється ще й тим, що для автоматизованої змістовної обробки цифрових текстів необхідний комплекс методів, які зможуть забезпечити реалізацію алгоритмів і програмного забезпечення повного комп'ютерного лінгвістичного аналізу текстів природною мовою та автоматичної змістовної обробки інформації [2].

Для реалізації обчислювального процесу автоматичної смислової обробки інформації в комп'ютерних системах широкого застосування повинні бути розроблені способи організації високоефективного обчислювального процесу, що забезпечують формування результатів

пошуку та аналітичної обробки інформації в реальному масштабі часу. Виходячи з наведеного, розробка методів і засобів формування семантичного ядра цифрових текстів у вигляді обмеженої множини ключових слів та словосполучень є актуальною задачею інформаційних технологій.

Мета і задачі роботи. Мета кваліфікаційної роботи магістра полягає у розробці методу автоматизованого формування семантичного ядра цифрових текстів, що призначений для автоматизованого формування множини ключових семантичних одиниць за цифровим текстом та множинами слів і словосполучень цього тексту з показниками їх важливості, одержуючи сформовані семантичні ядра за різними показниками важливості, зокрема із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах. Для досягнення поставленої мети розробки методу автоматизованого формування семантичного ядра цифрових текстів були поставлені й вирішені наступні завдання:

1. Проведено аналіз предметної області семантичного аналізу цифрових текстів та відомих підходів до автоматизації формування семантичного ядра цифрових текстів.

2. Вдосконалено метод автоматизованого формування семантичного ядра цифрових текстів.

3. Розроблено інформаційну технологію автоматизованого формування множини ключових семантичних одиниць.

4. Розроблено інформаційну систему автоматизованого формування множини ключових семантичних одиниць.

5. Проведено прикладне дослідження методу автоматизованого формування семантичного ядра цифрових текстів у складі інформаційної технології автоматизованого формування множини ключових семантичних одиниць і виконано аналіз результатів використання відповідної інформаційної системи.

Об’єкт дослідження – процес формування семантичного ядра цифрових текстів.

Предмет дослідження – інформаційні технології, моделі, методи та алгоритми для автоматизації процесів формування множин ключових слів і словосполучень цифрових текстів.

Методи дослідження, застосовані для вирішення поставлених завдань: для розв’язання поставлених задач використовуються основні положення методів аналізу даних і теорії множин, а для реалізації інформаційної системи автоматизованого формування множини ключових семантичних одиниць – методології проектування інформаційних систем і об’єктно-орієнтований підхід.

Наукова новизна одержаних результатів. В результаті роботи були отримані інновації і положення наукової новизни:

1. Вдосконалено метод автоматизованого формування семантичного ядра цифрових текстів, що дозволяє за множиною слів і словосполучень цифрового тексту з зіставленими значеннями показників їх семантичної важливості, а також необхідним відсотком щільності ключових слів у тесті автоматизовано формувати відповідні множини ключових слів та словосполучень тексту за різними показниками важливості.

2. Розроблено нову інформаційну технологію автоматизованого формування множини ключових семантичних одиниць, що дозволяє з використанням створеного методу автоматизованого формування

семантичного ядра цифрових текстів за вхідними даними у вигляді цифрового тексту та множина слів і словосполучень цього тексту з показниками їх важливості одержувати вихідні дані у вигляді сформованих семантичних ядер із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Практичне значення одержаних результатів. В роботі виконано розробку експериментальної інформаційної системи автоматизованого формування множини ключових семантичних одиниць, реалізовано модулі та складові для інформаційної системи. Інформаційна система використовує створений метод автоматизованого формування семантичного ядра цифрових текстів й забезпечує можливість одержувати дані у вигляді сформованих семантичних ядер із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Було проведено ряд досліджень, коли для різних цільових відсотків щільності ключових слів у тесті обраховувались компоненти семантичного ядра цифрових текстів за різними способами обрахунку. Середня точність формування семантичного ядра тексту із слів при обрахунку порогу щільності у символах склала 82,93%, середня точність формування семантичного ядра тексту із словосполучень при обрахунку порогу щільності у символах склала 86,19%, середня точність формування семантичного ядра тексту із слів при обрахунку порогу щільності у словах склала 79,63%, а середня точність формування семантичного ядра тексту із словосполучень при обрахунку порогу щільності у словах склала 83,55%.

Загалом формування семантичного ядра тексту із словосполучень (при цьому одне слово тут розглядається як різновид словосполучення) виявило більш точний результат ніж формування семантичного ядра тексту із слів, оскільки дозволило включити в актуальну множину ті слова, які не були визначені важливими але мали семантичну важливість у комбінації з іншими словами. Формування семантичного ядра тексту при обрахунку порогу щільності у символах виявилось більш ефективним за формування семантичного ядра при обрахунку порогу щільності у словах, оскільки дозволяє більш точно зафіксувати значення порогу щільності ключових слів у тексті.

Одержані результати свідчать, що розроблений метод автоматизованого формування семантичного ядра цифрових текстів, який використовує обмеження множин ключових слів і словосполучень за значенням порогу щільності, дозволяє одержати достатньо репрезентативні складові семантичного ядра тексту. Це дає привід стверджувати, що прикладне використання розробленого методу може бути ефективно застосоване при вирішення задач семантичного аналізу текстів відповідно до призначення.

Апробація результатів кваліфікаційної роботи магістра та публікації. Основні наукові та практичні результати кваліфікаційної роботи магістра доповідались на XIII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп’ютерних наук АПКН–2021» (15–16 жовтня 2021 року) у доповіді на тему «Інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів»; за темою роботи автором виконано наукову публікацію:

Войчишин О. О., Залуцька О. О., Попов Ю. М., Купрійчук В. О. Інформаційна технологія автоматизованого формування семантичного ядра

цифрових текстів. Збірник наукових праць за матеріалами XIII Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук АПКН-2021». Хмельницький, 2021. с. 298-305.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається із реферату, завдання, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань із 29 найменувань та 4 додатків. Загальний обсяг кваліфікаційної роботи магістра становить 102 сторінки, з них 86 сторінок основного тексту та 16 сторінок додатків. У роботі наведено 28 рисунків та 13 таблиць.

Розділ 1

Аналіз предметної області семантичного аналізу цифрових текстів

1.1 Дослідження актуальності використання засобів формування семантичного ядра цифрових текстів

Щоденний розвиток інформаційних технологій та стрімке збільшення обсягів інформації, безсумнівно, дають людині велику множину переваг, знання стають все більш доступними. Однак виникає і наступна проблема: за такої кількості технічної літератури, документів, технологічних даних та статей стає складно виокремити дійсно корисний та потрібний матеріал. Часто можна зіткнутись із тим, що вищеописаний матеріал не має змісту чи зазначеної мети, а це, в свою чергу, змушує виконувати зайву роботу, витрачаючи час, котрий міг бути використаний на безпосередню роботу із власним матеріалом чи проєктом, а не на пошук та аналіз інформації.

Аналізуючи граматичні структури й конструкції речень та відслідковуючи відношення між словами у визначеному контексті, методи семантичного аналізу текстів дозволяють комп’ютеру розпізнавати, розуміти та інтерпретувати речення, абзаци та повноцінні документи – усю цифрову інформацію, з якою людина працює кожного дня [1].

Однією із найбільших проблем на IT ринку, досі не вирішених, є відсутність або ж недосконалість технологій та сервісів автоматичної смислової обробки неструктурованої текстової інформації. Ця проблема ускладнюється ще й тим, що для автоматизованої змістовної обробки цифрових текстів необхідний комплекс методів, які зможуть забезпечити реалізацію алгоритмів і програмного забезпечення повного комп'ютерного лінгвістичного аналізу текстів природною мовою та автоматичної

змістовної обробки інформації [2]. Для реалізації обчислювального процесу автоматичної смислової обробки інформації в комп'ютерних системах широкого застосування повинні бути розроблені способи організації високоефективного обчислювального процесу, що забезпечують формування результатів пошуку та аналітичної обробки інформації в реальному масштабі часу. Відсутність зазначеного науково-методичного базису і алгоритмів обумовлює наявність технологічних обмежень на рівень автоматизації процесів обробки інформації в електронних системах збереження інформації.

Інтелектуалізація процесу автоматичної обробки НТІ дозволить істотно розширити повноту оброблюваних даних, а також реалізувати принципово нові функції їх автоматичної смислової обробки.

Методи та системи семантичного аналізу можуть запропонувати значення слів та фраз, що можуть створити неточності як для розуміння читачем, так і для перекладу. Також це може допомогти у роботі різних компаній – автоматично вилучати інформацію, що може представляти цінність, наприклад із неструктурованих даних, таких як електронні листи, відгуки клієнтів чи заявки до технічної підтримки [3].

Окрім значного потенціалу для комерційних структур, робота таких інформаційних систем може допомогти в безлічі інших сфер діяльності людини – від аналізу наукових, технічних документів та літератури і до впровадження методів семантичного аналізу текстів у WEB-застосуваннях та пошукових системах.

Одним з перших важливих кроків використання інформаційних технологій в лінгвістиці є діджиталізація текстів – переведення матеріалу, існуючого в друкованому або усному вигляді, в цифрову форму. Саме в цьому випадку з'являється можливість залучення комп'ютерів для

виконання певних операцій над текстами природною мовою: їх перетворення, виділення з них окремих елементів і створення (синтезу) аналогічних текстів.

При автоматичному аналізі звукової мови вона перетворюється в друкований текст, над яким можна проводити подальші операції. Автоматичний синтез усної мови являє собою зворотний процес перетворення друкованого тексту, існуючого в цифровій формі, в звучний текст на природну людську мовур[4].

Створення надійної інформаційної системи для пошуку даних в науковій діяльності ставить завдання розробки систем, здатних виконувати такі дії, для яких зазвичай необхідне залучення людського розуму. Одне з найважливіших напрямків теоретичних розробок зі створення подібних систем пов’язане з моделюванням інтелектуальної діяльності людини. При цьому виникають різного роду проблеми, зумовлені тим, що розумова діяльність людини різноманітна і включає в себе значний пласт задач, з якими обчислювальні машини не здатні впоратись в повній мірі.

Електронний текст став феноменом, якому у сучасному науковому просторі приділяється велика кількість уваги. Саме він розглядається як основне джерело інформації. Існує кілька підходів до його аналізу. Можна, наприклад, визначати тему і ідею текстів, аналізувати, оцінювати смислове навантаження або виділяти сферу, з якою вони пов'язані (математика, комп'ютерні науки, література, соціологія) [5]. Однак комп'ютерні системи обробки даних, такі як пошукові або порівняльні системи, таких умінь не мають. Вони аналізують інформацію інакше. Тому актуальною є проблема розробки комп'ютерних методів і алгоритмів моделювання діяльності людського мозку. Для аналізу текстової інформації використовуються алгоритми семантичної, морфологічної та синтаксичної обробки текстів.

Саме ті галузі прикладної лінгвістики, які пов'язані з залученням комп'ютерів для вирішення практичних завдань використання мови, є предметом комп'ютерної лінгвістики, яка оформилася в 1960-і роки як особливий науковий напрямок. Комп'ютерну лінгвістику можна визначити як область використання комп'ютерних інструментів – програм, технологій організації та обробки даних для моделювання функціонування мови в тих чи інших умовах, а також сферу застосування комп'ютерних моделей мови в лінгвістиці та суміжних з нею дисциплінах [6].

У зв'язку з тим, що мова являє собою досить складне утворення, в комп'ютерній лінгвістиці склалися і розвиваються різні напрямки, приблизно порівнянні з окремими рівнями мови, з процесами породження і сприйняття мовленнєвих повідомлень або іншими видами людської діяльності, пов'язаної з мовою. Відповідно, до напрямів комп'ютерної лінгвістики належать [7]:

– автоматизований синтез текстів;

– автоматизований аналіз текстів;

– створення та підтримка автоматичних словників;

– створення автоматизованих інформаційно-пошукових систем;

– машинний переклад;

– створення автоматичних систем вивчення мови;

– автоматична атрибуція та дешифрування текстів;

– створення лінгвістичних баз даних;

– розробка програмних інструментів для рішення задач теоретичної та прикладної лінгвістики.

В умовах все зростаючого кількості текстів в навколишньому людини світі виникає проблема: як в морі інформації віднайти потрібні документи і познайомитися з їх змістом. Вирішенню цієї проблеми може

допомогти складання рефератів і анотацій повнотекстових документів. Вони дають читачеві уявлення про зміст вихідних документів і дозволяють оцінити ступінь необхідності звернення до повних текстів кожної роботи. Крім того, реферати та анотації акцентують увагу читача на нових відомостях, тобто дозволяють за невеликий проміжок часу дізнатися багато нової інформації.

Реферати та анотації складаються вручну, наприклад самим автором вихідного тексту або бібліографічним працівником, або автоматично, за допомогою спеціальних комп'ютерних програм. Найбільш якісним є перший вид рефератів і анотацій, оскільки в цьому випадку створюється новий текст, який називає основну думку висловлювання і відрізняється зв'язковим характером. Але для обробки великого масиву текстів за мінімальну кількість часу потрібне залучення автоматичних засобів для вирішення завдання реферування і анотування текстів [8].

Велика кількість наукових праць була спрямована на розробку математичних алгоритмів та комп'ютерних програм обробки текстів природною мовою. Для автоматизації цих процесів було створено різні моделі процесів обробки та аналізу текстів, а також структури та алгоритми для представлення результатів. У переважній більшості аналіз цифрових текстів було представлено наступною послідовністю: морфологічний аналіз тексту, синтаксичний аналіз та семантичний аналіз. Для кожного з цих етапів були створені відповідні моделі та алгоритми [9].

Найчастіше явища полісемії, омонімії та інші морфологічні явища приносять неоднозначність в мову та значно ускладнюють задачу встановлення коректного відображення семантично-синтаксичної структури тексту в його формальне логічне представлення.

Варто описати вищенаведені терміни та доповнити їх для чіткості розуміння предметної області. Полісемія – мовленнєве явище, що дослівно означає «багатозначність». Це означає, що мовна одиниця (слово, фразема, синтаксична форма) має декілька значень [10]. Омонімія – явище, за якого два слова, що звучать та пишуться однаково мають різне значення.

З іншого боку застосування ресурсоємних функцій логічно-семантичного аналізу робить програми обробки тексту занадто складними та повільними. Людина в процесі розуміння тексту не так часто застосовує логіку – лише по мірі виникнення логічних задач, в інших випадках відбувається асоціативний пошук семантичного концепту, що відповідає даному слову та є контекстно близьким до свого оточення. При цьому асоціативний пошук є значно швидшим та більш економічним засобом розв'язання неоднозначності інтерпретації тексту.

Незважаючи на всі наукові досягнення, існуючі лінгвістичні алгоритми не можуть поки що зрівнятися по якості з можливостями людини. Однією з головних причин цього є інформаційна ізольованість процесів обробки на кожному етапі аналізу – під час роботи процесу обмін даними з іншими процесами не відбувається. Процеси обмінюються даними лише при переході від попереднього етапу до наступного – тобто вихід попереднього процесу є входом для наступного. В той же час семантичний, синтаксичний та морфологічний аналізи природної мови, що здійснюються людиною, є паралельними взаємодіючими процесами. При визначенні структури речення один процес використовує результати інших [11].

Ключове слово є словом або словосполученням природної мови, яке використовують для вираження деякого аспекту змісту навчального матеріалу [12]. Елементи множини ключових термінів мають істотне

смислове навантаження і формують перелік розглянутих в навчальному матеріалі понять. Ключові терміни мають наступні властивості:

– є найбільш вживаними (частотними) найменуваннями, визначають ознаку предмета, стан або дію;

– представлені значущою лексикою, досить узагальнені за своєю семантикою (середнього ступеня абстракції), стилістично нейтральні й не оціночні;

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

– мінімальна кількість елементів у множині ключових термінів наближається до інваріанта змісту навчального матеріалу при їх логічному впорядкування;

– множина ключових термінів навчального матеріалу складається з 5-15 або 8-10 слів, що відповідає обсягу оперативної пам'яті людини [13].

Побудова семантичного ядра поділяється на наступні етапи:

– збір слів, що детально описують текст, з врахуванням його тематики й призначення, що можна зробити з допомогою SEO-аналізаторів тексту;

– кластеризація зібраних ключів по відповідних розділах і підрозділах тексту чи проєкту.

Окрім побудови семантичного ядра, не менш актуальною темою для дослідження є аналіз готового тексту та виокремлення ключових слів у ньому. Однак необхідно, щоб система розуміла, яке слово можна таким вважати, а яке необхідно відкинути при аналізі – тобто вміти знаходити межу між цими словами.

Семантичне ядро – це певний невпорядкований набір слів і словосполучень, що описують певний предмет, повністю розкриваючи його характеристики. Якщо розглянути термін з боку WEB-програмування, то це

слова, що відносяться до діяльності сайту чи діяльності компанії, що володіє сайтом. Коректно складене семантичне ядро має важливе значення для пошукової оптимізації, саме на його основі будується пошуковий механізм, без чого не обходиться проєктування сайту чи іншого WEB-застосування [14].

Із вищенаведеного матеріалу, можна із впевненістю сказати, що семантичний аналіз тексту та дослідження цифрових матеріалів в цілому – важлива та необхідна царина для досліджень та проєктів в галузі інформаційних технологій.

1.2 Аналіз семантичних властивостей тексту

Для того, щоб в повному обсязі виконати завдання кваліфікаційної роботи та створити систему формування семантичного ядра цифрових текстів, необхідно дослідити множину характеристик, які можуть бути застосовані при аналізі інформації. Також слід визначити трактування основній термінології, що зустрічається при дослідженні семантичних властивостей тексту.

Сучасні методи досліджень цифрових текстів передбачають аналіз та вивід інформації у різний спосіб. Для вирішення певних задач, наприклад, зручно відображати інформацію у вигляді «хмари слів»; кривою, що побудована за законом Ципфа чи просто у вигляді таблиці статистичних даних. Тому необхідно дослідити методи досліджень текстів та доступні способи відображення результатів.

Хмара тексту (хмара слів, хмара тегів) – це візуалізація частоти слів у тексті у вигляді зваженого списку. Що частіше слово повторюється в тексті, то більшим є розмір його відображення у «хмарі» (рисунок 1.1).

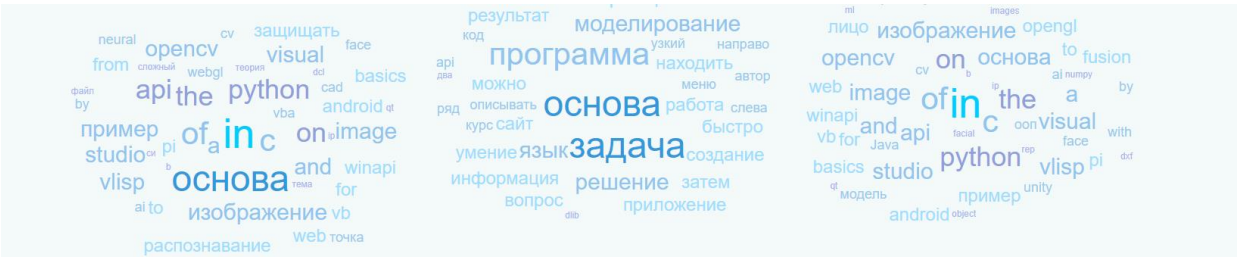


Рисунок 1.1 – Приклад подання інформації у вигляді хмари слів [7]

Іще одним методом представленням інформації є закон Ципфа. Закон полягає в тому, що при впорядкуванні всіх слів досліджуваного корпусу по спаданню частоти їх використання, то частота слова з порядковим номером *n* виявиться приблизно обернено пропорційною його порядковому номеру, тобто рангу слова [15]. В лінгвістиці закон використовують для опису різних даних: від доходу населення до розподілу росту міст. Закон користується попитом і під час семантичного дослідження тексту: часто його використовують задля того, щоб перевірити природність та оригінальність тексту. Вважається, що під час написання тексту людиною частота повтору слів відповідає цьому закону. Приклад подання результатів аналізу тексту за законом Ципфа представлено на рисунку 1.2.

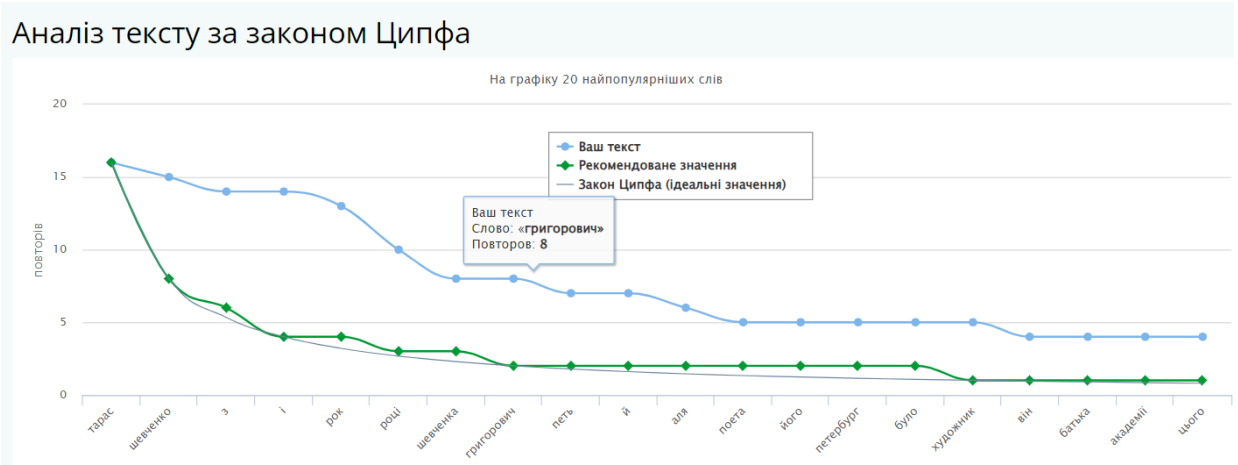


Рисунок 1.2 – Результати аналізу тексту за законом Ципфа

Поширеним явищем серед цифрових текстів стає переспам (нудота тексту). Звісно, граматично він не псує речення, однак його існування не лише псує читабельність тексту, а й псує SEO-релевантність сайтів, якщо тексти створюються для веб-ресурсів. Пошукові системи є посередниками між користувачем мережі Інтернет та контентом, який має задовольняти потреби людини. Саме для цього ПС використовують сучасні алгоритми, що дозволяють відібрати найбільш якісний та інформативний контент. Заспамлення тексту, штучне заповнення ключовими термінами, вважається недопустимим методом при створенні семантичного ядра веб-ресурсу.

У вищеописаному розділі розглянули основні характеристики тексту під час його семантичного дослідження. Вивчені явища допоможуть краще зрозуміти проблему поставленої задачі та якісно реалізувати функції аналізу тексту у власній інформаційній системі.

1.3 Аналіз існуючого програмного забезпечення предметної області

Сьогодні можна знайти чимало програмного забезпечення для семантичного аналізу тексту: від систем антиплагіату до онлайн-застосувань для оцінки тексту (ключові слова, частотність слів, словник ядра, відсоток води та перевірка нудоти статті). В цьому розділі розглянемо одні із найпотужніших та перспективних розробок.

Онлайн-сервіс «ISTIO» [16] розроблений для семантичного аналізу тексту та оцінює його насиченість ключовими словами, водянистість, заспамленість та багато інших параметрів.

Семантичний аналіз тексту від «ISTIO» оцінює рівень його насиченість ключовими словами, водянистість, заспамленності. Пошукові системи визначають якість і релевантність текстового контенту за словами і словосполученнями, з яких він складається.

Якщо в тексті досить тематичних ключових фраз, то пошукові системи оцінять його позитивно. Статті, в яких переважає вода і мало ключових слів, не потрапляють на перші сторінки видачі. Контент, перенасичений ключовими словами, відноситься до переспаму, його пошукові системи показують рідко (рисунок 1.3).

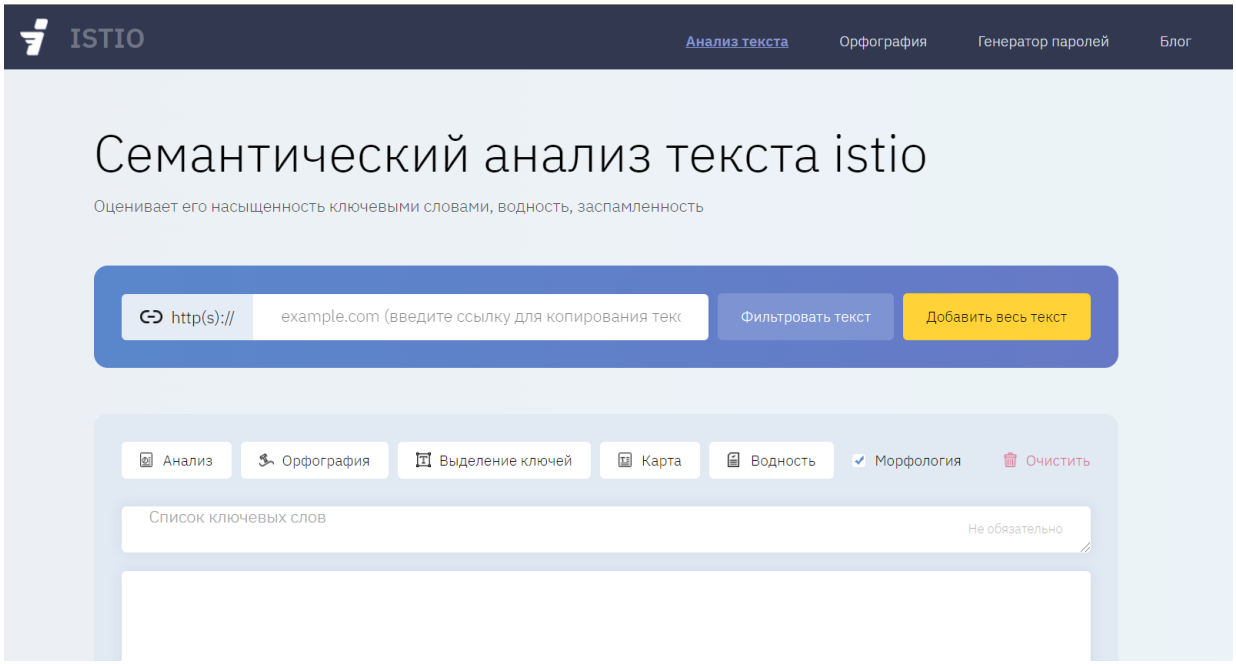


Рисунок 1.3 – Сервіс для семантичного аналізу тексту «ISTIO» [16]

SEO-аналіз тексту дозволяє оцінити статтю за показниками відсотка ключових слів, кількості стоп-слів. Сервіс показує:

– щільність ключових слів, їх процентне співвідношення в ядрі і в тексті;

– обсяг статті: кількість слів і символів (з пробілами і без);

– словник: загальна кількість одиниць, словник ядра;

– частотність слів, виводить топ-10 найбільш уживаних;

– мову статті і приблизну тематику;

– відсоток води.

