Магістерська кваліфікаційна робота

на тему:

«Автоматизована система контролю контенту студентських робіт»

Виконав:
ст. гр. 1КІ -19м
Чорний Д. С.
Науковий керівник:
к.т.н., доц. Захарченко С. М.

Вінниця 2020

Актуальність дослідження

- Перевірка контенту текстових робіт є досить розповсюдженою задачею, вона може застосовуватися у різних сферах діяльності людини.
- Перевірка контенту робіт на **плагіат** та **відповідність завданню** є актуальною задачею в навчальних закладах.
- Викладачі стикаються з труднощами при перевірці **великої кількості** студентських робіт, що містять описову (текстову) частину.
- Подібна розробка дозволить виявити списування, перевірити відповідність завдання варіанту та полегшити роботу викладача.

Метою дослідження є вдосконалення технології перевірки контенту студентських робіт.

Задачі магістерської роботи:

1) здійснити аналіз аналогічних систем перевірки контенту;

2) запропонувати математичну модель для представлення файлів;

3) запропонувати математичну модель для виявлення плагіату;

4) розробити алгоритми та програмні засоби для системи перевірки;

5) провести експериментальну перевірку розробленої системи контролю контенту студентських робіт.

Об’єктом дослідження є процес перевірки контенту у текстових файлах.

Предметом дослідження є методи, алгоритми та програмні засоби для створення автоматизованої системи перевірки контенту студентських робіт.

Наукова новизна

Вперше запропоновано комплексну автоматизовану систему перевірки студентських робіт, що дозволить не тільки перевіряти роботи студентів на наявність плагіату, а і контролювати наявність обов’язкових компонентів відповідно до технічного завдання, номеру варіанту тощо.

Практичне значення одержаних результатів

1. Реалізовано алгоритм виявлення плагіату у текстових файлах.
2. Реалізовано алгоритм пошуку наявності обов’язкових компонентів у текстових файлах.
3. Розроблено автоматизовану систему перевірки контенту студентських робіт.
4. Виконано програмну реалізацію основних модулів системи перевірки контенту студентських робіт.

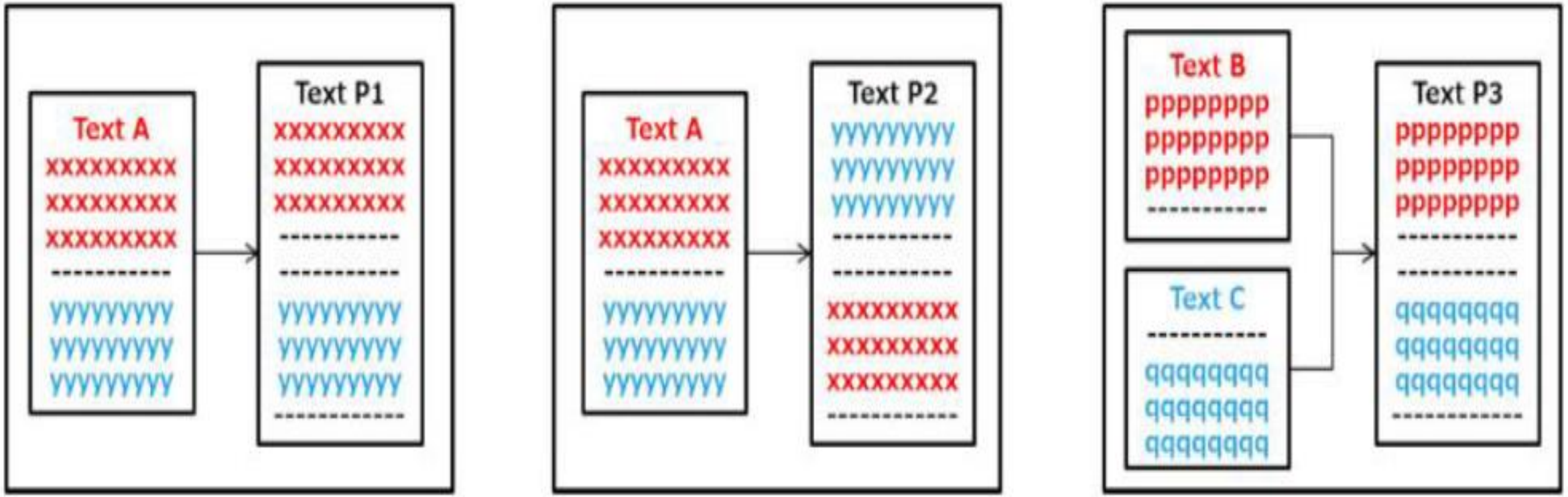
Публікації

«АВТОМАТИЗОВАНА СИСТЕМА КОНТРОЛЮ КОНТЕНТУ СТУДЕНТСКЬИХ РОБІТ»

