Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему *Метод визначення важливості семантичних одиниць у цифрових текстах*

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* _____ *О.О. Войчишин*
Курс, група виконавця — Підпис — Ініціали, прізвище

Керівник: *к.т.н., доцент кафедри ІПЗ* _____ *Ю.В. Форкун*
Науковий ступінь, посада — 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

09 грудня 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 визначення важливості семантичних одиниць у цифрових текстах»*

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

3. Керівник роботи *доцент кафедри ІПЗ Форкун Юрій Вікторович*
(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 визначення важливості семантичних одиниць у цифрових текстах, який дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок важливості семантичних одиниць тексту на основі дисперсійного оцінювання з урахуванням як внутрішніх відстаней між появами унікальних семантичних одиниць, так і початкових, кінцевих та кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту. Також потрібно провести прикладне дослідження ефективності створеного методу визначення важливості семантичних одиниць у цифрових текстах.

Реферат

Кваліфікаційна робота магістра розв’язує науково-технічну задачу автоматизованого визначення важливості семантичних одиниць у цифрових текстах за допомогою методу дисперсійного оцінювання та його модифікацій.

Актуальність теми. Інтелектуальний аналіз текстів займається категоризацією текстів, змінами в колекціях текстів та їх обробкою, пошуком інформації та розробкою засобів представлення інформації для користувача. Категоризація документів вважається процесом зіставлення документів однієї колекції з однією або декількома групами схожих між собою текстів.

Робота з надзвичайно великою кількістю текстової інформації завжди є дуже затратною в часі. Багато компаній і організацій покладаються на методи вилучення інформації для автоматизації ручної роботи за допомогою інтелектуальних алгоритмів, які можуть зменшити витрати та людські зусилля і зробити різні процеси із меншою кількістю помилок.

Семантичний аналіз широко застосовується при рішенні задач інформаційного пошуку, автоматичного перекладу, аналізу змісту, пошуку протиріч, реферування, аналізу інтересів користувачів інформаційної системи, авторства текстів тощо. Тому розробка і вдосконалення методів семантичного аналізу цифрових текстів є актуальним і перспективним напрямком прикладного застосування інформаційних технологій.

Мета і задачі роботи. *Мета кваліфікаційної роботи магістра* – створення методу визначення важливості семантичних одиниць у цифрових текстах, який дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок важливості семантичних одиниць тексту на основі дисперсійного оцінювання з урахуванням як внутрішніх відстаней між появами унікальних семантичних одиниць, так і початкових, кінцевих та кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту.

За результатом виконання роботи були *поставлені та вирішені наступні завдання*:

1. Проведено аналіз предметної області семантичного аналізу текстів, зокрема сучасних методів пошуку ключових семантичних одиниць у цифрових текстах.

2. Вдосконалено метод визначення важливості семантичних одиниць у цифрових текстах.

3. Розроблено інформаційну технологію автоматизованого пошуку семантичних одиниць у цифрових текстах.

4. Розроблено прикладну інформаційну систему для автоматизованого пошуку семантичних одиниць у цифрових текстах.

5. Проведено прикладне дослідження методу визначення важливості семантичних одиниць у цифрових текстах у складі інформаційної технології автоматизованого пошуку семантичних одиниць у цифрових текстах і виконано аналіз результатів використання відповідної інформаційної системи.

Об’єкт дослідження – процес семантичного аналізу цифрових текстів.

Предмет дослідження – інформаційні технології, моделі, методи та засоби для визначення важливості семантичних одиниць у цифрових текстах.

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

Наукова новизна одержаних результатів. В результаті роботи були отримані такі *інновації та положення наукової новизни*:

1. Вдосконалено метод визначення важливості семантичних одиниць у цифрових текстах на основі дисперсійного оцінювання, який відрізняється тим, що на відміну від існуючих дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а й з урахуванням початкових, кінцевих та похідних

кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту.

2. Розроблено нову інформаційну технологію автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, що дозволяє з використанням створеного методу визначення важливості семантичних одиниць у цифрових текстах за вхідними даними у вигляді вхідного цифрового тексту як відповідної впорядкованої множини символів та параметрами налаштувань одержувати вихідні дані у вигляді трьох множин ключових семантичних одиниць вхідного цифрового тексту за оцінками модифікацій дисперсійного оцінювання, які дозволяють виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а й із урахуванням початкових, кінцевих і кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту, а також сформованої зведеної таблиці оцінок семантичної важливості ключових семантичних одиниць за цими оцінками модифікацій дисперсійного оцінювання.

3. Розроблено нову інформаційну систему для автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, що дозволяє за створеною інформаційною технологією в результаті обробки вхідного цифрового тексту у вигляді відповідної впорядкованої множини символів виконувати автоматизоване визначення множин ключових семантичних одиниць за оцінками модифікацій дисперсійного оцінювання DE-BM1, DE-BM2 і DE-BM3, також формувати відповідну зведену таблицю оцінок семантичної важливості ключових семантичних одиниць вхідного цифрового тексту.

Практичне значення одержаних результатів. Для прикладного дослідження розроблених методу визначення важливості семантичних одиниць у цифрових текстах та інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах в розділі було створено інформаційну систему, що забезпечує відповідний функціонал. Інформаційна система дозволяє в результаті обробки вхідного цифрового тексту у вигляді

відповідної впорядкованої множини символів виконувати автоматизоване визначення множин ключових семантичних одиниць за оцінками модифікацій дисперсійного оцінювання DE-BM1, DE-BM2 і DE-BM3, також формувати відповідну зведену таблицю оцінок семантичної важливості ключових семантичних одиниць вхідного цифрового тексту. Інформаційна система автоматизованого пошуку семантичних одиниць у цифрових текстах не потребує використання бази даних й складається із чотирьох модулів: модуля попередньої обробки тексту та формування текстового вектору, модуля визначення відстаней між появами семантичних одиниць, модуля дисперсійного оцінювання важливості семантичних одиниць і модуля формування зведеної таблиці ключових семантичних одиниць.

Дослідження ефективності методу визначення важливості семантичних одиниць у цифрових текстах виявило, що при пошуку ключових семантичних одиниць у текстах обсягом від 300 до 500 слів найвищу ефективність (61,82%) продемонстрував метод дисперсійного оцінювання модифікації DE-BM2, проте метод дисперсійного оцінювання модифікації DE-BM3 теж виявив спів ставня результати (57,14%). Водночас класичний метод дисперсійного оцінювання модифікації DE-BM1 виявив значно гірші результати (39,15%). Це пояснюється великою кількістю семантичних одиниць у таких текстах, що мають низьку кількість появ, і відповідно низьку кількість відстаней між появами семантичних одиниць для дисперсійного обрахунку класичним методом.

При пошуку ж ключових семантичних одиниць у текстах обсягом від 500 до 2000 слів всі модифікації дисперсійного оцінювання продемонстрували подібні результати (DE-BM1 68,25%, DE-BM2 74,59% DE-BM3 73,92%). Це пояснюється тим, що потенційні ключові семантичні одиниці у таких текстах мають достатньо велику кількість появ, і відповідно велику кількість відстаней між появами семантичних одиниць для ефективного дисперсійного обрахунку навіть класичним методом.

Проведені дослідження дозволяють зробити висновок про ефективність використання розробленого методу визначення важливості семантичних

одиниць в цифрових текстах для пошуку ключових семантичних одиниць модифікацією DE-BM3 при семантичному аналізі цифрових текстів, особливо – невеликих за обсягом. Одержані результати можуть бути практично використані при вирішенні прикладних завдань визначення важливості семантичних одиниць та пошуку ключових семантичних одиниць у цифрових текстах, наприклад, при вирішенні задачі адаптивної пропозиції товарів у інтернет-магазині за семантичними ознаками.

Апробація результатів кваліфікаційної роботи магістра та публікації. Основні наукові і практичні результати кваліфікаційної роботи магістра доповідались у доповіді за темою «Інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів» на XIII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2021» (15-16 жовтня 2021 року); 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 завдання, реферату, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань із 36 найменувань та 5 додатків. Загальний обсяг кваліфікаційної роботи магістра становить 103 сторінки, з них 83 сторінки основного тексту та 20 сторінок додатків. У роботі наведено 35 рисунків та 2 таблиці.

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

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
DE	Disperce Evaluation
MS	Microsoft
SEO	Search engine optimization
PCA	Principal component analysis
TF	Term Frequency
TF-IDF	Term Frequency – Inverse Document Frequency
ЛСА	Латентно-семантичний аналіз
ІС	Інформаційна система
ІТ	Інформаційні технології
КН	Комп’ютерні науки
ПЗ	Пояснювальна записка
ПП	Програмний продукт
ОС	Операційна система
ПК	Персональний комп’ютер

Вступ

Кваліфікаційна робота магістра розв’язує науково-технічну задачу автоматизованого визначення важливості семантичних одиниць у цифрових текстах за допомогою методу дисперсійного оцінювання та його модифікацій.

Актуальність теми. Інтелектуальний аналіз текстів займається категоризацією текстів, змінами в колекціях текстів та їх обробкою, пошуком інформації та розробкою засобів представлення інформації для користувача. Категоризація документів вважається процесом зіставлення документів однієї колекції з однією або декількома групами схожих між собою текстів.

Робота з надзвичайно великою кількістю текстової інформації завжди є дуже затратною в часі. Багато компаній і організацій покладаються на методи вилучення інформації для автоматизації ручної роботи за допомогою інтелектуальних алгоритмів, які можуть зменшити витрати та людські зусилля і зробити різні процеси із меншою кількістю помилок.

Семантичний аналіз широко застосовується при рішенні задач інформаційного пошуку, автоматичного перекладу, аналізу змісту, пошуку протиріч, реферування, аналізу інтересів користувачів інформаційної системи, авторства текстів тощо. Тому розробка і вдосконалення методів семантичного аналізу цифрових текстів є актуальним і перспективним напрямком прикладного застосування інформаційних технологій.

Мета і задачі роботи. *Мета кваліфікаційної роботи магістра* – створення методу визначення важливості семантичних одиниць у цифрових текстах, який дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок важливості семантичних одиниць тексту на основі дисперсійного оцінювання з урахуванням як внутрішніх відстаней між появами унікальних семантичних одиниць, так і початкових, кінцевих та кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту.

За результатом виконання роботи були *поставлені та вирішені наступні завдання*:

1. Проведено аналіз предметної області семантичного аналізу текстів, зокрема сучасних методів пошуку ключових семантичних одиниць у цифрових текстах.

2. Вдосконалено метод визначення важливості семантичних одиниць у цифрових текстах.

3. Розроблено інформаційну технологію автоматизованого пошуку семантичних одиниць у цифрових текстах.

4. Розроблено прикладну інформаційну систему для автоматизованого пошуку семантичних одиниць у цифрових текстах.

5. Проведено прикладне дослідження методу визначення важливості семантичних одиниць у цифрових текстах у складі інформаційної технології автоматизованого пошуку семантичних одиниць у цифрових текстах і виконано аналіз результатів використання відповідної інформаційної системи.

Об’єкт дослідження – процес семантичного аналізу цифрових текстів.

Предмет дослідження – інформаційні технології, моделі, методи та засоби для визначення важливості семантичних одиниць у цифрових текстах.

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

Наукова новизна одержаних результатів. В результаті роботи були отримані такі *інновації та положення наукової новизни*:

1. Вдосконалено метод визначення важливості семантичних одиниць у цифрових текстах на основі дисперсійного оцінювання, який відрізняється тим, що на відміну від існуючих дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а й з урахуванням початкових, кінцевих та похідних

кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту.

2. Розроблено нову інформаційну технологію автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, що дозволяє з використанням створеного методу визначення важливості семантичних одиниць у цифрових текстах за вхідними даними у вигляді вхідного цифрового тексту як відповідної впорядкованої множини символів та параметрами налаштувань одержувати вихідні дані у вигляді трьох множин ключових семантичних одиниць вхідного цифрового тексту за оцінками модифікацій дисперсійного оцінювання, які дозволяють виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а й із урахуванням початкових, кінцевих і кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту, а також сформованої зведеної таблиці оцінок семантичної важливості ключових семантичних одиниць за цими оцінками модифікацій дисперсійного оцінювання.

3. Розроблено нову інформаційну систему для автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, що дозволяє за створеною інформаційною технологією в результаті обробки вхідного цифрового тексту у вигляді відповідної впорядкованої множини символів виконувати автоматизоване визначення множин ключових семантичних одиниць за оцінками модифікацій дисперсійного оцінювання DE-BM1, DE-BM2 і DE-BM3, також формувати відповідну зведену таблицю оцінок семантичної важливості ключових семантичних одиниць вхідного цифрового тексту.

Практичне значення одержаних результатів. Для прикладного дослідження розроблених методу визначення важливості семантичних одиниць у цифрових текстах та інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах в розділі було створено інформаційну систему, що забезпечує відповідний функціонал. Інформаційна система дозволяє в результаті обробки вхідного цифрового тексту у вигляді

відповідної впорядкованої множини символів виконувати автоматизоване визначення множин ключових семантичних одиниць за оцінками модифікацій дисперсійного оцінювання DE-BM1, DE-BM2 і DE-BM3, також формувати відповідну зведену таблицю оцінок семантичної важливості ключових семантичних одиниць вхідного цифрового тексту. Інформаційна система автоматизованого пошуку семантичних одиниць у цифрових текстах не потребує використання бази даних й складається із чотирьох модулів: модуля попередньої обробки тексту та формування текстового вектору, модуля визначення відстаней між появами семантичних одиниць, модуля дисперсійного оцінювання важливості семантичних одиниць і модуля формування зведеної таблиці ключових семантичних одиниць.

Дослідження ефективності методу визначення важливості семантичних одиниць у цифрових текстах виявило, що при пошуку ключових семантичних одиниць у текстах обсягом від 300 до 500 слів найвищу ефективність (61,82%) продемонстрував метод дисперсійного оцінювання модифікації DE-BM2, проте метод дисперсійного оцінювання модифікації DE-BM3 теж виявив спів ставня результати (57,14%). Водночас класичний метод дисперсійного оцінювання модифікації DE-BM1 виявив значно гірші результати (39,15%). Це пояснюється великою кількістю семантичних одиниць у таких текстах, що мають низьку кількість появ, і відповідно низьку кількість відстаней між появами семантичних одиниць для дисперсійного обрахунку класичним методом.

При пошуку ж ключових семантичних одиниць у текстах обсягом від 500 до 2000 слів всі модифікації дисперсійного оцінювання продемонстрували подібні результати (DE-BM1 68,25%, DE-BM2 74,59% DE-BM3 73,92%). Це пояснюється тим, що потенційні ключові семантичні одиниці у таких текстах мають достатньо велику кількість появ, і відповідно велику кількість відстаней між появами семантичних одиниць для ефективного дисперсійного обрахунку навіть класичним методом.

Проведені дослідження дозволяють зробити висновок про ефективність використання розробленого методу визначення важливості семантичних

одиниць в цифрових текстах для пошуку ключових семантичних одиниць модифікацією DE-BM3 при семантичному аналізі цифрових текстів, особливо – невеликих за обсягом. Одержані результати можуть бути практично використані при вирішенні прикладних завдань визначення важливості семантичних одиниць та пошуку ключових семантичних одиниць у цифрових текстах, наприклад, при вирішенні задачі адаптивної пропозиції товарів у інтернет-магазині за семантичними ознаками.

Апробація результатів кваліфікаційної роботи магістра та публікації. Основні наукові і практичні результати кваліфікаційної роботи магістра доповідались у доповіді за темою «Інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів» на XIII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2021» (15-16 жовтня 2021 року); 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 завдання, реферату, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань із 36 найменувань та 5 додатків. Загальний обсяг кваліфікаційної роботи магістра становить 103 сторінки, з них 83 сторінки основного тексту та 20 сторінок додатків. У роботі наведено 35 рисунків та 2 таблиці.

Розділ 1
Характеристика предметної області семантичного аналізу текстів

1.1 Аналіз предметної області

Впродовж еволюції людства текст став одним із способів подання інформації та формою людської комунікації. Витвір, утворений внаслідок мовленнєвого процесу, який є завершеним та об’єктивованим, має вигляд письмового документа та літературно опрацьований відповідно до типу документа називається текстом. Текст складається із заголовка та низки особливих одиниць, які об'єднані різними типами граматичного, логічного, стилістичного і лексичного зв'язку, він має певну цілеспрямованість та прагматичне становлення [30] .

Сьогодні цифрові технології стрімко впроваджуються в усі сфери життя людини по всьому світі: від промислових виробництв до взаємодії між людьми, від предметів побуду до збереження інформації [2]. Одним із ключових інструментів цифровізації є цифрова платформа, вона забезпечує інформаційний обмін та транзакції між великою кількістю користувачів і є сукупністю технологічних рішень [3] .

Таким чином, текст, який можна прочитати на різних цифрових платформах називається цифровим текстом.

Семантика надає правила для інтерпретації синтаксису, які не надають значення напряму, але обмежують можливі інтерпретації задекларованого [4]. Семантична модель включає в себе слово, визначення слова, поєднання його з іншими словами та утворення з нього фраз і речень. Семантичний аналіз тексту є одним із способів аналізу тексту. На відміну від лексичного та синтаксичного аналізу, семантика орієнтована на змістовну інтерпретацію, пов'язана з внутрішнім представленням сенсу описаних об'єктів [5]. Семантичний аналіз тексту є складною математичною задачею, рішення якої можна застосувати при створенні штучного інтелекту, при цьому ускладняється необхідність обробки природної мови [6].

За своєю суттю семантичний аналіз схожий на інтелектуальний аналіз тексту, який є одним з напрямків штучного інтелекту та інтелектуального аналізу даних (Data Mining). Відмінність наведених вище напрямів полягає не тільки в кінцевих методах, а і в тому, що інтелектуальний аналіз даних працює із сховищами та базами даних, а не електронними бібліотеками та корпусами текстів, як це робить інтелектуальний аналіз текстів.

Інтелектуальний аналіз текстів займається категоризацією текстів, змінами в колекціях текстів та їх обробкою, пошуком інформації та розробкою засобів представлення інформації для користувача. Категоризація документів вважається процесом зіставлення документів однієї колекції з однією або декількома групами схожих між собою текстів.

У випадку з класифікацією документів, інформаційний аналіз тексту відносить тексти до раніше визначених класів. Для цього користувач має надати системі інформаційного аналізу текстів переліки класів та зразки документів, які належать цим класам.

Інший випадок категоризації документів – кластеризація документів. Система самостійно визначає множину кластерів, за якими вона може розподілити тексти. Таким чином, користувачу необхідно задати лише кількість кластерів, на яку він бажає розбити оброблювану колекцію [7].

Робота з надзвичайно великою кількістю текстової інформації завжди є дуже затратною в часі. Багато компаній і організацій покладаються на методи вилучення інформації для автоматизації ручної роботи за допомогою інтелектуальних алгоритмів, які можуть зменшити витрати та людські зусилля і зробити різні процеси із меншою кількістю помилок [8].

Спрощення текстів і є проблемою застосування вилучення інформації. Його мета полягає у створенні структурованого уявлення про інформацію, яка є у тексті. Для вилучення інформації характерні наступні задачі:

– визначення термінології: характерний для даного тексту ключових слів та словосполучень;

– розпізнавання іменованих сутностей: географічні назви, імена людей, тощо;

– виявлення відносин між суб’єктами;

– кореферентний аналіз: виявлення кореферентності та анафоричних зв’язків між текстовими сутностями [9].

До статистичних показників семантичного аналізу відносять кількість символів, враховуючи і не враховуючи пробіли, кількість слів, у тому числі унікальних та значимих, кількість граматичних помилок, відсоток академічної та класичної нудоти, кількість води та стоп-слів, слів які не мають ніякого смислового змісту [6].

Синтаксична одиниця, утворена поєднанням двох або більше повнозначних слів, які поєднуються підрядним зв’язком, називається словосполученням [10]. Словосполучення завжди складається з головного і залежного слова. Воно має різний ступінь семантичної єдності, тому й виділяють синтаксично вільні та синтаксично нечленовані словосполучення. Синтаксично вільні мають компонент який виступає як окремий член речення, а синтаксично нечленовані є граматичною єдністю, яка виконує роль одного члена речення [11].

Слово або словосполучення, яке означає чітко окреслене спеціальне поняття якої-небудь галузі науки, техніки, мистецтва, суспільного життя тощо називається терміном [12]. Семантичне творення термінів передбачає використання загальновживаної лексики у термінологічному значенні або надання термінологічній лексиці інших галузей нових термінологічних значень [13].

Слово в тексті, яке в сукупності з іншими ключовими словами, дає високорівневий опис змісту текстового документу та виявляє його тематику, називається ключовим словом. Ключові слова характеризуються тим, що:

– є найбільш вживаними і визначають ознаку предмета, стан або дію;

– представлені значимою лексикою;

– пов’язані між собою сіткою семантичних зв’язків, перетину визначень;

– набір з 5-15 або 8-10 слів, який відповідає об’єму оперативної пам’яті людини [14].

Семантичний аналіз широко застосовується при рішенні задач інформаційного пошуку, автоматичного перекладу, аналізу змісту, пошуку протирічь, реферування, аналізу інтересів користувачів інформаційної системи, авторства текстів тощо.

1.2 Аналіз існуючого програмного забезпечення предметної області

У зв'язку з тим, що працівникам різних сфер все частіше доводиться працювати з великими обсягами неструктурованих текстів, попит на програмне забезпечення, яке надає інструменти для семантичного аналізу тексів, росте.

Одними з таких відомим програмних забезпечень для семантичного аналізу текстів є:

– Advego;
– Serpstat;
– Istio;
– Wordstat.

Семантичний аналізатор тексту Advego [15] є професійним інструментом для оцінки якості текстів, seo-оптимізації статей та пошуку ключових слів. Він перевіряє кількість символів, нудоту, кількість води і щільність ключових слів та фраз в онлайн-режимі. На рисунку 1.1 показано інтерфейс аналізатору Advego для введення тексту для аналізу.

Результат аналізу тексту виводиться у вигляді таблиці з статистичними показниками (рис. 1.2) і таблиці з ключовими словами (рис.1.3).

Таким чином, якщо в тексті багато стоп-слів, води і мало ключових слів, то якість статті визначиться як низька. Проте якщо кількість ключових слів буде перевищена, то є можливість, що дана стаття навряд чи буде показана у результаті пошуку.

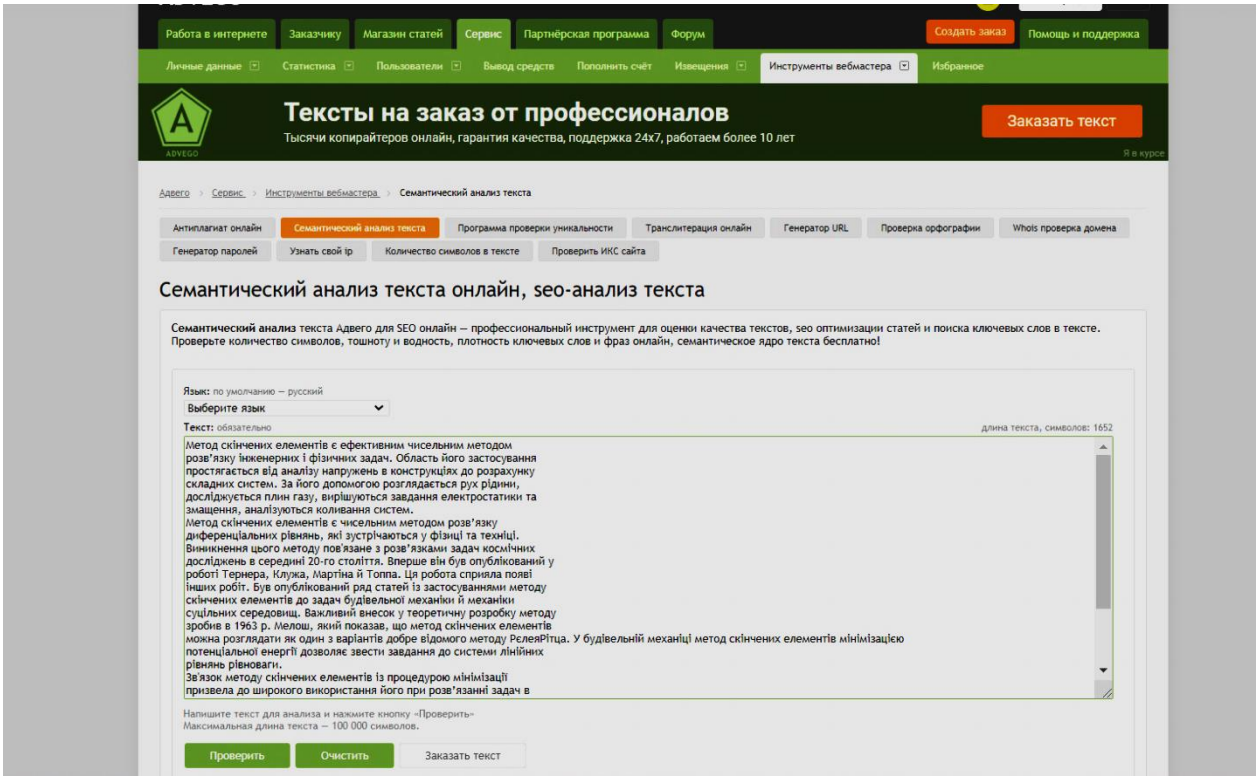
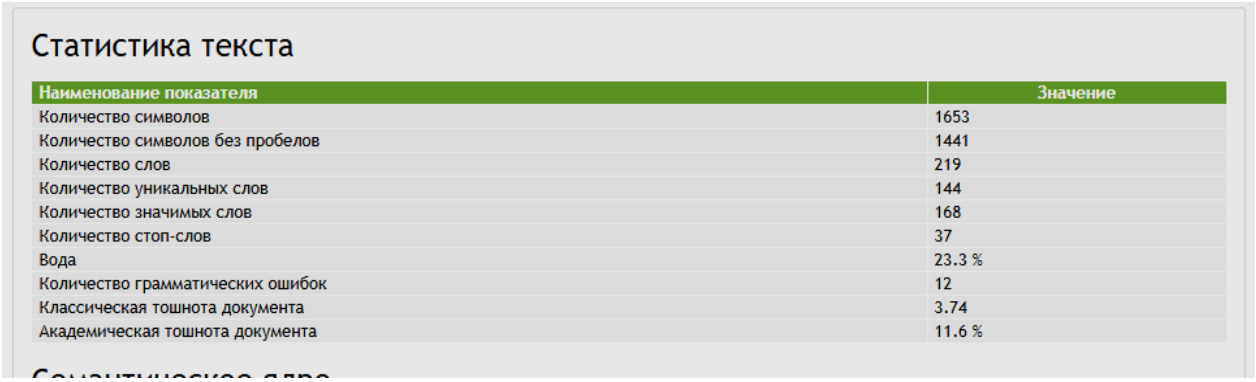


Рисунок 1.1 – Введення цифрового тексту в аналізатор тексту Advego для аналізу

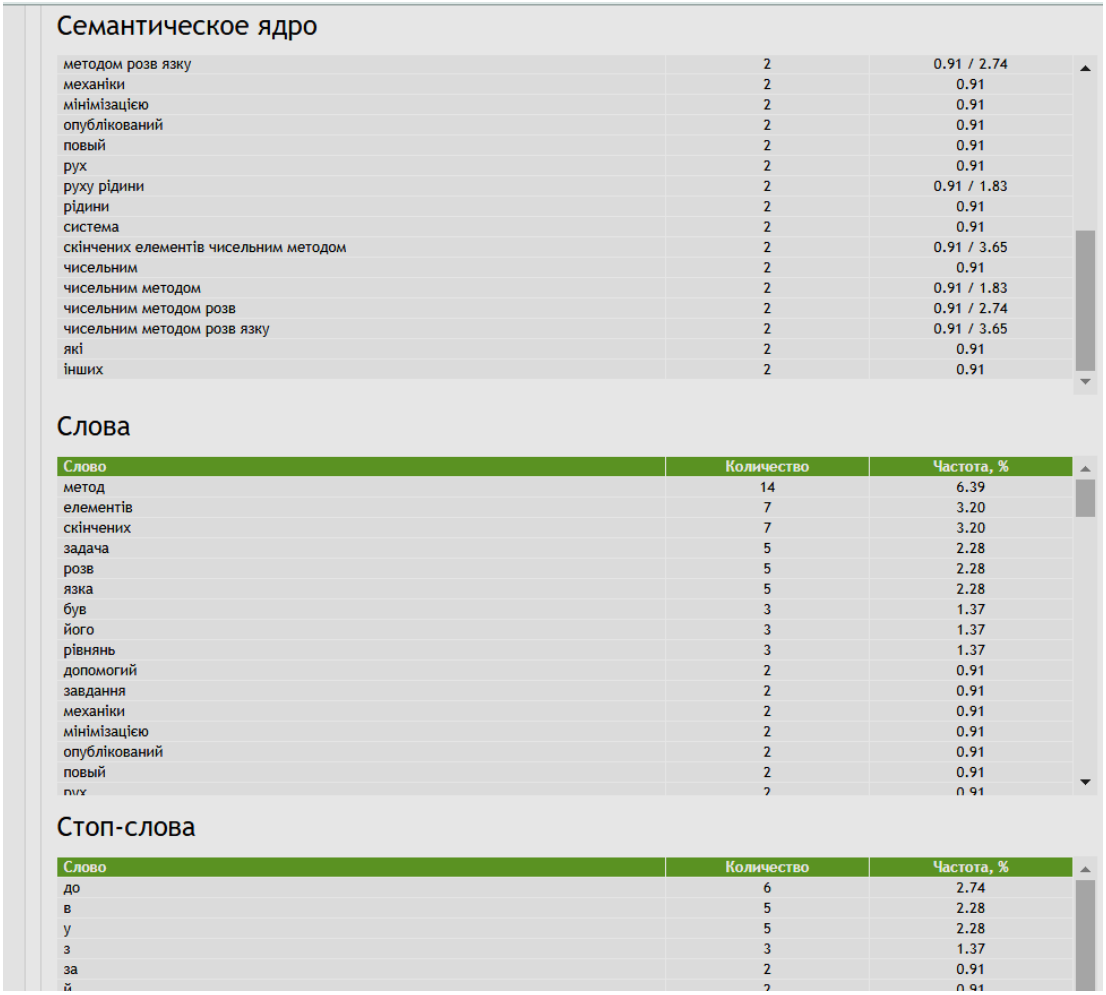


Статистика текста

Наименование показателя	Значение
Количество символов	1653
Количество символов без пробелов	1441
Количество слов	219
Количество уникальных слов	144
Количество значимых слов	168
Количество стоп-слов	37
Вода	23.3 %
Количество грамматических ошибок	12
Классическая тошнота документа	3.74
Академическая тошнота документа	11.6 %

Рисунок 1.2 – Статистичний результат [15]

Таким чином, SEO-аналізатор Advego є системою для семантичного аналізу текстів та визначає наступні показники: частотність слів, щільність ключових слів, відсоток ключових фраз, кількість стоп-слів; обсяг тексту: кількість символів із пробілами й без пробілів; кількість слів: унікальних, значимих, загальну кількість; відсоток води; нудоту тексту: класичну й академічну; кількість граматичних помилок [15].