Семантичний аналіз дозволяє автоматично порахувати кількість символів, оцінити нудоту і водянистість. Для зручності користувачів сервіс підсвічує ключові і воду, створює наочну карту частотних слів. Параметр нудоти тексту показує рівень його «забрудненості» ключовими словами. Чим їх більше, тим вище показник. Пошукові системи розцінюють таку статтю як неякісну, в видачу вона не потрапляє (рисунок 1.4).

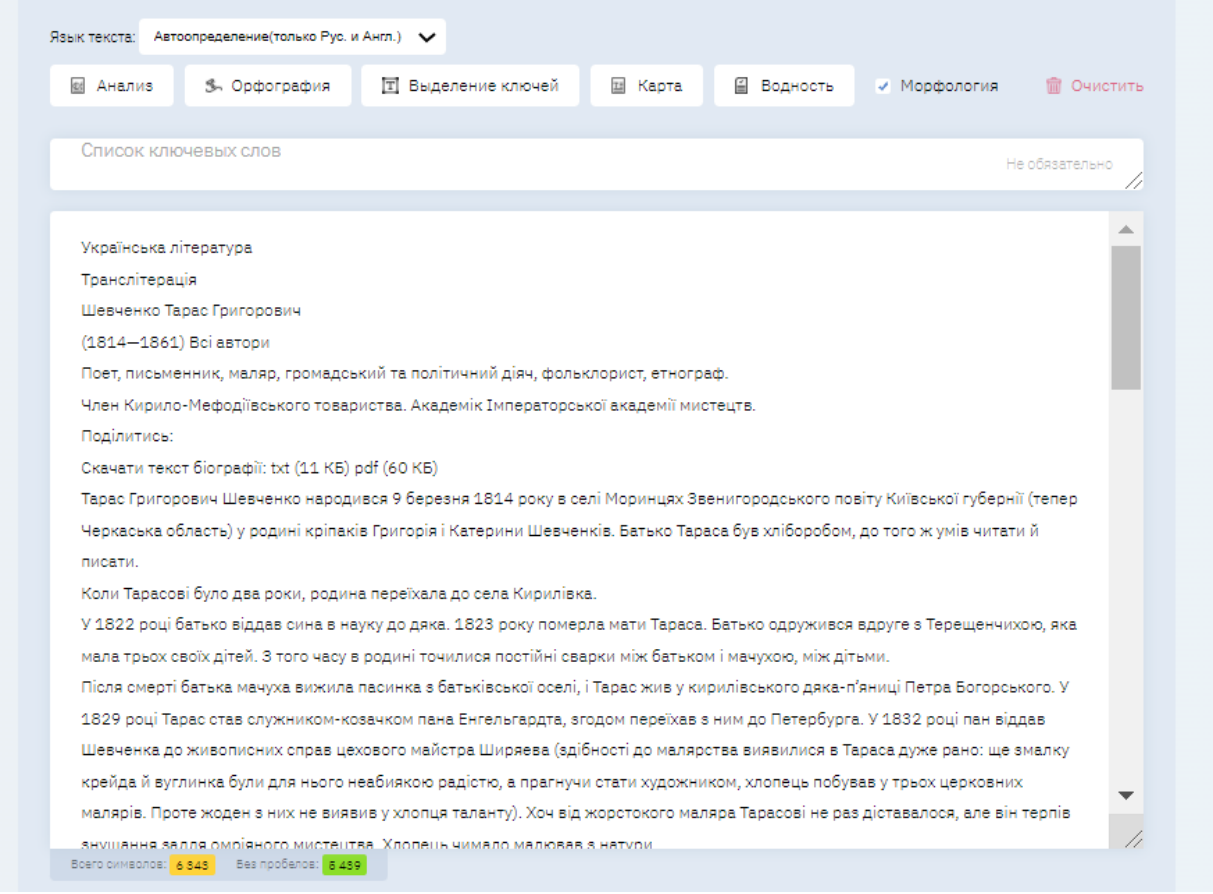


Рисунок 1.4 – Приклад заповнення форми для подальшого аналізу тексту платформою «ISTIO» [16]

Аналіз «води» тексту відображає наявність в статті стоп-слів, фразеологізмів, сполучних одиниць, які не несуть смислового навантаження. Якщо їх видалити, контент не втратить сенсу і стане якіснішим. Сервіс виділяє кольором слова без смислового навантаження. Це одиниці з необ'єктивною оцінкою, що не несуть конкретної інформації, а також підсилювачі. Виділені фрази рекомендується видалити або замінити.

Для прикладу візьмемо декілька абзаців із автобіографії Тараса Григоровича Шевченка [11].

Після вставки тексту можемо переглянути таблицю із результатами аналізу. Бачимо статистику найбільш вживаних слів у тексті, їх частоту, кількість (рисунок 1.5).

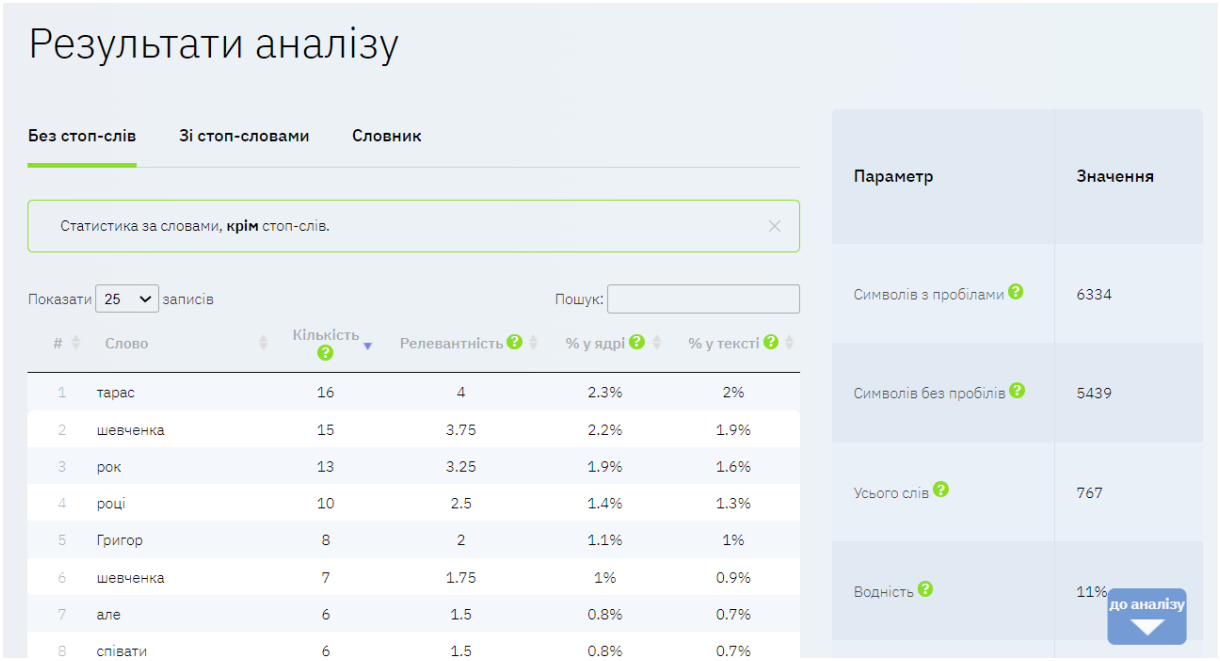


Рисунок 1.5 – Виведення результатів аналізу тексту платформою «ISTIO» на найбільш вживані слова, їх частоту та кількість [16]

Окрім цього, скористаємось іще деякими функціями цього сайту. Створимо карту виділених ключів тексту (рисунок 1.6).

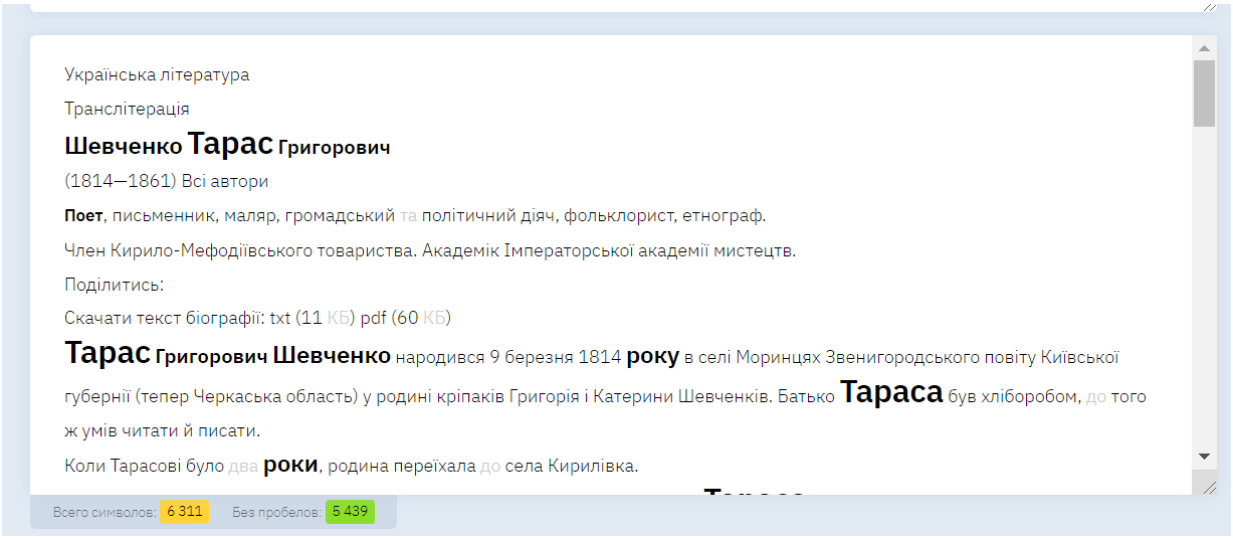


Рисунок 1.6 – Результат оформлення карти слів платформою «ISTIO» [16]

Переглянемо, як сайт виконує функцію виділення «води» у документах (рисунок 1.7)

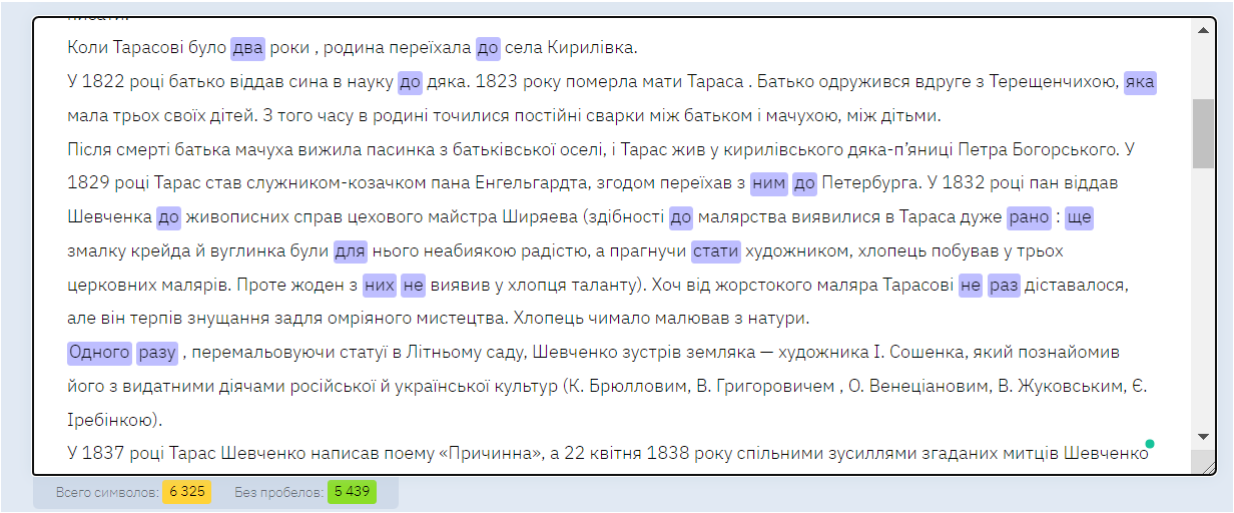


Рисунок 1.7 – Результат визначення «водянистості» тексту платформою «ISTIO» [16]

Окрім цього користувач може перевірити текст на «нудоту», тобто частоту повторення ключових слів у ньому. Для даного тексту обрали наступні ключові слова: Samsung, изогнутость, матрица. Перевіримо текст (рисунок 1.8).

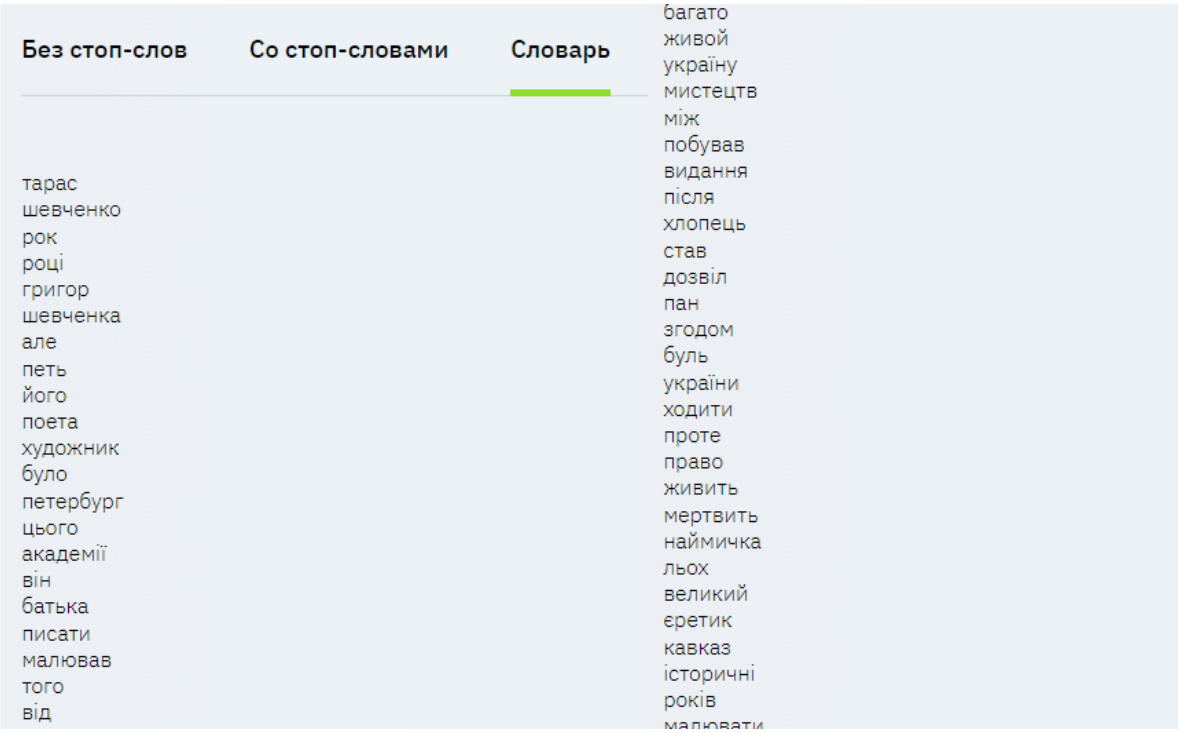


Рисунок 1.8 – Результат перевірки тексту платформою «ISTIO» на повторення ключових слів [16]

Окрім перевірки тексту, перевіримо і сайт. Оберемо той самий сайт, на котрому описується біографія Тараса Григоровича Шевченка [17]. Скористаємось сайтом «Miratext» [18]. Цей онлайн-сервіс пропонує проаналізувати текст на будь-якому сайті та вивести наступну інформацію:

– загальну статистику тексту;
– хмару слів: текст та посилання, зона тексту, зона посилань;
– найбільш вживані словосполучення;
– нудоту тексту;
– наповненість тексту водою;
– аналіз тексту по закону Ципфа;

На рисунку 1.9 представлено таблицю із загально статистикою.

Загальна статистика тексту

Параметр	Значення
Кількість символів із пробілами	6212
Кількість символів без пробілів	5287
Кількість слів	926
Кількість унікальних слів	534
"Тудота" тексту	4
"Водянистість" тексту	0%
Якість тексту згідно із законом Ципфа	36%

Рисунок 1.9 – Загальна статистика проаналізованого тексту платформою «Miratext» [18]

Також було сформовано хмари слів (рисунок 1.10), функція схожа до «Карти слів» попереднього онлайн-сервісу.

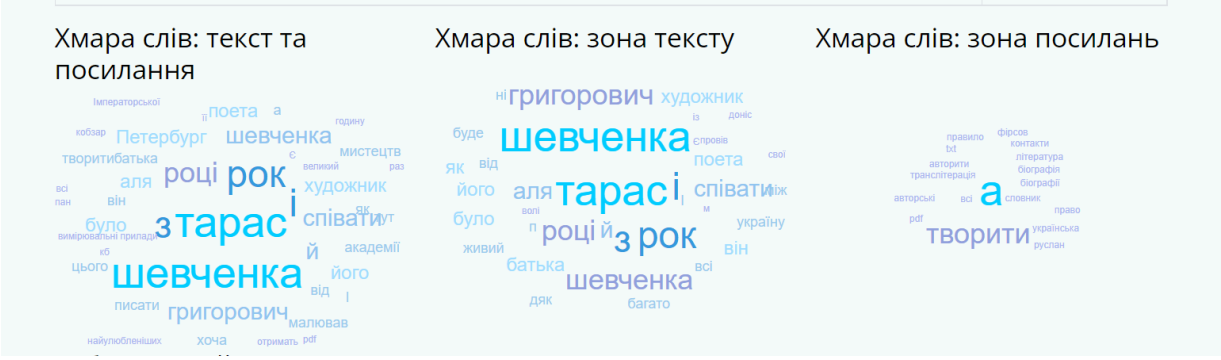


Рисунок 1.10 – Сформована платформою «Miratext» «хмара слів» на основі аналізу тексту, розміщеного на сайті [18]

На рисунку 1.11 представлено у вигляді таблиць перелік найбільш вживаних словосполучень з двох та трьох слів.

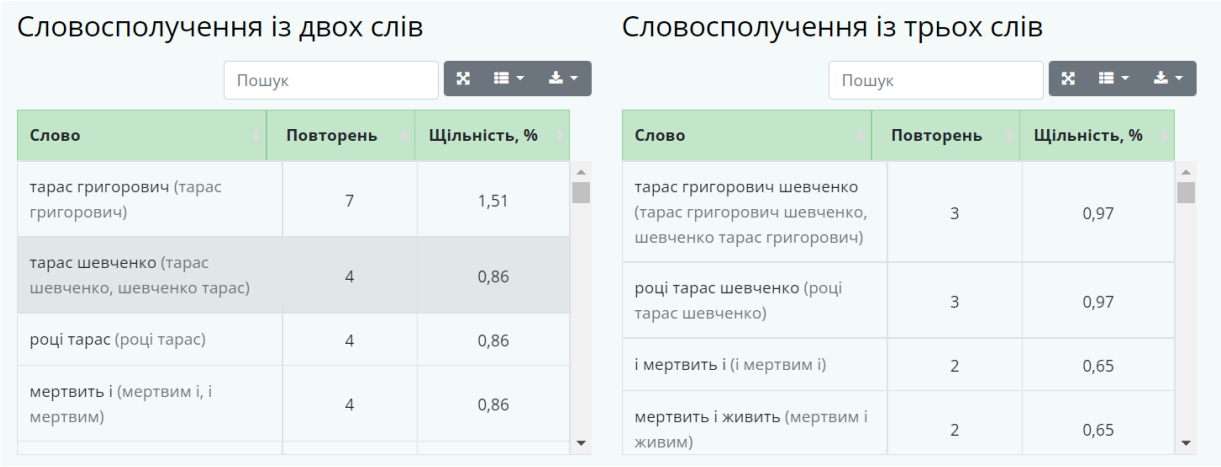


Рисунок 1.11 – Пошук найбільш вживаних словосполучень в тексті платформою «Miratext» [18]

Також важливо переглянути рівень нудоти та водянистості тексту, це представлено на рисунку 1.12.

Рисунок 1.12 – Перевірка тексту на нудоту та водянистість платформою «Miratext» [18]

І останньою функцією сайту є перегляд аналізу тексту за законом Ципфа (рисунок 1.13).

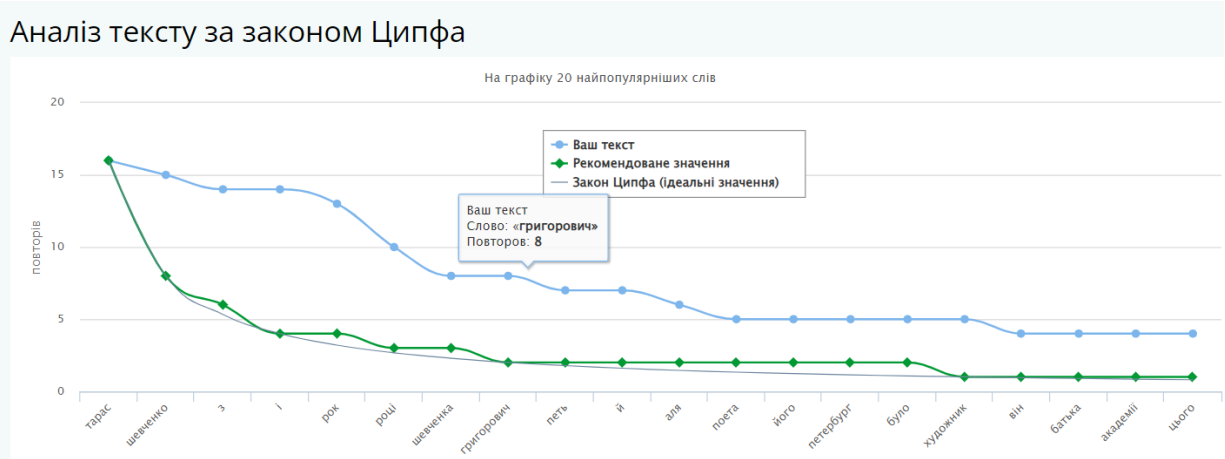


Рисунок 1.13 – Результат аналізу тексту платформою «Miratext» за законом Ципфа [18]

Таким чином, проаналізувавши сучасні сервіси для аналіз цифрової текстової інформації можна із впевненістю сказати, що семантичний аналіз тексту – потужна та корисна сфера інформаційних технологій. Саме тому необхідно досліджувати та розвивати нові потужні системи та методи обробки інформації.

Методи та системи семантичного аналізу можуть запропонувати значення слів та фраз, що можуть створити неточності як для розуміння читачем, так і для перекладу. Також це може допомогти у роботі різних компаній – автоматично вилучати інформацію, що може представляти цінність, наприклад із неструктурованих даних, таких як електронні листи, відгуки клієнтів чи заявки до технічної підтримки [3].

Окрім значного потенціалу для комерційних структур, робота таких інформаційних систем може допомогти в безлічі інших сфер діяльності людини – від аналізу наукових, технічних документів та літератури і до

впровадження методів семантичного аналізу текстів у WEB-застосуваннях та пошукових системах. Виходячи з наведеного, розробка методів і засобів формування семантичного ядра цифрових текстів у вигляді обмеженої множини ключових слів та словосполучень є актуальною задачею інформаційних технологій.

1.4 Постановка задачі

Мета кваліфікаційної роботи магістра полягає у розробці методу автоматизованого формування семантичного ядра цифрових текстів, що призначений для автоматизованого формування множини ключових семантичних одиниць за цифровим текстом та множинами слів і словосполучень цього тексту з показниками їх важливості, одержуючи сформовані семантичні ядра за різними показниками важливості, зокрема із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах. Для досягнення поставленої мети розробки методу автоматизованого формування семантичного ядра цифрових текстів необхідно розв’язати наступні *задачі дослідження*:

1. Провести аналіз предметної області семантичного аналізу цифрових текстів та відомих підходів до автоматизації формування семантичного ядра цифрових текстів.

2. Вдосконалити метод автоматизованого формування семантичного ядра цифрових текстів.

3. Розробити інформаційну технологію автоматизованого формування множини ключових семантичних одиниць.

4. Розробити інформаційну систему автоматизованого формування множини ключових семантичних одиниць.

5. Провести прикладне дослідження методу автоматизованого формування семантичного ядра цифрових текстів у складі інформаційної технології автоматизованого формування множини ключових семантичних одиниць і виконати аналіз результатів використання відповідної інформаційної системи.

Висновки до розділу 1

В розділі за результатом аналізу предметної області семантичного аналізу цифрових текстів й дослідження сучасного стану проблеми автоматизації формування семантичного ядра цифрових текстів встановлено, що однією із найбільших проблем на IT ринку, досі не вирішених, є відсутність або недосконаліть технологій і сервісів автоматичної смислової обробки неструктурованої текстової інформації. Ця проблема ускладнюється тим, що для автоматизованої змістовної обробки цифрових текстів необхідний комплекс методів, які зможуть забезпечити реалізацію алгоритмів й програмного забезпечення повного комп'ютерного лінгвістичного аналізу текстів природною мовою і автоматичної змістовної обробки інформації.

Для реалізації обчислювального процесу автоматичної смислової обробки інформації в комп'ютерних системах широкого застосування повинні бути розроблені способи організації високоефективного обчислювального процесу, що забезпечують формування результатів пошуку та аналітичної обробки інформації в реальному масштабі часу. Виходячи із наведеного, розробка методів й засобів формування

семантичного ядра цифрових текстів в вигляді обмеженої множини ключових слів та словосполучень є актуальною задачею інформаційних технологій.

Виходячи з наведеного, розробка методів і засобів формування семантичного ядра цифрових текстів у вигляді обмеженої множини ключових слів та словосполучень визначена актуальною задачею інформаційних технологій.