Молодь в науці: дослідження, проблеми, перспективи (МН-2021)

Основні етапи інтелектуальної перевірки документів

Приклади шаблонів плагіату



Токенізація

Початковий текст

На дворі стоїть червоний автомобіль. Це автомобіль Петра!

Видалення стоп-символів

На дворі стоїть червоний автомобіль Це автомобіль Петра

Видалення стоп-слів

дворі стоїть червоний автомобіль автомобіль Петра

Приведення вхідного тексту до нижнього регістру

дворі стоїть червоний автомобіль атомобіль петра

Обрахунок статистичних показників

$$\text{TF}\ (t,d) = \frac{f_{t,d}}{\text{кількість слів у } d}$$

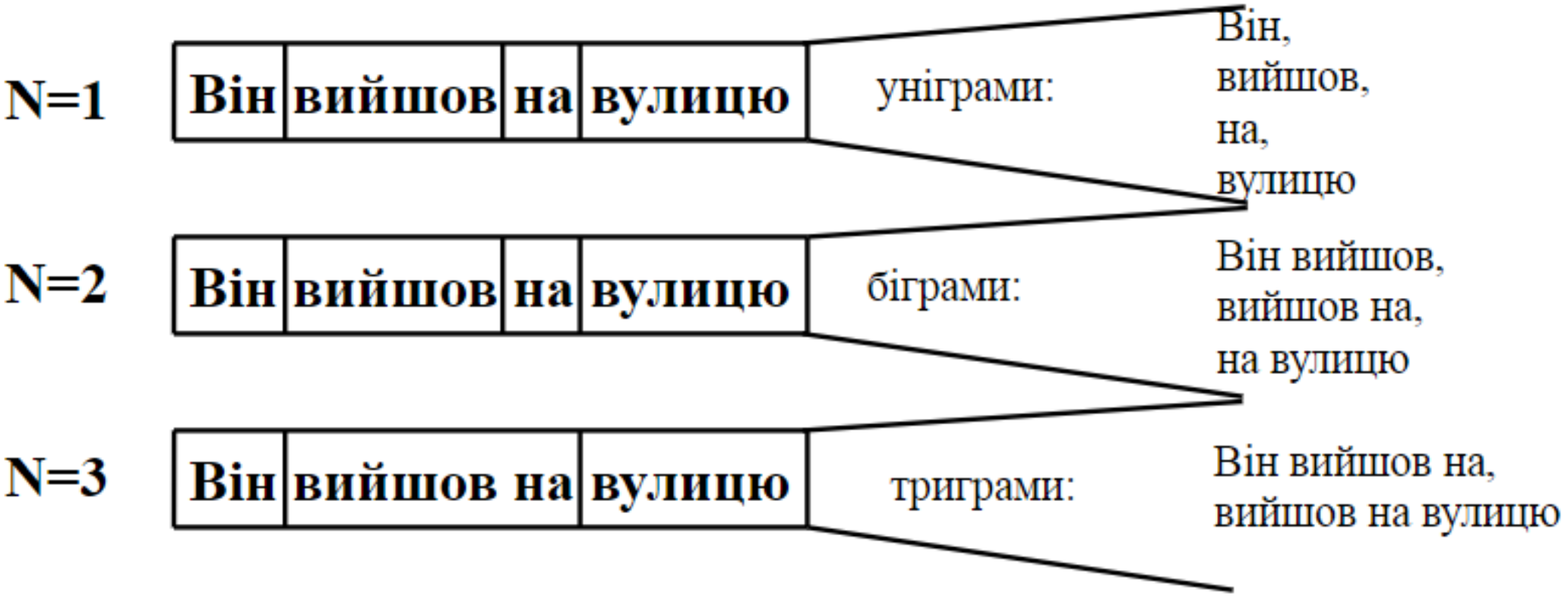
$$idf(t,D) = log\frac{N}{|\{d \in D : t \in d\}|}\ ,$$