Семантическое ядро

методом розв язку	2	0.91 / 2.74
механіки	2	0.91
мінімізацією	2	0.91
опублікований	2	0.91
повый	2	0.91
рух	2	0.91
руху рідини	2	0.91 / 1.83
рідини	2	0.91
система	2	0.91
скінчених елементів чисельним методом	2	0.91 / 3.65
чисельним	2	0.91
чисельним методом	2	0.91 / 1.83
чисельним методом розв	2	0.91 / 2.74
чисельним методом розв язку	2	0.91 / 3.65
які	2	0.91
інших	2	0.91

Слова

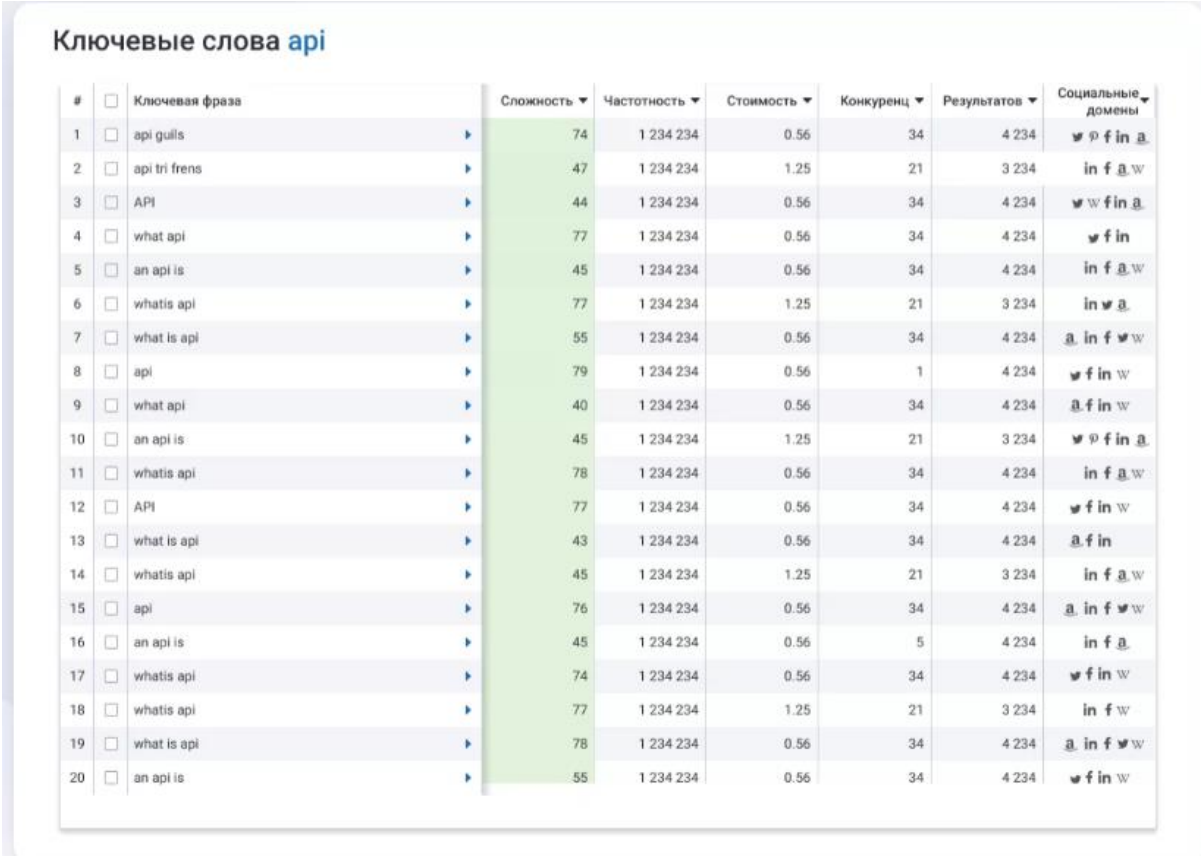
Слово	Количество	Частота, %
метод	14	6.39
елементів	7	3.20
скінчених	7	3.20
задача	5	2.28
розв	5	2.28
язка	5	2.28
був	3	1.37
його	3	1.37
рівнянь	3	1.37
допомогий	2	0.91
завдання	2	0.91
механіки	2	0.91
мінімізацією	2	0.91
опублікований	2	0.91
повый	2	0.91
рух	2	0.91

Стоп-слова

Слово	Количество	Частота, %
до	6	2.74
в	5	2.28
у	5	2.28
з	3	1.37
за	2	0.91
й	2	0.91

Рисунок 1.3 – Результат у вигляді переліку ключових слів [15]

Serpstat є багатофункціональною SEO-платформою, яка дає можливість автоматизувати та підвищити ефективність роботи спеціаліста з пошукового маркетингу. Одною з функцій даної платформи є аналіз тексту за URL. Платформа перевіряє які пошукові запити використовуються з ключовими словами, знаходить семантичні зв'язки, аналізує і пропонує найбільш прибуткові ключові слова для SEO, шукає найбільш підходящі ключові слова, приклади контенту по заданій темі, оцінює значимість ключових фраз та формує повномасштабний звіт про аналізи [16]. Один з результатів аналізу SEO-платформи Serpstat показаний на рисунку 1.4.



Ключевые слова api

#	Ключевая фраза	Сложность	Частотность	Стоимость	Конкуренц	Результатов	Социальные домены
1	api guils	74	1 234 234	0.56	34	4 234	
2	api tri frens	47	1 234 234	1.25	21	3 234	
3	API	44	1 234 234	0.56	34	4 234	
4	what api	77	1 234 234	0.56	34	4 234	
5	an api is	45	1 234 234	0.56	34	4 234	
6	whatis api	77	1 234 234	1.25	21	3 234	
7	what is api	55	1 234 234	0.56	34	4 234	
8	api	79	1 234 234	0.56	1	4 234	
9	what api	40	1 234 234	0.56	34	4 234	
10	an api is	45	1 234 234	1.25	21	3 234	
11	whatis api	78	1 234 234	0.56	34	4 234	
12	API	77	1 234 234	0.56	34	4 234	
13	what is api	43	1 234 234	0.56	34	4 234	
14	whatis api	45	1 234 234	1.25	21	3 234	
15	api	76	1 234 234	0.56	34	4 234	
16	an api is	45	1 234 234	0.56	5	4 234	
17	whatis api	74	1 234 234	0.56	34	4 234	
18	whatis api	77	1 234 234	1.25	21	3 234	
19	what is api	78	1 234 234	0.56	34	4 234	
20	an api is	55	1 234 234	0.56	34	4 234	

Рисунок 1.4 – Результат пошуку ключових слів платформи Serpstat [16]

Сервіс Istio використовується для семантичного аналізу тексту, сервіс оцінює насиченість тексту ключовими словами, кількість води та спаму. На рисунку 1.5 показано інтерфейс для введення тексту.

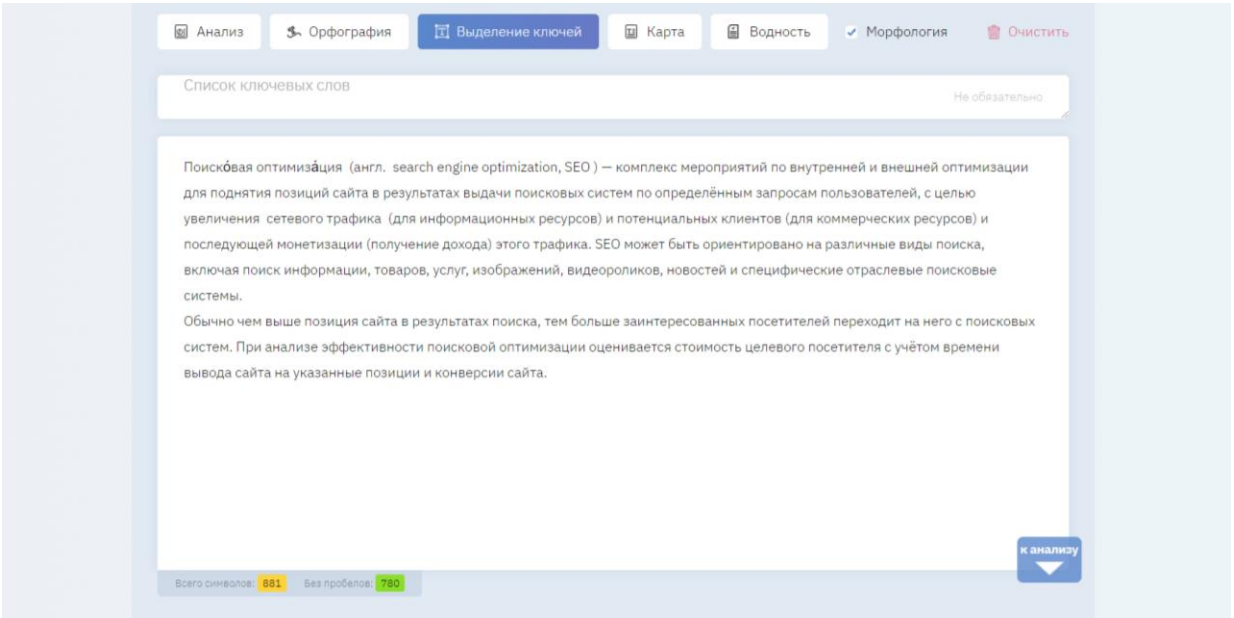


Рисунок 1.5 – Введення тексту для аналізу на сервісі Istio [16]

Семантичний аналіз тексту від Istio дозволяє переглянути результат аналізу у вигляді переліку ключових слів. Сервіс відкидає усі стоп-слова, але також є можливість переглянути результат аналізу разом із стоп-словами. У результаті аналізу виводиться кількість слів, релевантність, відсоток у ядрі та відсоток у тексті. Також вираховуються кількість символів з пробілами та без, кількість слів, води, класична нудота, словник, словник ядра, тематика і топ 10 слів (рис. 1.6).

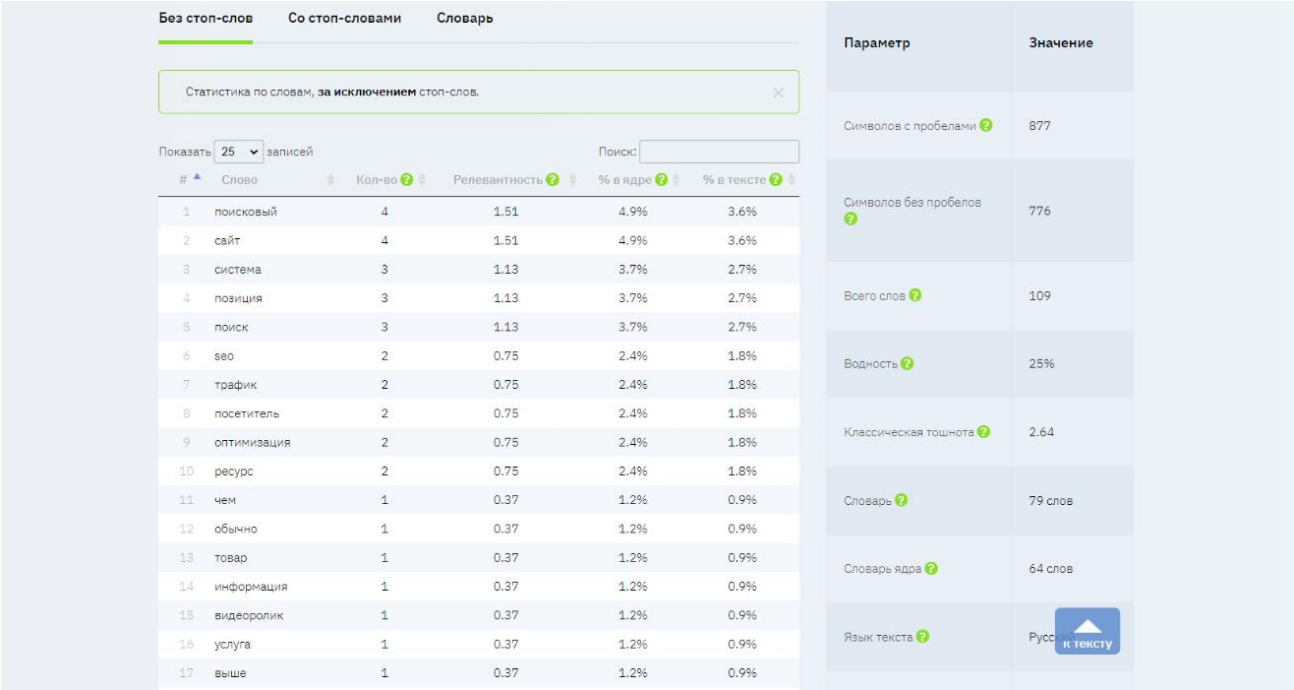


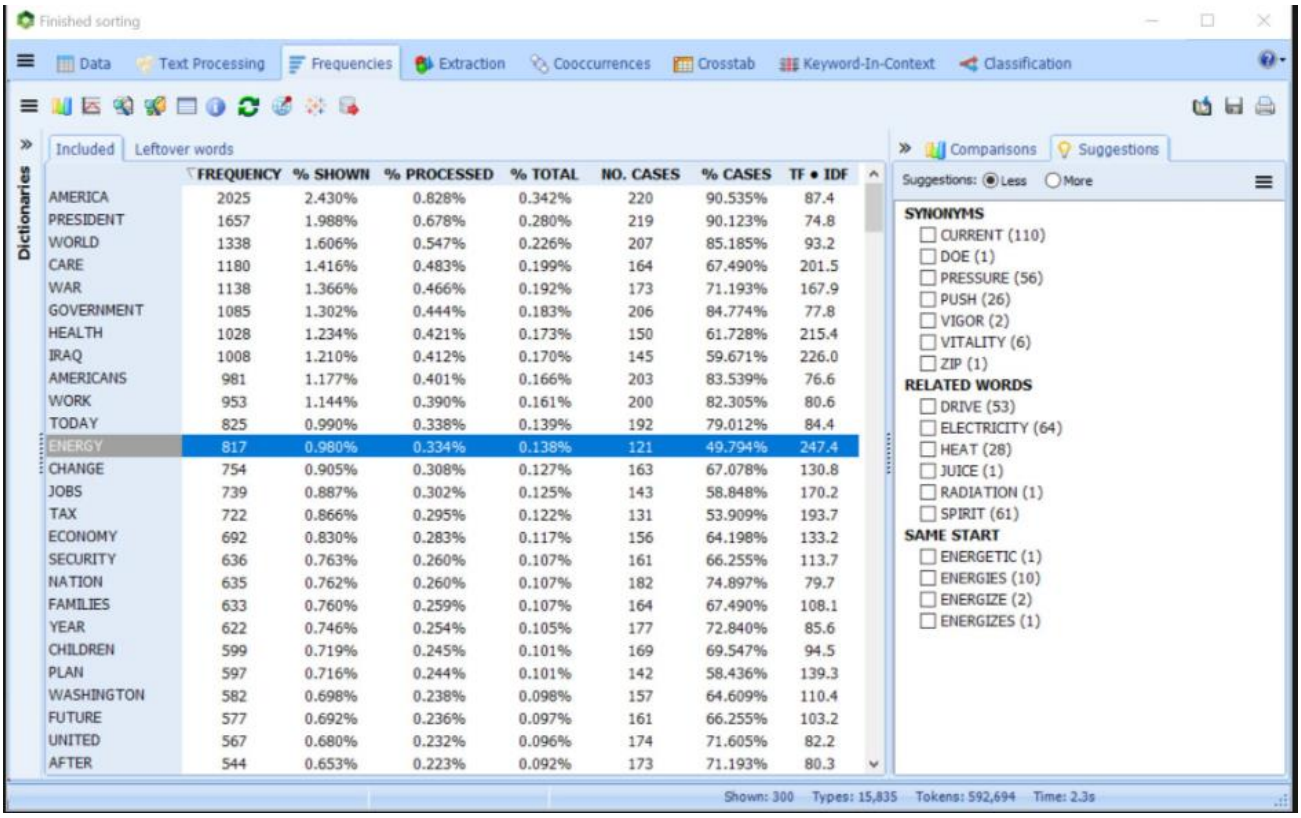
Рисунок 1.6 – Результат аналізу тексту сервісом Istio [16]

Загалом Istio є класичним сервісом для семантичного аналізу текстів, який визначає ключові слова та інші параметри.

Wordstat є гнучким та простим у використанні програмним забезпеченням для аналізу тексту. Програма дає можливість провести статистичний аналіз появи спільних термінів у корпусах текстів, дозволяє проводити кластерний та інші види аналізу текстів, проводиться класифікацію вмісту за допомогою словників, які визначаються користувачем, автоматично вилучає тему з використанням ієрархічної кластеризації, дозволяє пов’язувати

неструктурований текст із структурованими даними і має широкий спектр засобів візуалізації та інтерпретації результатів аналізу текстів.

На рисунку 1.7 показано результат виконання аналізу тексту інформаційною системою. Виводиться перелік ключових слів з статистикою по кожному слову [17].



	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF • IDF
AMERICA	2025	2.430%	0.828%	0.342%	220	90.535%	87.4
PRESIDENT	1657	1.988%	0.678%	0.280%	219	90.123%	74.8
WORLD	1338	1.606%	0.547%	0.226%	207	85.185%	93.2
CARE	1180	1.416%	0.483%	0.199%	164	67.490%	201.5
WAR	1138	1.366%	0.466%	0.192%	173	71.193%	167.9
GOVERNMENT	1085	1.302%	0.444%	0.183%	206	84.774%	77.8
HEALTH	1028	1.234%	0.421%	0.173%	150	61.728%	215.4
IRAQ	1008	1.210%	0.412%	0.170%	145	59.671%	226.0
AMERICANS	981	1.177%	0.401%	0.166%	203	83.539%	76.6
WORK	953	1.144%	0.390%	0.161%	200	82.305%	80.6
TODAY	825	0.990%	0.338%	0.139%	192	79.012%	84.4
ENERGY	817	0.980%	0.334%	0.138%	121	49.794%	247.4
CHANGE	754	0.905%	0.308%	0.127%	163	67.078%	130.8
JOBS	739	0.887%	0.302%	0.125%	143	58.848%	170.2
TAX	722	0.866%	0.295%	0.122%	131	53.909%	193.7
ECONOMY	692	0.830%	0.283%	0.117%	156	64.198%	133.2
SECURITY	636	0.763%	0.260%	0.107%	161	66.255%	113.7
NATION	635	0.762%	0.260%	0.107%	182	74.897%	79.7
FAMILIES	633	0.760%	0.259%	0.107%	164	67.490%	108.1
YEAR	622	0.746%	0.254%	0.105%	177	72.840%	85.6
CHILDREN	599	0.719%	0.245%	0.101%	169	69.547%	94.5
PLAN	597	0.716%	0.244%	0.101%	142	58.436%	139.3
WASHINGTON	582	0.698%	0.238%	0.098%	157	64.609%	110.4
FUTURE	577	0.692%	0.236%	0.097%	161	66.255%	103.2
UNITED	567	0.680%	0.232%	0.096%	174	71.605%	82.2
AFTER	544	0.653%	0.223%	0.092%	173	71.193%	80.3

Рисунок 1.7 – Результат аналізу тексту сервісом Wordstat [17]

Сьогодні семантичний аналіз тексту переважно проводиться на онлайн-платформах, усі вони діють за стандартним алгоритмом, але той чи інший відрізняються своїм результатом, хоча як визначено основою семантичного аналізу є визначення переліку ключових слів цифрових текстів. Наведене програмне забезпечення дозволяє вирішити актуальні задачі семантичного аналізу текстів.

1.3 Аналіз сучасних наукових публікацій з семантичного аналізу текстів

В ряді робіт [18] пропонується використання методу латентно-семантичного аналізу (ЛСА) для вилучення семантики з тексту та її представлення. Метод є теорією і методом екстракції і представлення контекстно-залежного змісту слів за допомогою статистичної обробки великого корпусу текстів. Суть методу полягає у сукупності усіх контекстів, в яких певне слово вживається або, навпаки, не вживається, обумовлює набір обмежень, які визначають подібність значень слів або множини слів. Таким чином, між словами і контекстом, в якому вони вживаються, існують приховані (латентні) зв’язки.

Метод ЛСА дозволяє визначити асоціативну і семантичну близькість та вирахувати кореляції між двома термами, двома документами, або між термом і документом. Ефективність застосування методу ЛСА в сфері знань людини підтверджена різноманітними прикладами його роботи. Зокрема, вперше зазначений метод був застосований з метою автоматичного індексування текстів та виявлення їх асоціативно-семантичної структури. Використання методу ЛСА знайшло своє відображення у системах вилучення, представлення семантичної інформації з тексту.

Відомий підхід до використання EM-алгоритму та методу головних компонент PCA (Principal Component Analysis) для лінгво-статистичного аналізу, тобто методи, у яких тексти представляються у вигляді векторів у багатомірному просторі ознак. У найпростішому випадку кожна з ознак відповідає наявності в тексті однієї з словоформ, які зустрічаються в тексті. При цьому компонента може дорівнювати або нулю, або одиниці, але в складніших випадках за кількістю випадків зустрічальності терміну в тексті формується вектор частот. Дані вектори можуть нормуватись.

Через занадто велику розмірність простору ознак застосування цих алгоритмів стає проблематичним тому, що алгоритми кластеризації оперують

матрицями. Таким чином, ЕМ-алгоритм оперує імовірнісною моделлю відношення документу до відповідного кластеру. Він базується на представленні реалізації багатовимірної випадкової величини. Щодо метод PCA, тут простір зменшеної розмірності будується на власних векторах коваріаційної матриці, яка відповідає декільком найбільшим власним числам [19].

У дослідженні [20] було визначено, що для вирішення задачі автоматизації пошуку ключових слів у навчальних матеріалах, доцільним є реалізувати автоматизацію побудови семантичної моделі навчальних курсів і в свою чергу її використання у відповідних інформаційних технологіях. Аналіз термінологічної бази навчальних матеріалів є одним із способів визначення оцінки семантичної відповідності.

1.4 Аналіз методів пошуку ключових слів

Для семантичного аналізу текстів та визначення ключових слів існує велика кількість методів, проте найбільше використовуються:

– TF-IDF;

– BM25;

– DE.

Метод TF-IDF полягає у використанні статистичного показника, який використовується для оцінки значимості слів у контексті документа, що є частиною колекції документів чи корпусу текстів. TF (term frequency – частота слова) – відношення числа входжень обраного слова до загальної кількості слів документа, таким чином, оцінюється важливість слова t_i в межах обраного документа:

$$Tf_i = \frac{n(i)}{\sum_k n_{ik}} \tag{1.1}$$

де n_i є число входжень слова в документ, а в знаменнику – загальна кількість слів в документі.

IDF (inverse document frequency – обернена частота документа) – інверсія частоти, з якою слово зустрічається в документах колекції, використання IDF зменшує вагу найбільш вживаних слів:

$$Idf_i = \log \frac{D}{d_i} \tag{1.2}$$

де D – кількість документів колекції; d_i – кількість документів, в яких зустрічається дане слово.

Вибір основи логарифму у формулі не має значення, адже зміна основи призведе до зміни ваги кожного слова на постійний множник, тобто вагове співвідношення залишиться незмінним.

Іншими словами, показник TF-IDF це добуток двох множників: TF та IDF:

$$TfIdf = Tf * Idf \tag{1.3}$$

Більшу вагу TF-IDF отримають слова з високою частотою появи в межах документа та низькою частотою вживання в інших документах колекції.

Показник TF-IDF використовується в задачах семантичного аналізу текстів та інформаційного пошуку. Його можна застосовувати як один з критеріїв релевантності документа до пошукового запиту, а також при визначенні міри спорідненості документів при кластеризації. Найпростішу функцію ранжування можна визначити як суму TF-IDF кожного ключового слова в запиті. Більшість просунутих функцій ранжування ґрунтуються на цій простій моделі [21].

BM25 є пошуковою функцією на неврегульованих безлічі термінів и безлічі документів, які вона оцінює на основі частоти використання слів із запиту у кожному документі, без урахування взаємовідношення між ними. Метод не обмежується однією функцією. Найбільш відомий, коли дається запит Q , який містить слова $q_1,...,q_n$, тоді функція BM25 дає наступну оцінку релевантності D документа запита Q:

$$score(D,Q) = \sum_{i=1}^{n} Idf(q_i) * \frac{f(q_i, D) * (k_1 + 1)}{f(q_i, D) + k_1 * (1 - b + b * \frac{|D|}{avgdl})} \quad (1.4)$$

де $f(q_i,D)$ є частотою слова q_i в документі D, $|D|$ є довжиною документа, $avgdl$ – середня довжина документа в колекції, k_1 і b – вільні коефіцієнти.

IDF(q_i) є оберненою частотою документа слова q_i. Зазвичай воно визначається так:

$$\log \frac{N}{n(q_i)} \quad (1.5)$$

де N є загальною кількістю документів в колекції, а $n(q_i)$ – це кількість документів, які містять q_i [21].

DE (Disperse Evaluation) розраховується наступним чином:

$$\sigma_A = \frac{\sqrt{(\Delta A^2) - (\Delta A)^2}}{(\Delta A)}, \quad (1.6)$$

де $\triangle A$ – середнє значення послідовності $\triangle A_1$, $\triangle A_2$, $\triangle A_k$, $(\triangle A^2)$ – середнє значення послідовності $\triangle A_1^2$, $\triangle A_2^2$, $\triangle A_k^2$.

A є словом і позначається як A_k^n , де індекс k = 1, 2,…, k є номером появи цього слова в тексті, а n – номером даного слова в тексті.

Таким чином, метод дисперсійного оцінювання DE є найбільше перспективним для семантичного аналізу текстів, оскільки він не залежить від мови документу і з його допомогою можна ефективно провести аналіз текстів написаних флективними мовами.

1.5 Постановка задачі

Мета кваліфікаційної роботи магістра – створення методу визначення важливості семантичних одиниць у цифрових текстах, який дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок важливості семантичних одиниць тексту на основі дисперсійного оцінювання з урахуванням як внутрішніх відстаней між появами унікальних

семантичних одиниць, так і початкових, кінцевих та кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту.

Для досягнення поставленої мети створення методу визначення важливості семантичних одиниць у цифрових текстах потрібно розв’язати наступні *задачі дослідження*:

1. Провести аналіз предметної області семантичного аналізу текстів, зокрема сучасних методів пошуку ключових семантичних одиниць у цифрових текстах.

2. Вдосконалити метод визначення важливості семантичних одиниць у цифрових текстах.

3. Розробити інформаційну технологію автоматизованого пошуку семантичних одиниць у цифрових текстах.

4. Розробити прикладну інформаційну систему для автоматизованого пошуку семантичних одиниць у цифрових текстах.

5. Провести прикладне дослідження методу визначення важливості семантичних одиниць у цифрових текстах у складі інформаційної технології автоматизованого пошуку семантичних одиниць у цифрових текстах.

Висновки до розділу 1

В розділі проведено аналіз проблеми визначення важливості семантичних одиниць у цифрових текстах в межах аналізу предметної області семантичного аналізу текстів, який виявив актуальність даного напрямку практичного застосування інформаційних технологій. Аналіз методів пошуку ключових слів встановив, що метод дисперсійного оцінювання важливості семантичних одиниць тексту є найбільш ефективним й його доцільно брати за основу при розробці методу визначення важливості семантичних одиниць у цифрових текстах.

В результаті, у розділі визначено мету кваліфікаційної роботи магістра як створення методу визначення важливості семантичних одиниць у цифрових

текстах, який дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок важливості семантичних одиниць тексту на основі дисперсійного оцінювання з урахуванням як внутрішніх відстаней між появами унікальних семантичних одиниць, так і початкових, кінцевих та кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту.

Розділ 2

Метод і засоби автоматизації визначення важливості семантичних одиниць у цифрових текстах

2.1 Формальне подання модифікацій дисперсійного методу визначення важливості семантичних одиниць

Відповідно до результатів п. 1.4, дисперсійна оцінка є оцінкою дискримінантної сили слів та дозволяє відділити з загальної множини широковживаних слів у тексті слова, які розташовані у тексті рівномірно [5].

Якщо деяке слово T (позначається як T_k^n, де індекс k – номер появи даного слова у тексті, а n – позиція даного слова у тексті) має інтервали між послідовними появами $\Delta T_k^m = T_{k+1}^m - T_k^n = m - n$, де на *m-й* та *n-й* позиції у тексті знаходиться слово *A*, яке зустрілось *k*+1-*й* та *k*-*й* рази, то дисперсійна оцінка розраховується як: $DE = \sqrt{(\Delta T^2) - (\Delta T)^2} \big/ (\Delta T)$, де (ΔT) – середнє значення послідовності $\Delta T_1, \Delta T_2, \Delta T_k$, (ΔT^2) – послідовність T_1^2, T_2^2, T_k^2, k – кількість появи слова T в тексті [6].

Дисперсійна оцінка дозволяє відокремити слова, що зустрічаються в тексті відносно рівномірно (для рівномірно розподілених слів ця оцінка дорівнює нулю), від слів, які розподілені нерівномірно. Тобто це оцінка дискримінантної сили слів, що важливо, зокрема, для інформаційного пошуку. Ідея даної оцінки близька до TF-IDF, проте більш коректно застосовується до цілих текстів, а не до масивів із великою кількістю документів.

Обрахунок відстаней між словами у тексті слугує підготовчим етапом до дисперсійного оцінювання слів, за якого визначаються для кожного слова з кількістю появ і тексті більше одного всі відстані між сусідніми появами слів.

У залежності від того, яким чином та у якій кількості визначаються відстані для кожного унікального слова тексту, розрізняють різні модифікації вихідного методу DE-BM пошуку ключових слів за дисперсійним оцінюванням DE-BM1, а саме: DE-BM2 та DE-BM3.

Метод пошуку ключових слів DE-BM1 для *n* появ слова враховує *n*-1 відстані. При цьому за відстань береться різниця між меншим порядковим номером наступного слова і більшим порядковим номером попереднього слова.

Метод пошуку ключових слів DE-BM2 для *n* появ слова враховує *n*+1 відстань. За відстань береться різниця між меншим порядковим номером наступного слова і більшим порядковим номером попереднього слова. Також додатково враховуються відстані: від початку тексту до першої появи слова в тексті, від останньої появи слова у тексті до кінця тексту.

Метод пошуку ключових слів DE-BM3 для *n* появ слова враховує *n* відстаней. За відстань береться різниця між меншим порядковим номером наступного слова і більшим порядковим номером попереднього слова. Також додатково враховується відстань, рівна сумі різниць між початком тексту до першої появи слова й між останньою появою слова до кінця тексту.

Таким чином, розглянуто модифікації вихідного методу DE-BM пошуку ключових слів за дисперсійним оцінюванням DE-BM1, а саме DE-BM2 та DE-BM3, що одержуються в залежності від того, яким чином та у якій кількості визначаються відстані для кожного унікального слова тексту. Обрахунок значень дисперсійного оцінювання за DE-BM1, DE-BM2 та DE-BM3 лежить в основі методу визначення важливості семантичних одиниць у цифрових текстах.

2.2 Схема методу визначення важливості семантичних одиниць у цифрових текстах

Метод визначення важливості семантичних одиниць у цифрових текстах на основі дисперсійного оцінювання забезпечує обрахунок значень дисперсійного оцінювання за модифікаціями DE-BM1, DE-BM2 та DE-BM3. Розроблюваний метод, на відміну від існуючих, дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а й з урахуванням початкових,

кінцевих та похідних кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту. На Рисунку 2.1 зображено схему кроків методу визначення важливості семантичних одиниць у цифрових текстах на основі дисперсійного оцінювання.

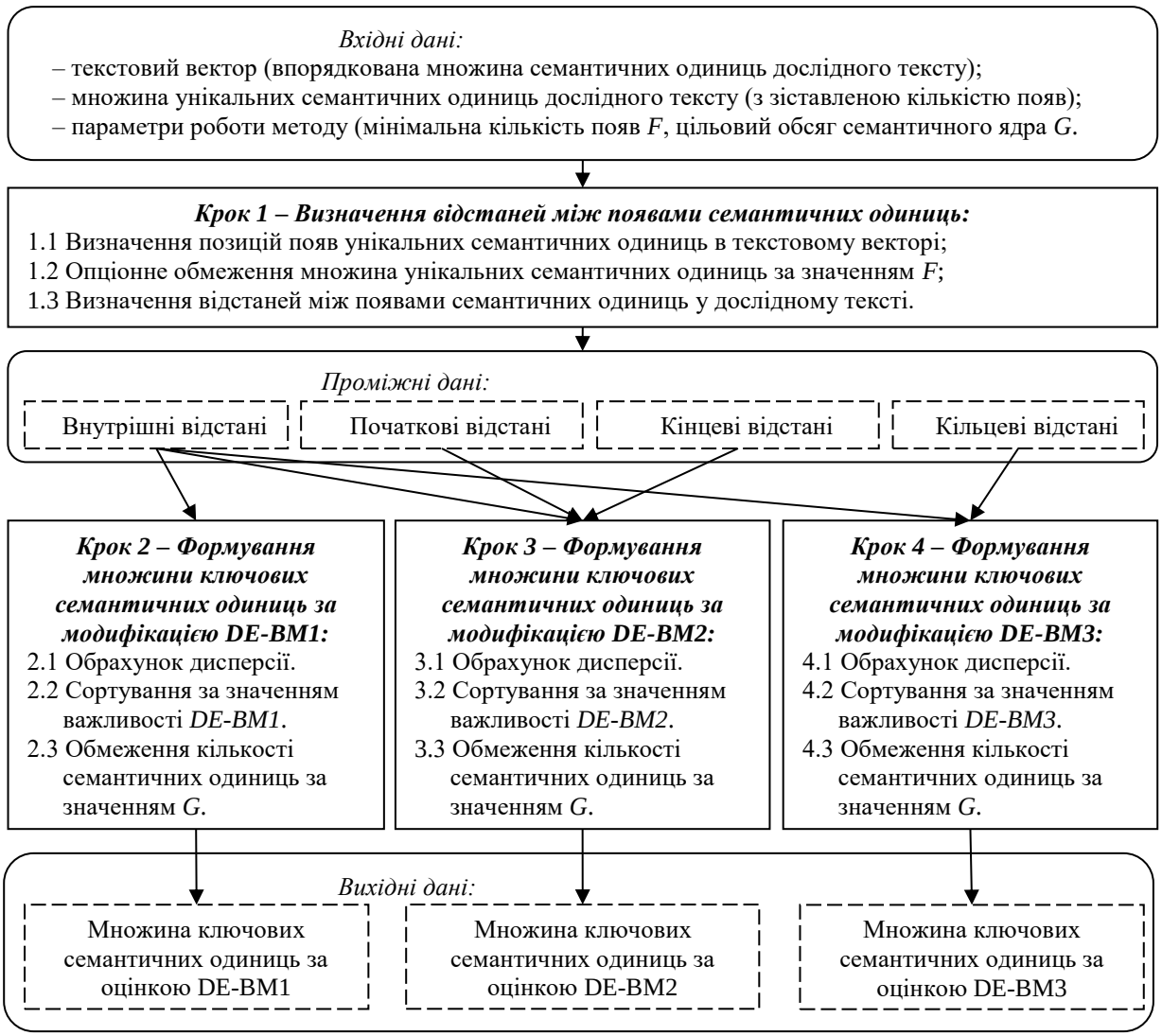


Рисунок 2.1 – Схема методу визначення важливості семантичних одиниць у цифрових текстах

Вхідні дані методу визначення важливості семантичних одиниць у цифрових текстах формують текстовий вектор у вигляді впорядкованої множини семантичних одиниць дослідного тексту, множина унікальних семантичних одиниць дослідного тексту з зіставленою кількістю їх появ та параметри роботи

методу, зокрема мінімальна кількість появ *F* та цільовий обсяг семантичного ядра *G*.