За результатом аналізу предметної області семантичного аналізу цифрових текстів, було визначено мету кваліфікаційної роботи магістра, яка полягає у розробці методу для автоматизованого формування множини ключових семантичних одиниць за цифровим текстом та множинами слів і словосполучень цього тексту з показниками їх важливості, одержуючи сформовані семантичні ядра за різними показниками важливості, зокрема із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Розділ 2

Метод і засоби автоматизованого формування семантичного ядра цифрових текстів

2.1 Аналіз підходу до обмеження за порогом щільності обсягу ключових семантичних одиниць

В ряді робіт [13, 17] пропонується використання дисперсійної оцінки для виокремлення ключових слів. Користуючись даною технологією, на основі введених даних у вигляді файлу автоматизовано формується структура цифрового документу для вибору елементу для аналізу, після чого проводиться сегментація по фразах і термінах, терміни лематизуються та їх множина компактифікується. На основі цього проводиться пошук та дисперсійне оцінювання важливості слів у вибраному фрагменті тексту, після чого оцінюється важливість термінів, а їх кількість обмежується відповідно до коефіцієнту щільності ключових слів. Вхідними даними інформаційної системи є цифровий документ, вихідними даними є відповідна множина ключових семантичних термінів.

Вхідними даними для оцінки ключових термінів є множина лемо-незалежних термінів M_{T1} із співставленою кожному з них кількістю зустрічей у досліджуваному тексті та впорядковану множину слів із співставленою кожному з них оцінкою його важливості (дисперсії) у досліджуваному тексті. Оцінка важливості V_n кожного терміна n із множини M_{T1} обчислюється за формулою:

$$v_n = \sum_{i=1}^{x_n} \frac{K_n \sigma_n}{k_n},$$

де K_n – кількість появ терміну n в множині M_{T1} ; k_n – кількість появ i-го слова терміну n в лематизованому текстовому контенті визначеного

фрагменту цифрового документу; σ_n – дисперсійна оцінка для *i*-го слова терміну n ; x_n – кількість слів у терміні n.

Відомий підхід до використання дисперсійної оцінки для виокремлення важливих слів у цифрових текстах проаналізовано у наступній статті [18].

В даній науковій роботі розглянуто наступний метод: елементи в множині M_T сортуються за значенням їх дисперсійної оцінки *DE*, після чого їх кількість обмежується відповідно до вхідного параметру граничної щільності ключових термінів у тексті *Q*. Щільність є відношенням кількості слів ключових термінів у цифровому матеріалі до загальної кількості слів [16] й становить 0,11-0,15. Відповідно, до множини ключових термінів M_{Term} автоматично будуть додані елементи з множини M_T з найбільшими значеннями *DE* доти, доки справджується рівність:

$$\sum_{i=1}^{n} \frac{K_n x_n}{X_{txt}} \le Q ,$$

де K_n – кількість появ терміну n в множині M_{TXT}; x_n – кількість слів у терміні n; X_{txt} – загальна кількість слів у тексті; n – поточна кількість термінів у множині M_{Term}.

Таким чином, проаналізувавши наукові публікації, можна підтвердити, що один із найпоширеніших та ефективних методів оцінки важливості ключових слів є метод дисперсійної оцінки, що дозволить із найбільшою вірогідністю знайти ключові слова в цифрових текстах. Проте в ряді випадків є ефективними і інші методи пошуку ключових семантичних одиниць, якими можуть вступати ключові слова, терміни та абревіатури. Проте складає проблему визначення порогу, який обмежуватиме обсяг семантичних одиниць.

За результатами аналізу існуючих джерел, можна прогнозувати велику ефективність методу обмеження обсягу семантичних одиниць за порогом їх щільності, проте для використання цього підходу потрібні прикладні дослідження, які стосуються як визначення кількісного порогу щільності семантичних одиниць, так і параметрів обрахунку щільності семантичних одиниць цим методом.

2.2 Схема методу автоматизованого формування семантичного ядра цифрових текстів

Метод автоматизованого формування семантичного ядра цифрових текстів призначений для формування множини ключових слів та словосполучень тексту згідно обраного цільового відсотку щільності ключових слів. Метод має дозволяти за множиною слів і словосполучень цифрового тексту з зіставленими значеннями показників їх семантичної важливості, а також необхідним відсотком щільності ключових слів у тесті автоматизовано формувати відповідні множини ключових слів та словосполучень тексту за різними показниками важливості.

На Рисунку 2.1 зображено схему кроків методу автоматизованого формування семантичного ядра цифрових текстів. Вхідними даними методу автоматизованого формування семантичного ядра цифрових текстів є текст у вигляді послідовної множини слів, множина слів та словосполучень тексту, показники важливості слів та словосполучень тексту, обраний для використання показник важливості та обраний цільовий відсоток щільності ключових слів у тесті.

На кроці 1 виконується обрахунок числа поява кожного унікального слова та словосполучення у тексті, після чого на кроці 2 проводиться

послідовний обрахунок порогового відсотку щільності для кожного унікального слова та словосполучення у тексті. За результатом цих дій, на кроці 3 виконується послідовне додавання до множини ключових слів та словосполучень, які мають пороговий відсоток щільності вищий за обраний цільовий відсоток для тексту. Формування результуючої множини ключових слів та словосполучень тексту на кроці 4 відбувається за результатом виконання кроку 3 та відповідає за формування вихідних даних.

Рисунок 2.1 – Схема методу автоматизованого формування семантичного ядра цифрових текстів

Вхідними даними методу автоматизованого формування семантичного ядра цифрових текстів є множина ключових слів та

словосполучень тексту згідно обраного цільового відсотку щільності ключових слів.

Вдосконалено метод автоматизованого формування семантичного ядра цифрових текстів, що дозволяє за множиною слів і словосполучень цифрового тексту з зіставленими значеннями показників їх семантичної важливості, а також необхідним відсотком щільності ключових слів у тесті автоматизовано формувати відповідні множини ключових слів та словосполучень тексту за різними показниками важливості.

Таким чином, метод автоматизованого формування семантичного ядра цифрових текстів призначений для формування множини ключових слів та словосполучень тексту згідно обраного цільового відсотку щільності ключових слів. Розроблений метод використовується як ключовий етап в межах інформаційної технології формування множини ключових семантичних одиниць й може бути застосований для інших прикладних задач.

2.3 Інформаційна технологія автоматизованого формування множини ключових семантичних одиниць

Інформаційна технологія автоматизованого формування множини ключових семантичних одиниць використовує розроблений метод автоматизованого формування семантичного ядра цифрових текстів й у якості вхідних даних має цифровий текст, множину слів тексту та показники їх важливості, а також множину словосполучень тексту та показники їх важливості. Відповідно, інформаційна технологія автоматизованого формування множини ключових семантичних одиниць має дозволяти з використанням створеного методу автоматизованого

формування семантичного ядра цифрових текстів за вхідними даними у вигляді цифрового тексту та множина слів і словосполучень цього тексту з показниками їх важливості одержувати вихідні дані у вигляді сформованих семантичних ядер із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах. На Рисунку 2.2 зображено схему етапів інформаційної технології автоматизованого формування множини ключових семантичних одиниць.

На Етапі 1 виконання інформаційної технології автоматизованого формування множини ключових семантичних одиниць виконується поелементна обробка тексту. Зокрема, проводиться обрахунок загальних параметрів тексту, таких як кількості слів, словосполучень і знаків. А після цього виконується очищення тексту від додаткових символів (знаків, цифр). Далі відбувається зменшення регістру тексту, за результатами чого виконується формування текстового вектору слів та текстового вектору словосполучень.

Етап 2 відповідає за пошук появ семантичних одиниць та перевірку текстового вектору. Спершу проводиться обрахунок позиції по словах для кожної появи кожного унікального слова, а також обрахунок позиції по словах для кожної появи кожного унікального словосполучення. Одночасно проводиться обрахунок позиції по символах для кожної появи кожного унікального слова і обрахунок позиції по символах для кожної появи кожного унікального словосполучення. Після цього виконується формування перевірочного тексту з текстового вектору слів і перевірочного тексту з текстового вектору словосполучень. За результатом, здійснюється

обрахунок кількості появ кожного унікального слова та кількості появ кожного унікального словосполучення.

Рисунок 3.2 – Схема інформаційної технології автоматизованого формування множини ключових семантичних одиниць

На Етапі 3 проводиться підготовка до застосування методу формування семантичного ядра. Для цього спершу виконується одержання з бази даних значень важливості унікальних слів тексту TF, TFIDF, DE. Також виконується одержання з БД значень важливості унікальних словосполучень тексту TF, TFIDF, DE. Після візуалізації цих даних, здійснюється сортування окремих переліків слів і словосполучень тексту за показниками важливості TF, TFIDF, DE. Останнім кроком виконується одержання від користувача цільового відсотку щільності для тексту.

Етап 4 безпосередньо відповідає за автоматизоване формування семантичного ядра цифрових текстів методом автоматизованого формування семантичного ядра цифрових текстів. Для цього незалежним чином виконується одержання семантичного ядра слів при обрахунку порогу щільності у символах, семантичного ядра словосполучень при обрахунку порогу щільності у символах, семантичного ядра слів при обрахунку порогу щільності у словах та семантичного ядра словосполучень при обрахунку порогу щільності у словах.

Відповідно, вихідні дані формуються як семантичне ядро тексту з таких складових: семантичне ядро тексту із слів при обрахунку порогу щільності у символах, семантичне ядро тексту із словосполучень при обрахунку порогу щільності у символах, семантичне ядро тексту із слів при обрахунку порогу щільності у словах, семантичне ядро тексту із словосполучень при обрахунку порогу щільності у словах.

Таким чином, інформаційна технологія автоматизованого формування множини ключових семантичних одиниць використовує розроблений метод автоматизованого формування семантичного ядра цифрових текстів й дозволяє перетворювати вхідні дані у вигляді цифрового

тексту, множини слів і словосполучень тексту з показниками їх семантичної важливості в вихідні дані у вигляді зразків семантичного ядра тексту в варіаціях із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Висновки до розділу 2

В розділі було розглянуто підхід до обмеження за порогом щільності обсягу ключових семантичних одиниць За результатами аналізу існуючих джерел встановлено, що можна прогнозувати велику ефективність методу обмеження обсягу семантичних одиниць за порогом їх щільності, проте для використання цього підходу потрібні прикладні дослідження, які стосуються як визначення кількісного порогу щільності семантичних одиниць, так і параметрів обрахунку щільності семантичних одиниць цим методом..

В розділі розроблено метод автоматизованого формування семантичного ядра цифрових текстів, який дозволяє за множиною слів й словосполучень цифрового тексту із зіставленими значеннями показників їх семантичної важливості, а також необхідним відсотком щільності ключових слів в тесті автоматизовано формувати відповідні множини ключових слів та словосполучень тексту за різними показниками важливості.

Також розроблено нову інформаційну технологію автоматизованого формування множини ключових семантичних одиниць, яка дозволяє з використанням створеного методу автоматизованого формування семантичного ядра цифрових текстів за вхідними даними у вигляді

цифрового тексту та множина слів й словосполучень цього тексту з показниками їх важливості одержувати вихідні дані у вигляді сформованих семантичних ядер із слів при обрахунку порогу щільності в символах, із словосполучень при обрахунку порогу щільності в символах, із слів при обрахунку порогу щільності в словах та із словосполучень при обрахунку порогу щільності в словах.

Розділ 3

Інформаційна система автоматизованого формування множини ключових семантичних одиниць

3.1 Структура та функціональне призначення складових системи

Інформаційна система автоматизованого формування множини ключових семантичних одиниць використовує створений метод автоматизованого формування семантичного ядра цифрових текстів й забезпечує можливість одержувати дані у вигляді сформованих семантичних ядер із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

На основі створеної структури інформаційної системи автоматизованого формування множини ключових семантичних одиниць відповідно до інформаційної технології було створено діаграму класів, на яку будемо опиратись при створенні програмного забезпечення. Розроблена діаграма класів зображена на рисунку 3.1.

Основним класом, що відповідає за обрахунок та визначення ключових слів є клас «DispersionEvolationOfSection». Його робота базується на результатах роботи класу «FindDispersion». Попередній клас групує його результати, адже «FindDispersion» проводить лише оцінювання термінів, чим формує перелік з результатами незалежно від кількості обрахованих слів в цифровому тексті.

Клас «OptionTerm» проводить процес обрахунку для збереження проміжних даних про терміни-кандидати, а саме значення терміну в тексті та його позицію в ньому.

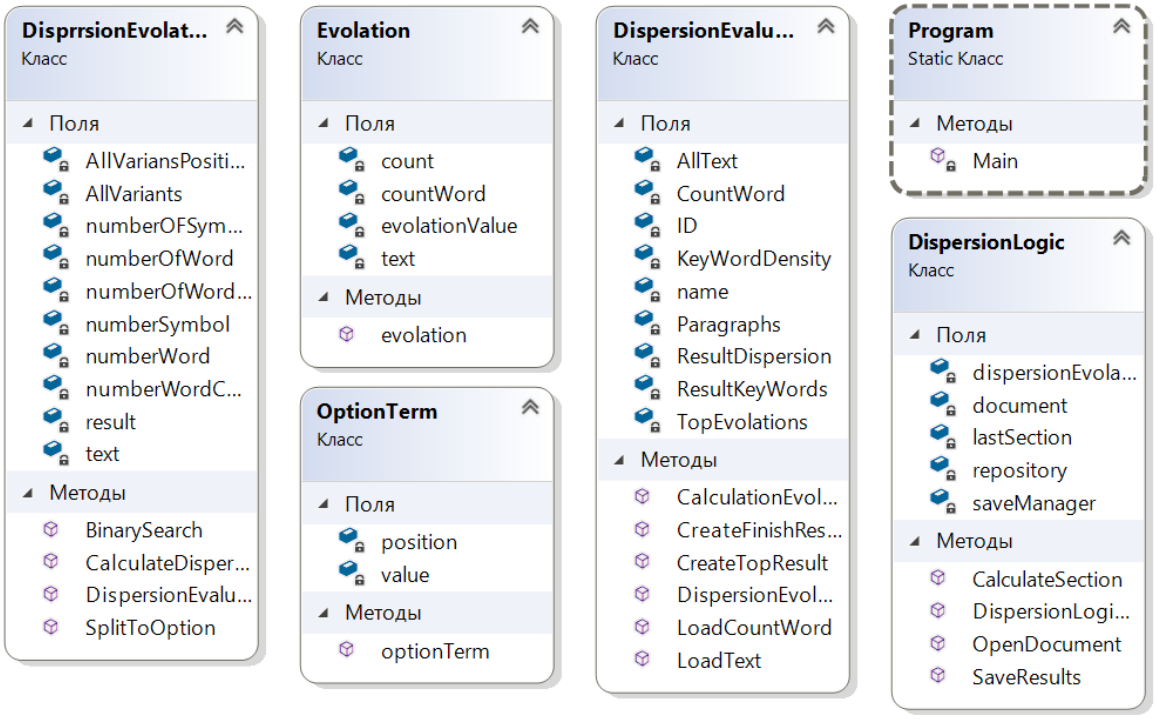


Рисунок 3.1 – Діаграма класів інформаційної системи автоматизованого формування множини ключових семантичних одиниць

Клас «Evolation» створених для збереження інформації про ключові терміни, а також обраховує кількість появ слів у тексті, загальну кількість слів та його символьне значення.

Клас «DispersionEvolation» призначений для пошуку та оцінки ключових словосполучень. Метод «CreateFinishResult» реалізовано задля обрахунку кінцевого результату із врахуванням щільності ключових термінів, ку користувач обирає самостійно.

Інші методи є другорядними та переважно призначені для опрацювання документів, забезпечуючи взаємозв’язок програми та документів .doc розширення.

Отож головним класом є «DispersionEvolationOfSection», саме на його основі проводяться усі необхідні для аналізу обчислення. Метод

«CreateTopResult» формує перелік з найвищими результатами незалежно від загальної кількості слів у тексті, чим формує базу параметрів для методу у попередньому головному класі.

Реалізовані класи «DispersionEvolationOfSection», «DispersionEvolation», «OptionTerm», «FindDispersion» та «DispersionLogic» повністю забезпечують правильність роботи інформаційної системи автоматизованого формування множини ключових семантичних одиниць, зокрема завантаження цифрових текстів, перегляд основних властивостей цих матеріалів, переглянути появу семантичних одиниць по словах та по словосполученнях, а в результаті – сформувати семантичне ядро тексту за TF, TFIDF та DE.

Отже, було описано структуру та функціональне призначення складових інформаційної системи автоматизованого формування множини ключових семантичних одиниць, яка використовує створений метод автоматизованого формування семантичного ядра цифрових текстів й забезпечує можливість за вхідними даними у вигляді цифрового тексту та множина слів і словосполучень цього тексту з показниками їх важливості одержувати вихідні дані у вигляді сформованих семантичних ядер із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

3.2 Розробка структури бази даних інформаційної системи

Жодний сучасний IT-проєкт не може повноцінно функціонувати без бази даних, тому її розробка – важливий крок при створенні інформаційної

технології формування семантичного ядра цифрових текстів. База даних – це організована структура, яка призначена для зберігання, зміни та обробки взаємозалежної інформації, переважно великих обсягів [21]. Для розуміння структури БД було створено діаграму даних, що відображає список таблиць, їх полів та зв'язків між ними.

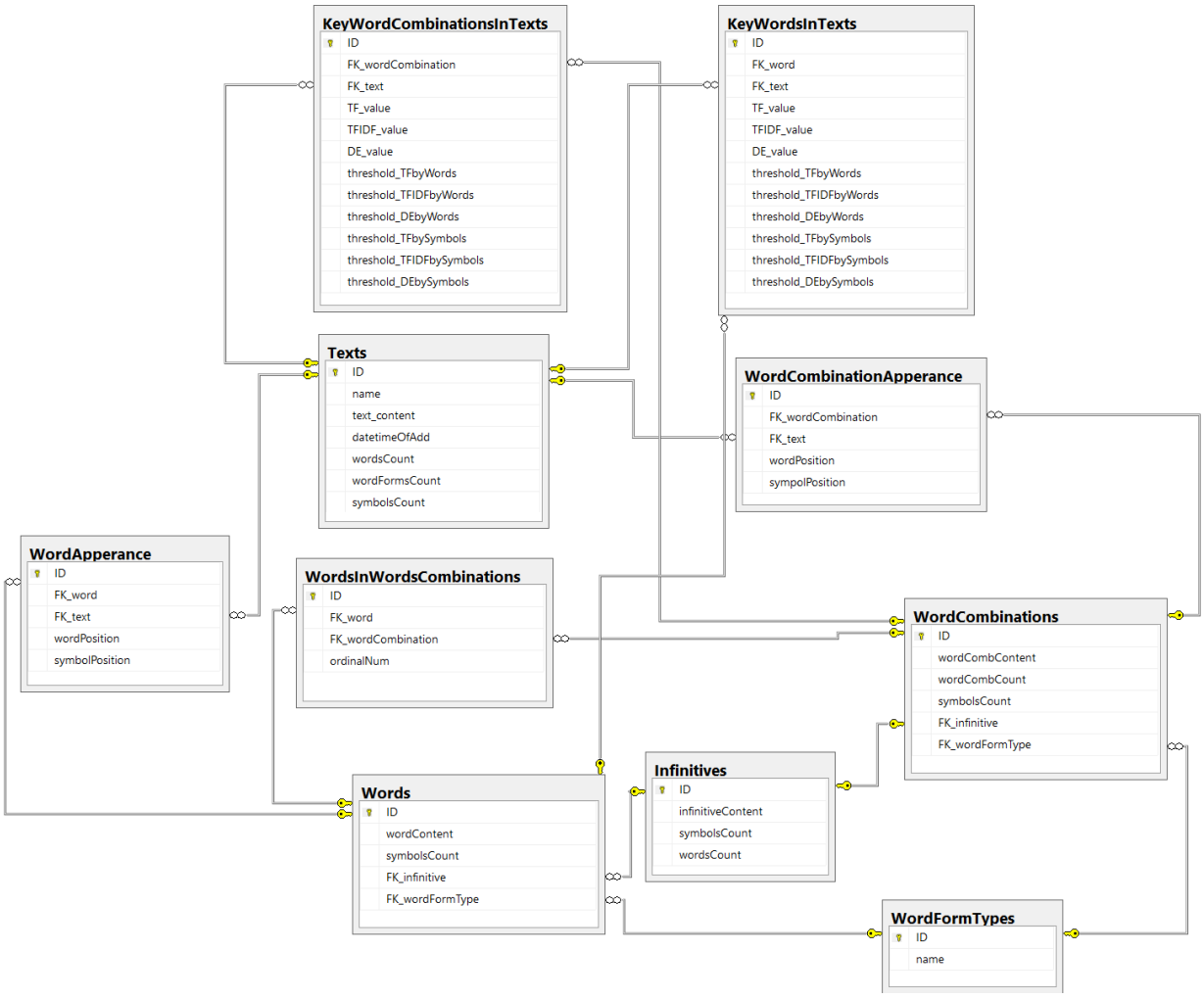


Рисунок 3.2 – Структура бази даних інформаційної системи автоматизованого формування множини ключових семантичних одиниць

Відповідно до поставленого завдання, а саме створення інформаційної системи автоматизованого формування множини ключових семантичних одиниць, було спроектовано структуру бази даних, що

дозволяє зберігати матеріал для аналізу, а саме тексти та їх компоненти, та супроводжувати ефективний обрахунок результатів аналізу тексту. Структура розробленої БД представлено на рисунку 3.2.

Опираючись на створену структуру БД створили відповідні таблиці та заповнили їх початковим даними. Таблиця «Texts» (таблиця 3.1) призначена для збереження текстів для наступного аналізу та основних параметрів цих матеріалів.

Таблиця 3.1 – Атрибути таблиці «Texts»

№П/П	Назва атрибуту	Тип даних	Опис
1	ID	int	Числовий ідентифікатор записів, первинний ключ
2	name	nvarchar(50)	Назва текстового матеріалу, що буде проаналізовано
3	text_content	nvarchar(MAX)	Текст матеріалу
4	datetimeOfAdd	smalldatetime	Дата й час додання тексту до БД
5	wordsCount	int	Кількість слів у текстовому матеріалі
6	wordFormsCount	int	Кількість словоформ у текстовому матеріалі
7	symbolsCount	int	Кількість символів у текстовому матеріалі

Таблиця «Words» створена для збереження інформації про слово, як семантичну одиницю, тобто інфінітив слова, словоформу та кількість символів, що містить обране слово.

Таблиця 3.2 – Атрибути таблиці «Words»

№П/П	Назва стовпця	Тип даних	Призначення
1	ID	int	Числовий ідентифікатор записів, первинний ключ
2	wordContent	varchar(50)	Текст слова
3	symbolsCount	int	Кількість символів у слові
4	FK_infinitive	int	Вторинний ключ, посилання на запис таблиці «Infinitives» для співставлення з відповідним інфінітивом слова.
5	FK_wordFormType	int	Вторинний ключ, посилання на запис таблиці «WordFormsTypes» для співставлення з відповідною словоформою

Таблиця «WordFormTypes» створена для збереження інформації про типи словоформ, а саме їх назви.

Таблиця 3.3 – Атрибути таблиці «WordFormTypes»

№П/П	Назва стовпця	Тип даних	Призначення
1	ID	int	Числовий ідентифікатор записів, первинний ключ
2	name	varchar(50)	Назва словоформи

Таблиця «WordCombinations» (таблиця 3.4) призначена для збереження інформації про словосполучення у тексті, а саме зміст словосполучення, їх кількість, кількість символів у ньому, інфінітив словосполучення, тим словоформи.

Таблиця 3.4 – Атрибути таблиці «WordCombinations»

№П/П	Назва стовпця	Тип даних	Призначення
1	ID	int	Числовий ідентифікатор записів, первинний ключ
2	wordCombContent	varchar(50)	Текстовий зміст словосполучення, повний текст
3	wordCombCount	int	Кількість повторів словосполучення в текстовому матеріалі
4	symbolsCount	int	Кількість символів у словосполученні
5	FK_infinitive	int	Вторинний ключ, посилання на запис таблиці «Infinitives» для співставлення з відповідним інфінітивом слова.
6	FK_wordFormType	int	Вторинний ключ, посилання на запис таблиці «WordFormsTypes» для співставлення з відповідною словоформою

Таблиця «Infinitives» (таблиця 3.5) призначена для збереження інформації про інфінітиви слів. Містить їх повний текст, кількість символів та кількість повторів у тексті.

Таблиця 3.5 – Атрибути таблиці «Infinitives»

№П/П	Назва стовпця	Тип даних	Призначення
1	ID	int	Числовий ідентифікатор записів, первинний ключ
2	infinitiveContent	varchar(50)	Текстовий зміст інфінітиву
3	symbolsCount	int	Кількість символів у слові
4	wordsCount	int	Кількість повторів слова у текстовому матеріалі

Таблиця «WordsInWordsCombinations» (таблиця 3.6) створена для збереження інформації щодо слів в словосполученнях. Містить дані про слова з яких утворюється словосполучення та порядковий номер словосполучення у тексті.

Таблиця 3.6 – Атрибути таблиці «WordsInWordsCombinations»

№П/П	Назва стовпця	Тип даних	Призначення
1	ID	int	Числовий ідентифікатор записів, первинний ключ
2	FK_word	int	Вторинний ключ, посилання на запис таблиці «Words» для співставлення з відповідним словом
3	FK_wordCombination	int	Вторинний ключ, посилання на запис таблиці «WordCombinations» для співставлення з відповідним словосполученням
4	ordinalNum	int	Порядковий номер словосполучення у тексті

Таблиця «WordCombinationApperance» (таблиця 3.7) містить інформацію щодо появи словосполучень в тексті: зміст словосполучення, номер позиції по словах в тексті та позицію по символах.

Таблиця 3.7 – Атрибути таблиці «WordCombinationApperance»

№П/П	Назва стовпця	Тип даних	Призначення
1	ID	int	Числовий ідентифікатор записів, первинний ключ
2	FK_wordCombination	int	Вторинний ключ, посилання на запис таблиці «WordCombinations» для співставлення з відповідним словосполученням
3	FK_text	int	Вторинний ключ, посилання на запис таблиці «Texts» для співставлення з відповідною назвою тексту
4	wordPosition	int	Позиція словосполучення по словах у тексті
5	sympolPosition	int	Позиція словосполучення по символах у тексті

Таблиця «KeyWordsInTexts» (таблиця 3.8) призначена для збереження інформації щодо ключових слів та значень основних математичних обчислень, здійснених при аналізі тексту.

Таблиця «KeyWordCombinationsInTexts» (таблиця 3.9) призначена для збереження інформації щодо ключових словосполучень та значень основних математичних обчислень, здійснених при аналізі тексту.