де N: загальна кількість документів у корпусі N=|D|;

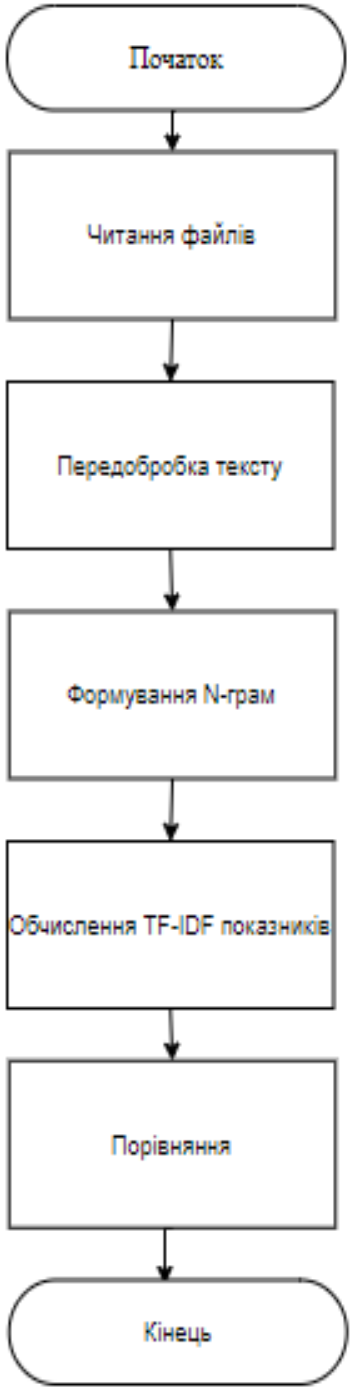
D|{d ∈ : t ∈ d}| : кількість документів, де з’являється термін (tf(t,d) ≠ 0).