На Кроці 1 виконується визначення відстаней між появами семантичних одиниць. Для цього спершу проводять визначення позицій появ унікальних семантичних одиниць в текстовому векторі, після чого здійснюється опціонне обмеження множина унікальних семантичних одиниць за значенням *F*. В результаті, виконується визначення відстаней між появами кожної зсемантичних одиниць у дослідному тексті.

Проміжні дані методу визначення важливості семантичних одиниць у цифрових текстах складають обраховані відстані між появами унікальних семантичних одиниць тексту, причому визначаються не тільки внутрішні, а й початкові, кінцеві та похідні кільцеві відстані між появами унікальних семантичних одиниць цифрового тексту.

Далі на Кроці 2 виконується формування множини ключових семантичних одиниць за модифікацією DE-BM1, для чого спершу проводиться обрахунок дисперсії семантичних одиниць цифрового тексту, потім проводиться сортування множини семантичних одиниць за значенням важливості *DE-BM1*, й після цього виконується обмеження кількості семантичних одиниць за значенням *G*. При цьому проведення обрахунку дисперсії семантичних одиниць цифрового тексту виконується з урахуванням тільки внутрішніх відстаней між появами унікальних семантичних одиниць цифрового тексту.

Паралельно на Кроці 3 виконується формування множини ключових семантичних одиниць за модифікацією DE-BM2, для чого спершу проводиться обрахунок дисперсії семантичних одиниць цифрового тексту, потім проводиться сортування множини семантичних одиниць за значенням важливості *DE-BM2*, і після цього виконується обмеження кількості семантичних одиниць за значенням *G*. При цьому проведення обрахунку дисперсії семантичних одиниць цифрового тексту виконується із урахуванням не тільки внутрішніх відстаней між появами унікальних семантичних одиниць цифрового тексту, а й початкових та кінцевих відстаней.

Також на Кроці 4 виконується формування множини ключових семантичних одиниць за модифікацією DE-BM3, для чого спершу проводиться обрахунок дисперсії семантичних одиниць цифрового тексту, потім проводиться сортування множини семантичних одиниць за значенням важливості *DE-BM3*, і після цього виконується обмеження кількості семантичних одиниць за значенням *G*. При цьому проведення обрахунку дисперсії семантичних одиниць цифрового тексту виконується із урахуванням як внутрішніх відстаней між появами унікальних семантичних одиниць цифрового тексту, так і кільцевих відстаней.

Вихідні дані методу визначення важливості семантичних одиниць у цифрових текстах складають 3 множини ключових семантичних одиниць, що визначені за оцінками модифікацій дисперсійного оцінювання DE-BM1, DE-BM2 й DE-BM3.

Таким чином, удосконалено метод визначення важливості семантичних одиниць у цифрових текстах на основі дисперсійного оцінювання, який на відміну від існуючих дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а і з урахуванням початкових, кінцевих і похідних кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту.

2.3 Інформаційна технологія автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

Інформаційна технологія автоматизованого пошуку ключових семантичних одиниць у цифрових текстах використовує створений метод визначення важливості семантичних одиниць у цифрових текстах для автоматизованого одержання за вхідними даними у вигляді вхідного цифрового тексту у вигляді відповідної впорядкованої множини символів та параметрами

налаштувань вихідних даних у вигляді трьох множин ключових семантичних одиниць вхідного цифрового тексту, одержаними за оцінками модифікацій дисперсійного оцінювання.

Використані модифікації дисперсійного оцінювання ключових семантичних одиниць дозволяють виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а й із урахуванням початкових, кінцевих і кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту.

Також інформаційна технологія автоматизованого пошуку ключових семантичних одиниць забезпечує формування зведеної таблиці оцінок семантичної важливості ключових семантичних одиниць за знайденими за оцінками модифікацій дисперсійного оцінювання ключових семантичних одиниць вхідного цифрового тексту.

Вхідні дані інформаційної технології автоматизованого пошуку семантичних одиниць у цифрових текстах формують цифровий дослідний текст (впорядкована множина символів тексту) та параметри налаштувань, до яких належать мінімальна кількість появ F і цільовий обсяг семантичного ядра G. На Рисунку 2.2 зображено відповідну схему етапів виконання інформаційної технології автоматизованого пошуку семантичних одиниць у цифрових текстах.

Етап 1 забезпечує попередню обробку цифрового дослідного тексту. За неї виконується приведення всіх літер цифрового дослідного тексту до нижнього регістру, після чого виконується видалення розділових знаків у цифровому дослідному тексті, крім апострофу і дефісу, а також видалення цифр у цифровому дослідному тексті, включаючи дужки.

Етап 2 інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах відповідає за формування текстового вектору. Спершу виконується розділення цифрового дослідного тексту на окремі семантичні одиниці, за якими й проводиться формування впорядкованої

множини семантичних одиниць цифрового дослідного тексту. Для перевірки коректності формування текстового вектору, на цьому етапі проводиться зворотна тестова збірка цифрового дослідного тексту за множиною семантичних одиниць.

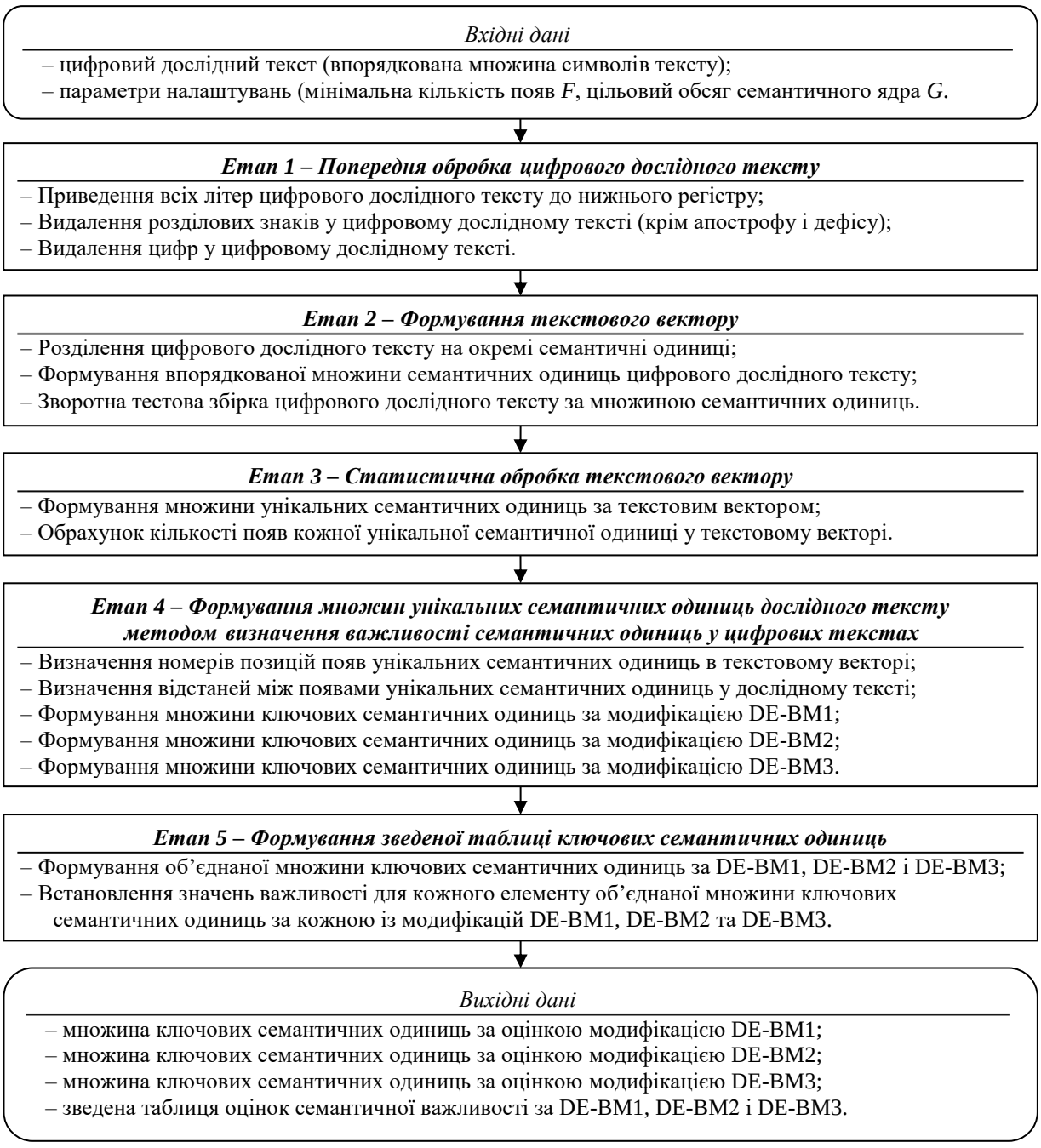


Рисунок 2.2 – Схема інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

На Етапі 3 проводиться статистична обробка текстового вектору. При цьому спершу виконується формування множини унікальних семантичних одиниць за текстовим вектором, а після цього здійснюється обрахунок кількості появ кожної унікальної семантичної одиниці у текстовому векторі.

Етап 4 інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах забезпечує використання розробленого в п.2.2 методу визначення важливості семантичних одиниць у цифрових текстах для формування множин унікальних семантичних одиниць дослідного тексту. При цьому виконується визначення номерів позицій появ унікальних семантичних одиниць в текстовому векторі, за визначенням номерів позицій появ унікальних семантичних одиниць слідує визначення відстаней між появами унікальних семантичних одиниць у дослідному тексті. Причому визначаються не тільки внутрішні, а і початкові, кінцеві та похідні кільцеві відстані між появами унікальних семантичних одиниць цифрового тексту. За визначеними відстанями виконується формування множини ключових семантичних одиниць за модифікаціями DE-BM1, DE-BM2 та DE-BM3.

На Етапі 5 інформаційної технології виконується формування зведеної таблиці ключових семантичних одиниць. Для цього спершу проводиться формування об’єднаної множини ключових семантичних одиниць за модифікаціями DE-BM1, DE-BM2 і DE-BM3, а на основі неї відбувається встановлення значень важливості для кожного елементу об’єднаної множини ключових семантичних одиниць за кожною з модифікацій DE-BM1, DE-BM2 та DE-BM3.

Вихідні дані інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах складають три множини ключових семантичних одиниць, що визначені за оцінками трьох модифікацій дисперсійного оцінювання, та зведена таблиця оцінок семантичної важливості за DE-BM1, DE-BM2 і DE-BM3.

Таким чином, розроблено інформаційну технологію автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, що дозволяє з

використанням створеного методу визначення важливості семантичних одиниць у цифрових текстах за вхідними даними у вигляді вхідного цифрового тексту як відповідної впорядкованої множини символів та параметрами налаштувань одержувати вихідні дані у вигляді трьох множин ключових семантичних одиниць вхідного цифрового тексту за оцінками модифікацій дисперсійного оцінювання, які дозволяють виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а й з врахуванням початкових, кінцевих та кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту, а також формувати зведену таблицю оцінок семантичної важливості ключових семантичних одиниць за обрахованими оцінками модифікацій дисперсійного оцінювання.

2.4 Функції інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

Для прикладного дослідження розроблених методу визначення важливості семантичних одиниць у цифрових текстах та інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах необхідно створити інформаційну систему, що забезпечує відповідний функціонал.

Відповідно до п.2.3, розроблена згідно інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах інформаційна система мусить виконувати наступні функції:

– приведення всіх літер цифрового дослідного тексту до нижнього регістру;

– видалення розділових знаків у цифровому дослідному тексті (крім апострофу і дефісу);

– видалення цифр у цифровому дослідному тексті;

– розділення цифрового дослідного тексту на окремі семантичні одиниці;

– формування впорядкованої множини семантичних одиниць цифрового дослідного тексту;

– зворотна тестова збірка цифрового дослідного тексту за множиною семантичних одиниць;

– формування множини унікальних семантичних одиниць за текстовим вектором;

– обрахунок кількості появ кожної унікальної семантичної одиниці у текстовому векторі;

– визначення номерів позицій появ унікальних семантичних одиниць в текстовому векторі;

– визначення відстаней між появами унікальних семантичних одиниць у дослідному тексті;

– формування множини ключових семантичних одиниць за модифікацією DE-BM1;

– формування множини ключових семантичних одиниць за модифікацією DE-BM2;

– формування множини ключових семантичних одиниць за модифікацією DE-BM3;

– формування об’єднаної множини ключових семантичних одиниць за DE-BM1, DE-BM2 і DE-BM3;

– встановлення значень важливості для кожного елементу об’єднаної множини ключових семантичних одиниць за кожною із модифікацій DE-BM1, DE-BM2 та DE-BM3.

За коректного виконання інформаційною системою наведених функцій, можна робити висновок про її відповідність до створеної інформаційної технології автоматизованого пошуку семантичних одиниць у цифрових текстах й придатність до дослідження її ефективності.

Висновки до розділу 2

В розділі розглянуто модифікації вихідного методу DE-BM пошуку ключових слів за дисперсійним оцінюванням DE-BM1, а саме DE-BM2 та DE-BM3, що одержуються у залежності від того, яким чином та у якій кількості визначаються відстані для кожного унікального слова тексту. Обрахунок значень дисперсійного оцінювання за DE-BM1, DE-BM2 й DE-BM3 лежить у основі методу визначення важливості семантичних одиниць у цифрових текстах.

Також в розділі наведено метод визначення важливості семантичних одиниць у цифрових текстах на основі дисперсійного оцінювання, який дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а і з урахуванням початкових, кінцевих і похідних кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту.

Розроблено інформаційну технологію автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, що дозволяє з використанням створеного методу визначення важливості семантичних одиниць у цифрових текстах за вхідними даними у вигляді вхідного цифрового тексту як відповідної впорядкованої множини символів та параметрами налаштувань одержувати вихідні дані у вигляді трьох множин ключових семантичних одиниць вхідного цифрового тексту за оцінками модифікацій дисперсійного оцінювання, які дозволяють виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а й з врахуванням початкових, кінцевих та кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту, а також формувати зведену таблицю оцінок семантичної важливості ключових семантичних одиниць за обрахованими оцінками модифікацій дисперсійного оцінювання.

Для прикладного дослідження розроблених методу визначення важливості семантичних одиниць у цифрових текстах й інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах необхідно створити інформаційну систему, що забезпечує відповідний функціонал. Тому в розділі наведено перелік функцій, за коректного виконання яких інформаційною системою можна робити висновок про її відповідність до створеної інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах й придатність до дослідження її ефективності.

Розділ 3

Інформаційна система автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

3.1 Структура інформаційної системи

Інформаційна система автоматизованого пошуку ключових семантичних одиниць у цифрових текстах дозволяє за створеною інформаційною технологією в результаті обробки вхідного цифрового тексту у вигляді відповідної впорядкованої множини символів виконувати автоматизоване визначення множин ключових семантичних одиниць за оцінками модифікацій дисперсійного оцінювання DE-BM1, DE-BM2 і DE-BM3, також формувати відповідну зведену таблицю оцінок семантичної важливості ключових семантичних одиниць вхідного цифрового тексту.

Структуру інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах зображено на Рисунку 3.1, інформаційна система не потребує використання бази даних та складається з чотирьох модулів: модуля попередньої обробки тексту та формування текстового вектору, модуля визначення відстаней між появами семантичних одиниць, модуля дисперсійного оцінювання важливості семантичних одиниць та модуля формування зведеної таблиці ключових семантичних одиниць.

Модуль попередньої обробки тексту та формування текстового вектору інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах забезпечує виконання функцій автоматизованого приведення всіх літер цифрового дослідного тексту до нижнього регістру, автоматизованого видалення розділових знаків у цифровому дослідному тексті, автоматизованого видалення цифр у цифровому дослідному тексті та автоматизованого розділення цифрового дослідного тексту на окремі семантичні одиниці. Також модуль забезпечує виконання функцій автоматизованого формування впорядкованої множини семантичних одиниць цифрового тексту та

автоматизованої зворотної тестової збірки цифрового тексту за множиною семантичних одиниць.

Рисунок 3.1 – Структура інформаційної системи автоматизованого пошуку семантичних одиниць у цифрових текстах

Модуль визначення відстаней між появами семантичних одиниць інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах забезпечує виконання функцій автоматизованого формування множини унікальних семантичних одиниць за текстовим вектором, автоматизованого обрахунку кількості появ кожної семантичної одиниці у текстовому векторі, автоматизованого визначення номерів позицій появ семантичних одиниць в текстовому векторі та автоматизованого визначення відстаней між появами семантичних одиниць у дослідному тексті.

Модуль дисперсійного оцінювання важливості семантичних одиниць інформаційної системи автоматизованого пошуку ключових семантичних

одиниць у цифрових текстах забезпечує виконання функцій автоматизованого формування множини ключових семантичних одиниць за модифікацією DE-BM1, автоматизованого формування множини ключових семантичних одиниць за модифікацією DE-BM2 й автоматизованого формування множини ключових семантичних одиниць за модифікацією DE-BM3.

Модуль формування зведеної таблиці ключових семантичних одиниць інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах забезпечує виконання функцій автоматизованого формування об’єднаної множини ключових семантичних одиниць та автоматизованого встановлення значень важливості для кожного елементу об’єднаної множини ключових семантичних одиниць за кожною із модифікацій DE-BM1, DE-BM2 та DE-BM3.

Таким чином, розроблено структуру інформаційної системи для автоматизованого пошуку семантичних одиниць у цифрових текстах, що дозволяє за створеною інформаційною технологією в результаті обробки вхідного цифрового тексту у вигляді відповідної впорядкованої множини символів виконувати автоматизоване визначення множин ключових семантичних одиниць за оцінками модифікацій дисперсійного оцінювання DE-BM1, DE-BM2 і DE-BM3, також формувати відповідну зведену таблицю оцінок семантичної важливості ключових семантичних одиниць вхідного цифрового тексту.

Інформаційна система автоматизованого пошуку семантичних одиниць у цифрових текстах не потребує використання бази даних та складається з чотирьох модулів: модуля попередньої обробки тексту та формування текстового вектору, модуля визначення відстаней між появами семантичних одиниць, модуля дисперсійного оцінювання важливості семантичних одиниць та модуля формування зведеної таблиці ключових семантичних одиниць. Кожен з зазначених модулів забезпечує окремий інтерфейс для автоматизованого виведення результатів своєї роботи користувачеві.

3.2 Проєктування інтерфейсу інформаційної системи автоматизованого пошуку ключових семантичних одиниць

Інформаційна система автоматизованого пошуку семантичних одиниць у цифрових текстах визначення важливості семантичних одиниць у цифрових текстах має виконувати функції перегляду, обробки інформації, формування та обмеження за певними ознаками множин слів та ін. В результаті роботи системи має бути отримано зведену таблицю оцінок семантичної важливості, тому необхідно забезпечити доступ до усіх необхідних функцій за допомогою елементів інтерфейсу.

Рисунок 3.2 – Схема інтерфейсу вкладки «Формування текстового вектору»

Відповідно до п.3.1, інформаційна система автоматизованого пошуку семантичних одиниць у цифрових текстах не потребує використання бази даних та складається з чотирьох модулів: модуля попередньої обробки тексту та формування текстового вектору, модуля визначення відстаней між появами семантичних одиниць, модуля дисперсійного оцінювання важливості

семантичних одиниць та модуля формування зведеної таблиці ключових семантичних одиниць. Кожен з цих модулів забезпечує окремий інтерфейс для автоматизованого виведення результатів своєї роботи користувачеві.

Робоча область першої вкладки «Формування текстового вектору» (рисунок 3.2) має вміщувати область перегляду тексту для дослідження, область перегляду обробленого тексту, елемент для роботи користувача з текстом та елемент для формування текстового вектора.

При натисканні на вкладку «Визначення відстаней» має відкритись відповідна робоча область (рисунок 3.3), яка містить область перегляду множини унікальних слів та їх позицій у тексті, область перегляду відстаней між появами слів у тексті, елемент для формування переліку унікальних слів, елемент для визначення позицій слів, елемент для обмеження переліку до 2 появ слів у тексті, елемент для обмеження переліку до 3 появ слів у тексті та елемент для обрахунку відстаней між появами слів.

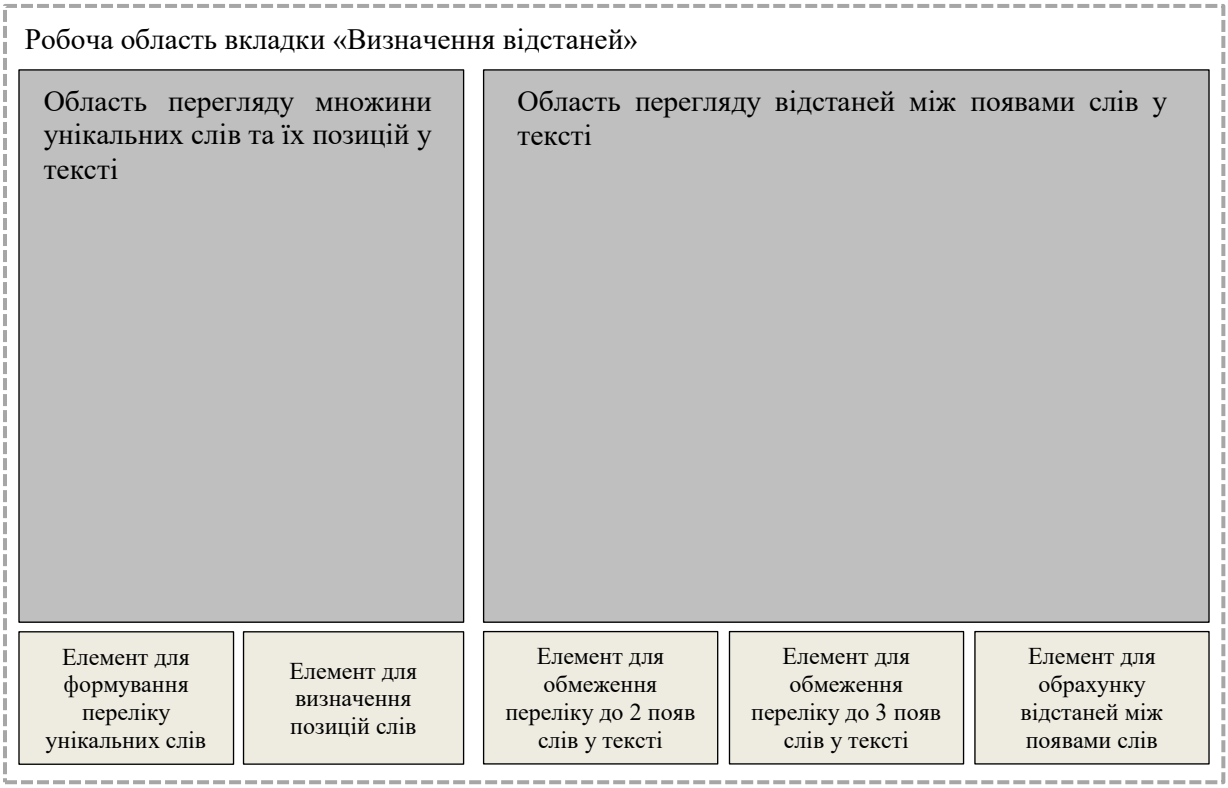


Рисунок 3.3 – Схема інтерфейсу вкладки «Визначення відстаней»

При натисканні на вкладку «Обрахунок дисперсії слів» має відкритись відповідна робоча область (рисунок 3.4), яка містить область перегляду множини

ключових слів за DE-BM1, область перегляду множини ключових слів за DE-BM2, область перегляду множини ключових слів за DE-BM3, елемент для обрахунку DE-BM1, елемент для сортування за значенням DE-BM1, елемент для обрахунку DE-BM2, елемент для сортування за значенням DE-BM2, елемент для обрахунку DE-BM3, елемент для сортування за значенням DE-BM3, область для введення кількості ключових слів та елемент для обмеження ключових слів.

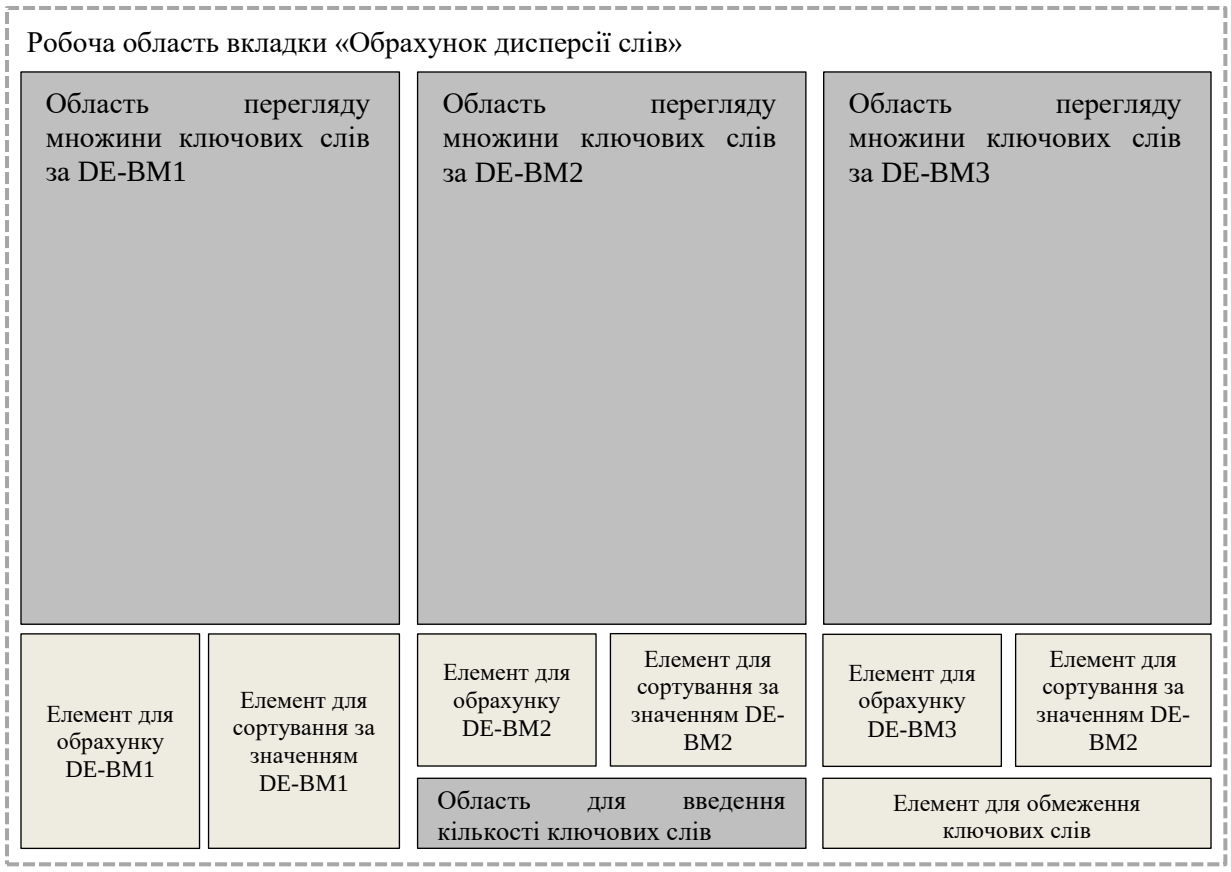


Рисунок 3.4 – Схема інтерфейсу вкладки «Обрахунок дисперсії слів»

Останньою та результативною вкладкою є - «Зведена таблиця оцінок семантичної важливості» (рисунок 3.5), робоча область має містити область перегляду зведеної таблиці оцінок семантичної важливості, елемент формування зведеної множини слів, елемент додавання даних важливості слів, елемент для сортування значень за DE-BM1, елемент для сортування значень за DE-BM2, елемент для сортування значень за DE-BM3 та елемент для експорту даних в Excel.

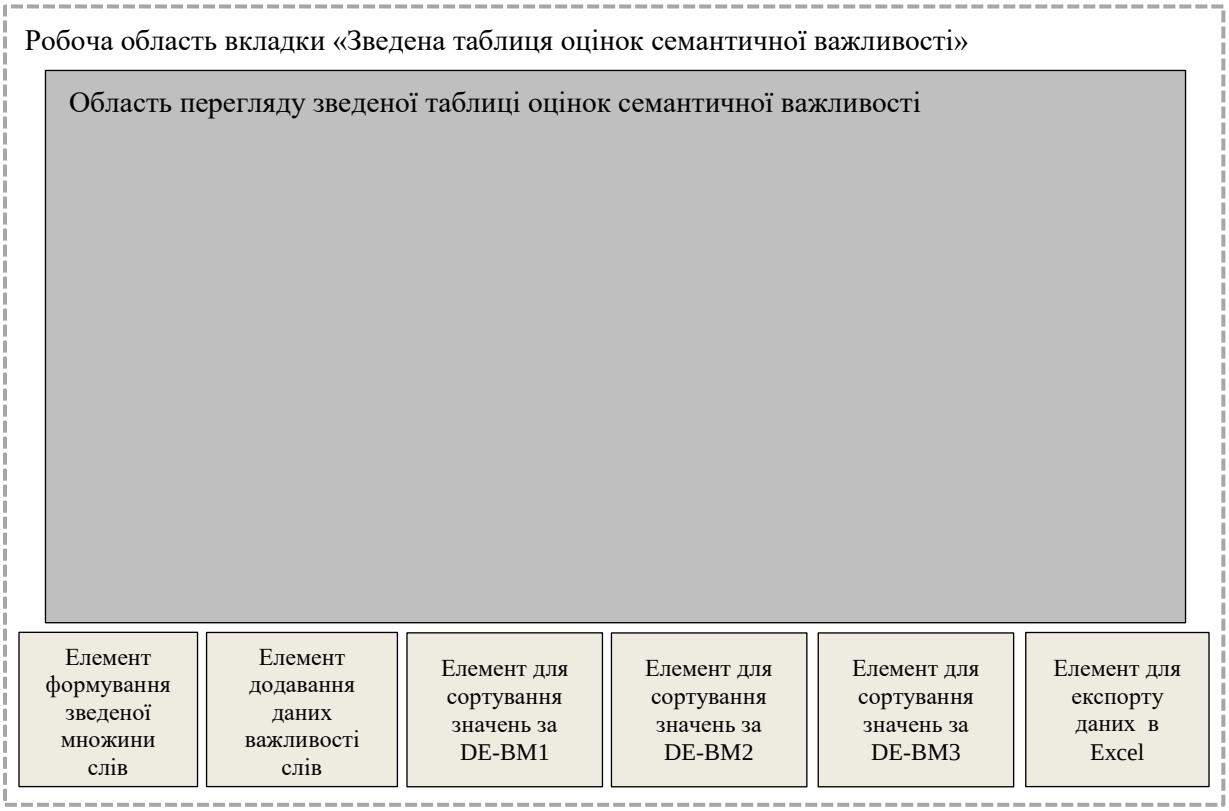


Рисунок 3.5 – Схема інтерфейсу вкладки «Зведена таблиця оцінок семантичної важливості»

Отже, згідно функцій і вимог до інтерфейсу інформаційної системи визначення важливості семантичних одиниць у цифрових текстах, її можна реалізувати у вигляді десктопного додатку з відповідними функціями.