Таблиця 3.8 – Атрибути таблиці «KeyWordsInTexts»

№П/П	Назва стовпця	Тип даних	Призначення
1	ID	int	Числовий ідентифікатор записів, первинний ключ
2	FK_word	int	Вторинний ключ, посилання на запис таблиці «Words» для співставлення з відповідним словом
3	FK_text	int	Вторинний ключ, посилання на запис таблиці «Texts» для співставлення з відповідною назвою тексту
4	TF_value	float	Обраховане значення term frequency
5	TFIDF_value	float	Обраховане значення term frequency-inverse document frequency
6	DE_value	float	Обраховане значення dispersion evaluation
7	threshold_TFbyWords	float	Поріг входження слова за значенню term frequency по словах
8	threshold_TFIDFbyWords	float	Поріг входження слова за term frequency-inverse document frequency по словах
9	threshold_DEbyWords	float	Поріг входження слова за dispersion evaluation по словах
10	threshold_TFbySymbols	float	Поріг входження слова за значенню term frequency по символах
11	threshold_TFIDFbySymbols	float	Поріг входження слова за term frequency-inverse document frequency по символах
12	threshold_DEbySymbols	float	Поріг входження слова за dispersion evaluation по символах

Таблиця «WordApperance» (таблиця 3.10) призначена для збереження інформації щодо появи слів у тексті, містить поля щодо назви тексту, текстового змісту слова, позицію слова в тексті по словах та по символах.

Таблиця 3.9 – Атрибути таблиці «KeyWordCombinationsInTexts»

№П/П	Назва стовпця	Тип даних	Призначення
1	ID	int	Числовий ідентифікатор записів, первинний ключ
2	FK_wordCombination	int	Вторинний ключ, посилання на запис таблиці «WordCombinations» для співставлення з відповідним словосполученням
3	FK_text	int	Вторинний ключ, посилання на запис таблиці «Texts» для співставлення з відповідною назвою тексту
4	TF_value	float	Обраховане значення term frequency
5	TFIDF_value	float	Обраховане значення term frequency-inverse document frequency
6	DE_value	float	Обраховане значення dispersion evaluation
7	threshold_TFbyWords	float	Поріг входження слова за значенню term frequency по словах
8	threshold_TFIDFbyWords	float	Поріг входження слова за term frequency-inverse document frequency по словах
9	threshold_DEbyWords	float	Поріг входження слова за dispersion evaluation по словах
10	threshold_TFbySymbols	float	Поріг входження слова за значенню term frequency по символах
11	threshold_TFIDFbySymbols	float	Поріг входження слова за term frequency-inverse document frequency по символах
12	threshold_DEbySymbols	float	Поріг входження слова за dispersion evaluation по символах

В результаті виконання розділу було створено структуру бази даних інформаційної системи автоматизованого формування множини ключових семантичних одиниць та відповідні таблиці. Створена БД дозволяє зберігати усю необхідну інформацію щодо текстів, їх змісту та результатів аналізу, потрібну для створення інформаційної системи автоматизованого формування множини ключових семантичних одиниць.

Таблиця 3.10 – Атрибути таблиці «WordApperance»

№П/П	**Назва стовпця**	**Тип даних**	**Призначення**
1	ID	int	Числовий ідентифікатор записів, первинний ключ
2	FK_word	int	Вторинний ключ, посилання на запис таблиці «Words» для співставлення з відповідним словом
3	FK_text	int	Вторинний ключ, посилання на запис таблиці «Texts» для співставлення з відповідною назвою тексту
4	wordPosition	int	Позиція слова по словах у тексті
5	sympolPosition	int	Позиція слова по символах у тексті

Подана структура бази даних дозволяє проводити маніпуляції усієї необхідної інформації, потрібної для коректної роботи інформаційної системи автоматизованого формування множини ключових семантичних одиниць.

3.3 Аналіз рекомендованих засобів розробки інформаційної системи автоматизованого формування множини ключових семантичних одиниць

Інформаційна система автоматизованого формування множини ключових семантичних одиниць використовує створений метод автоматизованого формування семантичного ядра цифрових текстів й забезпечує можливість за вхідними даними у вигляді цифрового тексту та множина слів і словосполучень цього тексту з показниками їх важливості одержувати вихідні дані у вигляді сформованих семантичних ядер із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Сьогодні існує чимало середовищ розробки програмного забезпечення. Кожне з них пропонує ряд унікальних функцій та особливостей, зокрема підтримку різних мов програмування. Залежно від виду поставленої задачі обираємо відповідне середовище розробки, мову програмування та систему керування базами даних.

Для створення інформаційної системи формування семантичного ядра цифрових текстів використаємо платформу .NET, мову програмування C#, СКБД Microsoft SQL Server та середовище розробки Visual Studio 2017.

Microsoft .NET є платформою, що спрощує розробку додатків для розробників. Фреймворк підтримує розробку, а також підтримку та обслуговування сучасних додатків та WEB-служб XML. Також платформа пропонує розробникам середовище ООП та створення додатків, що можуть працювати на серверних платформах, таких як Windows, Linux та MAC [22].

C# - потужна об’єктно-орієнтована мова програмування, створена у 2000 році компанією Microsoft, вирізняється серед інших мов програмування своєю гнучкістю, адже використати цю мову можна при створенні додатків будь-якого типу призначення. Однією із найбільших переваг цієї мови програмування є її статична типізація, її легко читати та розуміти, що спрощує написання коду в цілому [23].

Перелік переваг мови програмування C#:[24]:

– підтримка ООП;

– статична типізація;

– мова підтримує поліморфізм;

– створені компоненти можуть використовуватись повторно;

Окрім цього необхідно обрати систему керування базами даних. Для справної роботи інформаційної системи необхідно забезпечити коректний зв'язок між базою та додатком. Тому, для надійної роботи інформаційнох системи проаналізуємо доступні середовища.

MySQL – одна із найпопулярніших систем керування базами даних SQL з відкритим вихідним кодом, розробляється, розповсюджується та підтримується корпорацією Oracle [25].

Програмне забезпечення MySQL – це система клієнтського серверу, що складається із багатопотокового SQL-серверу. Підтримує різні серверні частини, інструменти широкого інтерфейсу прикладного програмування.

Однак для виконання нашої задачі MySQL має багато недоліків, а саме:

– MySQL не так ефективно підтримує бази даних великого обсягу;

– не підтримує такі процедури як ROLE, COMMIT у версіях, що нижче 5.0;

– транзакції не виконується так швидко та ефективно, як у MS SQL Server;

– часті перебої роботи;

– значно гірше масштабування продуктивності, ніж у MS SQL Server.

Microsoft SQL Server – це система управління реляційними базами даних, розроблена компанією Microsoft. Він включає мову SQL і T-SQL, власну мову Microsoft з можливостями обробки винятків, оголошення змінних і збережених процедур. Механізм бази даних поділено на два сегменти: реляційний механізм, який використовується для обробки команд та запитів. Другий – механізм зберігання, призначений керувати різними функціями бази даних, як-от таблиці, сторінки, файли, індекси і транзакції.

Переваги використання Microsoft SQL Server [26]:

– відносно легко встановити на комп’ютер та розпочати роботу;

– висока швидкодія та продуктивність;

– високі показники безпеки;

– відмінний механізм відновлення даних.

Не менш важливим кроком є вибір середовища розробки для нашого програмного продукту. Розглянемо рейтинг топ-середовищ серед розробників ПЗ. Як продемонстровано на рисунку, бачимо, що Microsoft Visual Studio є провідною серед інших та має найбільше позитивних відгуків.

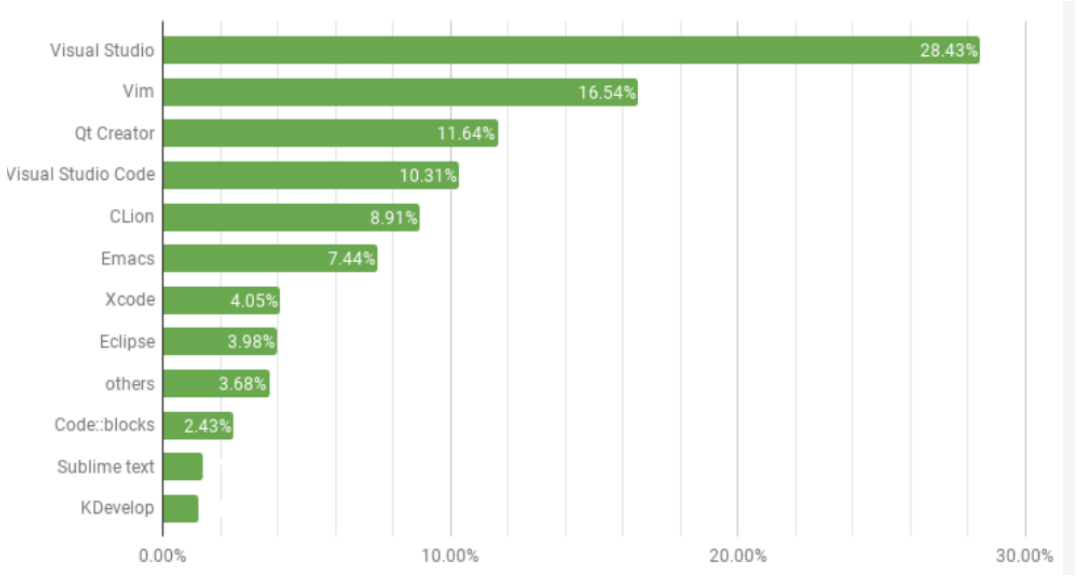


Рисунок 4.1 – Рейтинг найбільш популярних IDE станом на 2021 рік [27]

В якості середовища розробки було обрано Visual Studio. VS забезпечить коректну роботу інформаційної системи та бази даних. Зокрема, існує ряд переваг [28]:

– чітке структурування проектів та файлів;

– дозволяє створювати не лише консольні програми, а й програми з графічним інтерфейсом та WEB-додатки;

– можливість розробки для різних платформ.

Також можна розглянути середовище розробки Eclipse, адже воно є популярним серед C# розробників. Однак Eclipse має ряд недоліків:

Незважаючи на те, що Eclipse є якісним середовищем розробки для програмування C# та Java він використовує значно більше системних та процесорних ресурсів, ніж Visual Studio

– займає велику кількість пам’яті;

– часті помилки з із сумісністю систем;

– складний для вивчення, не інтуїтивний інтерфейс;

– функція налагодження не така досконала, як у інших IDE інструментів.

Отож для розробки інформаційної системи автоматизованого формування множини ключових семантичних одиниць рекомендовано використати засоби розробки: платформу .NET, мову програмування C#, СКБД Microsoft SQL Server та середовище розробки Visual Studio 2017.

Висновки до розділу 3

У розділі було описано структуру та функціональне призначення складових інформаційної системи автоматизованого формування множини ключових семантичних одиниць, яка використовує створений метод автоматизованого формування семантичного ядра цифрових текстів й забезпечує можливість за вхідними даними у вигляді цифрового тексту та множина слів і словосполучень цього тексту з показниками їх важливості одержувати вихідні дані у вигляді сформованих семантичних ядер із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Відповідно до поставленого завдання, а саме створення інформаційної системи автоматизованого формування множини ключових семантичних одиниць, було спроектовано структуру бази даних, що дозволяє зберігати матеріал для аналізу, а саме тексти і їх компоненти, та супроводжувати ефективний обрахунок результатів аналізу тексту.

Також в розділі було проведено аналіз рекомендованих засобів розробки інформаційної системи автоматизованого формування множини ключових семантичних одиниць й визначено рекомендовані засоби розробки інформаційної системи.

Розділ 4

Дослідження ефективності методу автоматизованого формування семантичного ядра цифрових текстів

4.1 Розробка прикладних компонентів інформаційної системи автоматизованого формування множини ключових семантичних одиниць

Основним завданням роботи інформаційної системи автоматизованого формування множини ключових семантичних одиниць є виведення списку ключових слів тексту та їх значущість в ньому. Для цього в програмі використовуються методи та класи, що описані та проілюстровані вище. Етапи роботи користувача із системою будуть поділятись на наступні етапи: «Тексти», «Появи семантичних одиниць», «Лінгвістична база» та «Формування семантичного ядра», що є основним у роботі.

Перейдемо до початку нашої роботи. У створену БД завантажуємо необхідні матеріали, котрі необхідно дослідити. Після цього можемо взаємодіяти із вкладками на формі. Вкладка «Тексти» призначена для того, щоб користувач міг обрати необхідний текст із бази даних, натиснувши на поле необхідної назви та кнопку «Обрати текст» (рисунок 4.1).

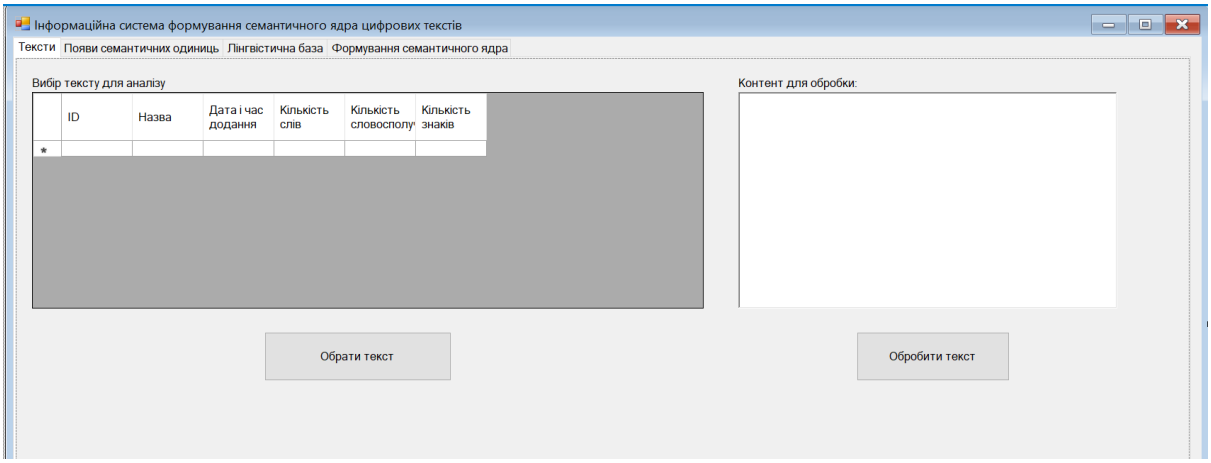


Рисунок 4.1 – Вкладка «Тексти»

Матеріали, котрі необхідно проаналізувати, мають завантажуватись із записом відповідних полів, а саме: назва, дата й час додання до БД. Кількість слів, словосполучень та кількість знаків система обраховує автоматично, відповідно заповнюються всі поля таблиці. Спершу необхідно завантажити дані із БД для представлення у вигляд таблиці. Реалізація даного модуля представлена у наступному програмному коді:

```
void LoadData()
        {
            string connectString = "Server=.\\SQLEXPRESS;Initial Catalog=data;Integrated
Security=true;";
            SqlConnection myConnection1 = new SqlConnection(connectString);
            myConnection1.Open();
            string query1 = "SELECT * FROM Texts ORDER BY ID";
            SqlCommand command = new SqlCommand(query1, myConnection1);
            SqlDataReader reader = command.ExecuteReader();
            List<string[]> data1 = new List<string[]>();
            while (reader.Read())
            {
                data1.Add(new string[2]);
                data1[data1.Count - 1][0] = reader[0].ToString();
                data1[data1.Count - 1][1] = reader[1].ToString();
            }
            reader.Close();
            myConnection1.Close();
            foreach (string[] a in data1)
                dataGridView3.Rows.Add(a);
             }
```

Для виконання наступних операцій користувачеві необхідно натиснути «Обробити текст», після чого відповідні методи класу обчислюють необхідні параметри для подальшої роботи. Перейдемо до реалізації роботи наступної вкладки «Появи семантичних одиниць». Дана вкладка створена для того, щоб користувач зміг переглянути основні статистичні дані щодо слів. Інформація розділена для перегляду по словах тексту та по його словосполученнях (рисунок 4.3).

Призначення цієї вкладки в наступному: саме тут користувач може переглянути появи слів тексту, їх позиції по словах та по символах, а також відтворити текст за заданими параметрами. Програмний код, що відповідає за реалізацію елементу наведено нижче.

```
Section section = new Section
{
Sections = new List<ISection>(),
Paragraphs = new List<IParagraph>(),
Level = currentParagraph.Level,
Guid = Guid.NewGuid(),
Name = headerText
};
section.Paragraphs.Add(currentParagraph);

lastParagraphIndex++;

MakeSections(section, paragraphs);
currentSection.Sections.Add(section);
}
else if (currentParagraph.Level <
currentSection.Level)
{
break;
}
}
}
```

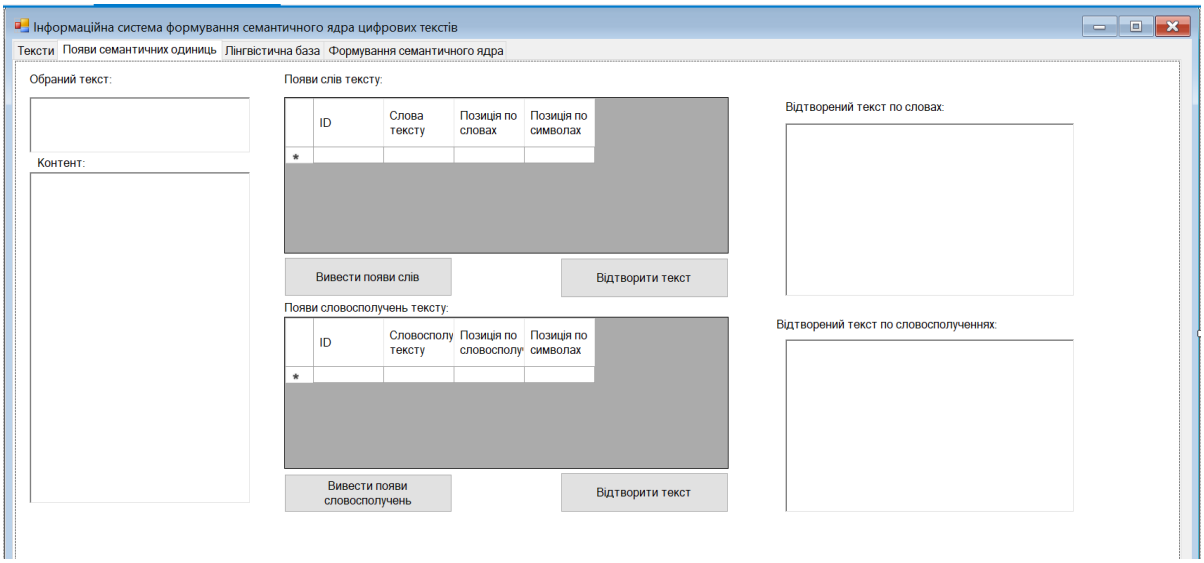


Рисунок 4.3 – Вкладка «Появи семантичних одиниць»

Наступною вкладкою є «Формування семантичного ядра» (рисунок 4.4), вона є основною в нашій роботі. Саме на ній відображатиметься інформація про TF, TFIDF та DE, їх значення та поріг входження кожного значення по словах. Користувач зможе сформувати семантичне ядро декількома способами:

– важливість по словах, обрахунок по символах;

– важливість по словах, обрахунок по словах;

– важливість по словосполученнях, обрахунок по символах;

– важливість по словосполученнях, обрахунок по словосполученнях.

Вищенаведену вкладку формує наступний програмний код:

```
 private void DetermineKeyword(INewSection
section)
 {
 List<Term> keyTerms = new List<Term>();
 for (int i = 1; i <= this.wordCountInTerm;
i++)
 {
 List<Term> result =
this.evaluator.Evaluate(section, i);
 keyTerms.AddRange(result);
 }

 this.keyTerms.Add(section.Guid, keyTerms);

 section.KeyTerms = keyTerms;
 }

 private void SetWordInfo(INewSection section)
 {
 List<Word> allWords = new List<Word>();
 foreach (var term in section.KeyTerms)
 {
 allWords.AddRange(term.Words);
 }

 List<string> uniqueWords = allWords
 .ToLookup(word => word.Text)
 .Select(a => a.First().Text)
 .ToList();
```

Окрім цього, користувач матиме змогу відсортувати слова за значенням параметрів та власноруч обрати значення щільності ключових слів.

Програмний код, що відповідає за реалізацію попереднього модуля наведено нижче:

```
 if (term.Words.First().SpeechPart !=
ESpeechPart.Noun &&
term.Words.First().SpeechPart !=
ESpeechPart.Adjective)
 continue;

 if (term.Words.Last().SpeechPart !=
ESpeechPart.Noun &&
term.Words.Last().SpeechPart !=
ESpeechPart.Adjective)
 continue;

 int n = 0;
 for (int i = 0; i < wordCount; i++)
 {
 if (term.Words[i].SpeechPart ==
ESpeechPart.Noun
 || term.Words[i].SpeechPart ==
ESpeechPart.Adjective
 || term.Words[i].SpeechPart ==
ESpeechPart.Conjunction
 || term.Words[i].SpeechPart ==
ESpeechPart.Particle)
 {
 n++;
 }
 }
```

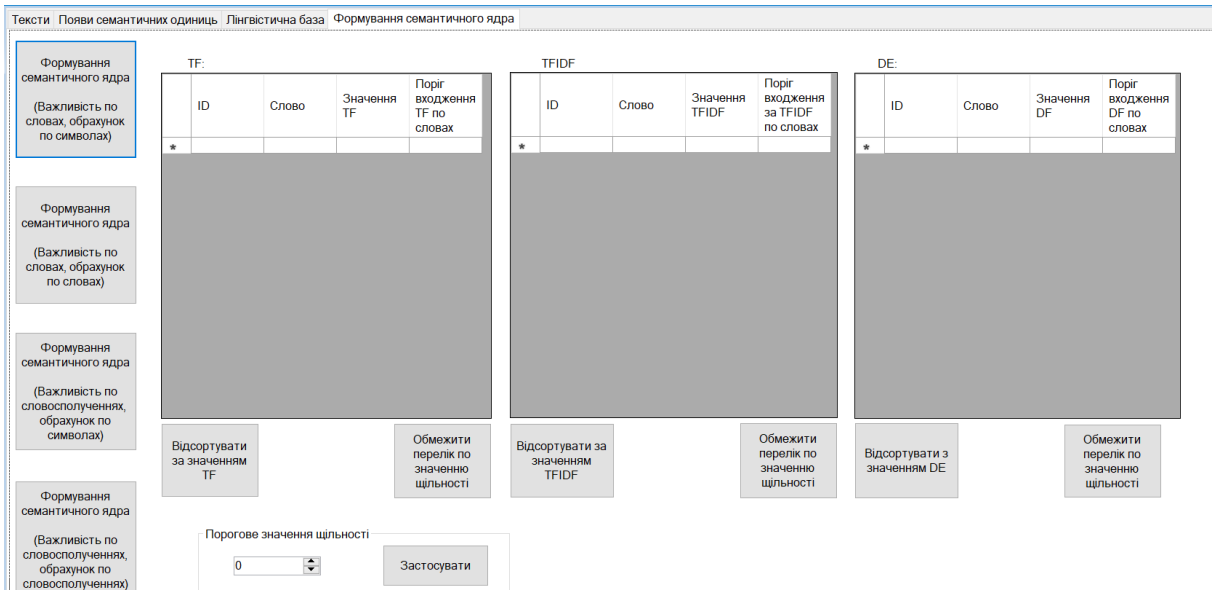


Рисунок 4.4 – Вкладка «Формування семантичного ядра»

Окремий метод реалізовано для параметру щільності пошуку ключових слів. Результат роботи програмного коду наведено нижче (рисунок 4.5).

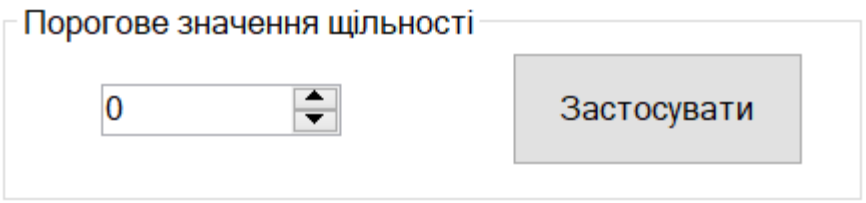


Рисунок 4.5 – Поле для введення порогового значення щільності пошуку термінів

Програмний код, що відповідає за реалізацію попереднього модуля наведено нижче:

```
 foreach (var infinitive in uniqueInfinitive)
 {
 var keyTermForms = section.KeyTerms.Where(x =>
x.Infinitive == infinitive).ToList();
 double maxEvaluation = keyTermForms.Max(x =>
x.Evaluation);
 if (maxEvaluation == 0)
 {
 continue;
 }

 int count = keyTermForms.Sum(x => x.Count);
 Term readableForm =
GetReadableKeyTermForm(keyTermForms);
 if (readableForm == null)
 {
 readableForm = keyTermForms.First(x =>
x.Evaluation == maxEvaluation);
 }
```

Окрім цього, виведені результати аналізу цифрового тексту можна відсортувати залежно від рішення користувача. Доступні сортування за , TFIDF, DE та обмеження переліку по значенню щільності. Відповідні розділи користувацького меню зображено на рисунку 4.6.