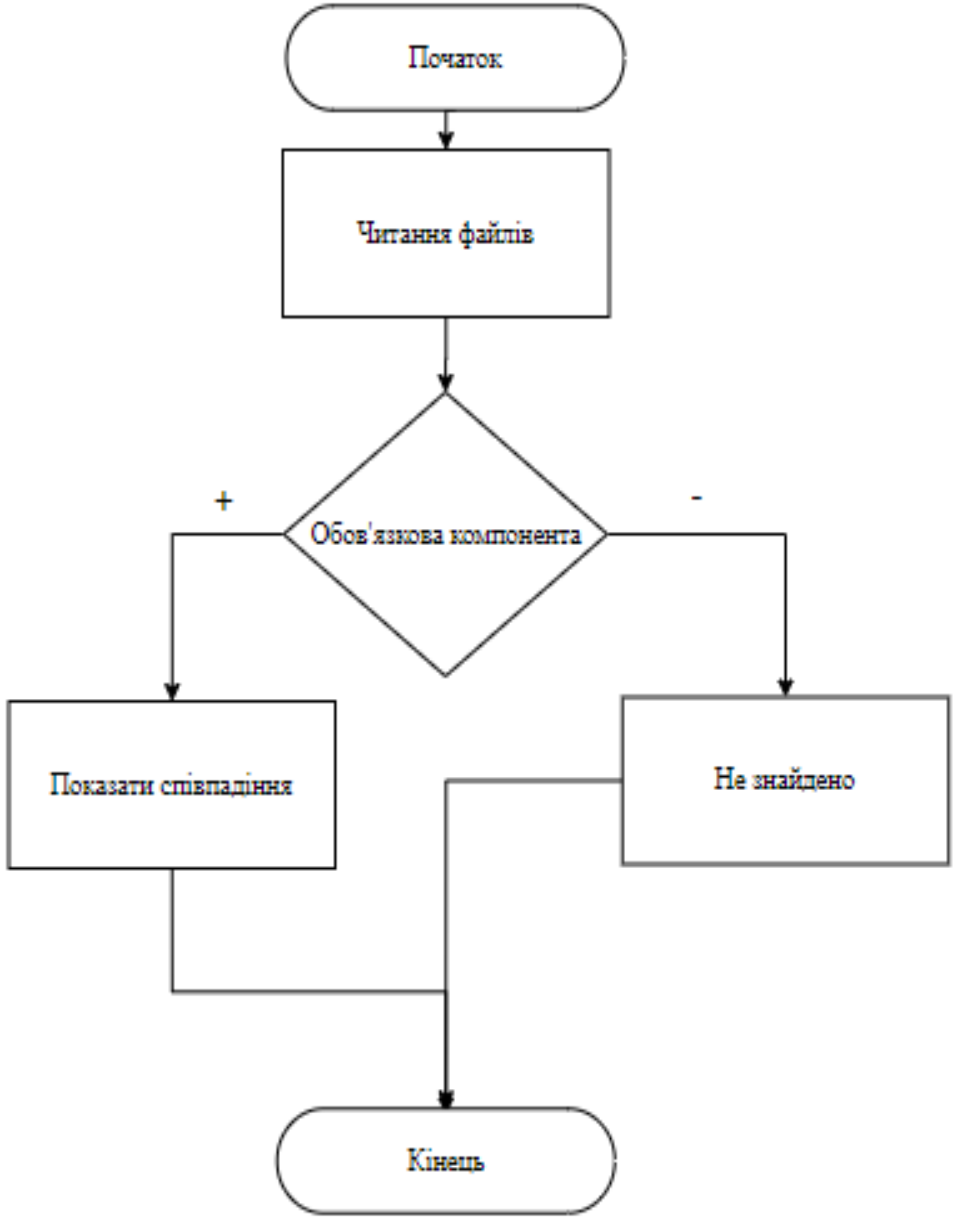
$$tfidf(t,d,D) = tf(t,d)idf(t,D)$$

Використання N-грам, для вирішення проблеми втрати порядку



Алгоритм роботи





Тестування

Стек містить 6 файлів курсових робіт з «Комп’ютерних мереж» (файли від 1 до 6) та 2 курсові з інших предметів (файли 7 та 8). Файли будуть позначатися номерами, без вказування прізвищ.

Файли для тестів у середньому мають розмір від 45 до 55 сторінок та містять від 6000 до 7500 тисяч слів.