Розроблено окремі інтерфейси інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах для автоматизованого виведення користувачеві результатів роботи модулів інформаційної системи: модуля попередньої обробки тексту та формування текстового вектору, модуля визначення відстаней між появами семантичних одиниць, модуля дисперсійного оцінювання важливості семантичних одиниць та модуля формування зведеної таблиці ключових семантичних одиниць.

3.3 Аналіз рекомендованих засобів розробки інформаційної системи

Для розробки інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах використовується ряд засобів. Для початку було обрано середовище розробки Visual Studio 2019. За допомогою даного середовища досить зручно створювати не лише класичні десктопні додатки але і мобільні, Web-додатки та ін [27]. Також, Visual Studio практичний у тому, що підтримує велику кількість мов програмкування.

На сьогодні існує велика кільки засобів створення програмного забезпечення, та найбільше відомими є:

– .NET;

– Java;

– PHP.

.NET – це платформа від компанії Microsoft, яка дозоляє створювати програмні додатки для операційних систем родини Microsoft Windows [22].

Розробляти програмне забезпечення на платформі .NET можна за допомогою трьох мов програмування : C#, F# та Visual Basic. Різні реалізації даної платформи вирішують складні задачі:

– .NET Core – платформа крос-платформна реалізація для розробки веб-сайтів, серверів та консольних програм на операційних системах Windows, Linux та macOS;

– .NET Framework підтримує веб-сайти, сервіси, настільні додатки і багато іншого на Windows;

– Xamarin/Mono – .NET реалізація для запуску програм у всіх основних мобільних операційних системах.

Також для розширення функціональності платформа підтримує здорову екосистему пакетів, побудованих на стандарті .NET [23].

Платформа Java є набором програм, які допомагають розробникам програмного забезпечення ефективно розробляти та запускати програми на мові

програмування Java. Платформа включає в себе віртуальну машину, компілятор та набір бібліотек.

Мова програмування Java є багатоплатформенною, об'єктно-орієнтованою та орієнтованою на мережу. Вона вважається однією швидких, надійних та безпечних мов програмування, також вона використовується для:

– розробки додатків під Android;
– створення корпоративного програмного забезпечення;
– аналізу великих даних;
– програмування апаратних пристроїв Java;
– серверних технологій, таких як Apache, Jboss, GlassFish тощо.

JDK (Java Development Kit) – одна зі складових платформи, середовище розробки програмного забезпечення, яке використовується для створення аплетів та програм Java. JDK містить інструменти, які необхідні для написання програм Java і JRE (Java Runtime Enviroment) і для їх виконання. Він включає в себе компілятор, панель запуску програм та Appletviewer. Компілятор перетворює код, що написаний на Java, у байт-код, панель запуску програм в Java відкриває JRE, завантажує необхідний клас і виконує його основний метод.

Одним з механізмів JRE є Java Virtual Machine (JVM) – механізм, який забезпечує середовище для керування кодом Java або програмами. Компілятор Java створює код для віртуальної машини – JVM [24].

PHP (Hypertext Pre-processor) є мовою сценаріїв, яка виконується на стороні сервера, які використовуються для розробки статичних або динамічних веб-сайтів або веб-додатків. PHP сценарії інтерпретуються лише на сервері, на якому встановлено PHP [25].

PHP здатна працювати на всіх основних операційних системах та використовуватись з усіма провідними веб-серверами, такими як Nginx, OpenBSD і Apache. Деякі хмарні середовища також підтримують PHP, а саме Microsoft Azure and Amazon AWS. PHP не обмежується лише обробкою HTML, вона має вбудовану підтримку для створення файлів PDF, GIF, JPEG та

зображень PNG. Однією з переваг даної мови є широкий перелік підтримуваних баз даних: MySQL, PostgreSQL, MS SQL, db2, Oracle Database і MongoDB [26].

Отож було обрано мову програмування C#. Досить проста та зручна у використанні мова програмування, об’єктно-орієнтовна та ідеально підходить для комбінування з платформою .Net [28].

Інтерфейсом програмування було обрано Windows Forms, який включає в себе .Net Framework та спрощує створення додатків за рахунок графічного інтерфейсу [29].

Отже, для розробки інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах рекомендовано використати засоби розробки: Visual Studio 2019, мова програмування C#, інтерфейс програмування Windows Forms та платформу .NET.

3.4 Архітектура модулів інформаційної системи

Інформаційна система визначення важливості семантичних одиниць у цифрових текстах реалізована за допомогою класів, структура яких представлена на рисунку 3.6.

Робота програми розпочинається з класу «Program», який вміщує в собі метод «Main». Клас «Form1» - форма на якій відбуваються усі маніпуляції в інформаційній системі.

Клас «Convert_text» включає методи, які обробляють вхідний текст системи, забирає розділові знаки, знаки пунктуації та ін. Даний клас обраховує та визначає унікальні слова в тексті.

Клас «Distance» вміщує методи, які обраховують дистанції між словами в тексті. Клас «Dispers_eval» складається з методів, які визначають дисперсну оцінку, в даному додатку таких методів використовується 3 (ВМ1, ВМ2, ВМ3).

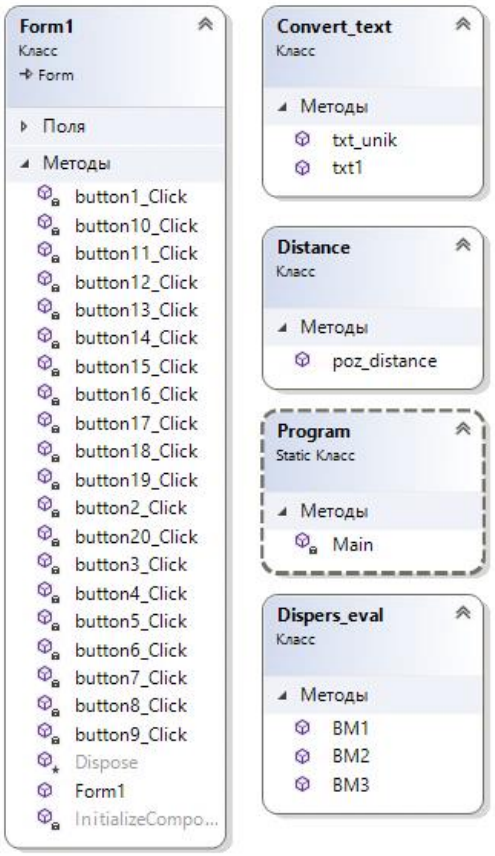


Рисунок 3.6 – Діаграма класів інформаційної системи

Таким чином, наведена діаграма класів дозволяє реалізовує наведену в п.3.1 архітектуру модулів інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах.

Висновки до розділу 3

У розділі розроблено структуру інформаційної системи для автоматизованого пошуку семантичних одиниць у цифрових текстах, що дозволяє за створеною інформаційною технологією в результаті обробки вхідного цифрового тексту у вигляді відповідної впорядкованої множини символів виконувати автоматизоване визначення множин ключових семантичних одиниць за оцінками модифікацій дисперсійного оцінювання DE-BM1, DE-BM2 і DE-BM3, також формувати відповідну зведену таблицю оцінок

семантичної важливості ключових семантичних одиниць вхідного цифрового тексту.

Виконано формування рекомендованої комбінації засобів розробки інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, для розробки інформаційної системи було рекомендовано використати засоби розробки: Visual Studio 2019, мова програмування C#, інтерфейс програмування Windows Forms та платформу .NET.

Інформаційна система автоматизованого пошуку семантичних одиниць у цифрових текстах не потребує використання бази даних й складається із чотирьох модулів: модуля попередньої обробки тексту та формування текстового вектору, модуля визначення відстаней між появами семантичних одиниць, модуля дисперсійного оцінювання важливості семантичних одиниць і модуля формування зведеної таблиці ключових семантичних одиниць. Кожен із зазначених модулів забезпечує окремий інтерфейс для автоматизованого виведення результатів своєї роботи користувачеві.

Також у розділі розроблено окремі інтерфейси інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах для автоматизованого виведення користувачеві результатів роботи модулів інформаційної системи: модуля попередньої обробки тексту й формування текстового вектору, модуля визначення відстаней між появами семантичних одиниць, модуля дисперсійного оцінювання важливості семантичних одиниць і модуля формування зведеної таблиці ключових семантичних одиниць.

Розділ 4

Дослідження ефективності методу визначення важливості семантичних одиниць у цифрових текстах

4.1 Розробка прикладних компонентів інформаційної системи автоматизованого пошуку ключових семантичних одиниць

Для прикладного дослідження розроблених методу визначення важливості семантичних одиниць у цифрових текстах та інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах було створено інформаційну систему, що забезпечує відповідний функціонал. За коректного виконання інформаційною системою наведених функцій, можна робити висновок про її відповідність створеній інформаційної технології автоматизованого пошуку семантичних одиниць у цифрових текстах й придатність до дослідження її ефективності та відповідно ефективності методу визначення важливості семантичних одиниць у цифрових текстах. Відтак, розроблена інформаційна система визначення важливості семантичних одиниць у цифрових текстах має забезпечувати виконання функцій перегляду та редагування необхідних даних для роботи програми.

Визначення дистанції між словами у тексті відбувається за допомогою наступного програмного коду:

```
public int[,] poz_distance(List<string> mas1, int[] kilk_pojav, string[] poz_vse,int [,] mas,int zagal_kilk,
double[,] mas_kvadrat,string [,]poz_vsered)
        {
            int z = 0;
            for (int i = 0; i < mas1.Count; i++)
            {               if (mas1[i] == "") { }
                else            {
                    for (int j = 0; j < kilk_pojav[i]; j++)       {
                     if(kilk_pojav[i]==2 )
                        {   if (j == 0)
                            {      mas[i, 0] = Convert.ToInt32(poz_vse[z]);
                                mas_kvadrat[i, 1] += Math.Pow(Convert.ToInt32(poz_vse[z]), 2);//bm2 поч
                                z++;
                            }
                            else if(j == kilk_pojav[i] - 1)
                            {
                                mas[i, 1] = zagal_kilk - Convert.ToInt32(poz_vse[z]);
                                mas[i, 2] += Convert.ToInt32(poz_vse[z]) - Convert.ToInt32(poz_vse[z - 1]);
                                poz_vsered[i, 0] += (Convert.ToInt32(poz_vse[z]) - Convert.ToInt32(poz_vse[z -
1])).ToString() + " ";

                                //число в квадраті
```

```
                                mas_kvadrat[i, 0] += Math.Pow(Convert.ToInt32(poz_vse[z]),2) -
Math.Pow(Convert.ToInt32(poz_vse[z - 1]),2); //bm1 внутр
                                mas_kvadrat[i, 2] += Math.Pow(zagal_kilk - Convert.ToInt32(poz_vse[z]),
2);//bm2 кінц

                                z++;        }     }
                        else
                        {
                            if (kilk_pojav[i] == 1)
                            {
                                poz_vsered[i, 0] = "0";
                                mas[i, 0] = Convert.ToInt32(poz_vse[z]);
                                mas[i, 1] = zagal_kilk - Convert.ToInt32(poz_vse[z]);
                                mas_kvadrat[i, 1] += Math.Pow(Convert.ToInt32(poz_vse[z]), 2);//bm2 поч
                                mas_kvadrat[i, 2] += Math.Pow(zagal_kilk - Convert.ToInt32(poz_vse[z]),
2);//bm2 кінц
                                z++;
                            }
                            else
                                if (kilk_pojav[i] > 2)
                            {
                                if (j == 0)
                                {
                                    mas[i, 0] = Convert.ToInt32(poz_vse[z]);
                                    mas_kvadrat[i, 1] += Math.Pow(Convert.ToInt32(poz_vse[z]), 2);//bm2 поч
                                    z++;
                                }
                                else if (j == kilk_pojav[i] - 1)
                                {
                                    mas[i, 1] = zagal_kilk - Convert.ToInt32(poz_vse[z]);
                                    mas[i, 2] += Convert.ToInt32(poz_vse[z]) - Convert.ToInt32(poz_vse[z - 1]);
                                    poz_vsered[i, 0] += (Convert.ToInt32(poz_vse[z]) -
Convert.ToInt32(poz_vse[z - 1])).ToString() + " ";

                                    mas_kvadrat[i, 0] += Math.Pow(Convert.ToInt32(poz_vse[z]), 2) -
Math.Pow(Convert.ToInt32(poz_vse[z - 1]), 2);
                                    mas_kvadrat[i, 2] += Math.Pow(zagal_kilk - Convert.ToInt32(poz_vse[z]),
2);//bm2 кінц
                                    z++;   }
                                else     {
                                    mas[i, 2] += Convert.ToInt32(poz_vse[z]) - Convert.ToInt32(poz_vse[z - 1]);
                                    poz_vsered[i, 0] += (Convert.ToInt32(poz_vse[z]) -
Convert.ToInt32(poz_vse[z - 1])).ToString() + " ";

                                    mas_kvadrat[i, 0] += Math.Pow(Convert.ToInt32(poz_vse[z]), 2) -
Math.Pow(Convert.ToInt32(poz_vse[z - 1]), 2);
                                    z++;
                                }     }
                            else { z++;  }        }
                        mas[i, 3] = mas[i, 0] + mas[i, 1];
                        mas_kvadrat[i, 3] += Math.Pow(mas[i, 0],2) + Math.Pow(mas[i, 1], 2);//bm3 кільц

                    } }   }       return mas;
        }
```

Результат виконання даного програмного коду зображено на рисунку 4.1.

Обрахунок дисперсії слів за методом BM1 відбувається за допомогою наступного програмного коду:

```
public double[] BM1(List<string> mas1, double[] mas_bm1,int [,]mas,int []kilk_pojav,double [,] kvadrat)
        {
            for (int i = 0; i < mas1.Count; i++)
            {
                if (mas1[i] == "") { }
                else
                {
                   if(kilk_pojav[i] <=2 ) { }
                        else
                    mas_bm1[i] = Math.Sqrt((kvadrat[i,0]/kilk_pojav[i]) - (Math.Pow(mas[i,2]/kilk_pojav[i],2))) /
(mas[i, 2] / kilk_pojav[i]);
```

```
            }
        }

        return mas_bm1;
    }
```

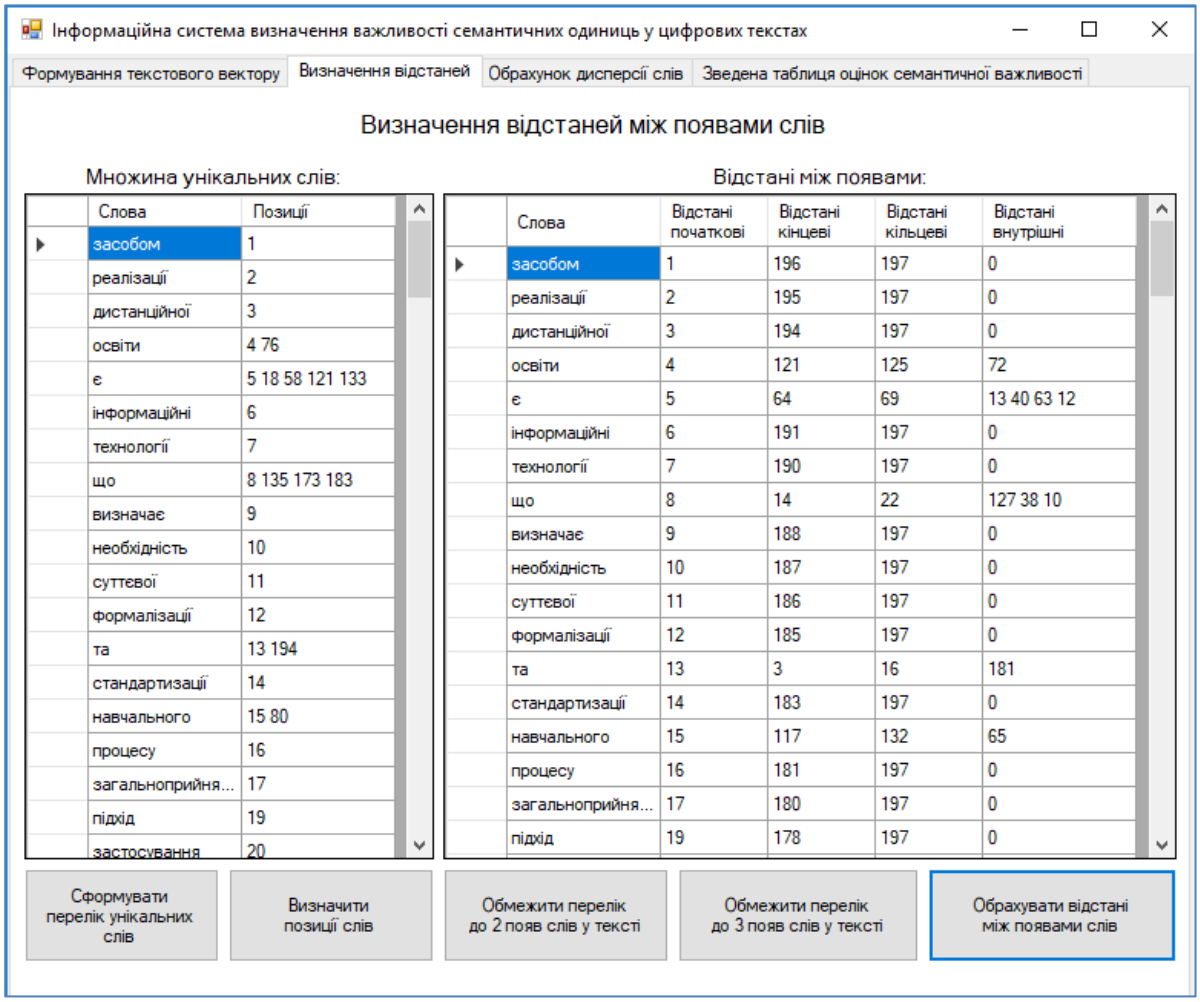


Рисунок 4.1 – Визначення відстаней між появами слів

Результат виконання даного програмного коду зображено на рисунку 4.2.

Таким чином, за допомогою програмних кодів формується інформаційна система автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, яка виконує усі необхідні функції. Інформаційну систему, що забезпечує відповідний функціонал, було створено для прикладного дослідження розроблених методу визначення важливості семантичних одиниць у цифрових текстах та інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах. За коректного виконання інформаційною системою наведених функцій, можна робити висновок про її

відповідність створеній інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах й придатність до дослідження її ефективності і відповідно ефективності методу визначення важливості семантичних одиниць у цифрових текстах.

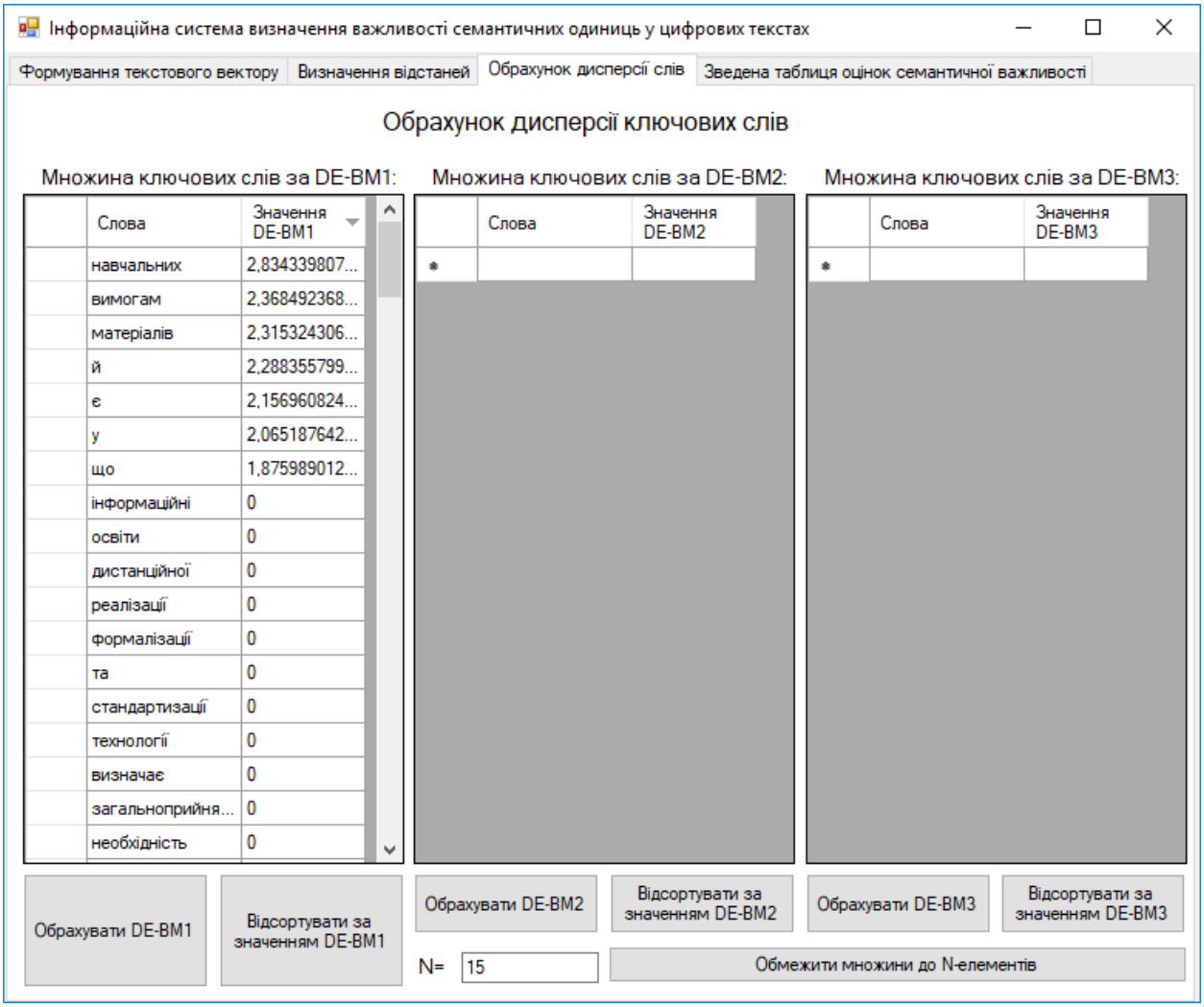


Рисунок 4.2 – Обрахунок дисперсії слів за методом BM1

Інформаційна система автоматизованого пошуку семантичних одиниць у цифрових текстах згідно п.2.4 має дозволяти за створеною інформаційною технологією в результаті обробки вхідного цифрового тексту у вигляді відповідної впорядкованої множини символів виконувати автоматизоване визначення множин ключових семантичних одиниць за оцінками модифікацій дисперсійного оцінювання DE-BM1, DE-BM2 і DE-BM3, також формувати відповідну зведену таблицю оцінок семантичної важливості ключових

семантичних одиниць вхідного цифрового тексту. Тому розроблена інформаційна система автоматизованого пошуку ключових семантичних одиниць у цифрових текстах має забезпечувати виконання функцій перегляду та редагування необхідних даних для роботи програми, для чого слід провести прикладне тестування інформаційної системи.

4.2 Прикладне тестування інформаційної системи

Для проведення тестового дослідження функціональності розробленої інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах було розроблено два тестові випадки.

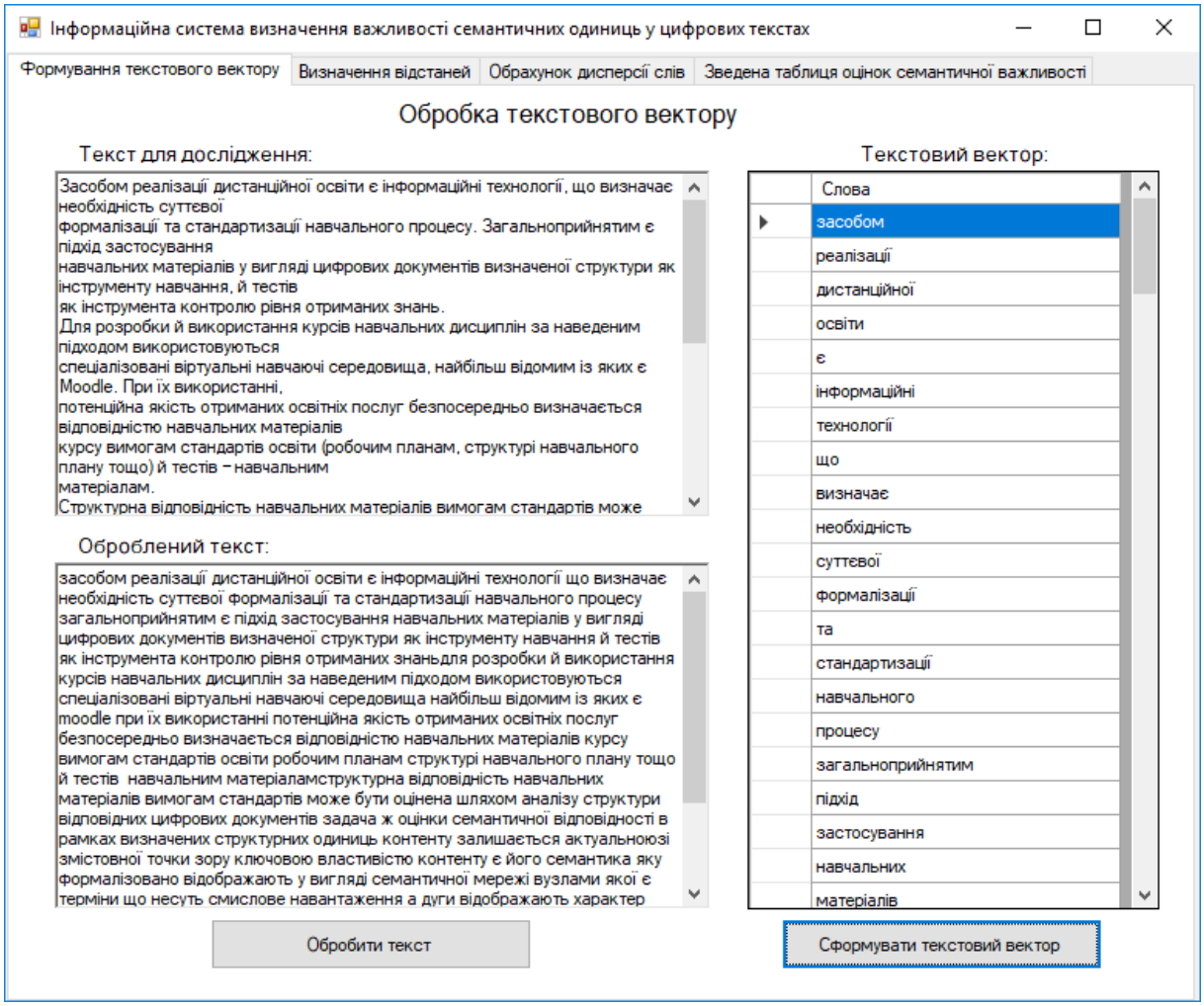
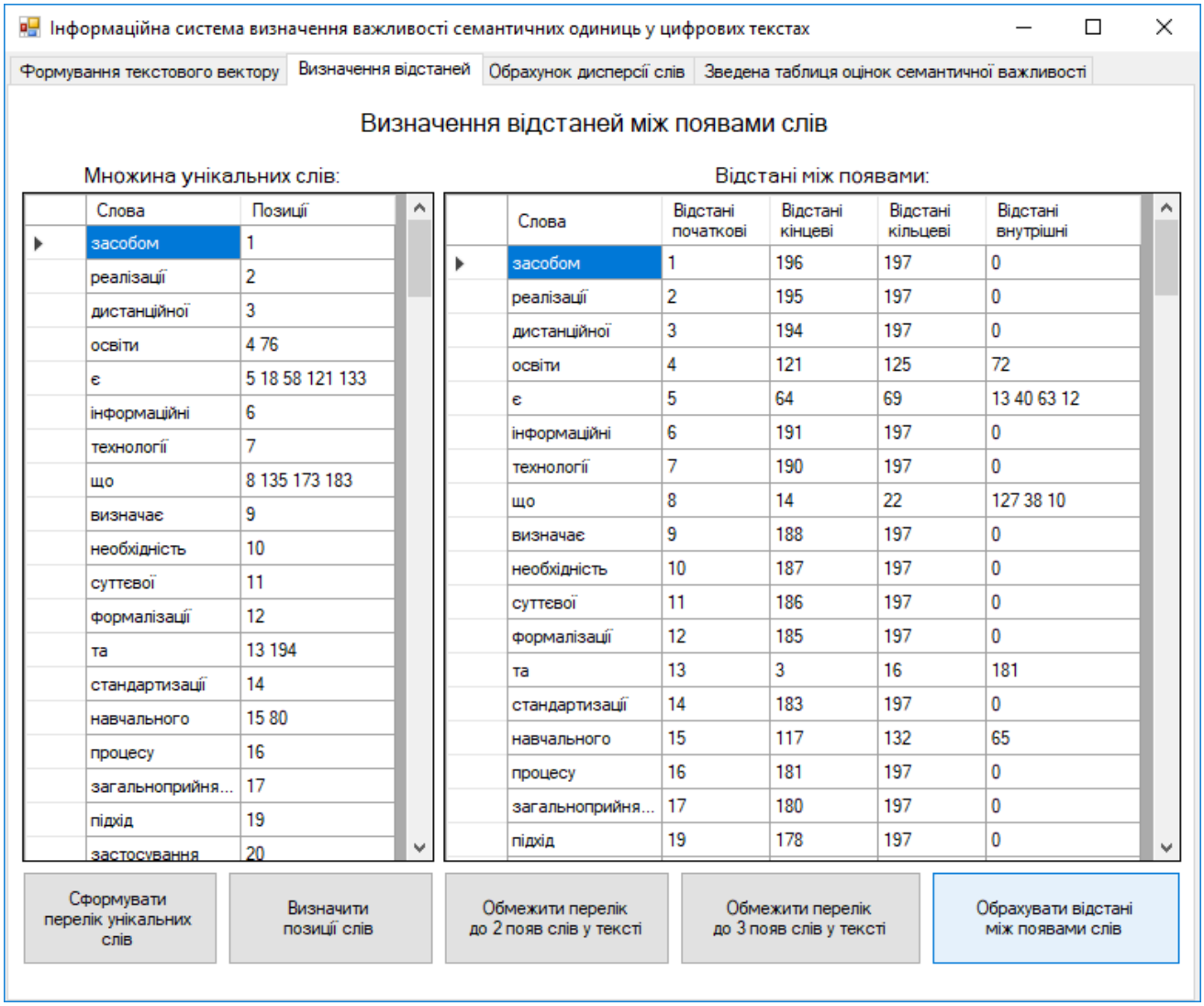


Рисунок 4.3 – Текстовий вектор, сформований у інформаційній системі автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

Таблиця 4.1 – Тест-кейс АК0001

Тест-кейс ID: AB0001	**Пріоритет:** 1	**Створено:** 28.10.2021, О.О.Войчишин
Назва: Перевіряється коректність визначення текстового вектору **Вхідні дані:** Оброблений текст		
Кроки		**Очікуваний результат**
1. Запустити додаток 2. Натиснути на кнопку «Обробити текст» 3. Натиснути на кнопку «Сформувати текстовий вектор» 4. Порівняти фактичний результат з очікуваним		Відображення коректних даних.
Результат виконання тест-кейсу: пройдено успішно		



Слова	Позиції
засобом	1
реалізації	2
дистанційної	3
освіти	4 76
є	5 18 58 121 133
інформаційні	6
технології	7
що	8 135 173 183
визначає	9
необхідність	10
суттєвої	11
формалізації	12
та	13 194
стандартизації	14
навчального	15 80
процесу	16
загальноприйня...	17
підхід	19
застосування	20

Слова	Відстані початкові	Відстані кінцеві	Відстані кільцеві	Відстані внутрішні
засобом	1	196	197	0
реалізації	2	195	197	0
дистанційної	3	194	197	0
освіти	4	121	125	72
є	5	64	69	13 40 63 12
інформаційні	6	191	197	0
технології	7	190	197	0
що	8	14	22	127 38 10
визначає	9	188	197	0
необхідність	10	187	197	0
суттєвої	11	186	197	0
формалізації	12	185	197	0
та	13	3	16	181
стандартизації	14	183	197	0
навчального	15	117	132	65
процесу	16	181	197	0
загальноприйня...	17	180	197	0
підхід	19	178	197	0

Рисунок 4.4 – Виведення у інформаційній системі множини унікальних слів та відстаней між появами слів

У першому тест-кейсі (таблиця 4.1) перевіряється коректність визначення текстового вектору. При натисканні на кнопку «Сформувати текстовий вектор» в таблиці «Текстовий вектор» мають відобразитись коректні дані.