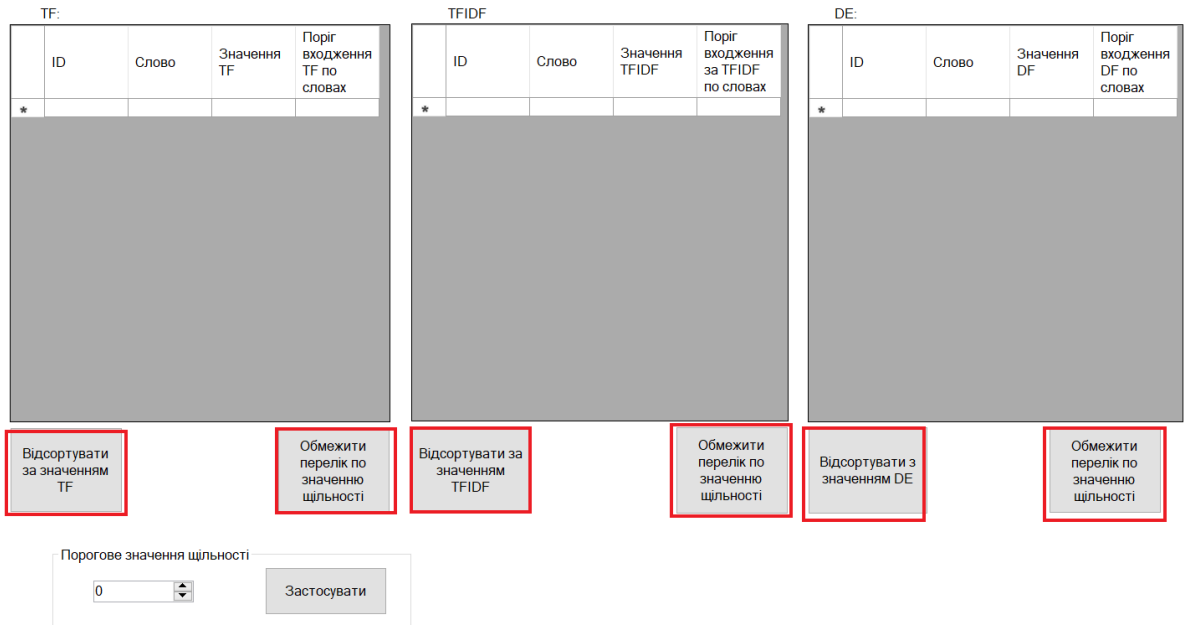


Рисунок 4.6 – Кнопки для сортування значень за різними параметрами та обмеження переліку за значенням щільності

Отже, реалізовано модулі та складові для інформаційної системи автоматизованого формування множини ключових семантичних одиниць. Інтерфейс користувача створено зрозумілим та інтуїтивним, що в свою чергу створює ряд переваг над іншими програмними продуктами. Між складовими програми, її класами створено логічний зв'язок та забезпечене надійне з’єднання між базою даних та інформаційною системою.

4.2 Прикладне тестування інформаційної системи

Задля того, щоб переконатись у правильності та коректності роботи інформаційної системи формування семантичного ядра цифрових текстів, проведемо тестування системи. Для цього було створено ряд тест кейсів (тестові випадки). В першому тестовому випадку перевіримо роботу методів, що завантажують тексти в спеціально відведені секції на вкладці «Тексти» (рисунок 4.7).

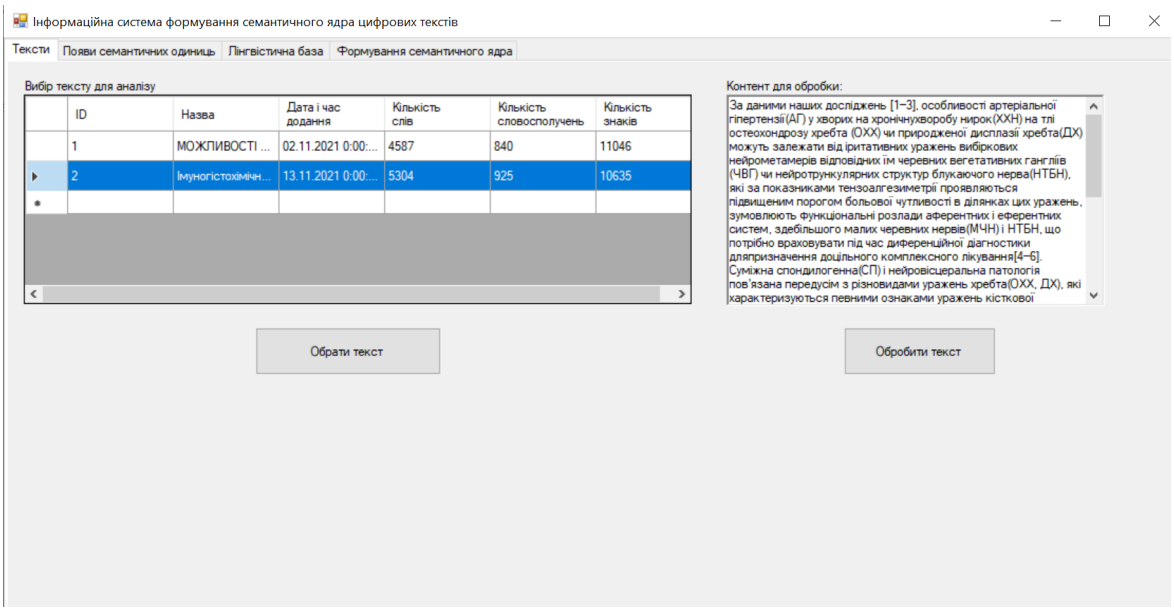


Рисунок 4.7 – Тестування функції завантаження текстів в додаток

В даному випадку ми спробували завантажити для аналізу матеріал, в якому задали лише назву, проте не завантажили сам цифровий документ, що унеможливлює подальшу роботу. Саме тому з'являється повідомлення про помилку (рисунок 4.8).

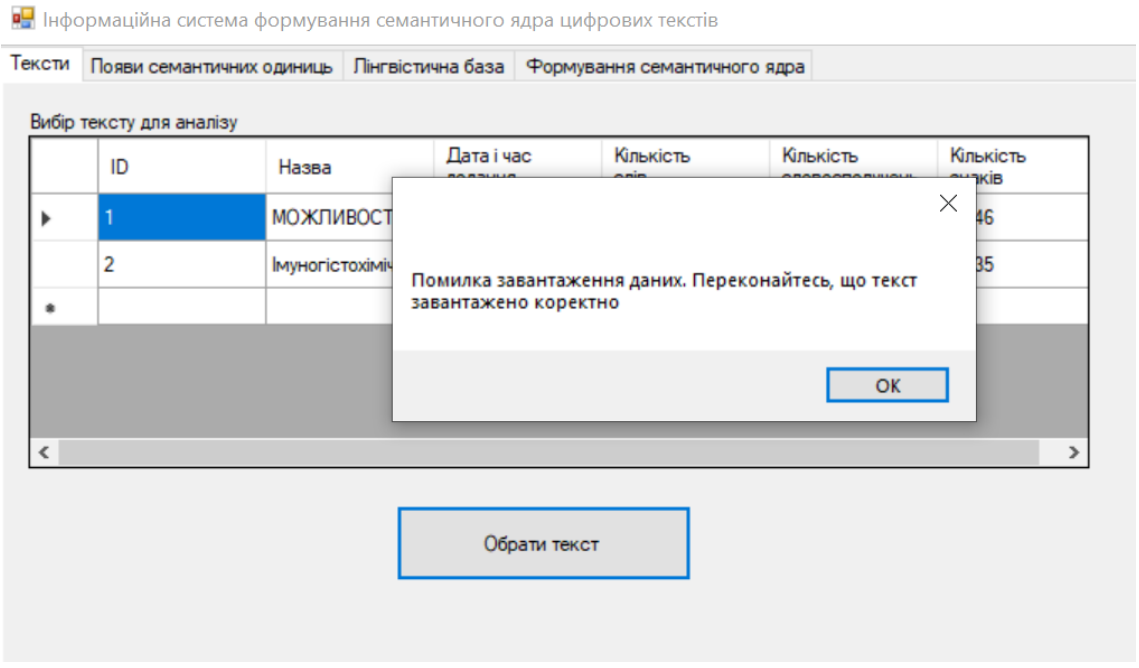


Рисунок 4.8 – Повідомлення про помилку завантаження даних

За вказаною помилкою сформовано наступний тест-кейс (таблиця 4.1).

Таблиця 4.1 – Тест-кейс AT0001

<table>
<tr><td>Тест-кейс ID: AT0001</td><td>Пріоритет: 3</td><td>Створено: 27.11.2021, Купрійчук В.О.</td></tr>
<tr><td colspan="3">Назва: Перевірка завантаження порожнього файлу
Вхідні дані: Файл .doc</td></tr>
<tr><td colspan="2">Кроки</td><td>Очікуваний результат</td></tr>
<tr><td colspan="2">Передумова: користувач на вкладці «Тексти» має обрати рядок в таблиці, що відповідає за обробку порожнього текстового файлу.
1. Запустити додаток
2. Перейти на вкладку «Тексти»
3. Натиснути на відповідний запис із порожнім файлом (2 рядок)
4. Натиснути кнопку «Обрати текст»
5. Порівняти фактичний результат з очікуваним
6. Закрити діалог із помилкою
7. Обрати рядок 1 із успішно завантаженим файлом
8. Натиснути кнопку «Обрати текст»
9. Порівняти фактичний результат з очікуваним</td><td>Діалог із помилкою: «Помилка завантаження даних. Переконайтесь, що текст завантажено коректно».

Текст оброблено системою, контент текстового файлу виведено в окреме поле «Контент»</td></tr>
<tr><td colspan="3">Результат виконання тест-кейсу: пройдено успішно</td></tr>
</table>

Наступний тест-випадок перевірятиме правильність роботи методу, що повертає порогове значення щільності на користувацькій вкладці «Формування семантичного ядра» (рисунок 4.9).

Рисунок 4.9 – Поле для задання порогового значення щільності на вкладці «Формування семантичного ядра»

Звіт про результати проведеного тестування наведено в таблиці 4.2

Таблиця 4.2 – Тест-кейс AT0002

Тест-кейс ID: AT0002	**Пріоритет: 2**	**Створено:** 27.11.2021, Купрійчук В.О.
Назва: Перевірка коректності значення порогової щільності **Вхідні дані:** число типу int		
Кроки	**Очікуваний результат**	
Передумова: користувач на вкладці «Формування семантичного ядра» задає порогове значення щільності. 1. Запустити додаток 2. Перейти на вкладку «Формування семантичного ядра» 3. Задати число «999» в полі «Порогове значення щільності» 4. Натиснути кнопку «Застосувати» 5. Порівняти фактичний результат з очікуваним 6. Закрити діалог із помилкою 7. Задати число «4» 8. Натиснути кнопку «Застосувати» 9. Порівняти фактичний результат з очікуваним	Діалог із помилкою: «Недопустиме значення порогової щільності. Оберіть інше значення». Значення оброблено системою, запущено методи пошуку ключових термінів із заданою щільністю.	
Результат виконання тест-кейсу: пройдено успішно		

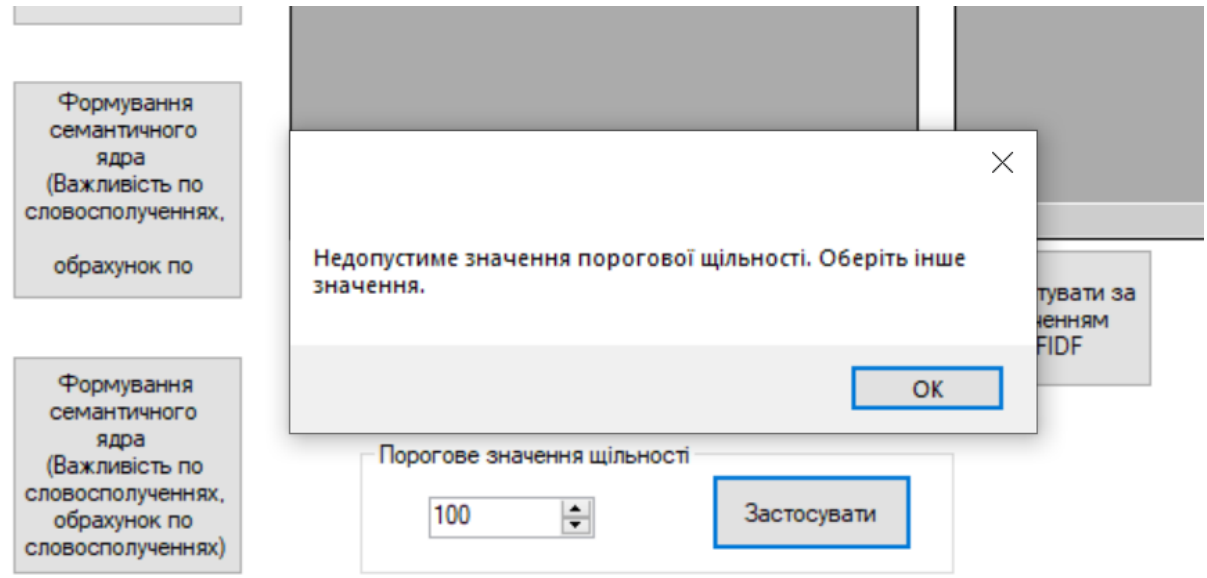


Рисунок 4.10 – Повідомленням про некоректне значення порогової щільності

Проведемо ще одне тестування системи, цього разу користувач спробує запустити пошук ключових термінів не обравши текст для обробки (рисунок 4.11).

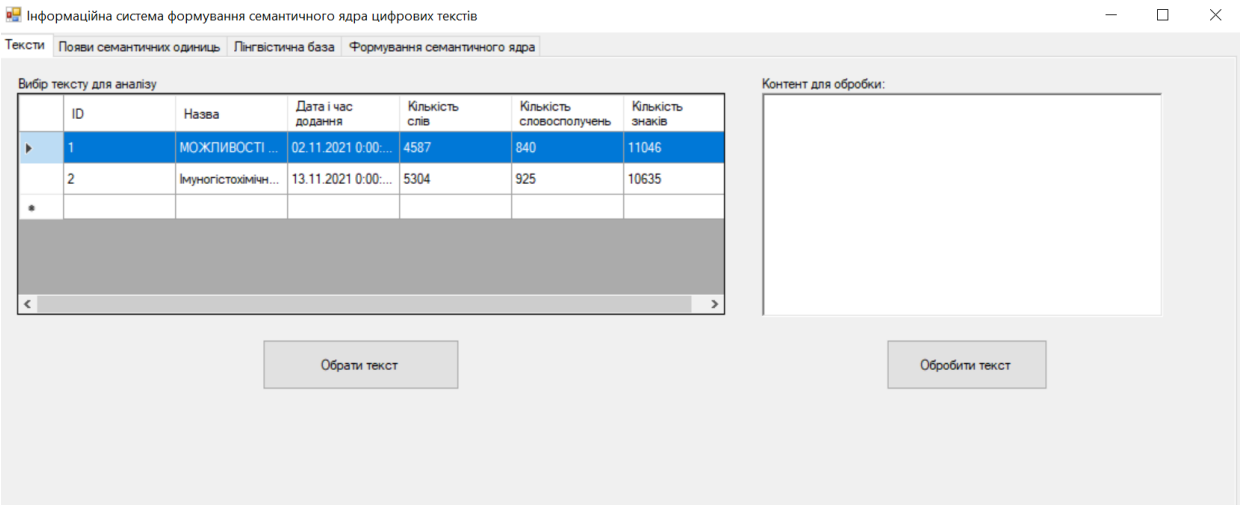


Рисунок 4.11 – Запуск пошуку ключових слів на вкладці «Тексти»

Результати проведеного тестування наведено у наступній таблиці (таблиця 4.3, рисунок 4.12)

Таблиця 4.3 – Тест-кейс AT0003

<table>
<tr><td>Тест-кейс ID: AT0003</td><td>Пріоритет: 1</td><td>Створено: 27.11.2021, Купрійчук В.О.</td></tr>
<tr><td colspan="3">Назва: Перевірка коректності завантаження тексту
Вхідні дані: файл розширення .doc</td></tr>
<tr><td colspan="2">Кроки</td><td>Очікуваний результат</td></tr>
<tr><td colspan="2">Передумова: користувач на вкладці «Тексти» не обирає текст для аналізу та запускає процес аналізу тексту.
1. Запустити додаток
2. Перейти на вкладку «Тексти»
3. Не обравши у полі «Тексти» матеріал, запустити процес аналізу тексту, натиснувши кнопку «Обробити текст»
4. Натиснути кнопку «Застосувати»
5. Порівняти фактичний результат з очікуваним
6. Закрити діалог із помилкою
7. Задати необхідний для аналізу текст
8. Натиснути кнопку «Обробити текст»
9. Порівняти фактичний результат з очікуваним</td><td>Діалог із помилкою: «Операція неможлива, спершу оберіть текст для аналізу».

Текст було задано, результати обробки тексту відображаються у відповідних полях</td></tr>
<tr><td colspan="3">Результат виконання тест-кейсу: пройдено успішно</td></tr>
</table>

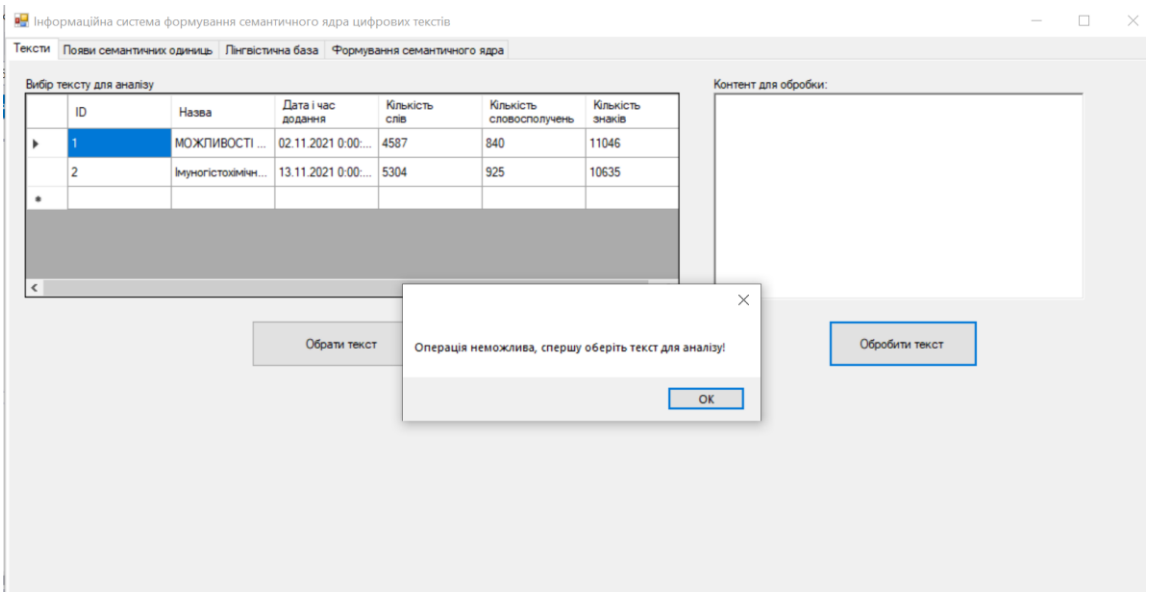


Рисунок 4.12 – Сповіщення про некоректний вибір тексту

Отже, дослідження коректності виконання функцій інформаційної системи автоматизованого формування семантичного ядра цифрових текстів успішно пройдено. Наочно доведено, що функції користувача було реалізовано в інформаційній системі. Успішне виконання прикладного тестування інформаційної системи автоматизованого формування семантичного ядра цифрових текстів визначає її придатність для коректного дослідження ефективності розробленого методу автоматизованого формування семантичного ядра цифрових текстів.

4.3 Функціональне дослідження інформаційної системи та визначення вимог до її розгортання

Розроблена інформаційна система автоматизованого формування множини ключових семантичних одиниць використовує створений метод автоматизованого формування семантичного ядра цифрових текстів та забезпечує можливість одержувати дані в вигляді сформованих семантичних ядер з слів при обрахунку порогу щільності у символах, з словосполучень при обрахунку порогу щільності в символах, з слів при обрахунку порогу щільності у словах та з словосполучень при обрахунку порогу щільності в словах.

Після того, як користувач запускає програму першим його кроком має бути вибір тексту для аналізу. Для цього необхідно перейти на вкладку «Тексти» (рисунок 4.7) та із таблиці обрати один із запропонованих для аналізу.

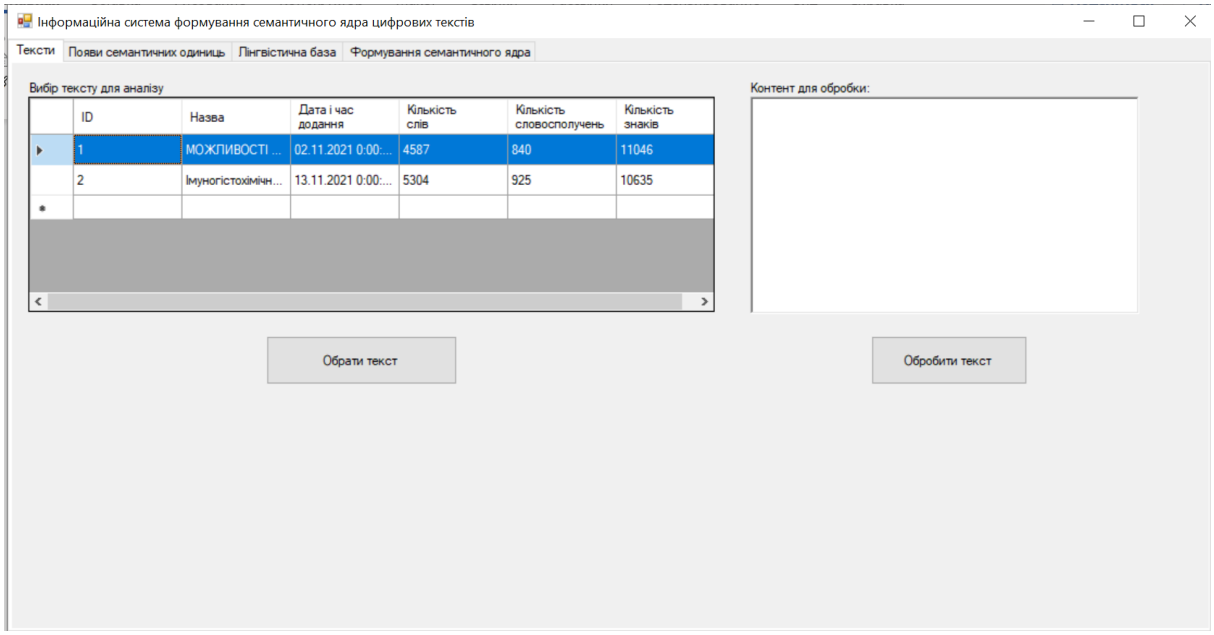


Рисунок 4.7 – Вкладка «Тексти» для вибору тексту для подальшого аналізу

Після вибору матеріалу необхідно підтвердити дію та натиснути кнопку «Обрати текст». Якщо операція успішна – користувач побачить текст в полі «Текст для обробки». Для аналізу тексту натискаємо кнопку «Обробити текст» (рисунок 4.8).

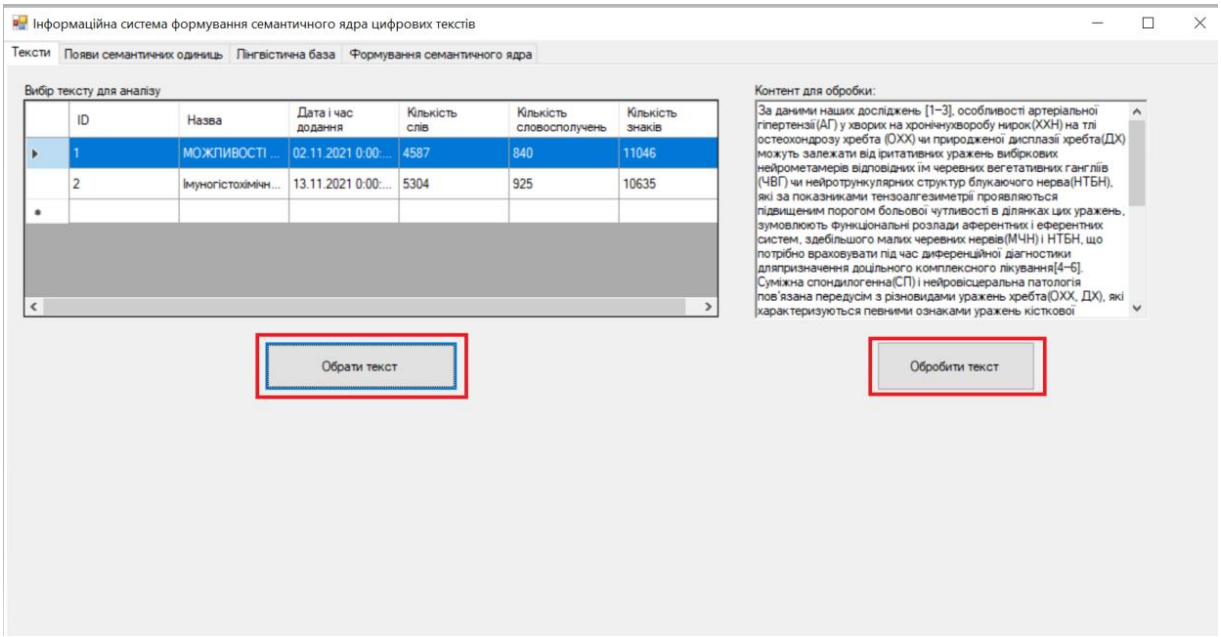


Рисунок 4.8 – Вибір тексту для обробки та підтвердження дії

Після вибору тексту переходимо до початкової обробки тексту. Для цього переходимо на вкладку «Появи семантичних одиниць» (рисунок 4.9), де можемо переглянути появи слів тексту та появи словосполучень тексту. В полі «Обраний текст» відображатиметься назва обраного тексту, а в полі «Контент» - його зміст, відповідно.

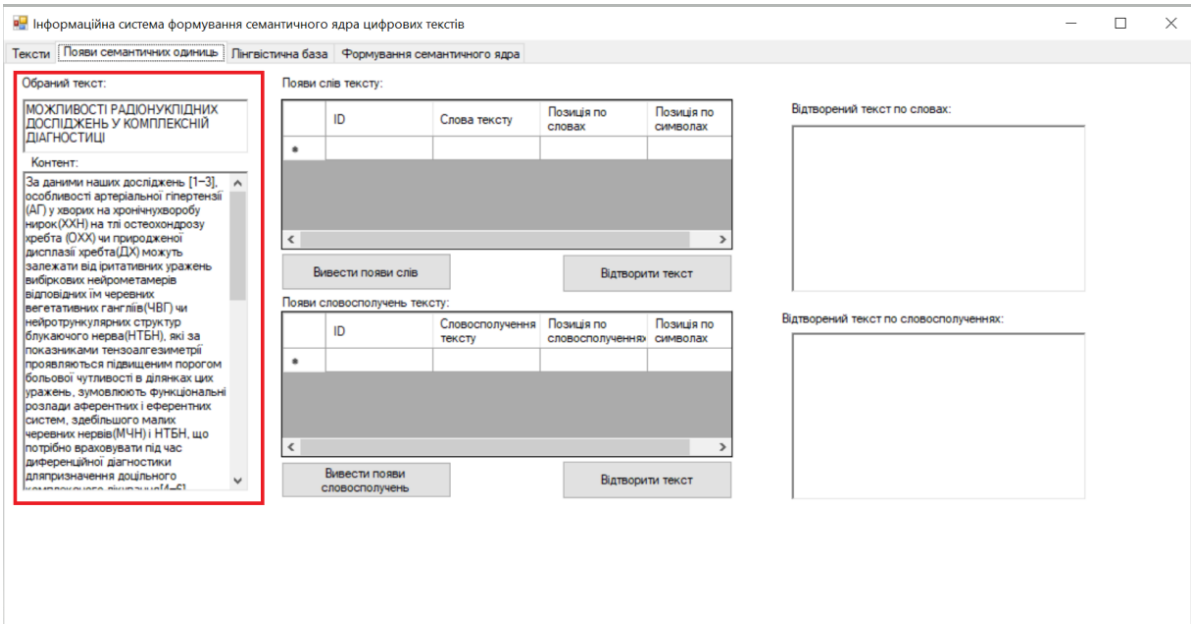


Рисунок 4.9 – Відображення назви обраного тексту та його контенту на вкладці «Появи семантичних одиниць»

Для відображення результатів роботи програми, а саме виведення появи слів тексту, словосполучень та основних характеристик тексту натискаємо кнопки «Вивести появи слів» та «Вивести появи словосполучень» (рисунок 4.10), відповідно. Після цього на екрані відображаються результати.