Результати для 5 N-грам

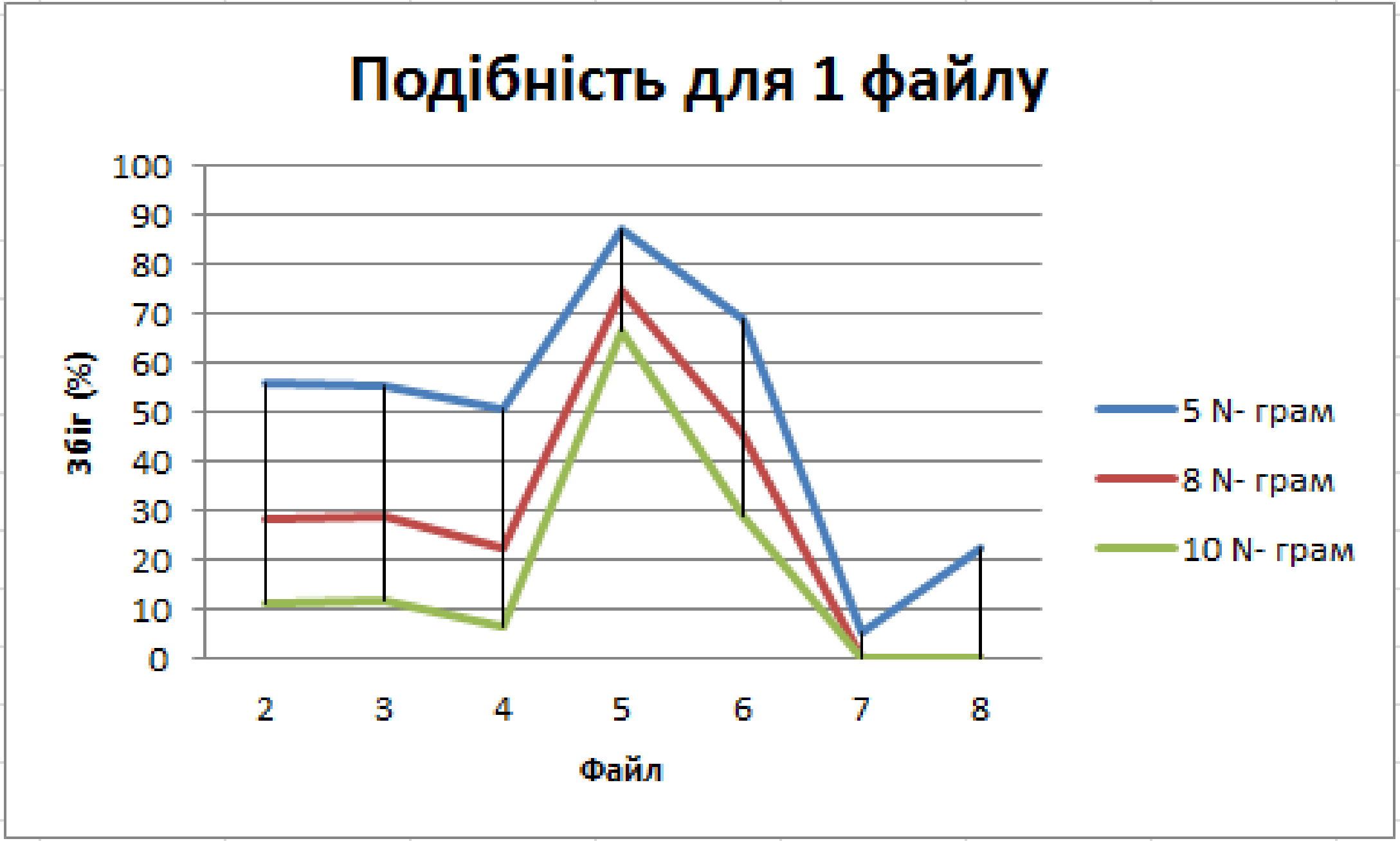
	1	2	3	4	5	6	7	8
1		55,58	55,22	50,44	86,6	68,83	5,26	22,17
2	55,58		63,58	67,67	58	57,77	5,48	23,19
3	55,22	63,58		57,67	51,71	54,49	5,34	22,61
4	50,44	67,67	57,67		51,71	54,49	5,34	22,61
5	86,6	58	55,91	51,71		69,22	5,35	22,67
6	68,83	57,77	61,21	54,49	69,22		5,76	24,22
7	5,26	5,48	5,24	5,34	5,35	5,76		15,88
8	22,17	23,19	22,13	22,61	22,67	24,22	15,88	

Результати для 8 N-грам

	1	2	3	4	5	6	7	8
1		28,06	28,62	22,17	74,61	45,23	0	0
2	28,06		40,58	45,2	30,56	29,49	0	0,01
3	28,62	40,58		33,9	27,99	36,3	0	0
4	22,17	45,2	33,9		23,33	26,03	0	0
5	74,61	30,56	27,99	23,33		45,75	0	0
6	45,23	29,49	36,3	26,03	45,75		0	0
7	0	0	0	0	0	0		3,8
8	0,01	0	0	0	0	0	3,8	

Результати для 10 N-грам

	1	2	3	4	5	6	7	8
1		11,09	11,63	6,18	66,1	28,93	0	0
2	11,09		25,78	31,62	12,55	10,91	0	0
3	11,63	25,78		20,07	9,95	19,32	0	0
4	6,18	31,62	20,07		6,8	8,47	0	0
5	66,1	12,55	9,95	6,8		29,19	0	0
6	28,93	10,91	19,32	8,47	29,19		0	0
7	0	0	0	0	0	0		0
8	0	0	0	0	0	0	0	


Подібність для 1 файлу
100
90
80
70
60
50
40
30
20
10
0
Збіг (%)
2
3
4
5
6
7
8
Файл
5 N- грам
8 N- грам
10 N- грам

Демонстрація…

ВИСНОВКИ

У роботі було:

1. Розглянуто основні методи побудови систем аналізу контенту та проведено їх класифікацію за різними ознаками.

2. Розроблено теоретико-математичну модель системи.

3. Розроблено структуру програми, необхідне програмне забезпечення.

4. Проведено експериментальну перевірку розробленої автоматизованої системи контролю контенту студентський робіт, що показало коректність та правильність роботи програмних засобів.

Отже, поставлені задачі були виконані, а мета досягнута.

Дякую за увагу