У випадку коли натиснули на кнопку «Сформувати текстовий вектор», у інформаційній системі визначення важливості семантичних одиниць у цифрових текстах було відображено обраховані дані (рисунок 4.3).

Другий тестовий випадок (таблиця 4.2) перевіряє коректність виведення обрахованих даних. При натисканні на кнопку «Сформувати перелік унікальних слів», відображається множина унікальних слів, при натисканні на кнопку «Визначити позиції слів», відображаються позиції унікальних слів, при натисканні на кнопку «Обрахувати відстані між появами слів», відображаються відстані між появами слів у відповідній таблиці (рисунок 4.4).

Таблиця 4.2 – Тест-кейс AB0002

<table>
<tr><td>Тест-кейс ID: AK0002</td><td>Пріоритет: 1</td><td>Створено: 28.10.2021, О.О.Войчишин</td></tr>
<tr><td colspan="3">Назва: Перевірка коректності виведення обрахованих даних
Вхідні дані: Оброблений текст</td></tr>
<tr><td colspan="2">Кроки</td><td>Очікуваний результат</td></tr>
<tr><td colspan="2">1. Запустити додаток
2. Натиснути на кнопку «Обробити текст»
3. Натиснути на кнопку «Сформувати текстовий вектор»
4. Перейти на вкладку «Визначення відстаней»
5. Натиснути на кнопку «Сформувати перелік унікальних слів»
6. Натиснути на кнопку «Визначити позиції слів»
7. Натиснути на кнопку «Обрахувати відстані між появами слів»
8. Порівняти фактичний результат з очікуваним</td><td>Відображення коректних даних.</td></tr>
<tr><td colspan="3">Результат виконання тест-кейсу: пройдено успішно</td></tr>
</table>

Отже, прикладне тестування функціональності інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах було успішно пройдено.

4.3 Дослідження функціональності інформаційної системи автоматизованого пошуку ключових семантичних одиниць

Робота користувача з інформаційною системою автоматизованого пошуку ключових семантичних одиниць розпочинається з першої вкладки «Формування текстового вектору». Робоча область містить 2 області перегляду інформації та 2 елемента для роботи з нею (рисунок 4.5).

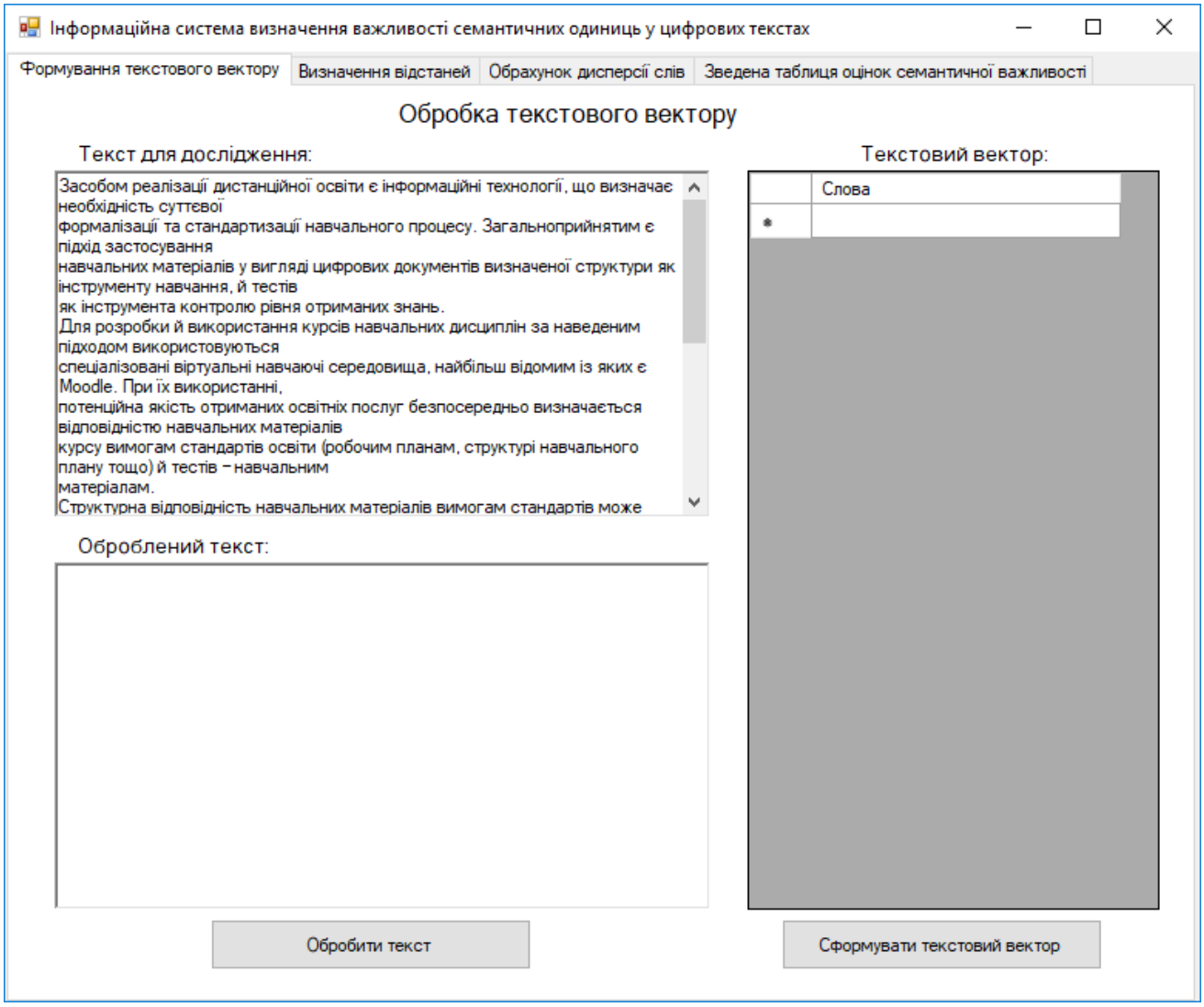


Рисунок 4.5 – Вкладка «Формування текстового вектору»

Для обробки тексту (видалення знаків пунктуації, зміна регістру та ін.) необхідно натиснути кнопку «Обробити текст», в результаті буде отримано оброблений текст (рисунок 4.6).

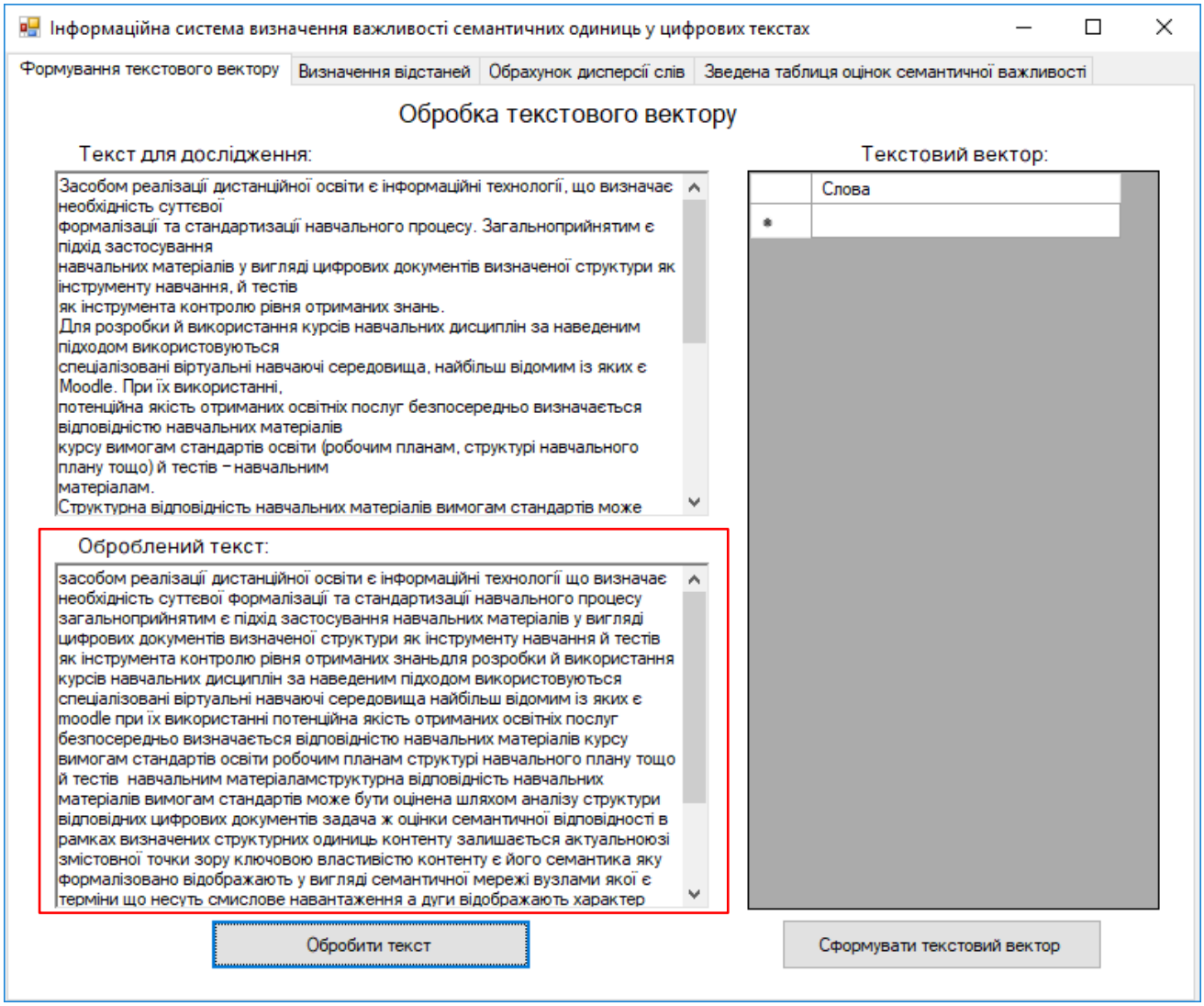


Рисунок 4.6 – Оброблений текст при роботі інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

По натисканню на кнопку «Сформувати текстовий вектор», формується та заповняється словами масив текстового вектору (рисунок 4.7).

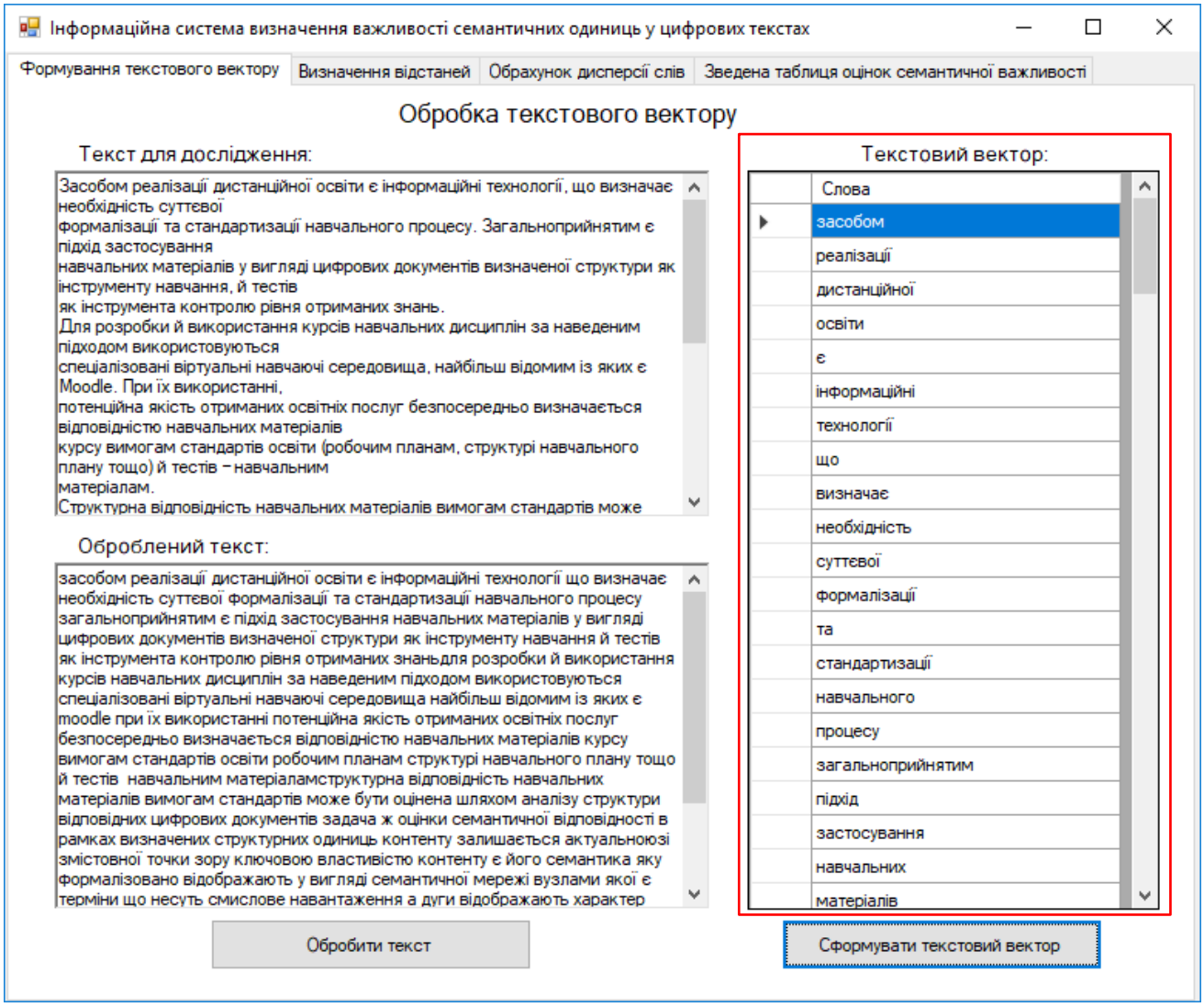


Рисунок 4.7 – Сформований текстовий вектор при роботі інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

Для визначення відстаней між появами слів, користувачу необхідно перейти на вкладку «Визначення відстаней». Робоча область вкладки містить 2 області перегляду інформації та 5 елементів для роботи з нею (рисунок 4.8).

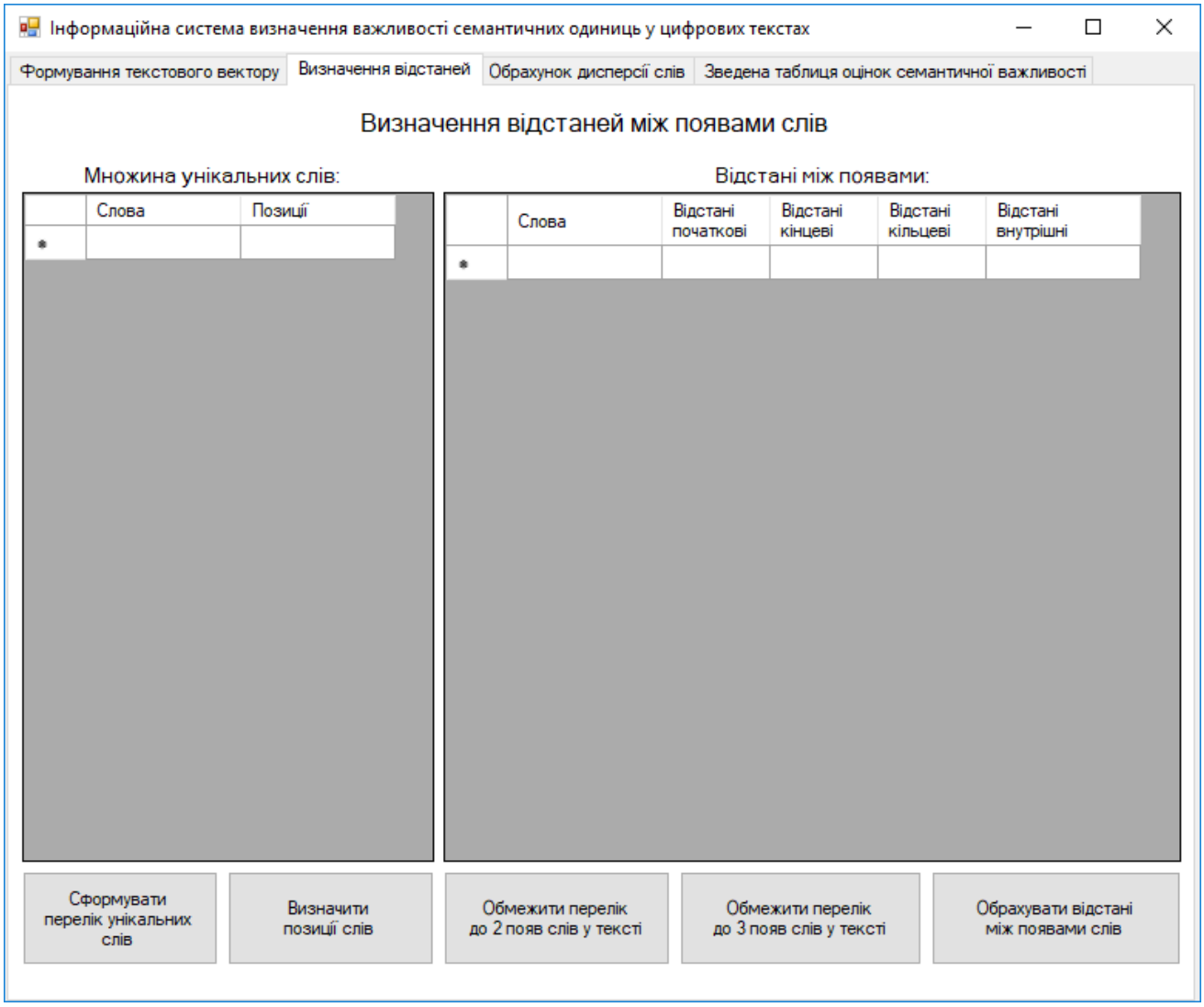


Рисунок 4.8 – Вкладка «Визначення відстаней» інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

Для формування переліку унікальних слів користувач має натиснути на відповідну кнопку, після чого відповідні дані запишуться у таблицю множини слів (рисунок 4.9).

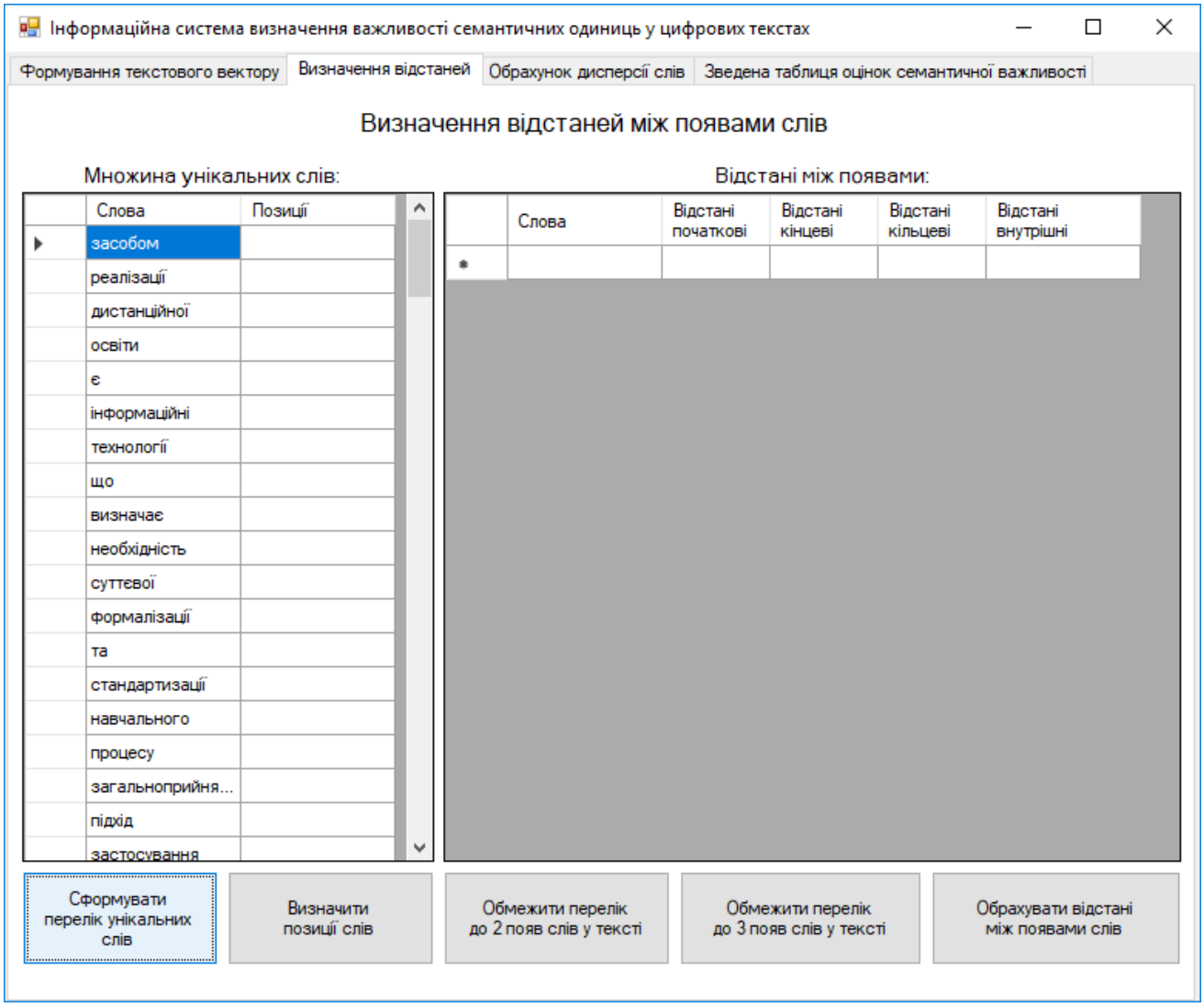


Рисунок 4.9 – Сформована множина унікальних слів при роботі інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

Після визначення унікальних слів, по натисканню на кнопку «Визначити позиції слів», для користувача обраховуються та виводяться дані позицій слів у тексті (рисунок 4.10).

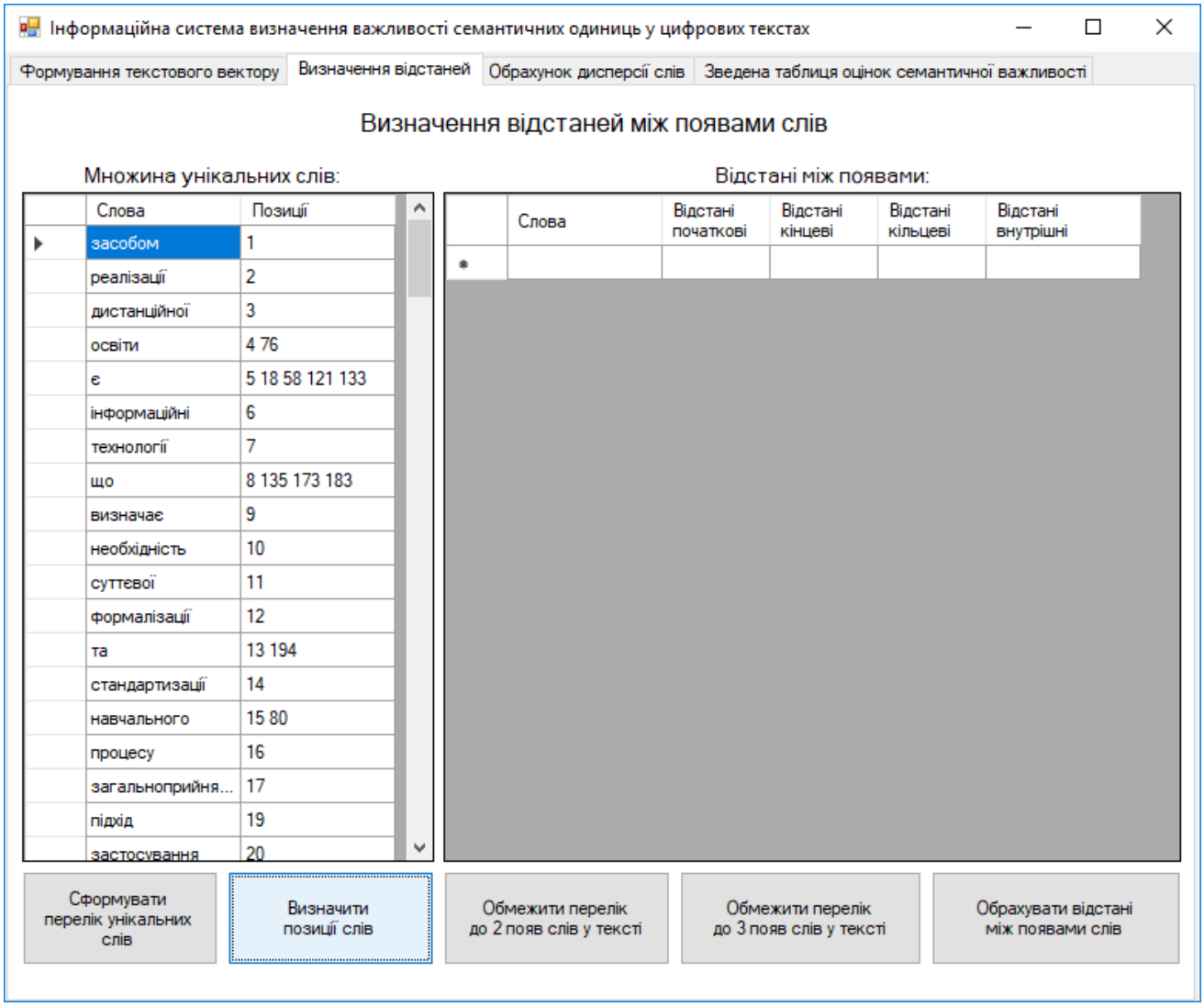


Рисунок 4.10 – Позиції слів у тексті при роботі інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

Користувач має можливість обмежити перелік слів до 2 або 3 його появ у тексті, для цього необхідно натиснути на кнопки «Обмежити перелік до 2 появ слів у тексті» або «Обмежити перелік до 3 появ слів у тексті» (рисунок 4.11-4.12).

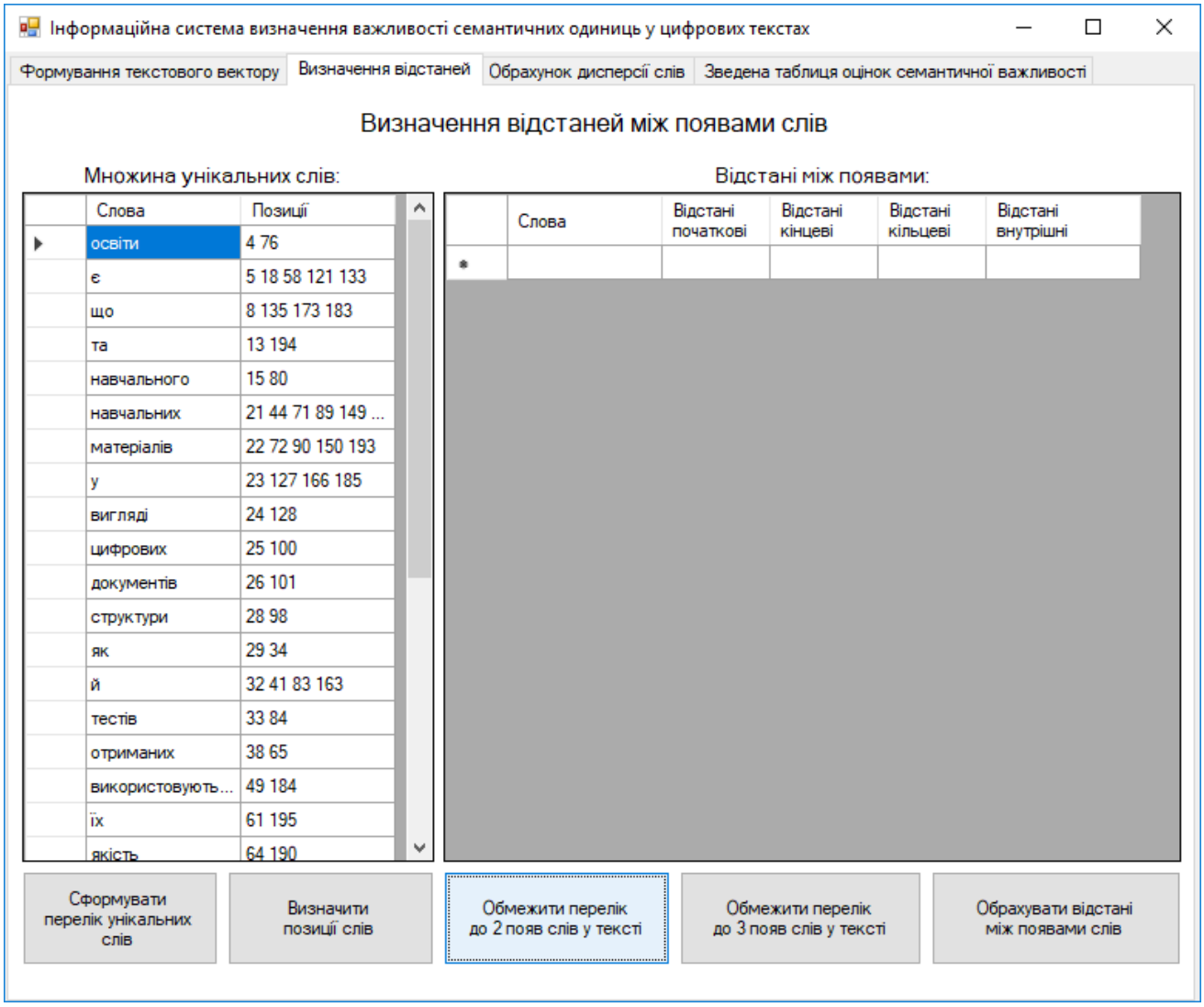


Рисунок 4.11 – Обмеження переліку унікальних слів до 2 появ у тексті при роботі інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