Наступною вкладкою є «Формування семантичного ядра» (рисунок 4.11). Користувачеві необхідно перейти на неї та виконати ряд дій для відображення відповідних ключових слів та їх значень. Перейдемо на зазначену вкладку та переконаємось, що значення автоматично виведені на

екран у відповідні таблиці. Користувач може відсортувати значення за TF, TFIDF та DE .

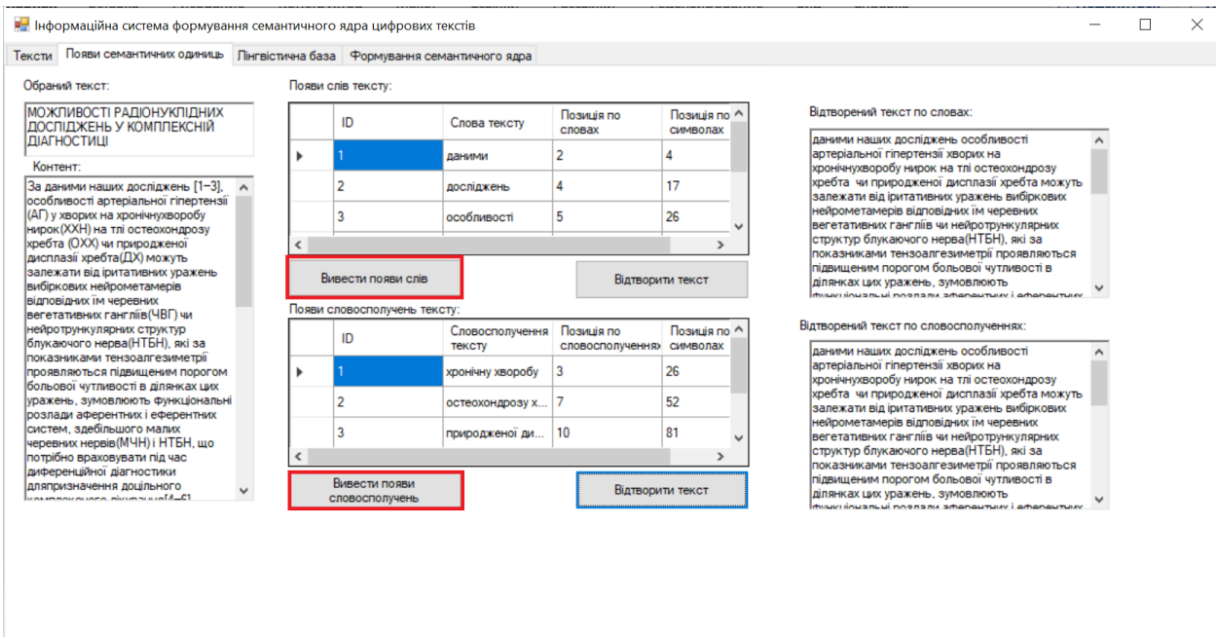


Рисунок 4.10 – Відображення результатів обчислень основних параметрів тексту на вкладці «Появи семантичних одиниць»

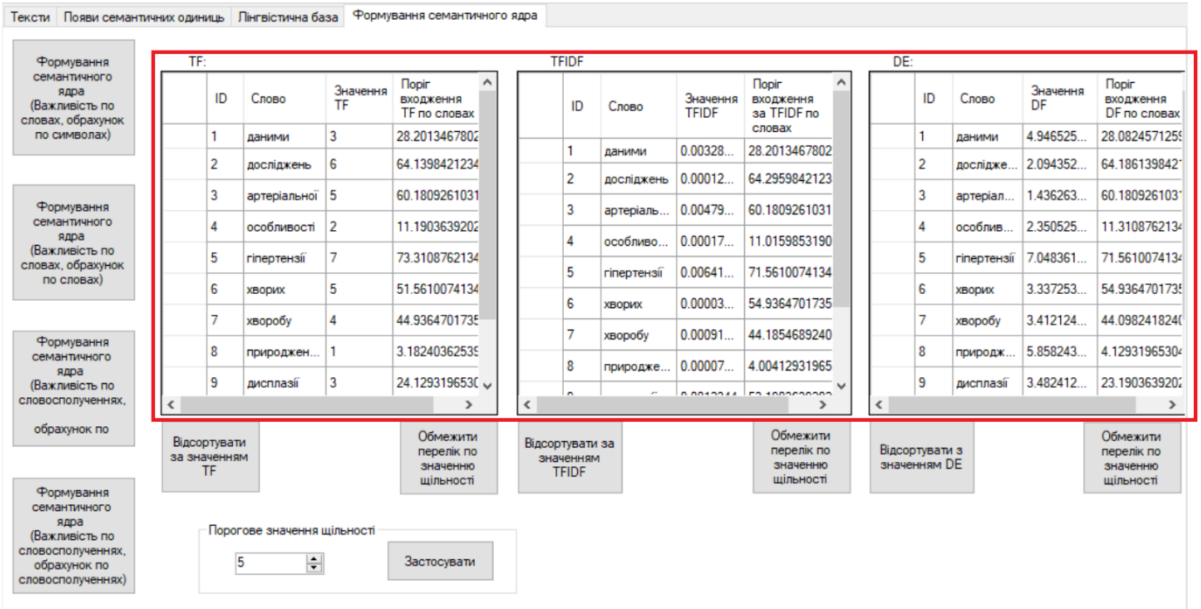


Рисунок 4.11 – Виведення результатів на вкладку «Формування семантичного ядра»

Окрім цього, користувач може обрати порогове значення щільності та застосувати його при обчисленні (рисунок 4.12). необхідно ввести бажане значення та натиснути кнопку «Застосувати».

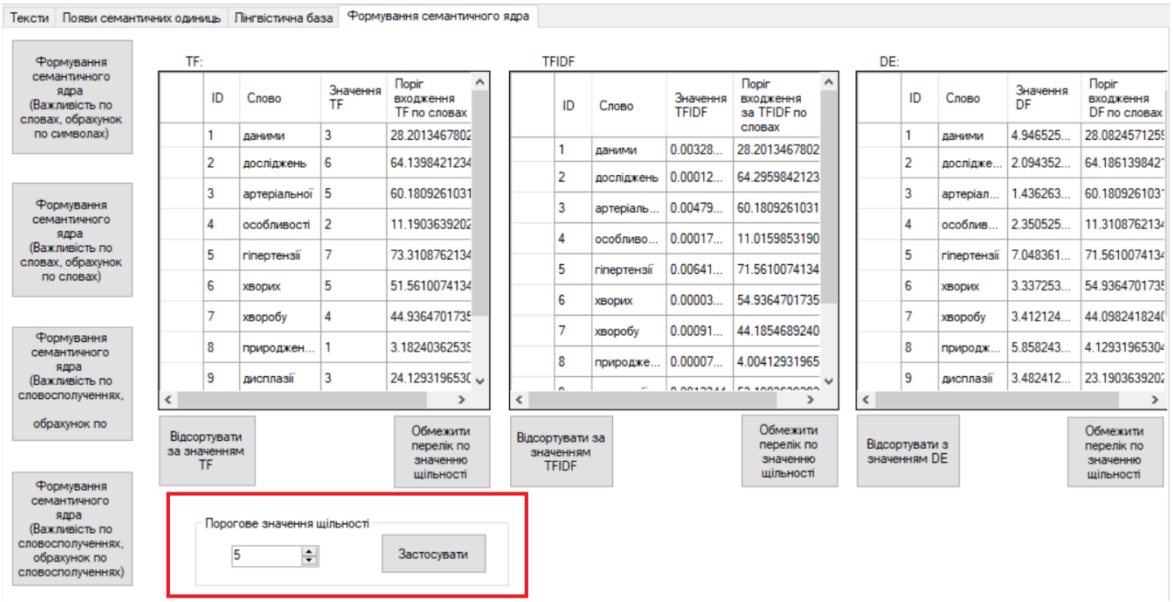


Рисунок 4.12 – Застосування поля «Порогове значення щільності»

Таким чином, програмний продукт було створено із максимально простим та інтуїтивним інтерфейсом. Окремі модулі інформаційної системи автоматизованого формування множини ключових семантичних одиниць розташовано на різних вкладках, що дає змогу користувачеві швидко зорієнтуватись та використовувати програму швидко та зручно.

Для забезпечення коректної роботи інформаційної системи автоматизованого формування множини ключових семантичних одиниць необхідно виконати певні технічні вимоги та пересвідчитись, що комп’ютер користувача має необхідні характеристики для запуску програми. Вимоги до апаратної частини:

– основним сервером повинна бути СКБД SQL Server версії 14 0 17289 0 та вище;

– Windows Server, від версії 2012 – поточна версія;

– Microsoft SQL Server, від версії 2012 – поточна версія;

– Microsoft .NET Framework, від версії 4.0 – поточна версія;

– дисковий простір – не менше 1000 Мб.

Вимоги до завантаження цифрових матеріалів для аналізу:

– цифрові тексти необхідно завантажити із розширенням .doc;

– необхідно зазначити назву тексту, що буде проаналізовано.

4.4 Дослідження ефективності методу автоматизованого формування семантичного ядра цифрових текстів

Розроблена інформаційна система автоматизованого формування множини ключових семантичних одиниць використовує створений метод автоматизованого формування семантичного ядра цифрових текстів для того, щоб одержувати вихідні дані:

– семантичне ядро тексту із слів при обрахунку порогу щільності у символах;

– семантичне ядро тексту із словосполучень при обрахунку порогу щільності у символах;

– семантичне ядро тексту із слів при обрахунку порогу щільності у словах;

– семантичне ядро тексту із словосполучень при обрахунку порогу щільності у словах.

При цьому обраний алгоритм визначення важливості слів і словосполучень не має значення, проте важливими є власне оцінки семантичної значущості цих слів і словосполучень.

Було проведено ряд досліджень, коли для різних цільових відсотків щільності ключових слів у тесті обраховувались компоненти семантичного ядра цифрових текстів за різними способами обрахунку.

Для оцінки якості формування семантичного ядра цифрових текстів використовуються оцінка точності [28]. Точність пошуку P є відношенням

числа релевантних ключових слів і словосполучень знайдених автоматично, до загальної кількості знайдених ключових слів і словосполучень в тексті:

$$P = \frac{|M_{TK}^{E} \cap M_{TK}|}{|M_{TK}|},$$

де M_{TK}^{E} – множина релевантних ключових слів і словосполучень, сформована експертом; M_{TK} – множина знайдених автоматично ключових слів і словосполучень.

Середня точність пошуку ключових слів і словосполучень $\overline{P}$ визначається так [28]:

$$\overline{P} = \frac{\sum_{i=1}^{k} P_k}{k},$$

де k – кількість текстів в вибірці.

До прикладу, за результатом аналізу 45 текстів за значень цільових відсотків щільності ключових слів у текстах 10-15% (Рисунок 4.13) середня точність формування семантичного ядра тексту із слів при обрахунку порогу щільності у символах склала 82,93%, середня точність формування семантичного ядра тексту із словосполучень при обрахунку порогу щільності у символах склала 86,19%, середня точність формування семантичного ядра тексту із слів при обрахунку порогу щільності у словах склала 79,63%, а середня точність формування семантичного ядра тексту із словосполучень при обрахунку порогу щільності у словах склала 83,55%.

Загалом формування семантичного ядра тексту із словосполучень (при цьому одне слово тут розглядається як різновид словосполучення) виявило більш точний результат ніж формування семантичного ядра тексту із слів, оскільки дозволило включити в актуальну множину ті слова, які не були визначені важливими але мали семантичну важливість у комбінації з іншими словами.

Формування семантичного ядра тексту при обрахунку порогу щільності у символах виявилось більш ефективним за формування семантичного ядра при обрахунку порогу щільності у словах, оскільки

дозволяє більш точно зафіксувати значення порогу щільності ключових слів у тексті.

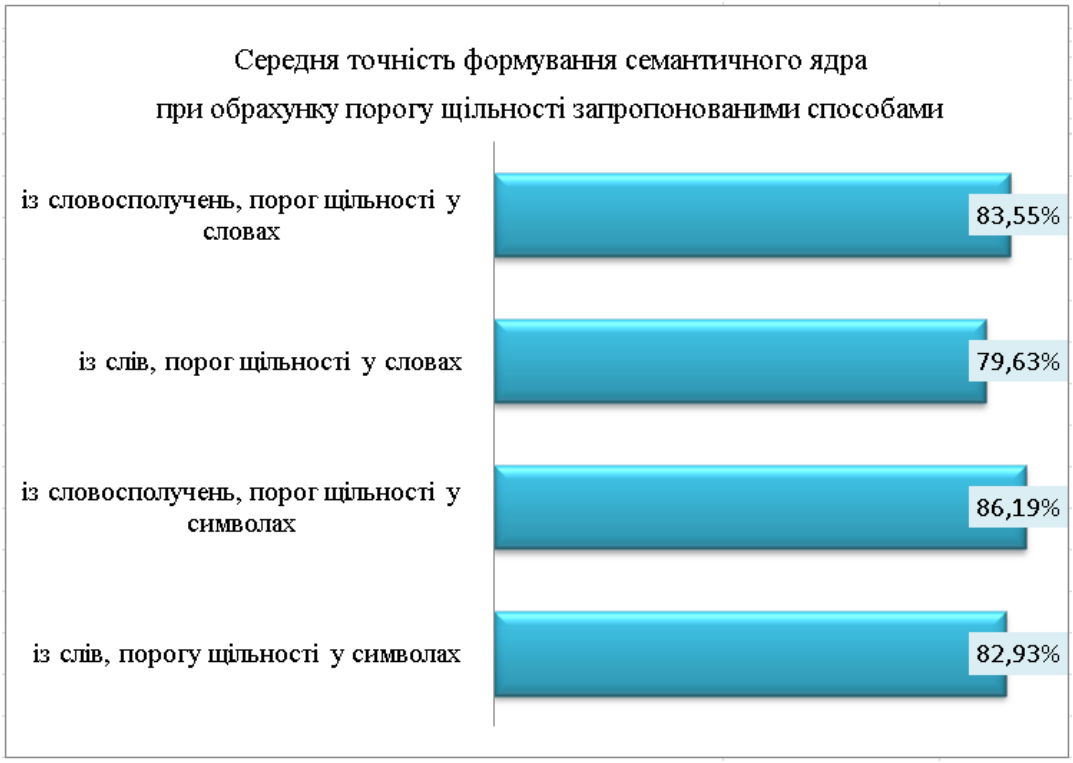


Рисунок 4.13 – Діаграма порівняння ефективності формування семантичного ядра тексту при обрахунку порогу щільності за різними способами

Одержані результати свідчать, що розроблений метод автоматизованого формування семантичного ядра цифрових текстів, який використовує обмеження множин ключових слів і словосполучень за значенням порогу щільності, дозволяє одержати достатньо репрезентативні складові семантичного ядра тексту. Це дає привід стверджувати, що прикладне використання розробленого методу може бути ефективно застосоване при вирішення задач семантичного аналізу текстів відповідно до призначення.

Висновки до розділу 4

В розділі було виконано розробку експериментальної інформаційної системи автоматизованого формування множини ключових семантичних одиниць, реалізовано модулі та складові для інформаційної системи. Між складовими програми, її класами створено логічний зв'язок та забезпечене з’єднання між базою даних та інформаційною системою.

Успішне виконання в розділі прикладного тестування інформаційної системи автоматизованого формування семантичного ядра цифрових текстів визначило її придатність для коректного дослідження ефективності розробленого методу автоматизованого формування семантичного ядра цифрових текстів.

Було проведено ряд досліджень, коли для різних цільових відсотків щільності ключових слів у тесті обраховувались компоненти семантичного ядра цифрових текстів за різними способами обрахунку. При цьому обраний алгоритм визначення важливості слів і словосполучень не має значення, проте важливими є власне оцінки семантичної значущості цих слів і словосполучень. Так, за результатом одного з досліджень, середня точність формування семантичного ядра тексту із слів при обрахунку порогу щільності у символах склала 82,93%, середня точність формування семантичного ядра тексту із словосполучень при обрахунку порогу щільності у символах склала 86,19%, середня точність формування семантичного ядра тексту із слів при обрахунку порогу щільності у словах склала 79,63%, а середня точність формування семантичного ядра тексту із словосполучень при обрахунку порогу щільності у словах склала 83,55%.

Загалом формування семантичного ядра тексту із словосполучень (при цьому одне слово тут розглядається як різновид словосполучення) виявило більш точний результат ніж формування семантичного ядра тексту із слів, оскільки дозволило включити в актуальну множину ті слова, які не були визначені важливими але мали семантичну важливість у комбінації з

іншими словами. Формування семантичного ядра тексту при обрахунку порогу щільності у символах виявилось більш ефективним за формування семантичного ядра при обрахунку порогу щільності у словах, оскільки дозволяє більш точно зафіксувати значення порогу щільності ключових слів у тексті.

Одержані результати свідчать, що розроблений метод автоматизованого формування семантичного ядра цифрових текстів, який використовує обмеження множин ключових слів і словосполучень за значенням порогу щільності, дозволяє одержати достатньо репрезентативні складові семантичного ядра тексту, що дає привід стверджувати, що прикладне використання розробленого методу може бути ефективно застосоване при вирішення задач семантичного аналізу текстів.

Загальні висновки

Кваліфікаційна робота магістра розв’язує науково-технічну задачу автоматизованого формування множини ключових семантичних одиниць за цифровим текстом та множинами слів і словосполучень цього тексту з показниками їх важливості, одержуючи сформовані семантичні ядра за різними показниками важливості, зокрема із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

За виконання роботи були поставлені й *вирішені наступні завдання*:

1. Проведено аналіз предметної області семантичного аналізу цифрових текстів та відомих підходів до автоматизації формування семантичного ядра цифрових текстів.

2. Вдосконалено метод автоматизованого формування семантичного ядра цифрових текстів.

3. Розроблено інформаційну технологію автоматизованого формування множини ключових семантичних одиниць.

4. Розроблено інформаційну систему автоматизованого формування множини ключових семантичних одиниць.

5. Проведено прикладне дослідження методу автоматизованого формування семантичного ядра цифрових текстів у складі інформаційної технології автоматизованого формування множини ключових семантичних одиниць і виконано аналіз результатів використання відповідної інформаційної системи.

В результаті роботи були отримані такі *інновації та положення наукової новизни*:

1. Вдосконалено метод автоматизованого формування семантичного ядра цифрових текстів, що дозволяє за множиною слів і словосполучень

цифрового тексту з зіставленими значеннями показників їх семантичної важливості, а також необхідним відсотком щільності ключових слів у тесті автоматизовано формувати відповідні множини ключових слів та словосполучень тексту за різними показниками важливості.

2. Розроблено нову інформаційну технологію автоматизованого формування множини ключових семантичних одиниць, що дозволяє з використанням створеного методу автоматизованого формування семантичного ядра цифрових текстів за вхідними даними у вигляді цифрового тексту та множина слів і словосполучень цього тексту з показниками їх важливості одержувати вихідні дані у вигляді сформованих семантичних ядер із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

В роботі виконано розробку експериментальної інформаційної системи автоматизованого формування множини ключових семантичних одиниць, реалізовано модулі та складові для інформаційної системи. Інформаційна система використовує створений метод автоматизованого формування семантичного ядра цифрових текстів й забезпечує можливість одержувати дані у вигляді сформованих семантичних ядер із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Було проведено ряд досліджень, коли для різних цільових відсотків щільності ключових слів у тесті обраховувались компоненти семантичного ядра цифрових текстів за різними способами обрахунку. Середня точність формування семантичного ядра тексту із слів при обрахунку порогу щільності у символах склала 82,93%, середня точність формування семантичного ядра тексту із словосполучень при обрахунку порогу

щільності у символах склала 86,19%, середня точність формування семантичного ядра тексту із слів при обрахунку порогу щільності у словах склала 79,63%, а середня точність формування семантичного ядра тексту із словосполучень при обрахунку порогу щільності у словах склала 83,55%.

Загалом формування семантичного ядра тексту із словосполучень (при цьому одне слово тут розглядається як різновид словосполучення) виявило більш точний результат ніж формування семантичного ядра тексту із слів, оскільки дозволило включити в актуальну множину ті слова, які не були визначені важливими але мали семантичну важливість у комбінації з іншими словами. Формування семантичного ядра тексту при обрахунку порогу щільності у символах виявилось більш ефективним за формування семантичного ядра при обрахунку порогу щільності у словах, оскільки дозволяє більш точно зафіксувати значення порогу щільності ключових слів у тексті.

Одержані результати свідчать, що розроблений метод автоматизованого формування семантичного ядра цифрових текстів, який використовує обмеження множин ключових слів і словосполучень за значенням порогу щільності, дозволяє одержати достатньо репрезентативні складові семантичного ядра тексту. Це дає привід стверджувати, що прикладне використання розробленого методу може бути ефективно застосоване при вирішення задач семантичного аналізу текстів відповідно до призначення.

Основні наукові та практичні результати кваліфікаційної роботи магістра доповідались на XIII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп’ютерних наук АПКН–2021» (15–16 жовтня 2021 року) у доповіді на тему «Інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів»; за темою роботи автором виконано наукову публікацію [29].

Перелік посилань

1. Крак, Ю. В.; Бармак, О. В.; Мазурець, О. В. Практична реалізація інформаційної технології автоматизованого визначення множини семантичних термінів в контенті навчальних матеріалів. Проблеми програмування, 2018, 2-3: 245-254.

2. Яхимович О. В. Визначення ключових слів з тексту повідомлень мікроблогів. 2016.

3. Nicolai C., Piazza M. The implicit commitment of arithmetical theories and its semantic core. Erkenntnis. 2019. Т. 84. №. 4. С. 913-937.

4. Zhang Y. et al. A Semantic Analysis and Community Detection-Based Artificial Intelligence Model for Core Herb Discovery from the Literature: Taking Chronic Glomerulonephritis Treatment as a Case Study. Computational and Mathematical Methods in Medicine. 2020.

5. Пономаренко І. В. Особливості побудови семантичного ядра сайту компанії .І. В. Пономаренко .Інфраструктура ринку. 2018. С. 104-108.

6. Соколовська С. Ф. Полісемія в тексті: типи реалізації та функції. Вісник Житомирського державного університету імені Івана Франка.2003. №. 11. С. 201-205.

7. Залуцька О. О. Інформаційний портрет ключових термінів у цифрових навчальних матеріалах. Матеріали III Міжнародної науково-практичної конференції «Сучасні інформаційні технології та інноваційні методики навчання: досвід, тенденції, перспективи». Тернопіль, 2019. С. 120-122.

8. Horan T. C., Emori T. G. Definitions of key terms used in the NNIS System. American journal of infection control. 2017. №. 2. С. 112-116.

9. Сервіс аналізу текстів онлайн «ISTIO». URL: https://istio.com

10. Сервіс повного семантичного онлайн-аналізу текстів «Miratext». URL: https://miratext.ru/seo_analiz_text

11. Біографія Тараса Григоровича Шевченка. URL: http://ukrlit.org/shevchenko_taras_hryhorovych

12. Крак Ю. В. Практична реалізація інформаційної технології автоматизованого визначення множини семантичних термінів в контенті навчальних матеріалів. Ю. В. Крак, О. В. Бармак, О. В. Мазурець. Науковий журнал «Проблеми програмування». Київ, 2018, №2-3. С.245-254.

13. Мазурець О. В. Інформаційна технологія автоматизованого визначення семантичних термінів в елементах навчальних матеріалів. О. В. Мазурець. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2018, №3. С.223-230.

14. Яхимович О. В. Формалізація задачі визначення ключових слів тексту. О. В. Яхимович, О. В. Бісікало. Матеріали XLVIII науково-технічної конференції підрозділів ВНТУ, Вінниця, 2019 р.

15. Рашин А. І. Створення бази даних водних об’єктів. 2021 р. С. 5.

16. Як обрати ключові слова з сайту. URL: https://serpstat.com/blog/how-to-collect-a-semantic-core-for-a-site/

17. Мазурець О.В. Коваль О.О. Інформаційна технологія рекурсивного пошуку ключових термінів у цифрових текстах.. Вісник Хмельницького національного університету. Технічні науки. 2019. №3. С. 188-196.

18. Бармак О. В., Мазурець О. В. Методи автоматизації визначення семантичних термінів у навчальних матеріалах. Вісник Хмельницького національного університету. Серія : Технічні науки. 2015. № 2. С. 209–213.