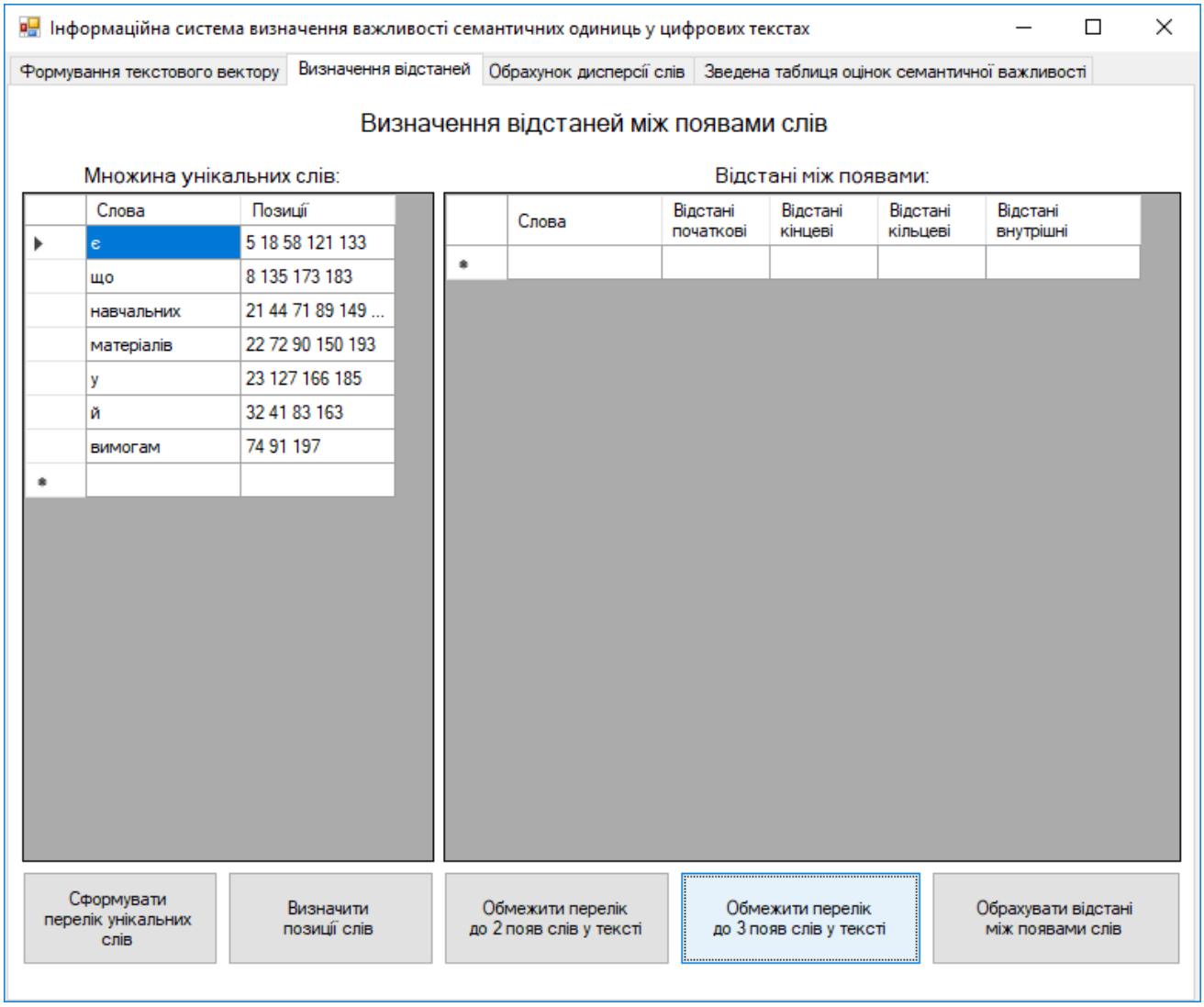
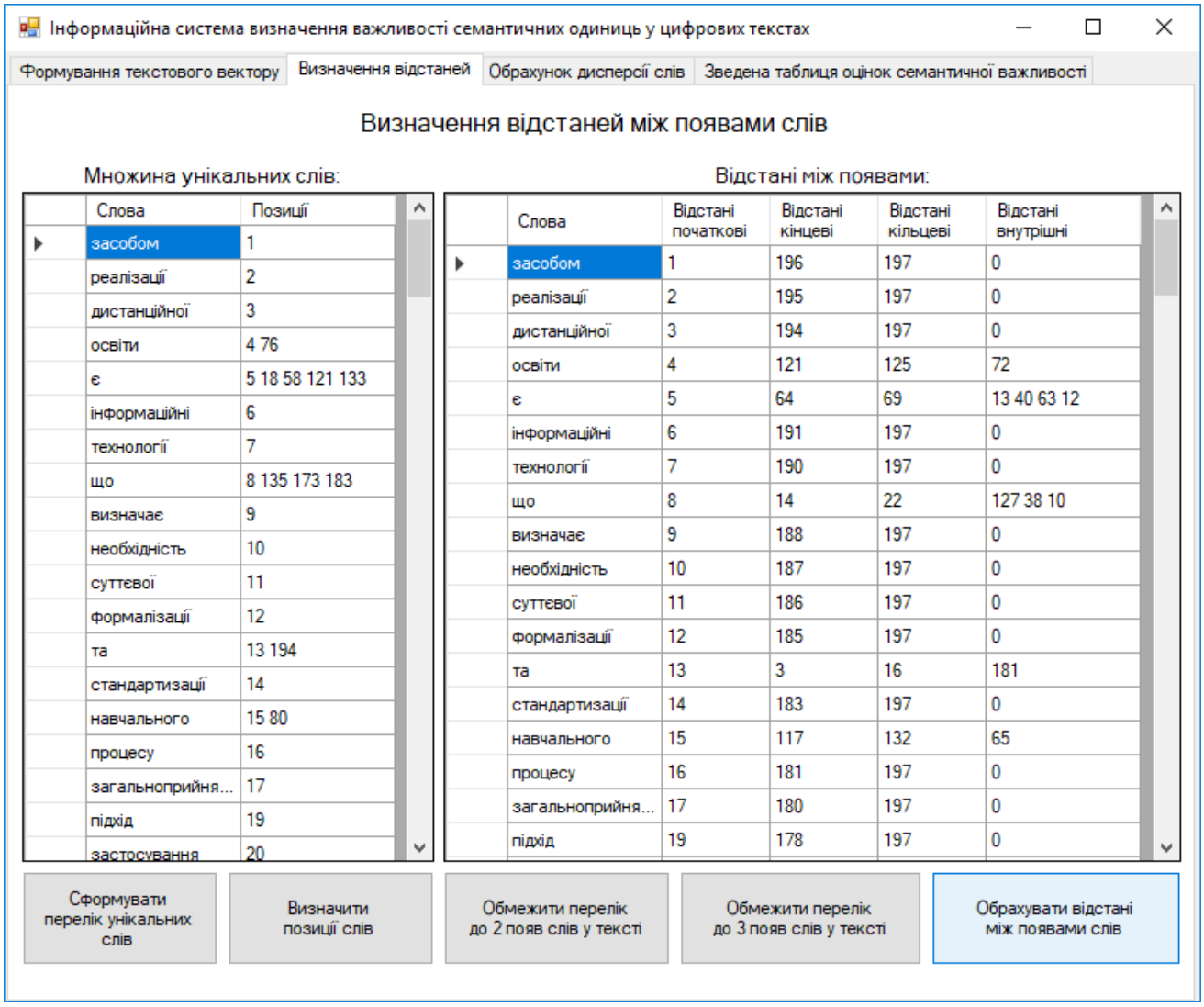


Рисунок 4.12 – Обмеження переліку унікальних слів до 3 появ у тексті при роботі інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

Для обрахування відстаней між появами слів, користувачу необхідно натиснути на відповідну кнопку на вкладці. Після натискання в таблицю виведуться обраховані дані (рисунок 4.13).



Слова	Позиції
засобом	1
реалізації	2
дистанційної	3
освіти	4 76
є	5 18 58 121 133
інформаційні	6
технології	7
що	8 135 173 183
визначає	9
необхідність	10
суттєвої	11
формалізації	12
та	13 194
стандартизації	14
навчального	15 80
процесу	16
загальноприйня...	17
підхід	19
застосування	20

Слова	Відстані початкові	Відстані кінцеві	Відстані кільцеві	Відстані внутрішні
засобом	1	196	197	0
реалізації	2	195	197	0
дистанційної	3	194	197	0
освіти	4	121	125	72
є	5	64	69	13 40 63 12
інформаційні	6	191	197	0
технології	7	190	197	0
що	8	14	22	127 38 10
визначає	9	188	197	0
необхідність	10	187	197	0
суттєвої	11	186	197	0
формалізації	12	185	197	0
та	13	3	16	181
стандартизації	14	183	197	0
навчального	15	117	132	65
процесу	16	181	197	0
загальноприйня...	17	180	197	0
підхід	19	178	197	0

Рисунок 4.13 – Виведення відстаней між появами слів при роботі інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

Для обрахунку дисперсії слів, користувачу необхідно перейти на відповідну вкладку (рисунок 4.14).

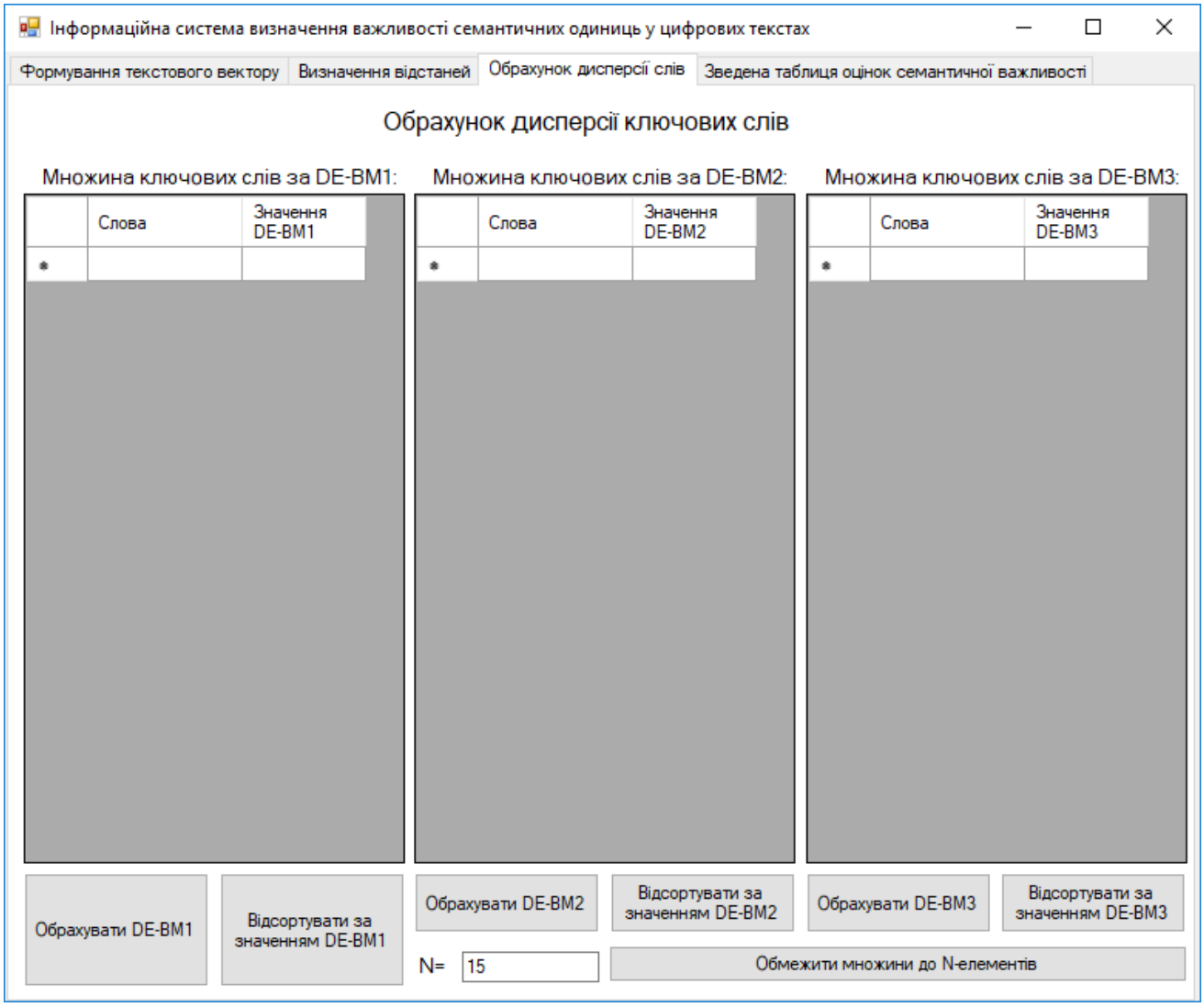


Рисунок 4.14 – Вкладка «Обрахунок дисперсії слів» інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

Дисперсія слів обраховується за трьома методами, для цього користувачу відведено 3 кнопки обрахування: «Обрахувати DE-BM1», «Обрахувати DE-BM2», «Обрахувати DE-BM3», та 3 кнопки для сортування обрахованих даних (рисунок 4.15). Також, є можливість обмежити множину слів до певної кількості елементів, для цього відведено окрему кнопку (рисунок 4.16).

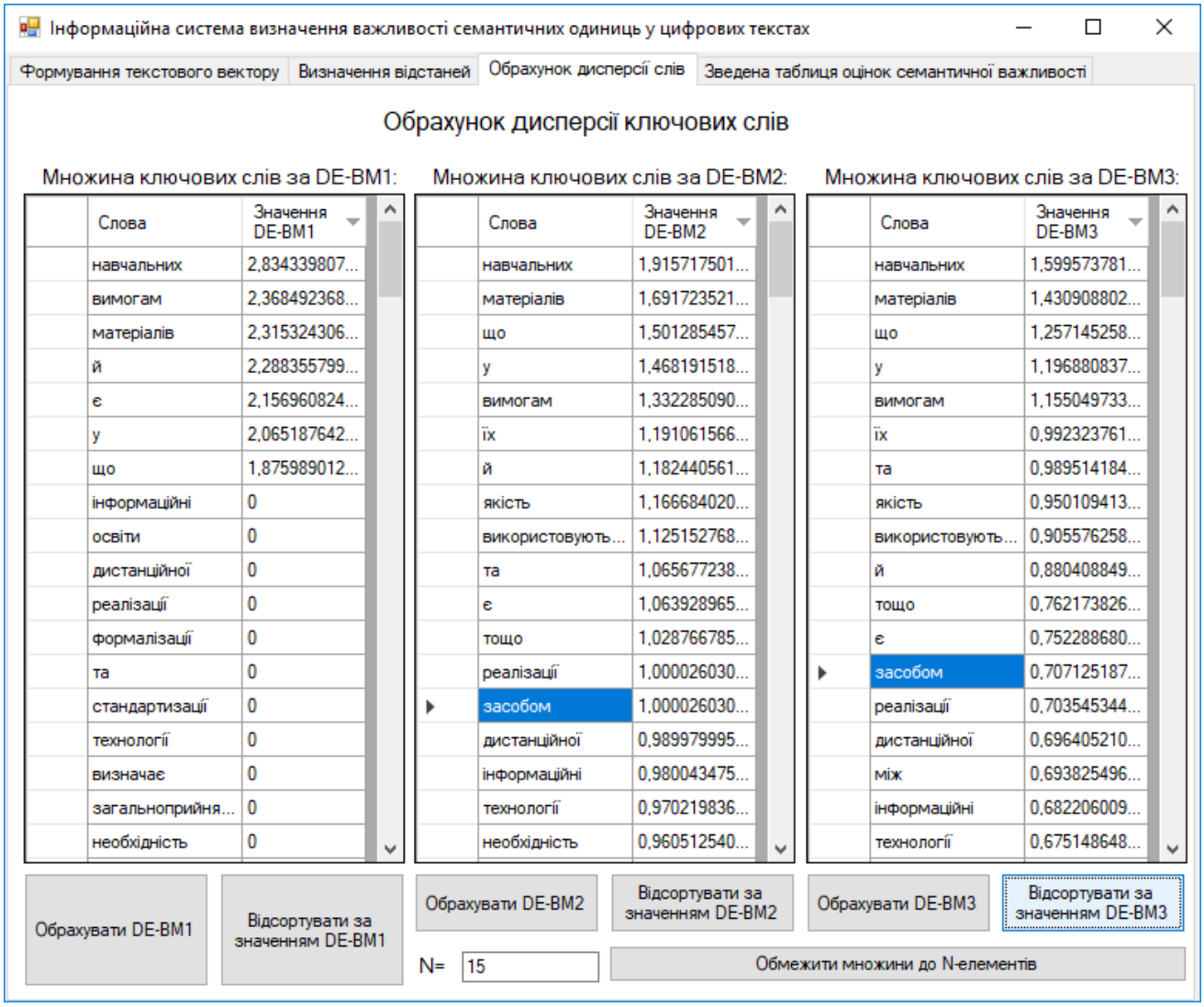


Рисунок 4.15 – Обраховані дисперсії слів при роботі інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

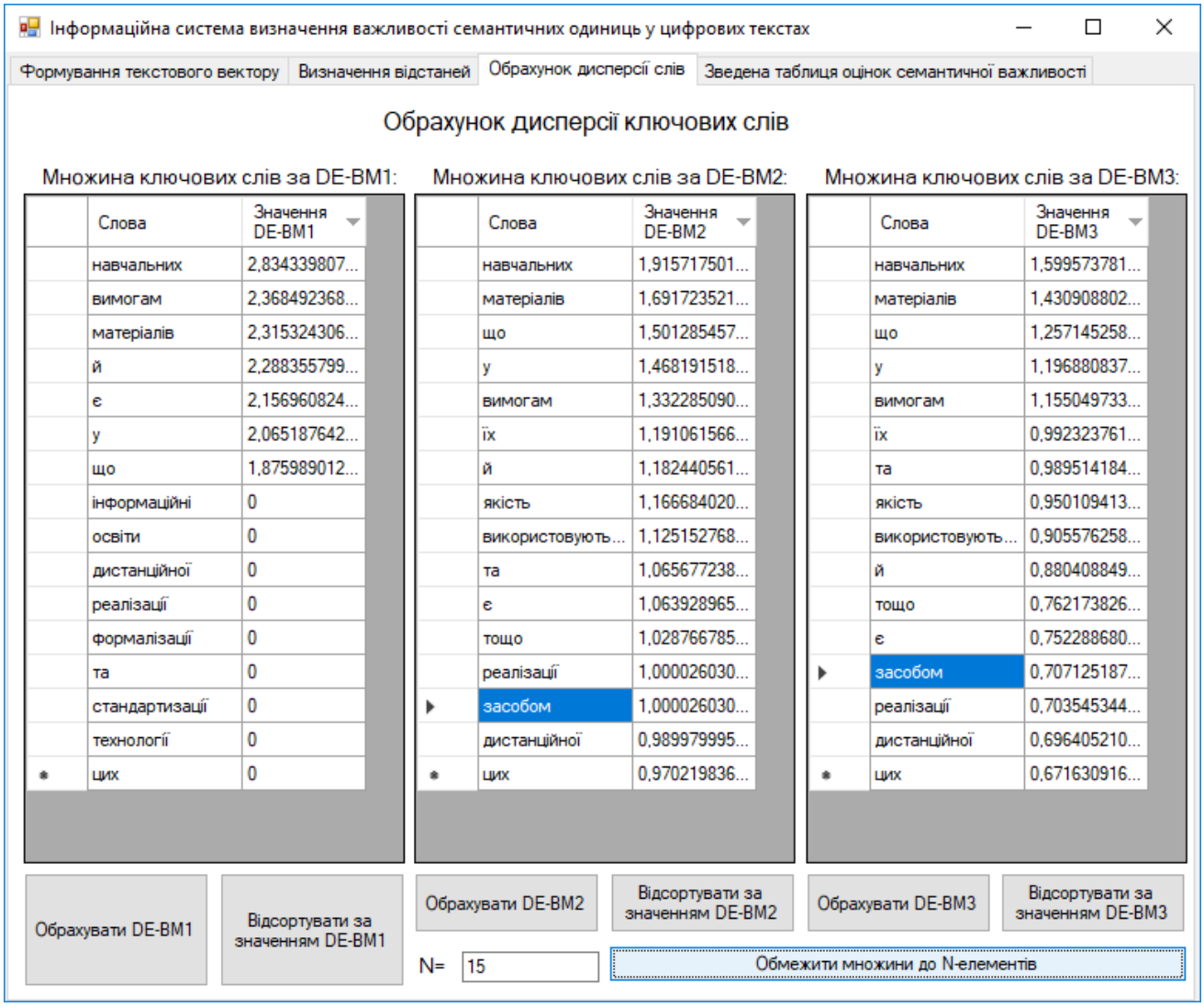
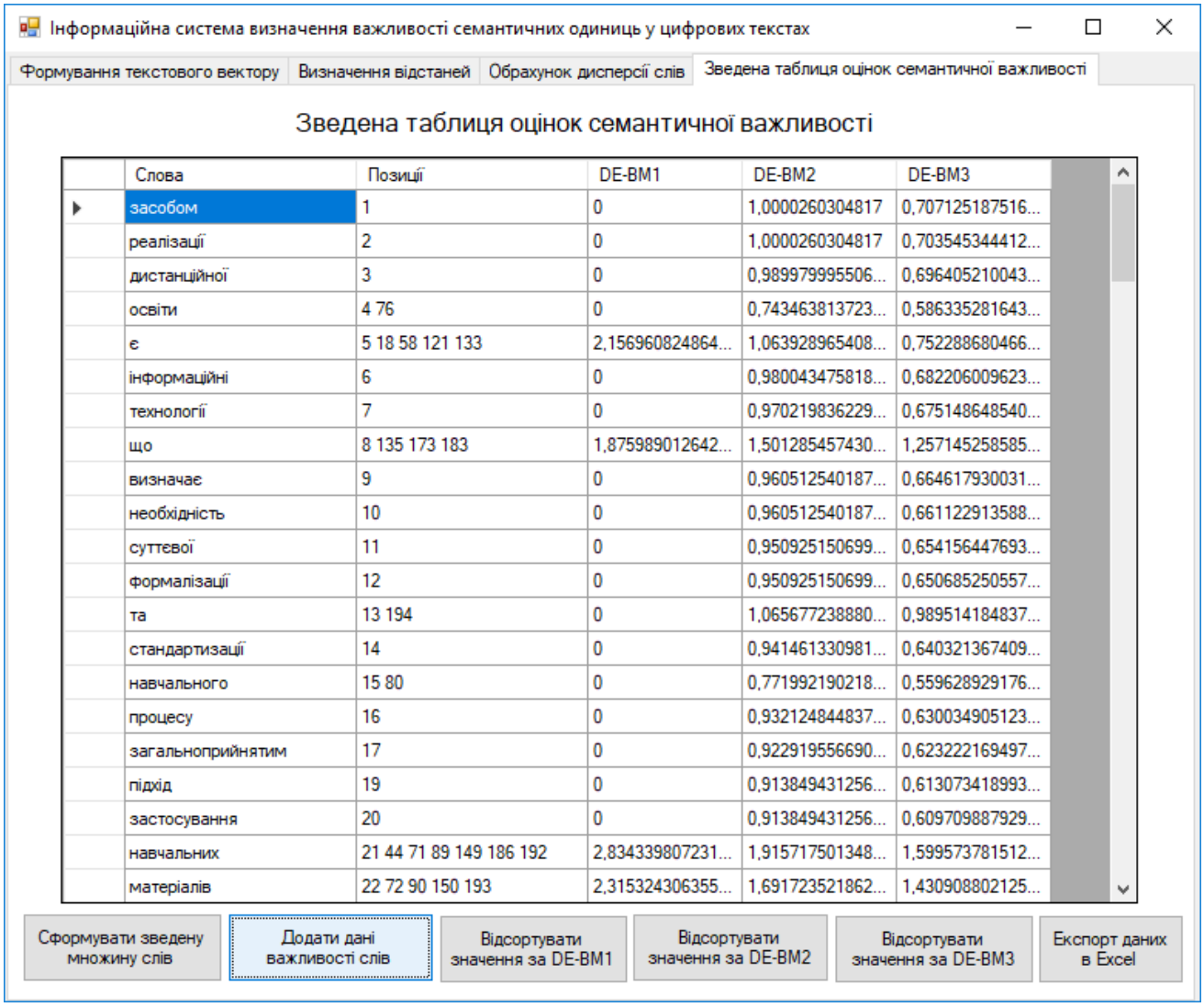


Рисунок 4.16 – Обмеження множини слів при роботі інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

Щоб зручно було переглядати обраховані дані, на відповідній вкладці створено зведену таблицю оцінок семантичної важливості та необхідні для виконання функцій кнопки (рисунок 4.17).



Слова	Позиції	DE-BM1	DE-BM2	DE-BM3
засобом	1	0	1,0000260304817	0,707125187516...
реалізації	2	0	1,0000260304817	0,703545344412...
дистанційної	3	0	0,989979995506...	0,696405210043...
освіти	4 76	0	0,743463813723...	0,586335281643...
є	5 18 58 121 133	2,156960824864...	1,063928965408...	0,752288680466...
інформаційні	6	0	0,980043475818...	0,682206009623...
технології	7	0	0,970219836229...	0,675148648540...
що	8 135 173 183	1,875989012642...	1,501285457430...	1,257145258585...
визначає	9	0	0,960512540187...	0,664617930031...
необхідність	10	0	0,960512540187...	0,661122913588...
суттєвої	11	0	0,950925150699...	0,654156447693...
формалізації	12	0	0,950925150699...	0,650685250557...
та	13 194	0	1,065677238880...	0,989514184837...
стандартизації	14	0	0,941461330981...	0,640321367409...
навчального	15 80	0	0,771992190218...	0,559628929176...
процесу	16	0	0,932124844837...	0,630034905123...
загальноприйнятим	17	0	0,922919556690...	0,623222169497...
підхід	19	0	0,913849431256...	0,613073418993...
застосування	20	0	0,913849431256...	0,609709887929...
навчальних	21 44 71 89 149 186 192	2,834339807231...	1,915717501348...	1,599573781512...
матеріалів	22 72 90 150 193	2,315324306355...	1,691723521862...	1,430908802125...

Рисунок 4.17 – Зведена таблиця даних інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

Таким чином, інтерфейс інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах є досить простим та зрозумілим, що дозволяє користувачу використовувати програму без жодних ускладнень. Інформаційну систему було створено для прикладного дослідження розроблених методу визначення важливості семантичних одиниць у цифрових текстах та інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах.

Дослідження функціональності інформаційної системи автоматизованого пошуку ключових семантичних одиниць виявило коректне виконання інформаційною системою функцій відповідно до інформаційної технології

автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, тому було зроблено висновок про її відповідність створеній інформаційної технології й придатність до дослідження її ефективності і відповідно ефективності методу визначення важливості семантичних одиниць у цифрових текстах.

4.4 Дослідження ефективності методу визначення важливості семантичних одиниць у цифрових текстах

Згідно з п.2.2, метод визначення важливості семантичних одиниць у цифрових текстах на основі дисперсійного оцінювання дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а й із врахуванням початкових, кінцевих та похідних кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту.

Обрахунок відстаней між словами тексту є підготовчим етапом до дисперсійного оцінювання ключових семантичних одиниць, за якого визначаються для кожного слова із кількістю появ у тексті більше одного всі відстані між сусідніми їх появами [30].

У залежності від того, яким чином та у якій кількості визначаються відстані для кожного унікального слова тексту [31], розрізняють різні модифікації вихідного методу DE-BM пошуку ключових слів за дисперсійним оцінюванням DE-BM1, а саме: DE-BM2 та DE-BM3.

Метод пошуку ключових семантичних одиниць початкової модифікації DE-BM1 для n появ слова враховує n-1 відстані. При цьому за відстань береться різниця між меншим порядковим номером наступного слова й більшим порядковим номером попереднього слова.

Метод пошуку ключових семантичних одиниць модифікації DE-BM2 для n появ слова враховує n+1 відстань. За відстань береться різниця між меншим

порядковим номером наступного слова й більшим порядковим номером попереднього слова. Також додатково враховуються відстані: від початку тексту до першої появи слова в тексті, від останньої появи слова в тексті до кінця тексту.

Метод пошуку ключових семантичних одиниць модифікації DE-BM3 для *n* появ слова враховує *n* відстаней. За відстань береться різниця між меншим порядковим номером наступного слова і більшим порядковим номером попереднього слова. Також додатково враховується відстань, рівна сумі різниць між початком тексту до першої появи слова і між останньою появою слова до кінця тексту.

Наприклад, в тексті (рис. 4.18) [32], що складається з 131 слова, й в якому слово «інтелект» зустрічається на позиціях 1, 11, 34, 60 і 83, для обрахунку дисперсійної оцінки ключових семантичних одиниць можна використати наступні відстані:

- відстані, рівні 10, 23, 26, та 23 за класичним підходом до обрахунку;
- додаткова відстань від початку тексту до першої появи слова у тексті 1;
- додаткова відстань від останньої появи слова у тексті до кінця тексту 48;
- додаткова відстань, рівна сумі різниць між початком тексту до першої появи слова та між останньою появою слова до кінця тексту 49.

Відповідно, для різновидів дисперсійного оцінювання семантичних одиниць буде використано наступні відстані:

- для вихідного методу пошуку ключових семантичних одиниць за дисперсійним оцінюванням DE-BM1: 10, 23, 26, 23;
- для методу пошуку ключових семантичних одиниць DE-BM2: 1, 10, 23, 26, 23, 48;
- для методу пошуку ключових семантичних одиниць DE-BM3: 10, 23, 26, 23, 49.

Дослідження ефективності розглянутих різновидів методів пошуку ключових семантичних одиниць може визначити їх особливості та рекомендовані області застосування.

Інтелект властивий людям, а також спостерігається у тварин.

Людина застосовує інтелект для обробки наявної інформації, наприклад, з метою побудови або вдосконалення розуміння, позиції, стратегії, методу, правила, комбінації, відношення, пояснення, рішення, плану чи цілі. Інтелект пов'язаний з іншими внутрішніми властивостями людини, такими як сприйняття, пам'ять, мова, уява, самосвідомість, самоконтроль, характер, володіння тілом, творчість, інтуїція і власне формується завдяки функціонуванню означених параметрів особистості. Інтелект найчастіше спрямовується на вирішення питань облаштування побуту і відпочинку, професійну діяльність, міжособистісні стосунки та самовдосконалення.

В повсякденному житті в сучасній розвинутій людині інтелект також проявляє себе у вигляді внутрішніх почуттів і образів мислення, таких як відчуття реальності, часу, простору, себе, ритму, гумору, відповідальності, ситуації, прекрасного, захищеності, небезпеки, такту, комфорту, міри, справедливості, довіри, свободи, поваги, власної гідності та інших, і у вигляді аналітичного, образного, практичного, абстрактного, тактичного або стратегічного образу мислення.

Рисунок 4.18 – Дослідний текст [32] для аналізу методом визначення важливості семантичних одиниць у цифрових текстах

В роботі для оцінки якості пошуку ключових семантичних одиниць використовуються оцінка точності [33]. Точність пошуку P є відношенням числа релевантних ключових семантичних одиниць знайдених автоматично, до загальної кількості знайдених ключових семантичних одиниць в тексті:

$$P = \frac{|M_{TK}^{E} \cap M_{TK}|}{|M_{TK}|}, \tag{4.1}$$

де M_{TK}^{E} – множина релевантних ключових семантичних одиниць, сформована експертом; M_{TK} – множина знайдених автоматично ключових семантичних одиниць.

Середня точність пошуку ключових семантичних одиниць $\overline{P}$ визначається так [33]:

$$\overline{P} = \frac{\sum_{i=1}^{k} P_k}{k}, \tag{4.2}$$

де k – кількість текстів у тестовій вибірці.

Наведені показники дозволять у подальшому оцінювати ефективність реалізованих методів пошуку ключових семантичних одиниць із використанням розробленої інформаційної системи.

Було проведено дослідження ефективності визначення важливості семантичних одиниць у цифрових текстах шляхом аналізу цифрових текстів інформаційною системою автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, яка дозволяє за створеною інформаційною технологією в результаті обробки вхідного цифрового тексту у вигляді відповідної впорядкованої множини символів виконувати автоматизоване визначення множин ключових семантичних одиниць за оцінками модифікацій дисперсійного оцінювання DE-BM1, DE-BM2 і DE-BM3.

Для аналізу було використано 81 текст обсягом від 300 до 500 слів і 35 текстів обсягом від 500 до 2000 слів з відкритих джерел із множинами ключових семантичних одиниць, визначеними їх авторами.

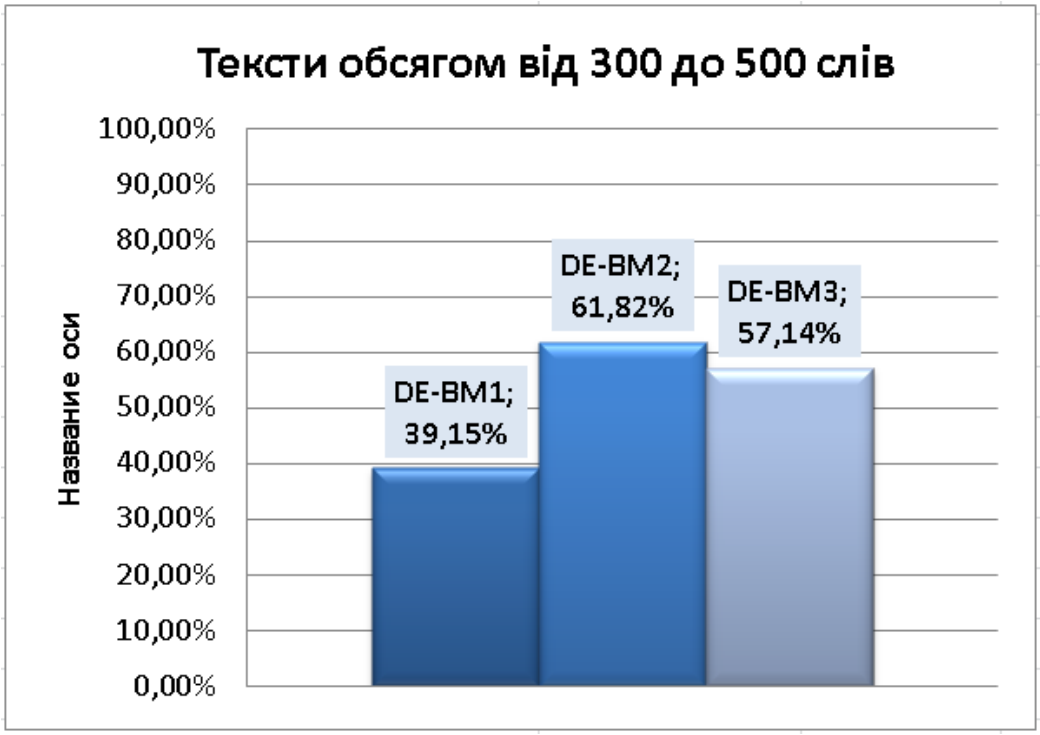


Рисунок 4.19 – Діаграма порівняння ефективності модифікацій методу визначення важливості семантичних одиниць у цифрових текстах при пошуку ключових семантичних одиниць у текстах обсягом від 300 до 500 слів

Дослідження ефективності методу визначення важливості семантичних одиниць у цифрових текстах при пошуку ключових семантичних одиниць у текстах обсягом від 300 до 500 слів виявили наступну середню точність для модифікацій дисперсійного оцінювання: DE-BM1 39,15%, DE-BM2 61,82% DE-BM3 57,14% (Рисунок 4.19).

Дослідження ефективності методу визначення важливості семантичних одиниць у цифрових текстах при пошуку ключових семантичних одиниць у текстах обсягом від 500 до 2000 слів виявили наступну середню точність для модифікацій дисперсійного оцінювання: DE-BM1 68,25%, DE-BM2 74,59% DE-BM3 73,92% (Рисунок 4.20).

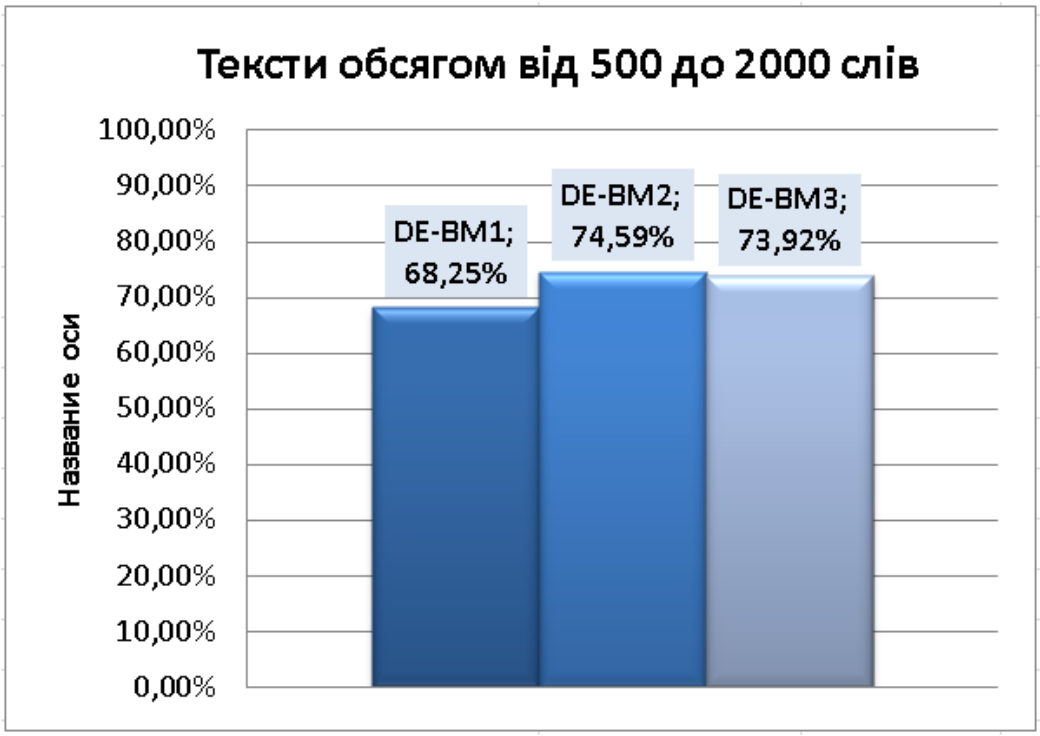


Рисунок 4.20 – Діаграма порівняння ефективності модифікацій методу визначення важливості семантичних одиниць у цифрових текстах при пошуку ключових семантичних одиниць у текстах обсягом від 500 до 2000 слів

Таким чином, дослідження ефективності методу визначення важливості семантичних одиниць у цифрових текстах виявило, що при пошуку ключових семантичних одиниць у текстах обсягом від 300 до 500 слів найвищу ефективність (61,82%) продемонстрував метод дисперсійного оцінювання

модифікації DE-BM2, проте метод дисперсійного оцінювання модифікації DE-BM3 теж виявив спів ставня результати (57,14%). Водночас класичний метод дисперсійного оцінювання модифікації DE-BM1 виявив значно гірші результати (39,15%). Це пояснюється великою кількістю семантичних одиниць у таких текстах, що мають низьку кількість появ, і відповідно низьку кількість відстаней між появами семантичних одиниць для дисперсійного обрахунку класичним методом.

Проте при пошуку ключових семантичних одиниць у текстах обсягом від 500 до 2000 слів всі модифікації дисперсійного оцінювання продемонстрували подібні результати (DE-BM1 68,25%, DE-BM2 74,59% і DE-BM3 73,92%). Це пояснюється тим, що потенційні ключові семантичні одиниці у таких текстах мають достатньо велику кількість появ, і відповідно велику кількість відстаней між появами семантичних одиниць для ефективного дисперсійного обрахунку навіть класичним методом.

Проведені дослідження дозволяють зробити висновок про ефективність використання розробленого методу визначення важливості семантичних одиниць у цифрових текстах для пошуку ключових семантичних одиниць модифікацією DE-BM3 при семантичному аналізі цифрових текстів, особливо – невеликих за обсягом.

Одержані результати корелюють з відомими науковими джерелами [30, 34, 35] та можуть бути практично використані при вирішенні прикладних завдань визначення важливості семантичних одиниць та пошуку ключових семантичних одиниць у цифрових текстах. Наприклад, при вирішенні задачі адаптивної пропозиції товарів у інтернет-магазині за семантичними ознаками [36].

Висновки до розділу 4

Для прикладного дослідження розроблених методу визначення важливості семантичних одиниць у цифрових текстах та інформаційної

технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах в розділі було створено інформаційну систему, що забезпечує відповідний функціонал.

Проведене в розділі дослідження функціональності інформаційної системи автоматизованого пошуку ключових семантичних одиниць виявило коректне виконання інформаційною системою функцій відповідно до інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, тому було зроблено висновок про її відповідність створеній інформаційної технології й придатність до дослідження її ефективності і відповідно ефективності методу визначення важливості семантичних одиниць у цифрових текстах.

Для дослідження ефективності методу визначення важливості семантичних одиниць у цифрових текстах було використано 81 текст обсягом від 300 до 500 слів і 35 текстів обсягом від 500 до 2000 слів з відкритих джерел із множинами ключових семантичних одиниць, визначеними їх авторами. Дослідження виявило, що при пошуку ключових семантичних одиниць у текстах обсягом від 300 до 500 слів найвищу ефективність (61,82%) продемонстрував метод дисперсійного оцінювання модифікації DE-BM2, проте метод дисперсійного оцінювання модифікації DE-BM3 теж виявив спів ставня результати (57,14%). Водночас класичний метод дисперсійного оцінювання модифікації DE-BM1 виявив значно гірші результати (39,15%). Це пояснюється великою кількістю семантичних одиниць у таких текстах, що мають низьку кількість появ, і відповідно низьку кількість відстаней між появами семантичних одиниць для дисперсійного обрахунку класичним методом.

При пошуку ключових семантичних одиниць у текстах обсягом від 500 до 2000 слів всі модифікації дисперсійного оцінювання продемонстрували подібні результати (DE-BM1 68,25%, DE-BM2 74,59% DE-BM3 73,92%). Це пояснюється тим, що потенційні ключові семантичні одиниці у таких текстах мають достатньо велику кількість появ, і відповідно велику кількість відстаней між

появами семантичних одиниць для ефективного дисперсійного обрахунку навіть класичним методом.

Проведені дослідження дозволяють зробити висновок про ефективність використання розробленого методу визначення важливості семантичних одиниць в цифрових текстах для пошуку ключових семантичних одиниць модифікацією DE-BM3 при семантичному аналізі цифрових текстів, особливо – невеликих за обсягом. Одержані результати можуть бути практично використані при вирішенні прикладних завдань визначення важливості семантичних одиниць та пошуку ключових семантичних одиниць у цифрових текстах. Наприклад, при вирішенні задачі адаптивної пропозиції товарів у інтернет-магазині за семантичними ознаками.

Загальні висновки

Кваліфікаційна робота магістра розв’язує науково-технічну задачу автоматизованого визначення важливості семантичних одиниць у цифрових текстах за допомогою методу дисперсійного оцінювання та його модифікацій. За результатом виконання роботи були поставлені та *вирішені наступні завдання*:

1. Проведено аналіз предметної області семантичного аналізу текстів, зокрема сучасних методів пошуку ключових семантичних одиниць у цифрових текстах.

2. Вдосконалено метод визначення важливості семантичних одиниць у цифрових текстах.

3. Розроблено інформаційну технологію автоматизованого пошуку семантичних одиниць у цифрових текстах.

4. Розроблено прикладну інформаційну систему для автоматизованого пошуку семантичних одиниць у цифрових текстах.

5. Проведено прикладне дослідження методу визначення важливості семантичних одиниць у цифрових текстах у складі інформаційної технології автоматизованого пошуку семантичних одиниць у цифрових текстах і виконано аналіз результатів використання відповідної інформаційної системи.

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

1. Вдосконалено метод визначення важливості семантичних одиниць у цифрових текстах на основі дисперсійного оцінювання, який відрізняється тим, що на відміну від існуючих дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а й з урахуванням початкових, кінцевих та похідних кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту.

2. Розроблено нову інформаційну технологію автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, що дозволяє з використанням створеного методу визначення важливості семантичних одиниць у цифрових текстах за вхідними даними у вигляді вхідного цифрового тексту як відповідної впорядкованої множини символів та параметрами налаштувань одержувати вихідні дані у вигляді трьох множин ключових семантичних одиниць вхідного цифрового тексту за оцінками модифікацій дисперсійного оцінювання, які дозволяють виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а й із урахуванням початкових, кінцевих і кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту, а також сформованої зведеної таблиці оцінок семантичної важливості ключових семантичних одиниць за цими оцінками модифікацій дисперсійного оцінювання.

3. Розроблено нову інформаційну систему для автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, що дозволяє за створеною інформаційною технологією в результаті обробки вхідного цифрового тексту у вигляді відповідної впорядкованої множини символів виконувати автоматизоване визначення множин ключових семантичних одиниць за оцінками модифікацій дисперсійного оцінювання DE-BM1, DE-BM2 і DE-BM3, також формувати відповідну зведену таблицю оцінок семантичної важливості ключових семантичних одиниць вхідного цифрового тексту.

Для прикладного дослідження розроблених методу визначення важливості семантичних одиниць у цифрових текстах та інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах в розділі було створено інформаційну систему, що забезпечує відповідний функціонал. Інформаційна система дозволяє в результаті обробки вхідного цифрового тексту у вигляді відповідної впорядкованої множини символів виконувати автоматизоване визначення множин ключових семантичних одиниць за оцінками модифікацій дисперсійного оцінювання DE-

BM1, DE-BM2 і DE-BM3, також формувати відповідну зведену таблицю оцінок семантичної важливості ключових семантичних одиниць вхідного цифрового тексту. Інформаційна система автоматизованого пошуку семантичних одиниць у цифрових текстах не потребує використання бази даних й складається із чотирьох модулів: модуля попередньої обробки тексту та формування текстового вектору, модуля визначення відстаней між появами семантичних одиниць, модуля дисперсійного оцінювання важливості семантичних одиниць і модуля формування зведеної таблиці ключових семантичних одиниць.

Дослідження ефективності методу визначення важливості семантичних одиниць у цифрових текстах виявило, що при пошуку ключових семантичних одиниць у текстах обсягом від 300 до 500 слів найвищу ефективність (61,82%) продемонстрував метод дисперсійного оцінювання модифікації DE-BM2, проте метод дисперсійного оцінювання модифікації DE-BM3 теж виявив спів ставня результати (57,14%). Водночас класичний метод дисперсійного оцінювання модифікації DE-BM1 виявив значно гірші результати (39,15%). Це пояснюється великою кількістю семантичних одиниць у таких текстах, що мають низьку кількість появ, і відповідно низьку кількість відстаней між появами семантичних одиниць для дисперсійного обрахунку класичним методом.

При пошуку ж ключових семантичних одиниць у текстах обсягом від 500 до 2000 слів всі модифікації дисперсійного оцінювання продемонстрували подібні результати (DE-BM1 68,25%, DE-BM2 74,59% DE-BM3 73,92%). Це пояснюється тим, що потенційні ключові семантичні одиниці у таких текстах мають достатньо велику кількість появ, і відповідно велику кількість відстаней між появами семантичних одиниць для ефективного дисперсійного обрахунку навіть класичним методом.

Проведені дослідження дозволяють зробити висновок про ефективність використання розробленого методу визначення важливості семантичних одиниць в цифрових текстах для пошуку ключових семантичних одиниць модифікацією DE-BM3 при семантичному аналізі цифрових текстів, особливо – невеликих за обсягом. Одержані результати можуть бути практично використані

при вирішенні прикладних завдань визначення важливості семантичних одиниць та пошуку ключових семантичних одиниць у цифрових текстах, наприклад, при вирішенні задачі адаптивної пропозиції товарів у інтернет-магазині за семантичними ознаками.

Основні наукові і практичні результати кваліфікаційної роботи магістра доповідались у доповіді за темою «Інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів» на XIII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп’ютерних наук АПКН-2021» (15-16 жовтня 2021 року); за темою роботи автором виконано наукову публікацію [36].

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. Науковий журнал «Проблеми програмування». Київ, 2018. №2-3. С.245-254.

2. Україна 2030е – країна з розвинутою цифровою економікою. URL: https://strategy.uifuture.org/kraina-z-rozvinutoyu-cifrovoyu-ekonomikoyu.html

3. Цифрові платформи: підходи до класифікації та визначення ролі в економічному розвитку. URL: http://bses.in.ua/journals/2018/35_2_2018/7.pdf

4. Wikipedia. Семантика. URL: https://uk.wikipedia.org/wiki/Семантика

5. Семантичний аналіз. URL: http://ermak.cs.nstu.ru/trans/Trans411.htm

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

7. Інтелектуальний аналіз тексту. URL: https://uk.wikipedia.org/wiki/Інтелектуальний_аналіз_тексту

8. Information Extraction. URL: https://nanonets.com/blog/information-extraction/

9. Витягування інформації. URL: https://uk.wikipedia.org/wiki/Витягування_інформації

10. Wikipedia. Словосполучення. URL: https://uk.wikipedia.org/wiki/Словосполучення

11. Синтаксис: структура, семантика, функція. URL: http://oldconf.neasmo.org.ua/node/2691

12. Термін. URL: http://sum.in.ua/s/termin

13. Марія Комова. Семантичне творення термінів на нзначення документів. URL: http://ena.lp.edu.ua:8080/bitstream/ntb/54990/2/2005n538_Komova_M-Semantychne_tvorennia_terminiv_78-83.pdf

14. Wikipedia. Ключове слово. URL: https://uk.wikipedia.org/wiki/Ключове_слово

15. Advego. URL: https://advego.com/text/seo/

16. Serpstat. URL: https://serpstat.com/ru/keyword-research/

17. Wikipedia. WordStat. URL: https://en.wikipedia.org/wiki/WordStat

18. Зеленський А.А. Актуальність дослідження програм семантичного аналізу текстів та огляд методів їх реалізації. URL: http://ds.knu.edu.ua/jspui/bitstream/123456789/2576/1/Актуальність_дослідження_програм_семантичного_аналізу_текстів_та_огляд_методів_їх_реалізації.pdf

19. Зембицька М. В. , Горошко А. В. Сучасний стан лінгво-статистичних методів семантичного аналізу текстів. URL: http://elar.khnu.km.ua/jspui/bitstream/123456789/7138/1/SE-2019-97-98.pdf

20. Кондаков О.В., Мазурець О.В., Скрипник Т.К. Математичні моделі для визначення семантичних термінів у контенті навчальних матеріалів. URL: http://elar.khnu.km.ua/jspui/bitstream/123456789/6936/1/093.pdf

21. TF-IDF. URL: https://uk.wikipedia.org/wiki/TF-IDF

22. Що таке .NET і чим займаються .NET-розробники? URL: https://training.epam.ua/#!/News/301?lang=ua

23. What is .NET?. URL: https://dotnet.microsoft.com/learn/dotnet/what-is-dotnet

24. What is Java? Definition, Meaning & Features of Java Platforms. URL: https://www.guru99.com/java-platform.html#2

25. What is PHP? Write your first PHP Program URL: https://www.guru99.com/what-is-php-first-php-program.html

26. What is PHP URL: https://www.phptutorial.net/php-tutorial/what-is-php/

27. Обзор і установка Visual Studio 2019 Community на Windows 10 URL: https://info-comp.ru/programmirovanie/739-install-visual-studio-2019-community.html

28. Вступ в C# URL: https://programm.top/uk/c-sharp/tutorial/introduction/

29. Windows Forms overview URL: https://docs.microsoft.com/ru-ru/dotnet/desktop/winforms/windows-forms-overview?view=netframeworkdesktop-4.8

30. Chen J. Smart Data Integration by Goal Driven Ontology Learning / J. Chen, D. Dosyn, V. Lytvyn, A. Sachenko // Advances in Big Data. – 2016. – Т. 529. – С. 283-292.

31. Wikipedia. Текст. URL: https://uk.wikipedia.org/wiki/Текст

32. Інтелект. Вікіпедія [Електронний ресурс] – Режим доступу: https://uk.wikipedia.org/wiki/Інтелект

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

34. 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. – С. 691-702.

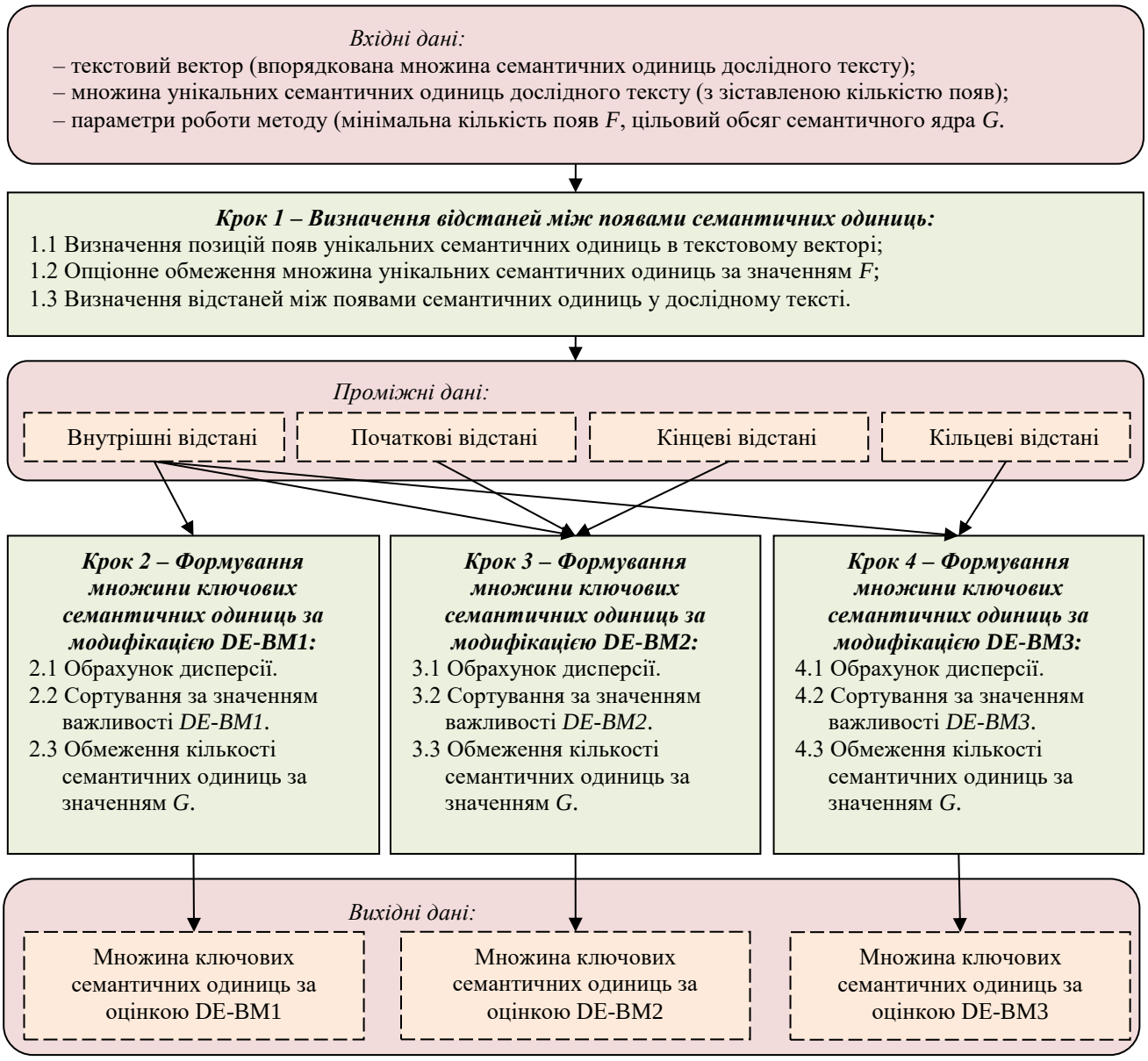
35. Ландэ Д. В. Компактифицированный горизонтальный граф видимости для сети слов / Д. В. Ландэ, А. А. Снарский // Труды Международной научной конференции «Интеллектуальный анализ информации ИАИ-2013. Знания и рассуждения» – КПИ, Киев: 2013. – С.158-164.

36. 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

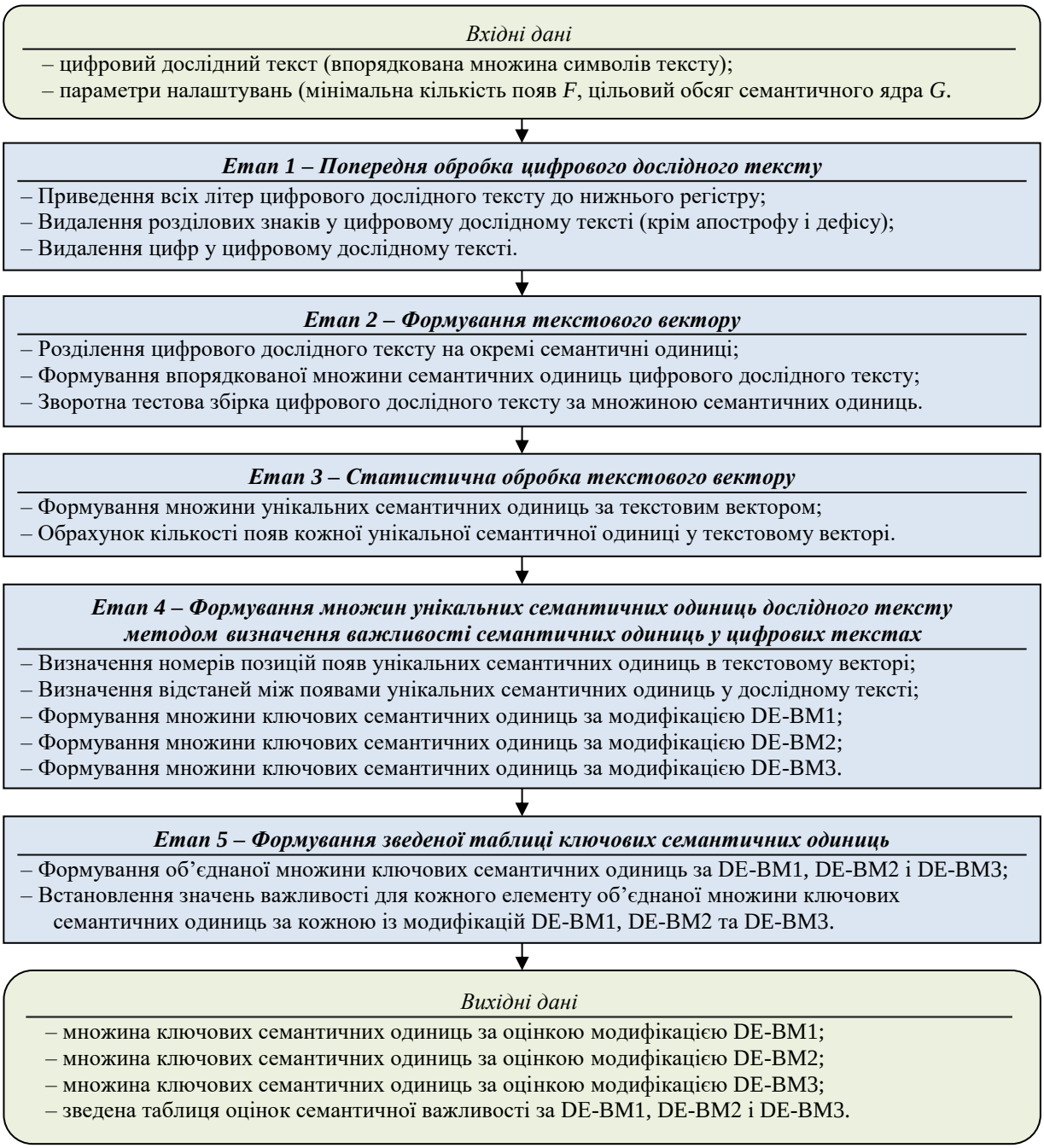
Додаток А

Схема методу визначення важливості семантичних одиниць у цифрових текстах



Додаток Б

Схема інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах



Додаток В

Структура інформаційної системи автоматизованого пошуку семантичних одиниць у цифрових текстах

Додаток Г

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., Затуцька 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.

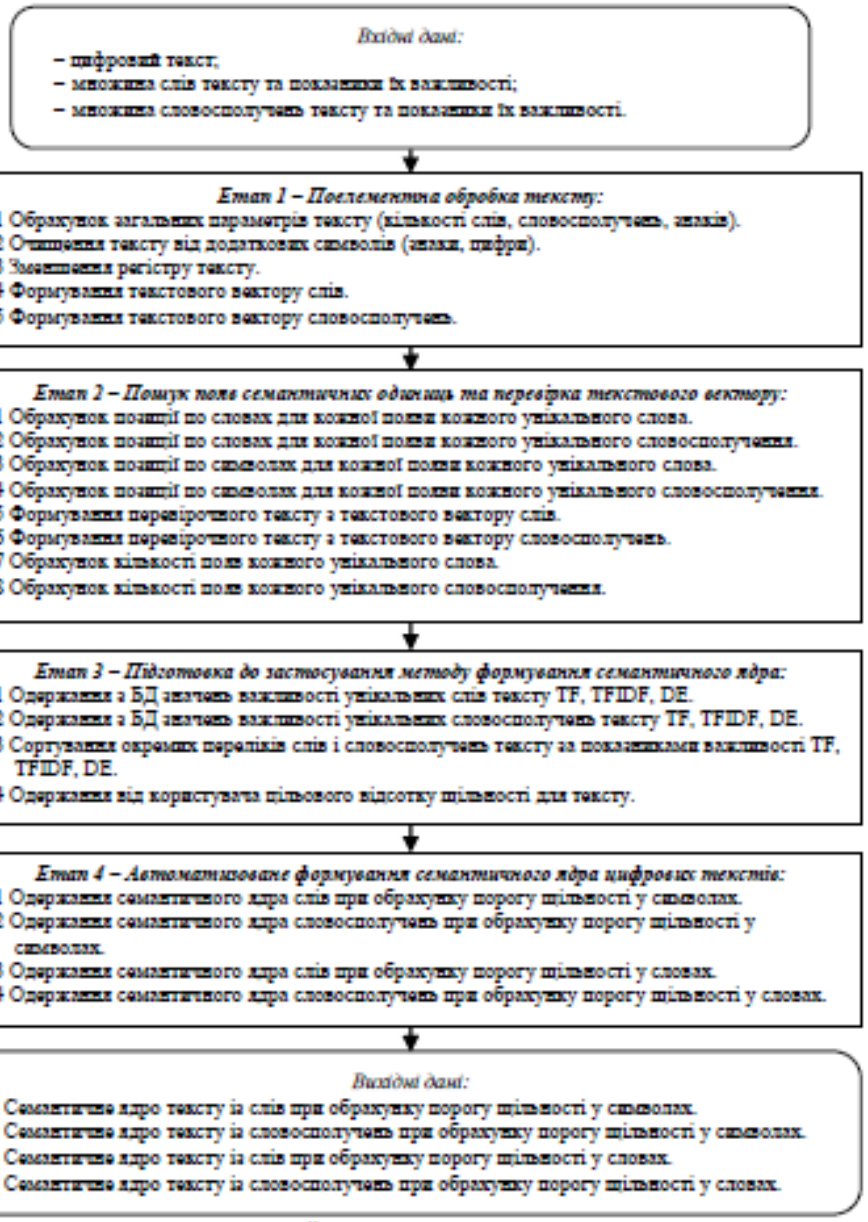


Рисунок 1 – Схема інформаційної технології формування множини ключових семантичних одиниць

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.

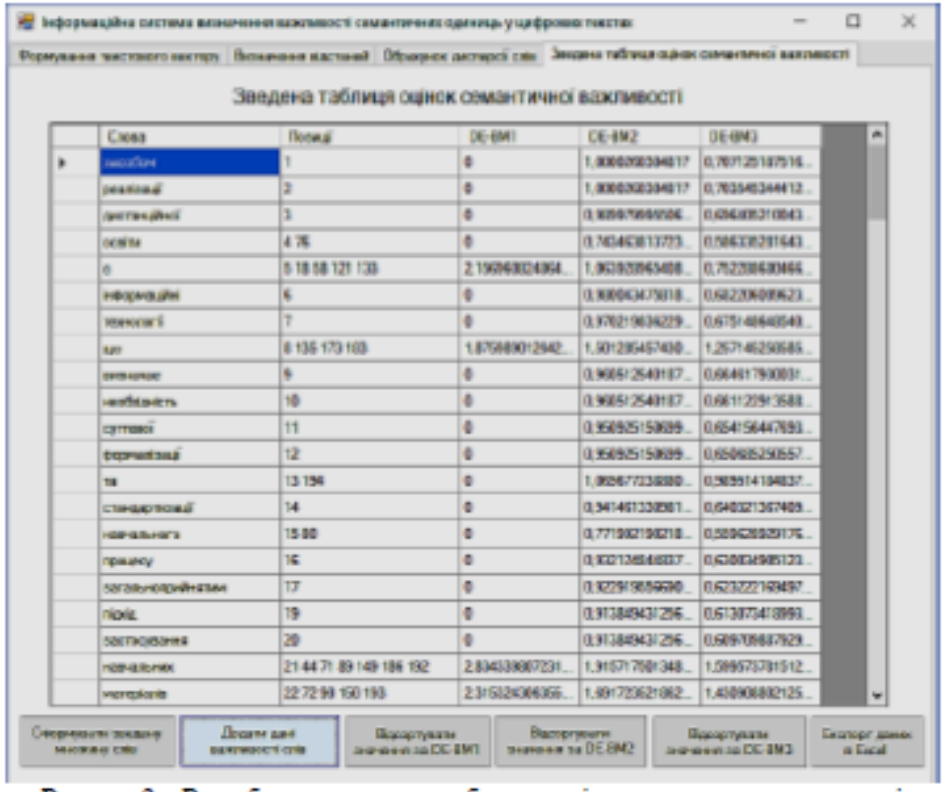


Рисунок 2 – Розроблене програмне забезпеченні для визначення важливості семантичних одиниць у цифрових текстах

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.

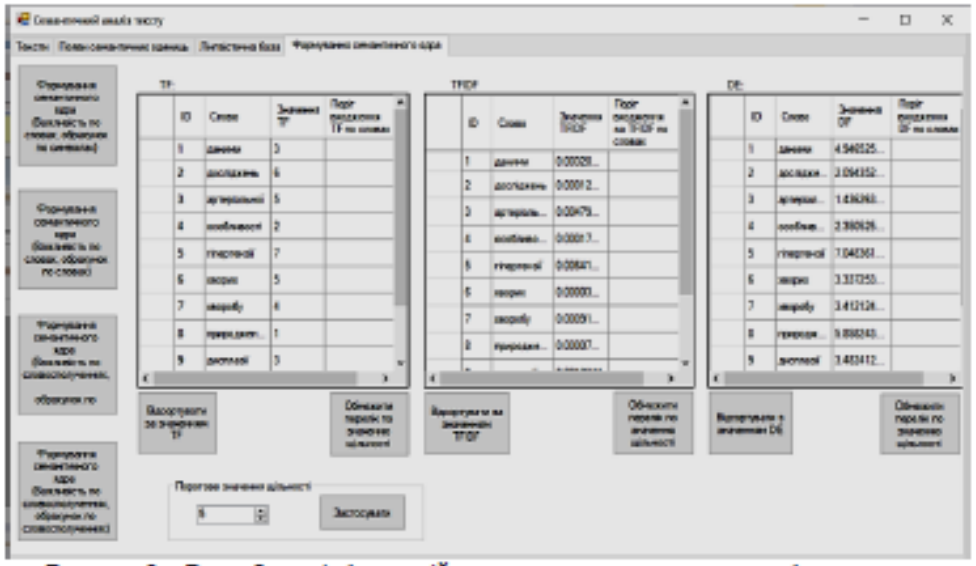


Рисунок 3 – Розроблена інформаційна система автоматизованого формування семантичного ядра цифрових текстів

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).