19. Переваги використання .NET. UR: https://graffersid.com/advantages-and-disadvantages-of-using-net/

20. Мова програмування C#. URL: https://www.codeguru.com/csharp/benefits-of-c/

21. Переваги використання MS SQL Server. URL: https://www.tek-tools.com/database/sql-server-best-practices-and-tools

22. Популярні середовища розробки. URL: https://proglib.io/p/ide-eclipse-za-i-protiv-ot-vedushchih-programmistov-2019-11-02

23. Eclispe: переваги та недоліки. URL: https://proglib.io/p/ide-eclipse-za-i-protiv-ot-vedushchih-programmistov-2019-11-02

24. Переваги та недоліки Visual Studio 2017. URL: https://www.trustradius.com/products/visual-studio-ide/reviews?qs=pros-and-cons

25. MS SQL Server: pros and cons. URL: https://bytescout.com/blog/2014/09/ms-sql-server-history-and-advantages.html

26. C# and JavaScript: pros and cons. URL: https://habr.com/ru/post/414593/

27. Disadvantages of MySQL. URL: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-sql/

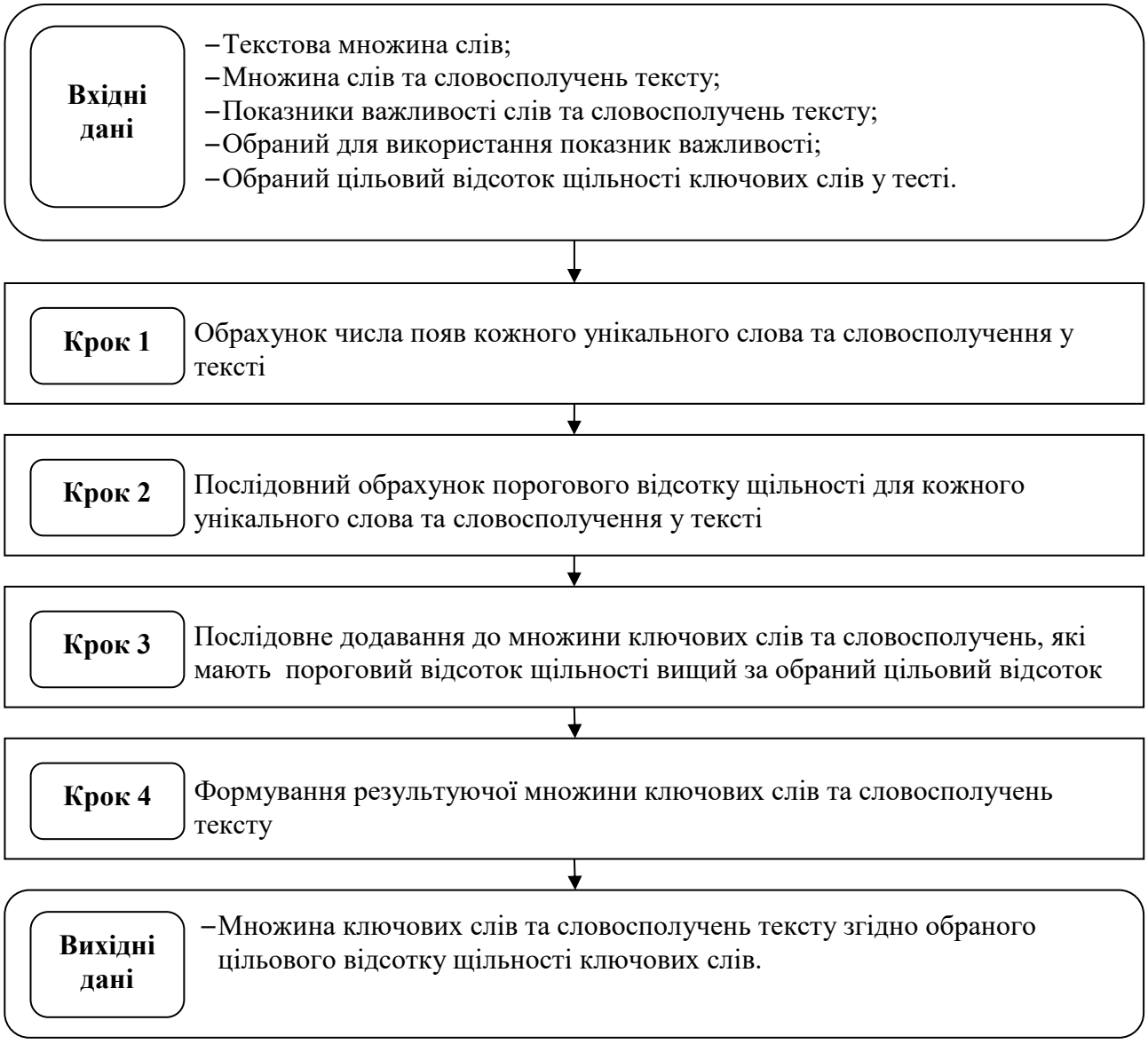
28. Мазурець О. В. Онтологічний підхід до побудови семантичної моделі навчальних матеріалів. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2017, №6. С. 223-229.

29. Войчишин О. О., Залуцька О. О., Попов Ю. М., Купрійчук В. О. Інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів. Збірник наукових праць за матеріалами XIII Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук АПКН-2021». Хмельницький, 2021. с. 298-305.

ДОДАТКИ

Додаток А

Схема методу автоматизованого формування семантичного ядра цифрових текстів



Додаток Б

Схема інформаційної технології автоматизованого формування множини ключових семантичних одиниць

Вхідні дані:
– цифровий текст;
– множина слів тексту та показники їх важливості;
– множина словосполучень тексту та показники їх важливості.

↓

Етап 1 – Поелементна обробка тексту:
1.1 Обрахунок загальних параметрів тексту (кількості слів, словосполучень, знаків).
1.2 Очищення тексту від додаткових символів (знаки, цифри).
1.3 Зменшення регістру тексту.
1.4 Формування текстового вектору слів.
1.5 Формування текстового вектору словосполучень.

↓

Етап 2 – Пошук появ семантичних одиниць та перевірка текстового вектору:
2.1 Обрахунок позиції по словах для кожної появи кожного унікального слова.
2.2 Обрахунок позиції по словах для кожної появи кожного унікального словосполучення.
2.3 Обрахунок позиції по символах для кожної появи кожного унікального слова.
2.4 Обрахунок позиції по символах для кожної появи кожного унікального словосполучення.
2.5 Формування перевірочного тексту з текстового вектору слів.
2.6 Формування перевірочного тексту з текстового вектору словосполучень.
2.7 Обрахунок кількості появ кожного унікального слова.
2.8 Обрахунок кількості появ кожного унікального словосполучення.

↓

Етап 3 – Підготовка до застосування методу формування семантичного ядра:
3.1 Одержання з БД значень важливості унікальних слів тексту TF, TFIDF, DE.
3.2 Одержання з БД значень важливості унікальних словосполучень тексту TF, TFIDF, DE.
3.3 Сортування окремих переліків слів і словосполучень тексту за показниками важливості TF, TFIDF, DE.
3.4 Одержання від користувача цільового відсотку щільності для тексту.

↓

Етап 4 – Автоматизоване формування семантичного ядра цифрових текстів:
4.1 Одержання семантичного ядра слів при обрахунку порогу щільності у символах.
4.2 Одержання семантичного ядра словосполучень при обрахунку порогу щільності у символах.
4.3 Одержання семантичного ядра слів при обрахунку порогу щільності у словах.
4.4 Одержання семантичного ядра словосполучень при обрахунку порогу щільності у словах.

↓

Вихідні дані:
– Семантичне ядро тексту із слів при обрахунку порогу щільності у символах.
– Семантичне ядро тексту із словосполучень при обрахунку порогу щільності у символах.
– Семантичне ядро тексту із слів при обрахунку порогу щільності у словах.
– Семантичне ядро тексту із словосполучень при обрахунку порогу щільності у словах.

Додаток В

Ксерокопії наукових публікацій, виконаних при роботі над кваліфікаційною роботою магістра

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

Перелік наукових публікацій:

1. Войчишин О. О., Залуцька О. О., Попов Ю. М., Купрійчук В. О. Інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів. Збірник наукових праць за матеріалами XIII Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2021». Хмельницький, 2021. с. 298-305.

Міністерство освіти і науки України
Хмельницький національний університет

ЗБІРНИК НАУКОВИХ ПРАЦЬ

за матеріалами XIII Всеукраїнської науково-практичної конференції
«Актуальні проблеми комп'ютерних наук АПКН-2021»

15-16 жовтня 2021

Хмельницький 2021

Перелік посилань:

1. Бойко О.М. Еволюційна технологія розв'язування задачі складання розкладів навчальних занять/ Бойко О.М. // Штучний інтелект. – 2006. – №3. – С. 341-348.
2. Бурнасов П.В. Математична постановка задачі складання розкладу занять // Вісник ІрГТУ. 2014. №4. С. 12-18.
3. Томашевський В.М., Новіков Ю.Л., Камінська П.А. Складання розкладів занять у дистанційних системах навчання // Вісник НТУУ «КПІ» Інформатика, управління та обчислювальна техніка, 2010. №. 52. С. 118-130.
4. Демчук М.В., Мазурець О.В. Автоматизація ведення розкладу занять у вузі. Збірник наукових праць за матеріалами восьмої міжнародної науково-технічної конференції «Актуальні проблеми комп'ютерних технологій 2014». Хмельницький, 2014. С.87-93.
5. Паралельний генетичний алгоритм побудови розкладу занять / М.М. Глибовець, Н.М. Гулаєва, М.М. Пасічник // Проблеми програмування. - 2015. - № 2. - С. 76-85.

УДК 004

Войчишин О. О., Залуцька О. О., Попов Ю. М., Купрійчук В. О.

Хмельницький національний університет

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АВТОМАТИЗОВАНОГО ФОРМУВАННЯ СЕМАНТИЧНОГО ЯДРА ЦИФРОВИХ ТЕКСТІВ

Розглянуто інформаційну технологію автоматизованого формування семантичного ядра цифрових текстів, яка дозволяє перетворювати вхідні дані у вигляді цифрового тексту, множини слів і словосполучень тексту з показниками їх семантичної важливості в вихідні дані у вигляді зразків семантичного ядра тексту. Зразки семантичного ядра тексту одержуються у варіаціях: із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Наведені в статті зразки програмного забезпечення, які дозволяють створювати множини термінів цифрових текстів, формувати семантичне ядро шляхом прикладного застосування розробленої інформаційної технології, а також практичне використання результатів для адаптивної пропозиції товарів у інтернет-магазині за семантичними ознаками, демонструють повний набір компонентів для практичного вирішення актуальної задачі інформаційних технологій.

Information technology for automated formation of semantic core of digital texts is considered, which allows to convert input data in form of digital text, sets of words and phrases of text with indicators of their semantic importance into source data in the form of samples of text semantic core. Samples of semantic core of text are obtained in variations: from words when calculating the density threshold in symbols, from phrases when calculating density threshold in symbols, from words when calculating the density threshold in words and from phrases when calculating density threshold in words.

The software samples presented in article, which allow to create sets of digital text terms, form a semantic core by applying developed information technology, as well as practical use of results for adaptive supply of goods in online store on semantic features, demonstrate a full set of components for practical solution.

Електронний текст став феноменом, якому у сучасному науковому просторі приділяється велика кількість уваги. Саме він розглядається як основне джерело інформації. Існує кілька підходів до його аналізу. Можна, наприклад, визначати тему і ідею текстів, аналізувати, оцінювати смислове навантаження або виділяти сферу, з якою вони пов'язані (математика, комп'ютерні науки, література, соціологія) [1].

У зв'язку з тим, що мова являє собою досить складне утворення, в комп'ютерній лінгвістиці склалися і розвиваються різні напрямки, приблизно порівнянні з окремими рівнями мови, з процесами породження і сприйняття

мовленнєвих повідомлень або іншими видами людської діяльності, пов'язаної з мовою. Відповідно, до напрямів комп'ютерної лінгвістики належать:

- автоматизований синтез текстів;
- автоматизований аналіз текстів;
- створення та підтримка автоматичних словників;
- створення автоматизованих інформаційно-пошукових систем;
- машинний переклад;
- створення автоматичних систем вивчення мови;
- автоматична атрибуція та дешифрування текстів;
- створення лінгвістичних баз даних;
- розробка програмних інструментів для рішення задач теоретичної та прикладної лінгвістики [2].

Велика кількість наукових праць була спрямована на розробку математичних алгоритмів та комп'ютерних програм обробки текстів природною мовою. Для автоматизації цих процесів було створено різні моделі процесів обробки та аналізу текстів, а також структури та алгоритми для представлення результатів. У переважній більшості аналіз цифрових текстів було представлено наступною послідовністю: морфологічний аналіз тексту, синтаксичний аналіз та семантичний аналіз. Для кожного з цих етапів були створені відповідні моделі та алгоритми [3].

Ключове слово є словом або словосполученням природної мови, яке використовують для вираження деякого аспекту змісту навчального матеріалу. Елементи множини ключових термінів мають істотне смислове навантаження і формують перелік розглянутих в навчальному матеріалі понять. Ключові терміни мають наступні властивості:

1) є найбільш вживаними (частотними) найменуваннями, визначають ознаку предмета, стан або дію;

2) представлені значущою лексикою, досить узагальнені за своєю семантикою (середнього ступеня абстракції), стилістично нейтральні й не оціночні;

3) пов'язані один з одним мережею семантичних зв'язків;

4) мінімальна кількість елементів у множині ключових термінів наближається до інваріанта змісту навчального матеріалу при їх логічному впорядкування [4].

Семантичне ядро – це певний невпорядкований набір слів і словосполучень, що описують певний предмет, повністю розкриваючи його характеристики [5]. Якщо розглянути термін з боку WEB-програмування, то це слова, що відносяться до діяльності сайту чи діяльності компанії, що володіє сайтом. Коректно складене семантичне ядро має важливе значення для пошукової оптимізації, саме на його основі будується пошуковий механізм, без чого не обходиться проєктування сайту чи іншого WEB-застосування [6].

В ряді робіт [7, 8] пропонується використання дисперсійної оцінки для виокремлення ключових слів. Користуючись даною технологією, на основі введених даних у вигляді файлу автоматизовано формується структура цифрового документу для вибору елементу для аналізу, після чого проводиться сегментація по фразах і термінах, терміни лематизуються та їх множина компактифікується. На основі цього проводиться пошук та дисперсійне оцінювання важливості слів у вибраному фрагменті тексту, після чого оцінюється важливість термінів, а їх кількість обмежується відповідно до коефіцієнту щільності ключових слів.

Метою роботи є розробка інформаційної технології, яка забезпечить автоматизоване формування множини ключових семантичних одиниць за множиною слів тексту та показників їх семантичної важливості.

Інформаційна технологія формування множини ключових семантичних одиниць використовує розроблений метод автоматизованого формування семантичного ядра цифрових текстів й у якості вхідних даних має цифровий текст, множину слів тексту та показники їх важливості, а також множину словосполучень тексту та показники їх важливості.

На Етапі 1 виконання інформаційної технології формування множини ключових семантичних одиниць виконується поелементна обробка тексту. Зокрема, проводиться обрахунок загальних параметрів тексту, таких як кількості слів, словосполучень і знаків. А після цього виконується очищення тексту від додаткових символів (знаків, цифр). Далі відбувається зменшення регістру тексту, за результатами чого виконується формування текстового вектору слів та текстового вектору словосполучень.

Етап 2 відповідає за пошук появ семантичних одиниць та перевірку текстового вектору. Спершу проводиться обрахунок позиції по словах для кожної появи кожного унікального слова, а також обрахунок позиції по словах для кожної появи кожного унікального словосполучення. Одночасно проводиться обрахунок позиції по символах для кожної появи кожного унікального слова і обрахунок позиції по символах для кожної появи кожного унікального словосполучення. Після цього виконується формування перевірочного тексту з текстового вектору слів і перевірочного тексту з текстового вектору словосполучень. За результатом, здійснюється обрахунок кількості появ кожного унікального слова та кількості появ кожного унікального словосполучення.

На Етапі 3 проводиться підготовка до застосування методу формування семантичного ядра. Для цього спершу виконується одержання з бази даних значень важливості унікальних слів тексту TF, TFIDF, DE. Також виконується одержання з БД значень важливості унікальних словосполучень тексту TF, TFIDF, DE. Після візуалізації цих даних, здійснюється сортування окремих переліків слів і словосполучень тексту за показниками важливості TF, TFIDF, DE. Останнім кроком виконується одержання від користувача цільового відсотку щільності для тексту.

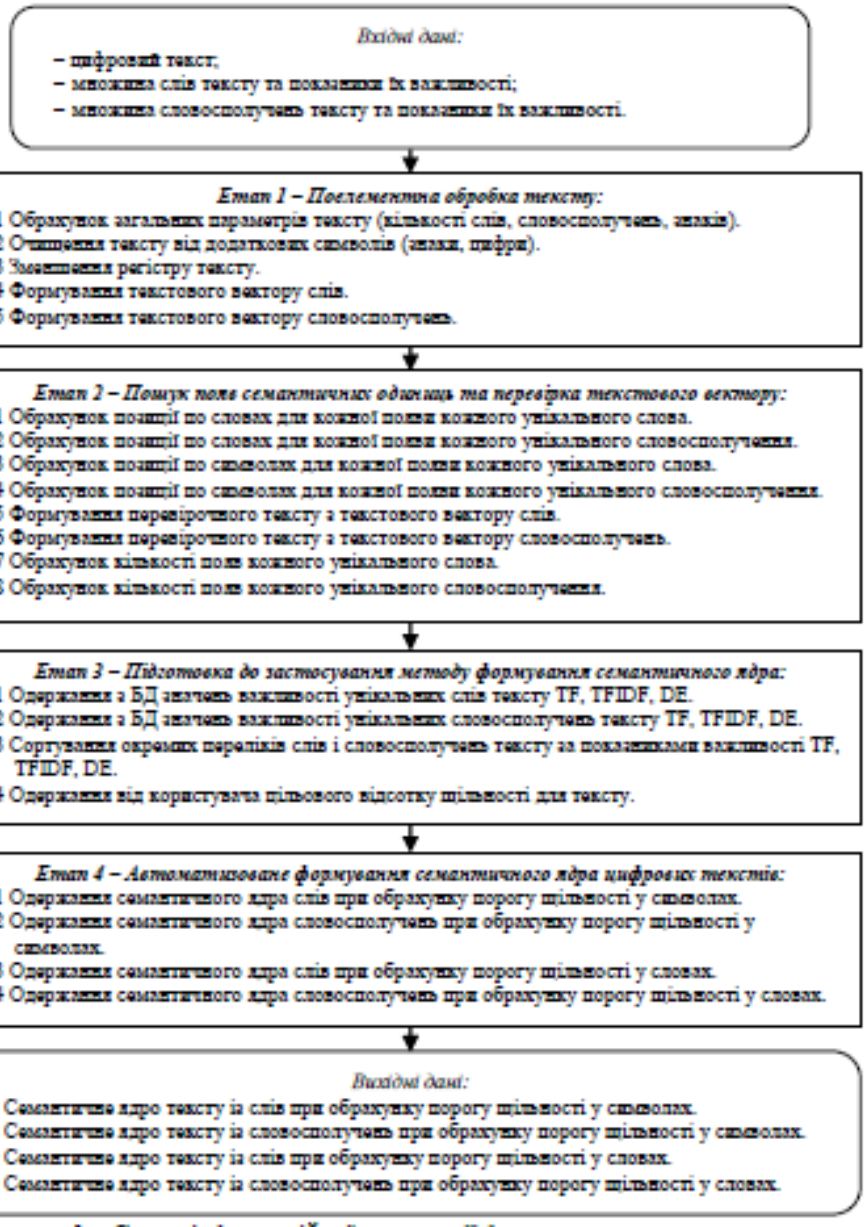


Рисунок 1 – Схема інформаційної технології формування множини ключових семантичних одиниць

Етап 4 безпосередньо відповідає за автоматизоване формування семантичного ядра цифрових текстів методом автоматизованого формування семантичного ядра цифрових текстів. Для цього незалежним чином виконується одержання семантичного ядра слів при обрахунку порогу щільності у символах, семантичного ядра словосполучень при обрахунку порогу щільності у символах, семантичного ядра слів при обрахунку порогу щільності у словах та семантичного ядра словосполучень при обрахунку порогу щільності у словах. Для цього спершу виконується обрахунок числа появ кожного унікального слова та словосполучення у тексті, після чого проводиться послідовний обрахунок порогового відсотку щільності для кожного унікального слова та словосполучення у тексті. За результатом цих дій, виконується послідовне додавання до множини ключових слів та словосполучень, які мають пороговий відсоток щільності вищий за обраний цільовий відсоток для тексту.

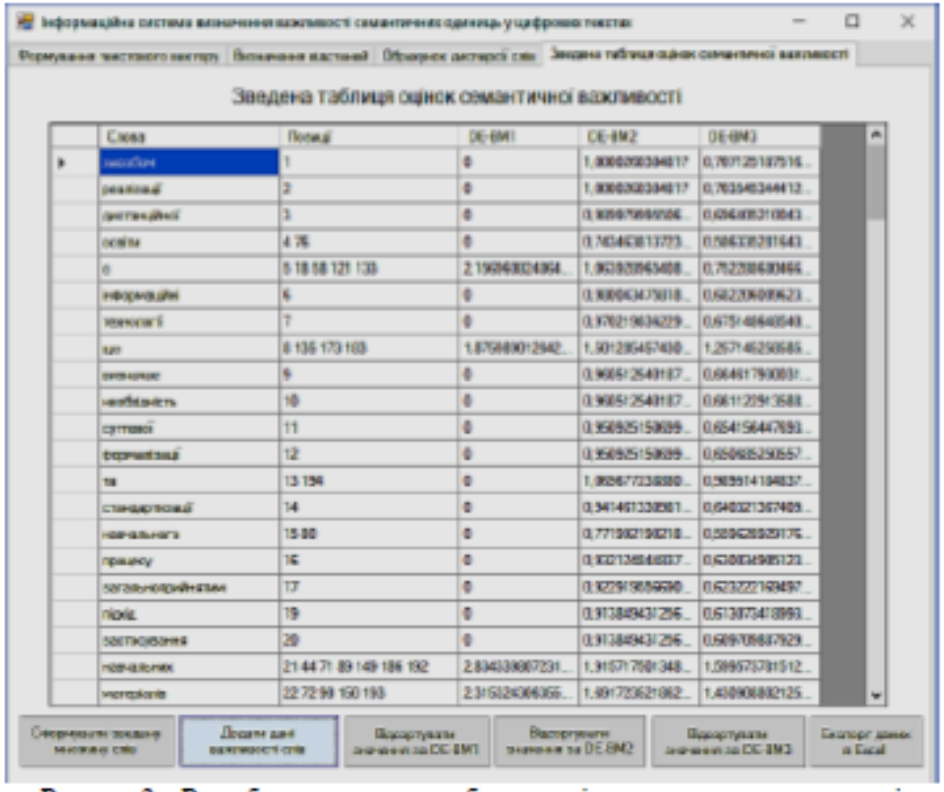


Рисунок 2 – Розроблене програмне забезпеченні для визначення важливості семантичних одиниць у цифрових текстах

Відповідно, вихідні дані формуються як семантичне ядро тексту з таких складових: семантичне ядро тексту із слів при обрахунку порогу щільності у символах, семантичне ядро тексту із словосполучень при обрахунку порогу щільності у символах, семантичне ядро тексту із слів при обрахунку порогу щільності у словах, семантичне ядро тексту із словосполучень при обрахунку порогу щільності у словах.

При застосуванні інформаційної технології автоматизованого формування семантичного ядра цифрових текстів авторами було використано множини термінів цифрових текстів, значення семантичної важливості яких обраховувалось з використанням методу дисперсійного оцінювання [9] шляхом використання відповідних розроблених програмних засобів (Рисунок 2).

В подальшому для формування семантичного ядра шляхом прикладного застосування інформаційної технології автоматизованого формування семантичного ядра цифрових текстів, наведеної вище, було розроблено відповідну програмну систему (Рисунок 3), вихідними даними якої є семантичне ядро тексту із слів і словосполучень.

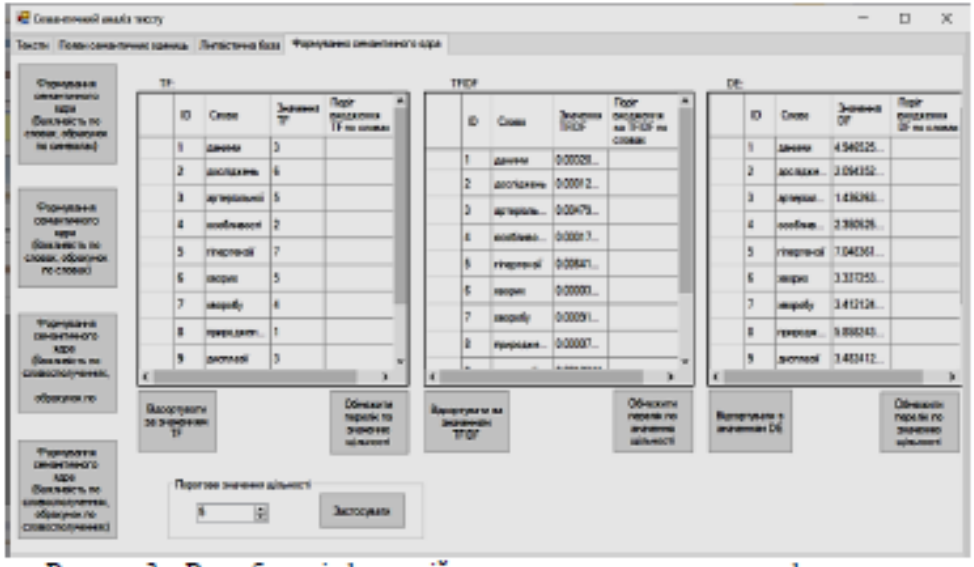


Рисунок 3 – Розроблена інформаційна система автоматизованого формування семантичного ядра цифрових текстів

Прикладом практичного використання створеної інформаційної технології автоматизованого формування семантичного ядра цифрових текстів є використання

адаптивна пропозиція товарів у інтернет-магазині за семантичними ознаками, реалізована авторами у відповідному створеному програмному забезпеченні (Рисунок 4).