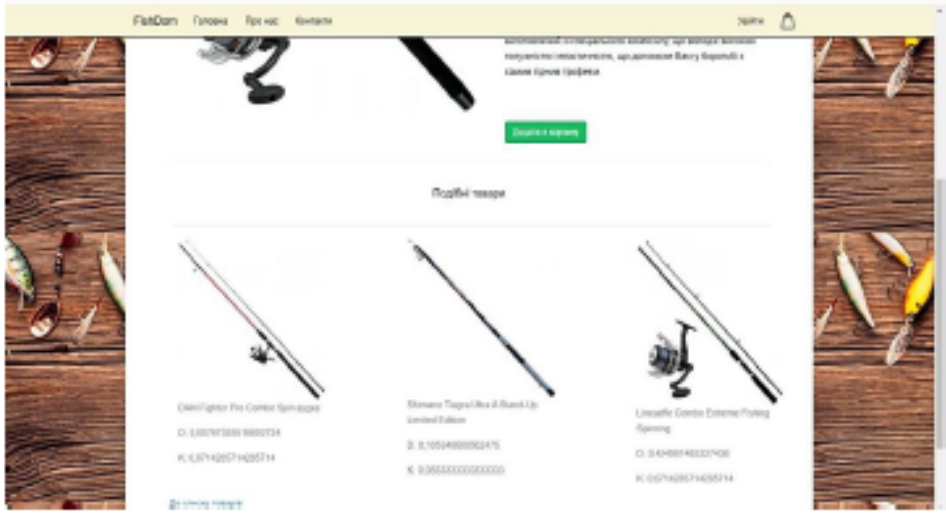


Рисунок 4 – Приклад практичного використання створеної інформаційної технології для адаптивної пропозиції товарів у інтернет-магазині за семантичними ознаками

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 Д

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

МЕТОД ВИЗНАЧЕННЯ ВАЖЛИВОСТІ СЕМАНТИЧНИХ ОДИНИЦЬ У ЦИФРОВИХ ТЕКСТАХ

Виконав:
студент 2 курсу, група КНм-20-1
Войчишин Олександр Олександрович

Керівник:
к.т.н., доцент кафедри ІПЗ
Форкун Юрій Вікторович

Мета роботи

Мета кваліфікаційної роботи магістра – створення методу визначення важливості семантичних одиниць у цифрових текстах, який дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок важливості семантичних одиниць тексту на основі дисперсійного оцінювання з урахуванням як внутрішніх відстаней між появами унікальних семантичних одиниць, так і початкових, кінцевих та кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту.

Для досягнення поставленої мети створення методу визначення важливості семантичних одиниць у цифрових текстах потрібно розв'язати наступні *задачі дослідження*:

1. Провести аналіз предметної області семантичного аналізу текстів, зокрема сучасних методів пошуку ключових семантичних одиниць у цифрових текстах.
2. Вдосконалити метод визначення важливості семантичних одиниць у цифрових текстах.
3. Розробити інформаційну технологію автоматизованого пошуку семантичних одиниць у цифрових текстах.
4. Розробити прикладну інформаційну систему для автоматизованого пошуку семантичних одиниць у цифрових текстах.
5. Провести прикладне дослідження методу визначення важливості семантичних одиниць у цифрових текстах у складі інформаційної технології автоматизованого пошуку семантичних одиниць у цифрових текстах.

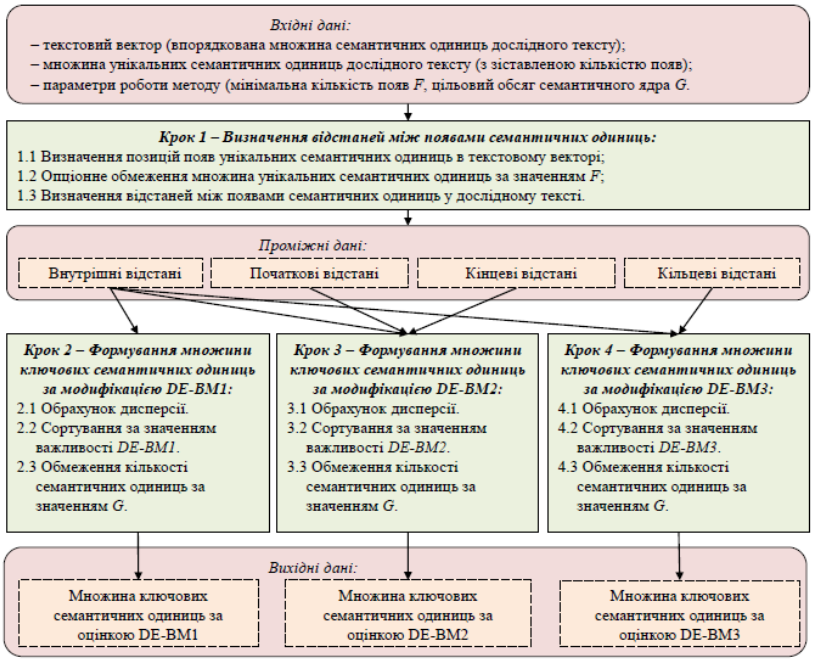


Схема методу визначення важливості семантичних одиниць у цифрових текстах

Модифікації дисперсійного методу визначення важливості семантичних одиниць

Обрахунок відстаней між словами у тексті слугує підготовчим етапом до дисперсійного оцінювання слів, за якого визначаються для кожного слова з кількістю появ і тексті більше одного всі відстані між сусідніми появами слів.

У залежності від того, яким чином та у якій кількості визначаються відстані для кожного унікального слова тексту, розрізняють різні модифікації вихідного методу DE-BM пошуку ключових слів за дисперсійним оцінюванням DE-BM1, а саме: DE-BM2 та DE-BM3.

- Метод пошуку ключових слів **DE-BM1** для *n* появ слова враховує *n*-1 відстані. При цьому за відстань береться різниця між меншим порядковим номером наступного слова і більшим порядковим номером попереднього слова.
- Метод пошуку ключових слів **DE-BM2** для *n* появ слова враховує *n*+1 відстань. За відстань береться різниця між меншим порядковим номером наступного слова і більшим порядковим номером попереднього слова. Також додатково враховуються відстані: від початку тексту до першої появи слова в тексті, від останньої появи слова у тексті до кінця тексту.
- Метод пошуку ключових слів **DE-BM3** для *n* появ слова враховує *n* відстаней. За відстань береться різниця між меншим порядковим номером наступного слова і більшим порядковим номером попереднього слова. Також додатково враховується відстань, рівна сумі різниць між початком тексту до першої появи слова й між останньою появою слова до кінця тексту.

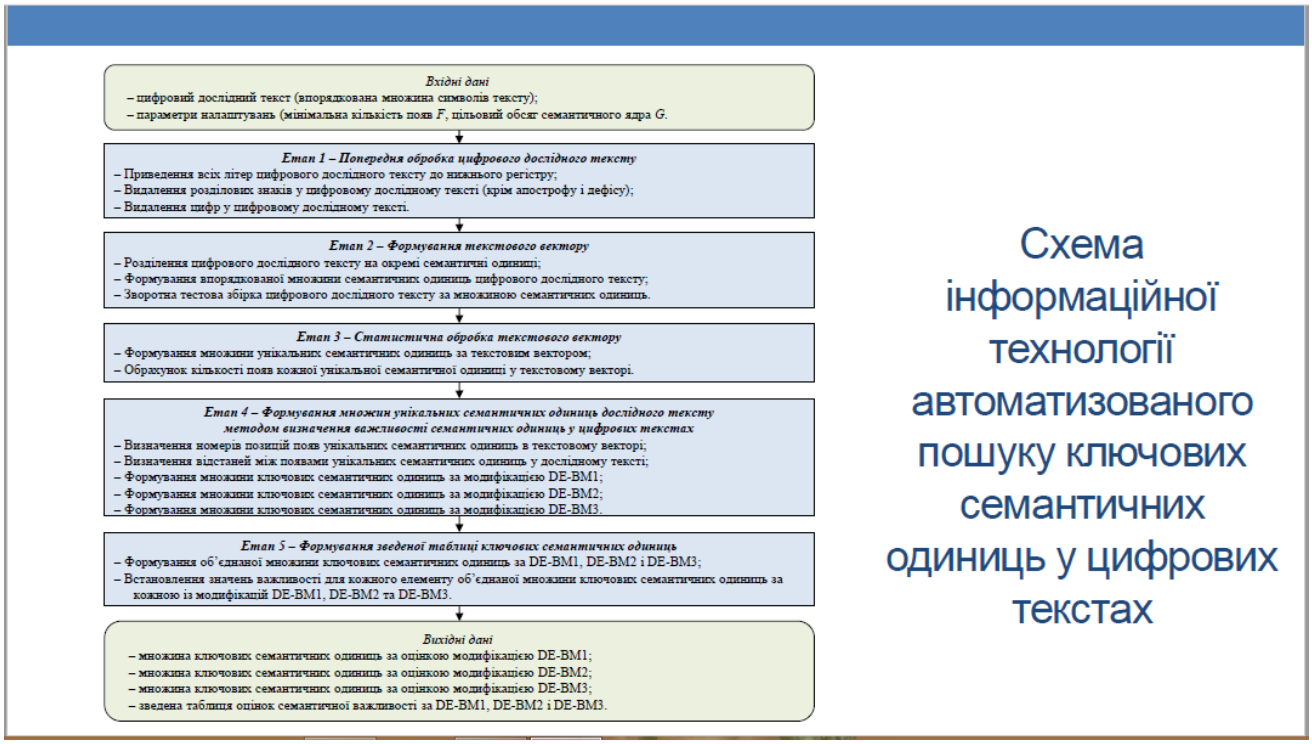
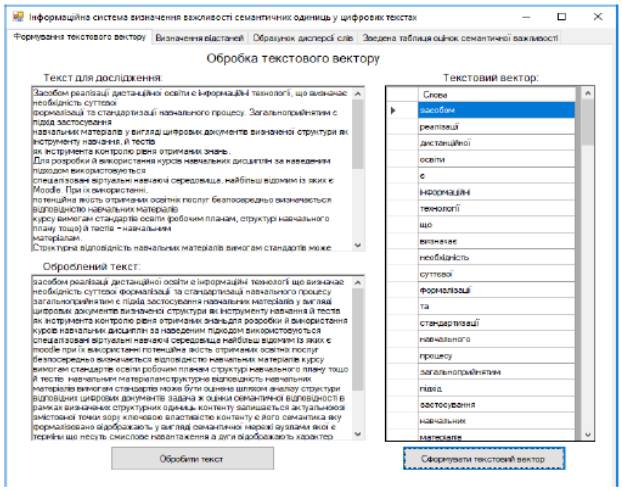
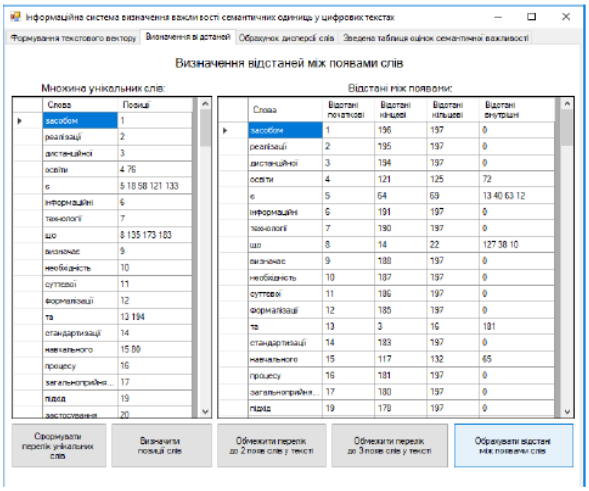


Схема інформаційної технології автоматизованого пошуку ключових семантичних одиниць у цифрових текстах

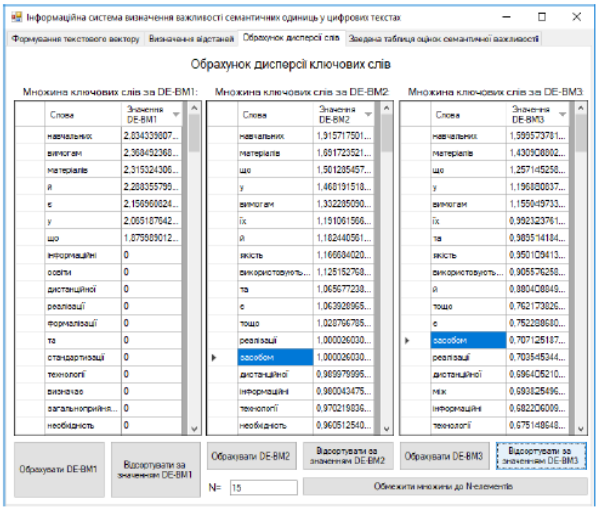
Структура інформаційної системи автоматизованого пошуку семантичних одиниць у цифрових текстах

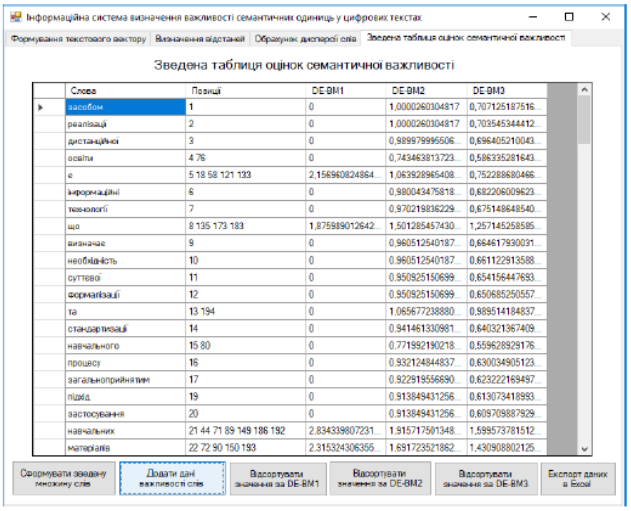
Інформаційна система автоматизованого пошуку семантичних одиниць у цифрових текстах





Інформаційна система автоматизованого пошуку семантичних одиниць у цифрових текстах





Дослідження ефективності методу визначення важливості семантичних одиниць

Інтелект властивий людям, а також спостерігається у тварин.

Людина застосовує інтелект для обробки наявної інформації, наприклад, з метою побудови або вдосконалення розуміння, позиції, стратегії, методу, правила, комбінації, відношення, пояснення, рішення, плану чи цілі. Інтелект пов'язаний з іншими внутрішніми властивостями людини, такими як сприйняття, пам'ять, мова, уява, самосвідомість, самоконтроль, характер, володіння тілом, творчість, інтуїція і власне формується завдяки функціонуванню означених параметрів особистості. Інтелект найчастіше спрямовується на вирішення питань облаштування побуту і відпочинку, професійну діяльність, міжособистісні стосунки та самовдосконалення.

В повсякденному житті в сучасній розвинутій людині інтелект також проявляє себе у вигляді внутрішніх почуттів і образів мислення, таких як відчуття реальності, часу, простору, себе, ритму, гумору, відповідальності, ситуації, прекрасного, захищеності, небезпеки, такту, комфорту, міри, справедливості, довіри, свободи, поваги, власної гідності та інших, і у вигляді аналітичного, образного, практичного, абстрактного, тактичного або стратегічного образу мислення.

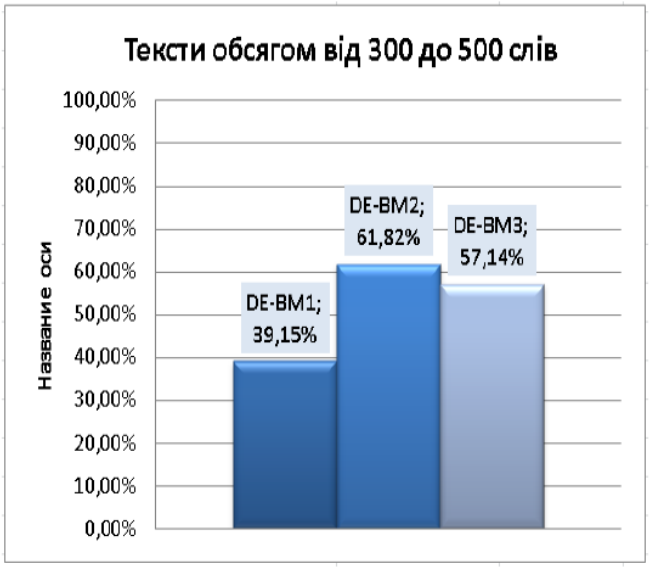
Наприклад, в тексті, що складається з 131 слова, й в якому слово «інтелект» зустрічається на позиціях 1, 11, 34, 60 і 83, для обрахунку дисперсійної оцінки ключових семантичних одиниць можна використати наступні відстані:

- відстані, рівні 10, 23, 26, та 23 за класичним підходом до обрахунку;
- додаткова відстань від початку тексту до першої появи слова у тексті 1;
- додаткова відстань від останньої появи слова у тексті до кінця тексту 48;
- додаткова відстань, рівна сумі різниць між початком тексту до першої появи слова та між останньою появою слова до кінця тексту 49.

Відповідно, для різновидів дисперсійного оцінювання семантичних одиниць буде використано наступні відстані:

- для вихідного методу пошуку ключових семантичних одиниць за дисперсійним оцінюванням DE-BM1: 10, 23, 26, 23;
- для методу пошуку ключових семантичних одиниць DE-BM2: 1, 10, 23, 26, 23, 48;
- для методу пошуку ключових семантичних одиниць DE-BM3: 10, 23, 26, 23, 49.

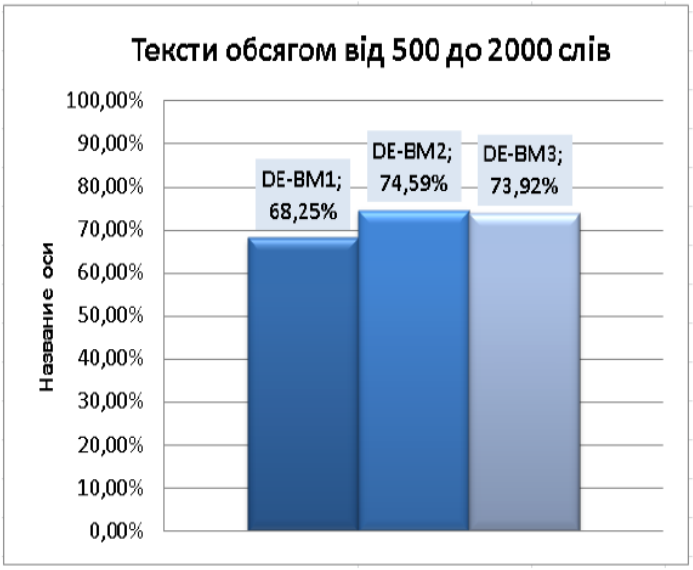
Дослідження ефективності методу визначення важливості семантичних одиниць



Дослідження ефективності методу визначення важливості семантичних одиниць у цифрових текстах виявило, що при пошуку ключових семантичних одиниць у текстах обсягом від 300 до 500 слів найвищу ефективність (61,82%) продемонстрував метод дисперсійного оцінювання модифікації DE-BM2, проте метод дисперсійного оцінювання модифікації DE-BM3 теж виявив спів ставня результати (57,14%). Водночас класичний метод дисперсійного оцінювання модифікації DE-BM1 виявив значно гірші результати (39,15%).

Це пояснюється великою кількістю семантичних одиниць у таких текстах, що мають низьку кількість появ, і відповідно низьку кількість відстаней між появами семантичних одиниць для дисперсійного обрахунку класичним методом.

Дослідження ефективності методу визначення важливості семантичних одиниць



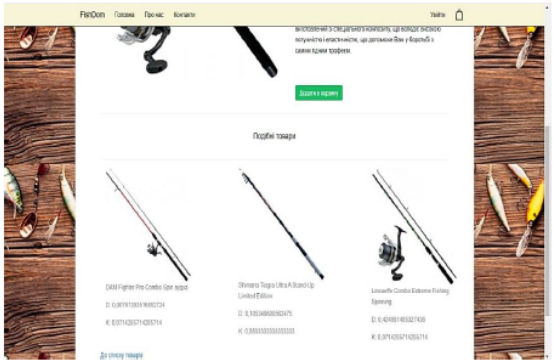
При пошуку ключових семантичних одиниць у текстах обсягом від 500 до 2000 слів всі модифікації дисперсійного оцінювання продемонстрували подібні результати (DE-BM1 68,25%, DE-BM2 74,59% і DE-BM3 73,92%).

Це пояснюється тим, що потенційні ключові семантичні одиниці у таких текстах мають достатньо велику кількість поява, і відповідно велику кількість відстаней між появами семантичних одиниць для ефективного дисперсійного обрахунку навіть класичним методом.

Практичне значення одержаних результатів

Проведені дослідження дозволяють зробити висновок про ефективність використання розробленого методу визначення важливості семантичних одиниць у цифрових текстах для пошуку ключових семантичних одиниць модифікацією DE-BM3 при семантичному аналізі цифрових текстів, особливо – невеликих за обсягом.

Одержані результати можуть бути практично використані при вирішенні прикладних завдань визначення важливості семантичних одиниць та пошуку ключових семантичних одиниць у цифрових текстах. Наприклад, було використано розроблений метод при вирішенні задачі адаптивної пропозиції товарів у інтернет-магазині за семантичними ознаками.



Положення новизни та інновації

- **Вдосконалено метод визначення важливості семантичних одиниць у цифрових текстах** на основі дисперсійного оцінювання, який відрізняється тим, що на відміну від існуючих дозволяє за впорядкованою множиною семантичних одиниць дослідного тексту виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а й з урахуванням початкових, кінцевих та похідних кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту.
- **Розроблено нову інформаційну технологію автоматизованого пошуку ключових семантичних одиниць у цифрових текстах**, що дозволяє з використанням створеного методу визначення важливості семантичних одиниць у цифрових текстах за вхідними даними у вигляді вхідного цифрового тексту як відповідної впорядкованої множини символів та параметрами налаштувань одержувати вихідні дані у вигляді трьох множин ключових семантичних одиниць вхідного цифрового тексту за оцінками модифікацій дисперсійного оцінювання, які дозволяють виконувати обрахунок семантичної важливості не тільки з урахуванням внутрішніх відстаней між появами унікальних семантичних одиниць, а й із урахуванням початкових, кінцевих і кільцевих відстаней між появами унікальних семантичних одиниць цифрового тексту, а також сформованої зведеної таблиці оцінок семантичної важливості ключових семантичних одиниць за цими оцінками модифікацій дисперсійного оцінювання.
- **Розроблено нову інформаційну систему для автоматизованого пошуку ключових семантичних одиниць у цифрових текстах**, що дозволяє за створеною інформаційною технологією в результаті обробки вхідного цифрового тексту у вигляді відповідної впорядкованої множини символів виконувати автоматизоване визначення множин ключових семантичних одиниць за оцінками модифікацій дисперсійного оцінювання DE-BM1, DE-BM2 і DE-BM3, також формувати відповідну зведену таблицю оцінок семантичної важливості ключових семантичних одиниць вхідного цифрового тексту.

Загальні висновки

Кваліфікаційна робота магістра розв'язує науково-технічну задачу автоматизованого визначення важливості семантичних одиниць у цифрових текстах за допомогою методу дисперсійного оцінювання та його модифікацій.

За результатом виконання роботи були поставлені та *вирішені наступні завдання*:

1. Проведено аналіз предметної області семантичного аналізу текстів, зокрема сучасних методів пошуку ключових семантичних одиниць у цифрових текстах.
2. Вдосконалено метод визначення важливості семантичних одиниць у цифрових текстах.
3. Розроблено інформаційну технологію автоматизованого пошуку семантичних одиниць у цифрових текстах.
4. Розроблено прикладну інформаційну систему для автоматизованого пошуку семантичних одиниць у цифрових текстах.
5. Проведено прикладне дослідження методу визначення важливості семантичних одиниць у цифрових текстах у складі інформаційної технології автоматизованого пошуку семантичних одиниць у цифрових текстах і виконано аналіз результатів використання відповідної інформаційної системи.

UNICHECK

Ім'я користувача:
Кафедра КН

ID перевірки:
1009466073

Дата перевірки:
02.12.2021 09:43:09 EET

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
02.12.2021 09:49:53 EET

ID користувача:
100005671

Назва документа: **2021_КРМ_Войчишин 20211201 late**

Кількість сторінок: **91** Кількість слів: **15419** Кількість символів: **124274** Розмір файлу: **5.77 MB** ID файлу: **1009480547**

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

6.77%
Схожість

Найбільша схожість: **4.05%** з джерелом з Бібліотеки (ID файлу: **1009433085**)

2.47% Джерела з Інтернету	36	Сторінка 93
4.66% Джерела з Бібліотеки	91	Сторінка 93

0.05% Цитат

Цитати	1	Сторінка 94
Посилання	1	Сторінка 94

0%
Вилучень

Немає вилучених джерел

Модифікації

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

Замінені символи	21
Підозріле форматування	28 сторінок

Thu Dec 02 08:47:54 EET 2021, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальное совпадение с одним документом 20.0%

Словари проверки: en_US, ru_RU, ua_UA. **Ошибок в документах: 5%**

	Документ		Суммарное совпадение по Базе Данных	
ID: 97774 Название: Метод визначення важливості семантичних одиниць у цифрових текстах Добавлено в БД: 2021-12-02 Авторы: О.О. Войчишин Руководители: О.В. Мазурець Консультанты: Опоненты:	Символы	Лексемы	Символы	Лексемы
	102967	501	24106 (23%)	198 (40%)

Источник плагиата

ID	Описание	Наличие плагиата в документе	
		Символы	Лексемы
95893	Название: ЗВІТ з науково-дослідної практики Добавлено в БД: 2021-09-29 Авторы: Войчишин О.О. Руководители: Скрипник Т.К. Консультанты: Опоненты:	20431 (20.0%)	176 (35.0%)

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, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: *Метод визначення важливості семантичних одиниць у цифрових текстах*
Автор: *Войчишин Олександр Олександрович*
Спеціальність: *122 – Компютерні науки*
Освітня програма: *освітньо-професійна*
Науковий керівник: *к.т.н., доц.каф.ІПЗ Форкун Юрій Вікторович*

Після аналізу звіту подібності зроблено такий висновок:

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

Підтвердження:

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

1) За програмою Anti-Plagiarism виявлені 20% запозичень вказують на документ автора роботи Войчишина О.О. та містять його Звіт з науково-дослідної практики.
2) За програмою UNICHECK виявлені 6,77%, які є фрагментарними, не більше 4,05% на джерело – містять поширені конструкції, загальновідомі терміни та визначення.

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ ідентичності/схожості, складає 23% і 6,77% відповідно, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи _____ Юрій Форкун

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 Войчишина Олександра Олександровича за темою: *Метод визначення важливості семантичних одиниць у цифрових текстах*

1. Актуальність обраної теми

Тема кваліфікаційної робота магістра є актуальною та належним чином обґрунтована. Стосується питання автоматизованого визначення важливості семантичних одиниць у цифрових текстах за допомогою методу дисперсійного оцінювання та його модифікацій. Семантичний аналіз широко застосовується при рішенні задач інформаційного пошуку, автоматичного перекладу, аналізу змісту, пошуку протиріч, реферування, аналізу інтересів користувачів інформаційної системи, авторства текстів тощо. Тому розробка і вдосконалення методів семантичного аналізу цифрових текстів є актуальним і перспективним напрямком прикладного застосування інформаційних технологій.

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

Обрана тема, 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 поставлені в межах теми завдання дослідження.

4. Наявність наукової новизни

В кваліфікаційній роботі представлена наукова новизна та інновації, відповідна спеціальності 122 «Комп'ютерні науки» в межах обраної області дослідження. Продемонстровано та 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 досліджено предметну область семантичного аналізу текстів, обґрунтовано актуальність та поставлені задачі дослідження. Другий розділ присвячено розробці методу і засобів автоматизації визначення важливості семантичних одиниць у цифрових текстах. У третьому розділі представлена розробка

інформаційної системи автоматизованого пошуку ключових семантичних одиниць у цифрових текстах. Четвертий розділ присвячено аналізу результатів прикладного дослідження функціональності інформаційної системи й відповідно досліджена можливість прикладного застосування методу визначення важливості семантичних одиниць у цифрових текстах, досліджено його ефективність.

6. Ступінь розкриття теми роботи

Тема роботи в повній мірі обґрунтована, проведено аналіз актуальності та відомих досліджень в межах обраної теми, поставлені завдання, які реалізовані та проведено аналіз результатів прикладного 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 за дослідження ефективності виконано порівняння ефективності модифікацій методу визначення важливості семантичних одиниць при пошуку ключових семантичних одиниць у текстах для випадків текстів обсягом від 300 до 500 слів і від 500 до 2000 слів, проте не роз'яснено причину вибору таких діапазонів. Зазначені недоліки не вплинули на загальну якість роботи та одержаний результат.

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 *добре*.

Опонент _____ к.т.н., доц. Корецька Л.О.

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

МОН УКРАЇНИ

кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу магістра

гр. КНм-20-1 Войчишина Олександра Олександровича за темою: *Метод визначення важливості семантичних одиниць у цифрових текстах*

1. Актуальність теми

Робота з надзвичайно великою кількістю текстової інформації завжди є дуже затратною в часі. Багато компаній і організацій покладаються на методи вилучення інформації для автоматизації ручної роботи за допомогою інтелектуальних алгоритмів, які можуть зменшити витрати та людські зусилля і зробити різні процеси із меншою кількістю помилок. Інтелектуальний аналіз текстів займається категоризацією текстів, змінами в колекціях текстів та їх обробкою, пошуком інформації та розробкою засобів представлення інформації для користувача. Тому розробка і удосконалення методів семантичного аналізу цифрових текстів є актуальним і перспективним напрямком прикладного застосування інформаційних технологій.

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

Поставлена у кваліфікаційній роботи магістра мета, пов'язана з розробкою методу визначення важливості семантичних одиниць у цифрових текстах, 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 Войчишин Олександр Олександрович проявив себе кваліфікованим фахівцем та дисциплінованим студентом, вчасно виконуючи поставлені етапи дослідження. В процесі наукових досліджень і при розробці програмного забезпечення проявив достатні для одержання успішного результату компетентності.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Магістрант самостійно виконував всі поставлені задачі. Одержані положення наукової новизни та інновації, викладені в роботі, є результатом особистої діяльності магістранта. Це дозволило провести створення нових та удосконалення вже існуючих теоретичних і прикладних засобів, створених та використаних у роботі.

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 визначення важливості семантичних одиниць у цифрових текстах на основі дисперсійного оцінювання, розроблено нову інформаційну технологію автоматизованого пошуку ключових семантичних одиниць у цифрових текстах, розроблено нову інформаційну систему для автоматизованого пошуку ключових семантичних одиниць у цифрових текстах. Результати роботи оприлюдненні на науково-практичній конференції.

6. Ступінь оволодіння методами дослідження

В роботі виявлено достатній ступінь оволодіння магістрантом необхідними методами дослідження.

7. Повнота та якість розкриття теми роботи

Тема роботи в повній мірі обґрунтована й розкрита, проведено аналіз актуальності та відомих досліджень в межах обраної теми, поставлені завдання, які у роботі виконані, та проведено аналіз результатів прикладного застосування запропонованих засобів визначення важливості семантичних одиниць у цифрових текстах на основі дисперсійного оцінювання.

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

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

9. Можливість практичного застосування кваліфікаційної роботи, окремих її частин

Проведені дослідження дозволяють зробити висновок про ефективність використання розробленого. Одержані результати можуть бути практично використані при вирішенні прикладних завдань визначення важливості семантичних одиниць та пошуку ключових семантичних одиниць у цифрових текстах, наприклад, при вирішенні задачі адаптивної пропозиції товарів у інтернет-магазині за семантичними ознаками.

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

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

Науковий керівник _____ к.т.н., доц.каф.ІПЗ Форкун Юрій Вікторович