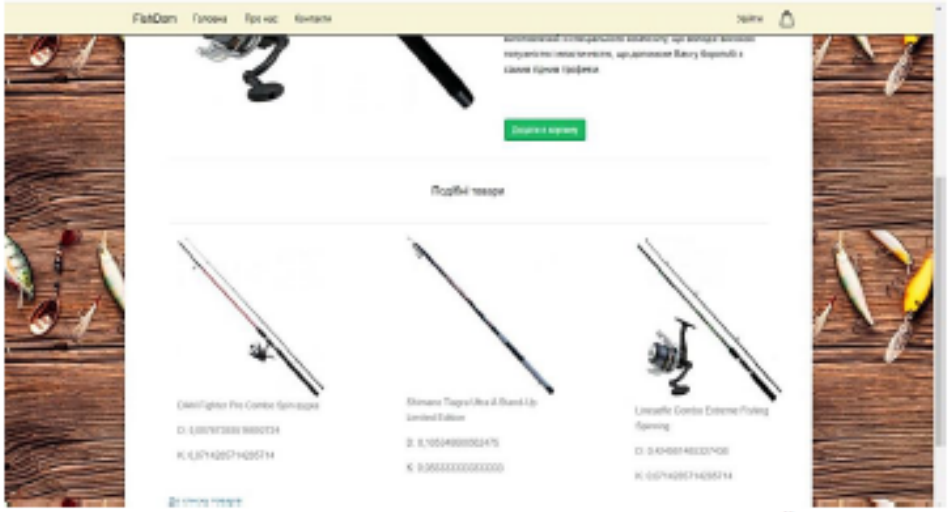


Рисунок 4 – Приклад практичного використання створеної інформаційної технології для адаптивної пропозиції товарів у інтернет-магазині за семантичними ознаками

Таким чином, інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів дозволяє перетворювати вхідні дані у вигляді цифрового тексту, множини слів і словосполучень тексту з показниками їх семантичної важливості в вихідні дані у вигляді зразків семантичного ядра тексту в варіаціях із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Наведені в статті зразки програмного забезпечення, які дозволяють створювати множини термінів цифрових текстів, значення семантичної важливості яких обраховується з використанням методу дисперсійного оцінювання, й формувати семантичне ядро шляхом прикладного застосування розробленої інформаційної технології, а також практичне використання створеної інформаційної технології для адаптивної пропозиції товарів у інтернет-магазині за семантичними ознаками, демонструють повний набір компонентів для практичного вирішення актуальної задачі інформаційних технологій.

Перелік посилань:

1. Keith A. Natural Language Semantics. Blackwell Publishers Ltd. Oxford, 2001. 251 p.
2. Cruse A. Meaning in Language. An Introduction to Semantics and Pragmatics. Second Edition. Oxford University Press. New York, 2004. 137 p.
3. Серажим К. С. Семантичний і семіотичний аспекти аналізу текстів. Вісник Київського національного університету імені Тараса Шевченка. Журналістика. Київ, 2013. № 20. С.34–36.
4. Ventura J. New Techniques for Relevant Word Ranking and Extraction / J. Ventura, J. Silva // Proceedings of the artificial intelligence 13th Portuguese conference on Progress in artificial intelligence, EPIA'07. – Berlin: Springer-Verlag, Berlin, Heidelberg, 2007. – P.691-702.
5. Бармак О. В. Методи автоматизації визначення семантичних термінів у навчальних матеріалах / О. В. Бармак, О. В. Мазурець // Вісник Хмельницького національного університету. Сер.: Технічні науки. Хмельницький. – 2015, №2(223). – С.209-213.
6. Ландэ Д. В. Компактифицированный горизонтальный граф видимости для сети слов / Д. В. Ландэ, А. А. Снарский // Труды Международной научной конференции «Интеллектуальный анализ информации ИАИ-2013. Знания и рассуждения» – КПИ, Киев: 2013. – С.158-164.
7. Залуцька О. О., Мазурець О. В. Інформаційний портрет ключових термінів у цифрових навчальних матеріалах. Матеріали III Міжнародної науково-практичної конференції «Сучасні інформаційні технології та інноваційні методики навчання: досвід, тенденції, перспективи». Тернопіль, 2019. С.120-122.
8. Крак Ю. В. Практична реалізація інформаційної технології автоматизованого визначення множини семантичних термінів в контенті навчальних матеріалів / Ю. В. Крак, О. В. Бармак, О. В. Мазурець // Науковий журнал «Проблеми програмування». Київ, 2018, №2-3. – С.245-254.
9. Мазурець О. В. Інформаційна технологія автоматизованого визначення семантичних термінів в елементах навчальних матеріалів / О. В. Мазурець // Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2018, №3. – С.223-230.

УДК 004

Галкіна Р. І., Багрій Р. О., Скрипник Т. К.

Хмельницький національний університет

ЗАСТОСУВАННЯ АДАПТИВНОГО ПІДХОДУ ДЛЯ РЕАЛІЗАЦІЇ СИСТЕМИ ОПИТУВАНЬ ТА ТЕСТУВАНЬ

У статті розглянуто основні положення традиційного тестування та передумови використання адаптивного тестування. Класична тестування не завжди може вирішити поставлені вимоги сучасним рівнем розвитку систем контролю знань. Тому у подібних випадках використовується адаптивний підхід тестування. Запропонована інформаційна система для проведення опитувань та тестувань, що дала можливість зменшити час, витрачений на проведення тестування, отримання більш точних результатів тестування та спрощення процесу перевірки результатів.

The article considers the main provisions of traditional testing and the prerequisites for the use of adaptive testing. Classical testing cannot always solve the requirements of the current level of development of knowledge control systems. Therefore, in such cases, an adaptive testing approach is used. An information system for conducting surveys and tests is proposed, which has made it possible to reduce the time spent on testing, obtain more accurate test results and simplify the process of verifying results.

На сьогодні автоматизація та комп'ютеризація торкаються майже всіх процесів, що оточують людину. В тому числі й процес збору інформації, а також оцінки якості її отримання. Ці зміни спричинені постійним вдосконаленням систем, що пов'язані з контролем процесу поширення та засвоєння знань.

Опитування – це метод збору соціологічної інформації про досліджуваний об'єкт під час безпосереднього (усне опитування, інтерв'ю) або опосередкованого (письмове опитування, анкетування) спілкування того хто опитує з респондентом [1].

Тестування – система формалізованих завдань, призначених для встановлення освітнього (кваліфікаційного) рівня особи. Педагогічне тестування – форма оцінювання знань учнів, студентів (абітурієнтів), основана на застосуванні педагогічних тестів.

Традиційний тест являє собою стандартизований метод оцінки рівня знань і структури підготовленості людини. При проведенні такого тестування всі відповідають на одні і ті ж завдання протягом однакового часу, в однакових умовах і з однаковими правилами оцінювання відповідей. Одне з головних питань теорії тестів – питання підбору оптимального за деякими критеріями тесту [2]. Кожен тест

Додаток Г

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

МЕТОД АВТОМАТИЗОВАНОГО ФОРМУВАННЯ СЕМАНТИЧНОГО ЯДРА ЦИФРОВИХ ТЕКСТІВ

Виконав:
студент 2 курсу, група КНм-20-1
Купрійчук Владислав Олександрович

Керівник:
викладач кафедри КН
Радюк Павло Михайлович

Мета роботи

Мета кваліфікаційної роботи магістра полягає у розробці методу автоматизованого формування семантичного ядра цифрових текстів, що призначений для автоматизованого формування множини ключових семантичних одиниць за цифровим текстом та множинами слів і словосполучень цього тексту з показниками їх важливості, одержуючи сформовані семантичні ядра за різними показниками важливості, зокрема із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Для досягнення поставленої мети розробки методу автоматизованого формування семантичного ядра цифрових текстів необхідно розв'язати наступні *задачі дослідження*:

1. Провести аналіз предметної області семантичного аналізу цифрових текстів та відомих підходів до автоматизації формування семантичного ядра цифрових текстів.
2. Вдосконалити метод автоматизованого формування семантичного ядра цифрових текстів.
3. Розробити інформаційну технологію автоматизованого формування множини ключових семантичних одиниць.
4. Розробити інформаційну систему автоматизованого формування множини ключових семантичних одиниць.
5. Провести прикладне дослідження методу автоматизованого формування семантичного ядра цифрових текстів у складі інформаційної технології автоматизованого формування множини ключових семантичних одиниць і виконати аналіз результатів використання відповідної інформаційної системи.

Положення новизни та інновації

- **Вдосконалено метод автоматизованого формування семантичного ядра цифрових текстів**, що дозволяє за множиною слів і словосполучень цифрового тексту з зіставленими значеннями показників їх семантичної важливості, а також необхідним відсотком щільності ключових слів у тесті автоматизовано формувати відповідні множини ключових слів та словосполучень тексту за різними показниками важливості.
- **Розроблено нову інформаційну технологію автоматизованого формування множини ключових семантичних одиниць**, що дозволяє з використанням створеного методу автоматизованого формування семантичного ядра цифрових текстів за вхідними даними у вигляді цифрового тексту та множина слів і словосполучень цього тексту з показниками їх важливості одержувати вихідні дані у вигляді сформованих семантичних ядер із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Структура вхідних даних методу побудови розкладів занять за генетичним алгоритмом

Щільність Q є відношенням кількості слів ключових термінів у цифровому матеріалі до загальної кількості слів й становить 0,11-0,15.

Відповідно, до множини ключових термінів автоматично будуть додані елементи з множини з найбільшими значеннями важливості доти, доки справджується рівність:

$$\sum_{i=1}^{n} \frac{K_n x_n}{X_{txt}} \le Q$$

де K_n – кількість появ терміну n в множині M_{TXT}; x_n – кількість слів у терміні n; X_{txt} – загальна кількість слів у тексті; n – поточна кількість термінів у множині ключових термінів .

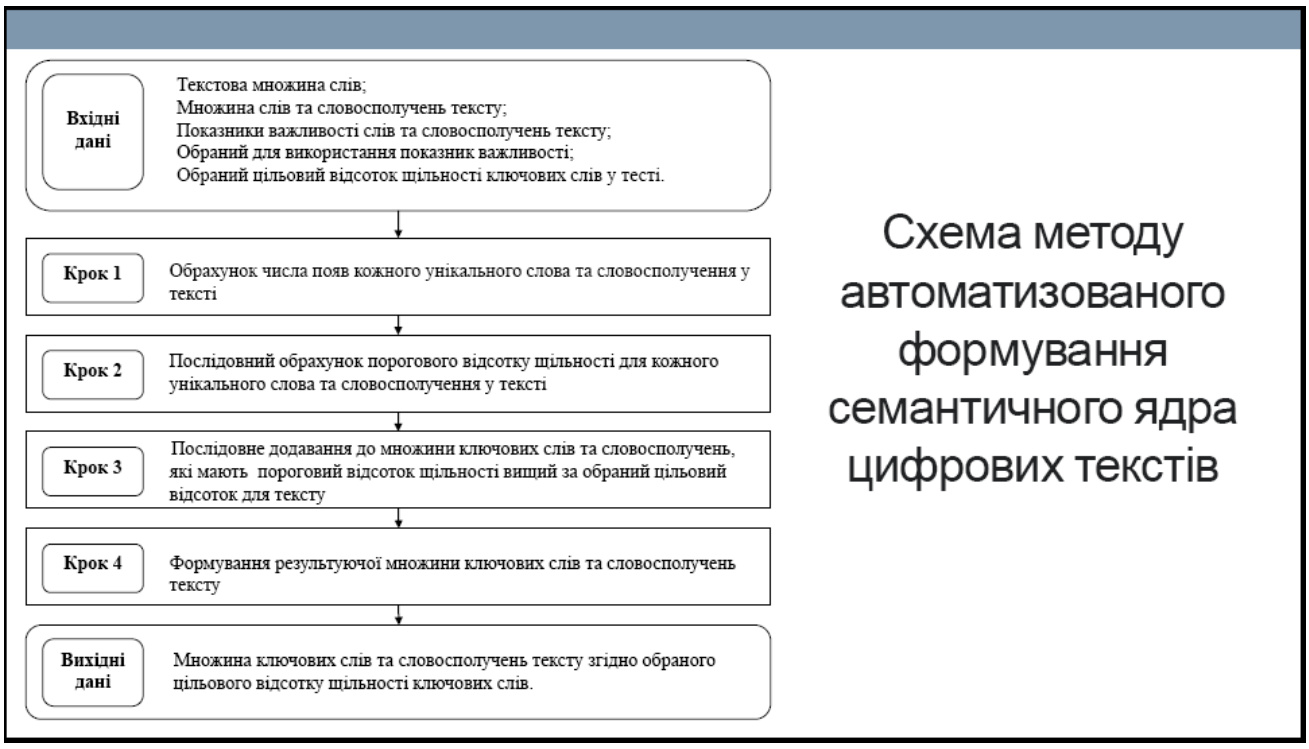
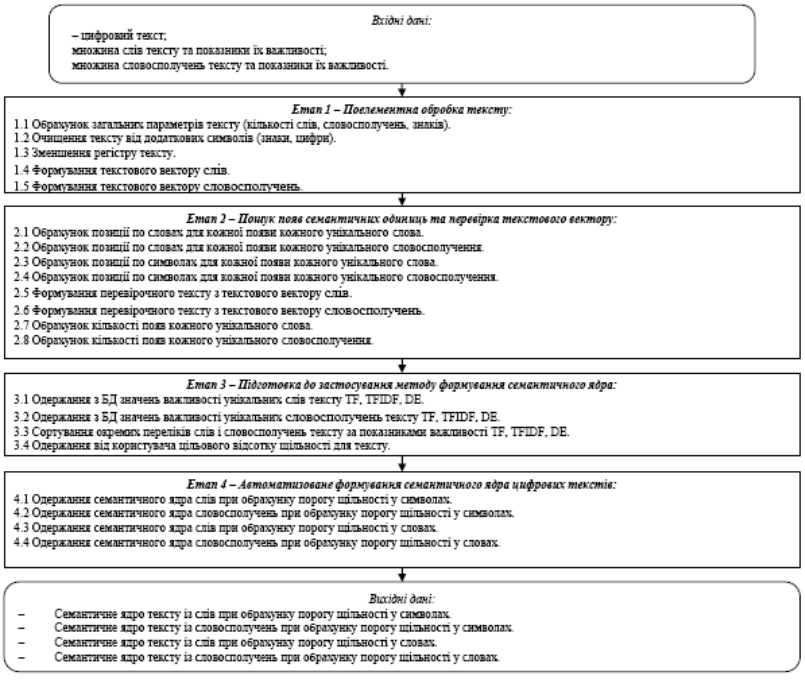
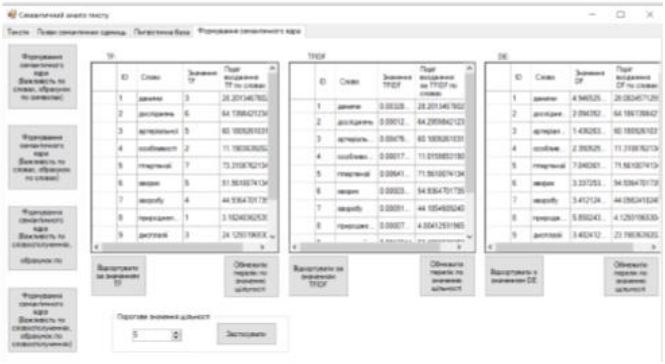


Схема методу автоматизованого формування семантичного ядра цифрових текстів

Схема інформаційної технології автоматизованого формування множини ключових семантичних одиниць



Інформаційна система автоматизованого формування множини ключових семантичних одиниць



Результати аналізу тексту.
Визначено TF, TFIDF, DE та пороги входження
цих значень

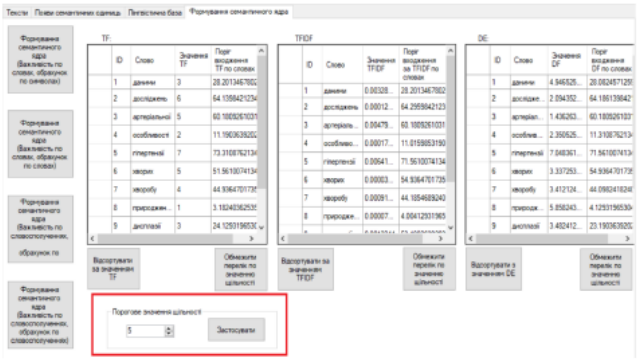
Обраний випадок: семантичне ядро тексту із слів при обрахунку порогу щільності у словах

Дослідження ефективності методу автоматизованого формування семантичного ядра цифрових текстів

До прикладу, за результатом аналізу 45 текстів за значень цільових відсотків щільності ключових слів у текстах 10-15%:

- середня точність формування семантичного ядра тексту із слів при обрахунку порогу щільності у символах склала 82,93%,
- середня точність формування семантичного ядра тексту із словосполучень при обрахунку порогу щільності у символах склала 86,19%,
- середня точність формування семантичного ядра тексту із слів при обрахунку порогу щільності у словах склала 79,63%,
- середня точність формування семантичного ядра тексту із словосполучень при обрахунку порогу щільності у словах склала 83,55%.

Дослідження ефективності методу автоматизованого формування семантичного ядра цифрових текстів



Одержані результати свідчать, що розроблений метод автоматизованого формування семантичного ядра цифрових текстів, який використовує обмеження множин ключових слів і словосполучень за значенням порогу щільності, дозволяє одержати достатньо репрезентативні складові семантичного ядра тексту. Це дає привід стверджувати, що прикладне використання розробленого методу може бути ефективно застосоване при вирішення задач семантичного аналізу текстів відповідно до призначення.

Загальні висновки

Кваліфікаційна робота магістра розв'язує науково-технічну задачу автоматизованого формування семантичного ядра цифрових текстів у вигляді множини ключових семантичних одиниць.

За виконання роботи були поставлені й *вирішені наступні завдання*:

1. Проведено аналіз предметної області семантичного аналізу цифрових текстів та відомих підходів до автоматизації формування семантичного ядра цифрових текстів.

2. Вдосконалено метод автоматизованого формування семантичного ядра цифрових текстів.

3. Розроблено інформаційну технологію автоматизованого формування множини ключових семантичних одиниць.

4. Розроблено інформаційну систему автоматизованого формування множини ключових семантичних одиниць.

5. Проведено прикладне дослідження методу автоматизованого формування семантичного ядра цифрових текстів у складі інформаційної технології автоматизованого формування множини ключових семантичних одиниць і виконано аналіз результатів використання відповідної інформаційної системи.

UNICHECK

Ім'я користувача:
Кафедра КН

ID перевірки:
1009629336

Дата перевірки:
09.12.2021 23:27:05 EET

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
09.12.2021 23:28:27 EET

ID користувача:
100005671

Назва документа: **2021_КРМ_Купрійчук 20211209 4 АНТИПЛАГІАТ**

Кількість сторінок: **78** Кількість слів: **12510** Кількість символів: **97432** Розмір файлу: **4.54 MB** ID файлу: **1009631272**

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

8.69%
Схожість

Найбільша схожість: **4.5%** з джерелом з Бібліотеки (ID файлу: **1009540885**)

2.66% Джерела з Інтернету	34	Сторінка 80
6.48% Джерела з Бібліотеки	102	Сторінка 80

0.19% Цитат

Цитати	3	Сторінка 81
Посилання	1	Сторінка 81

0%
Вилучень

Немає вилучених джерел

Модифікації

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

Замінені символи	9
Підозріле форматування	20 сторінок

Thu Dec 09 22:46:23 EET 2021, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальное совпадение с одним документом 21.0%

Словари проверки: en_US, ru_RU, ua_UA. **Ошибок в документах: 8%**

	Документ		Суммарное совпадение по Базе Данных	
ID: 98672 Название: Метод автоматизованого формування семантичного ядра цифрових текстів Добавлено в БД: 2021-12-09 Авторы: В.О. Купрійчук Руководители: П.М. Радюк Консультанты: Опоненты:	Символы	Лексемы	Символы	Лексемы
	88417	552	22634 (26%)	163 (30%)

Источник плагиата

ID	Описание	Наличие плагиата в документе	
		Символы	Лексемы
95904	Название: ЗВІТ з науково-дослідної практики Добавлено в БД: 2021-09-29 Авторы: Купрійчук В.О. Руководители: Скрипник Т.К. Консультанты: Опоненты:	18826 (21.0%)	130 (24.0%)

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, генерованоголікар системою виявлення текстових збігів/ідентичності/схожості:

Назва: *Метод автоматизованого формування семантичного ядра цифрових текстів*
Автор: *Купрійчук Владислав Олександрович*
Спеціальність: *122 – Комп'ютерні науки*
Освітня програма: *освітньо-професійна*
Науковий керівник: *викладач каф.КН Радюк Павло Михайлович*

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 21% з 26% запозичень вказують на документ автора роботи Купрійчука В.О. та містять Звіт з науково-дослідної практики.
2) За програмою UNICHECK виявлені 8,69%, які є фрагментарними, не більше 4,5% на джерело – містять поширені конструкції, загальновідомі терміни та визначення.

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ ідентичності/схожості, складає 26% і 8,69% відповідно, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи _____ Павло Радюк

Гарант ОП _____ Руслан Багрій

Завідувач кафедри КН _____ Олександр Бармак

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ
кафедра комп'ютерних наук

ВІДГУК ОПОНЕНТА

на кваліфікаційну роботу магістра

гр. КНм-20-1 Купрійчука Владислава Олександровича за темою: *Метод автоматизованого формування семантичного ядра цифрових текстів*

1. **Актуальність обраної теми**

Однією із найбільших проблем на IT ринку, досі не вирішених, є відсутність або ж недосконалість технологій та сервісів автоматичної смислової обробки неструктурованої текстової інформації. Ця проблема ускладнюється ще й тим, що для автоматизованої змістовної обробки цифрових текстів необхідний комплекс методів, які зможуть забезпечити реалізацію алгоритмів і програмного забезпечення повного комп'ютерного лінгвістичного аналізу текстів природною мовою та автоматичної змістовної обробки інформації. Виходячи з наведеного, розробка методів і засобів формування семантичного ядра цифрових текстів у вигляді обмеженої множини ключових слів та словосполучень є актуальною задачею інформаційних технологій.

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

Обрана тема автоматизованого формування семантичного ядра цифрових текстів, в межах якої реалізовані 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 повністю розкрито мету дослідження та поставленні в межах теми завдання дослідження.

4. Наявність наукової новизни

В кваліфікаційній роботі представлена наукова новизна та інновації, відповідні спеціальності 122 «Комп'ютерні науки» в межах обраної області дослідження. Продемонстровано й обґрунтовано результати, які мають наукове та інноваційне значення. Результати дослідження оприлюдненні на науково-практичній конференції.

5. Зміст кожного розділу роботи

Робота містить чотири розділи. У першому розділі проведено аналіз предметної області семантичного аналізу цифрових текстів, досліджено сучасний стан проблеми автоматизації формування семантичного ядра цифрових текстів та поставлені задачі дослідження. Другий розділ присвячено розробці методу автоматизованого формування

семантичного ядра цифрових текстів і інформаційної технології автоматизованого формування множини ключових семантичних одиниць. У третьому розділі виконано розробку інформаційної системи автоматизованого формування множини ключових семантичних одиниць, розглянуто відповідні діаграму класів та структуру бази даних. У четвертому розділі виконано дослідження ефективності розробленого методу автоматизованого формування семантичного ядра цифрових текстів.

6. Ступінь розкриття теми роботи

Тема роботи в достатній мірі обґрунтована й розкрита, 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

У роботі є незначні недоліки. Для дослідження ефективності методу автоматизованого формування семантичного ядра цифрових текстів використано замалу вибірку із 45 цифрових текстів. Наведені дослідження виконувались виключно за умов обрання значень цільових відсотків щільності ключових слів у текстах 10-15%, інших результатів не наведено. Є непослідовні посилання на джерела з переліку посилань. Наведене не вплинуло на практичний результат роботи та одержані положення наукової новизни та інновацій.

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

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

Опонент _____ к.т.н., доцент Оксана ЯШИНА

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ
кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу магістра

гр. КНм-20-1 Купрійчука Владислава Олександровича за темою: *Метод автоматизованого формування семантичного ядра цифрових текстів*

1. Актуальність теми

Методи семантичного аналізу текстів дозволяють комп'ютеру розпізнавати, розуміти та інтерпретувати речення, абзаци та повноцінні документи, усю цифрову інформацію, з якою людина працює кожного дня. Для реалізації обчислювального процесу автоматичної семантичної обробки інформації повинні бути розроблені способи організації високоефективного обчислювального процесу, що забезпечують формування результатів пошуку та аналітичної обробки інформації в реальному масштабі часу. Тому розробка методів і засобів формування семантичного ядра цифрових текстів є актуальною та перспективною задачею семантичного аналізу цифрових текстів.

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

Поставлена у кваліфікаційній роботи мета, пов'язана з автоматизацією формування семантичного ядра цифрових текстів, цілком відповідає предметній області спеціальності 122 «Комп'ютерні науки» та вимогам до кваліфікаційної роботи.

3. Професійні та особистісні якості магістранта

При роботі над кваліфікаційною роботою магістра Купрійчук Владислав Олександрович зарекомендував себе дисциплінованим студентом, який вчасно та пунктуально виконував необхідні етапи дослідження. Як в процесі наукових вишукувань, так і за розробки прикладного програмного забезпечення проявив достатні для одержання одержаного результату компетентності.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Магістрант самостійно виконував всі поставлені задачі. Одержані положення наукової новизни та інновації, означені в роботі, є результатом особистої діяльності магістранта.

5. Наукова новизна та оригінальність запропонованих підходів

В кваліфікаційній роботі магістра представлена наукова новизна та інновації, відповідні спеціальності 122 «Комп'ютерні науки» в межах обраної області дослідження. Продемонстровано й обґрунтовано результати, які мають наукове та інноваційне значення. Вдосконалено метод автоматизованого формування семантичного ядра цифрових текстів,

розроблено нову інформаційну технологію автоматизованого формування множини ключових семантичних одиниць. Результати роботи оприлюдненні на науково-практичній конференції.

6. Ступінь оволодіння методами дослідження

В роботі виявлено достатній рівень оволодіння магістрантом необхідними методами дослідження.

7. Повнота та якість розкриття теми роботи

Тема роботи в достатній мірі обґрунтована й розкрита, проведено аналіз актуальності та відомих досліджень в межах обраної теми, поставлені завдання, які у роботі виконані, а також проведено аналіз результатів прикладного застосування запропонованих засобів автоматизованого формування семантичного ядра цифрових текстів.

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

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

9. Можливість практичного застосування кваліфікаційної роботи, окремих її частин

В роботі виконано розробку експериментальної інформаційної системи автоматизованого формування множини ключових семантичних одиниць, реалізовано модулі та складові для інформаційної системи. Було проведено ряд досліджень, коли для різних цільових відсотків щільності ключових слів у тесті обраховувались компоненти семантичного ядра цифрових текстів за різними способами обрахунку. Одержані результати свідчать, що розроблений метод автоматизованого формування семантичного ядра цифрових текстів, який використовує обмеження множин ключових слів і словосполучень за значенням порогу щільності, дозволяє одержати достатньо репрезентативні складові семантичного ядра тексту. Це дає привід стверджувати, що прикладне використання розробленого методу може бути ефективно застосоване при вирішення задач семантичного аналізу текстів.

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

Враховуючи виявлений рівень виконання та факту забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка *задовільно*.

Науковий керівник _____ викладач каф.КН Радюк Павло Михайлович