НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ

МЕТОДИ КРИПТОАНАЛІЗУ 1

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №2

Статистичні критерії на відкритий текст

Варіант 18

Виконав:
Бєш Радомир ФІ-42мн
Перевірив:
Ядуха Д.В.

Київ — 2025

Зміст

1 Мета

Засвоєння статистичних методів розрізнення змістовного тексту від випадкової послідовності, порівняння їх, визначення похибок першого та другого роду.

2 Постановка задачі та варіант завдання

Номер варіанту завдань: 18
Завдання поставлені перед виконанням комп'ютерного парктикуму:

-1. Ознайомитись з порядком виконання комп'ютерного практикуму та відповідними вимогами до виконання роботи.
0. Уважно прочитати необхідні теоретичні відомості до комп'ютерного практикуму.
1. Створити новий репозиторій в системі контролю версій Git (бажано використовувати вебсервіс GitHub).
2. На великому тексті українською мовою (>1MB), необхідно розрахувати частоти літер і біграм, а також ентропію та індекс відповідності.
3. Отримати N текстів X українською мовою для довжин $L = 10, 100, 1000$ та 10000, для кожного з яких згенерувати спотворені тексти Y. Число N визначається відповідно до такої таблиці.

L	N
10 100 1000	10000
10000	1000

Спотворення тексту виконується такими способами:
(а) шляхом застосування шифру Віженера з випадковим ключем довжини r = 1, 5, 10.
(б) шляхом застосування шифру афінної та афінної біграмної підстановки з випадковими ключами.

А також тексти для аналізу формуються такими способами:
(а) y_i — рівномірно розподілена послідовність символів з $(Z_m)^l$
(б) y_i обчислюється відповідно до такого співвідношення:

$$y_i = (s_{i-1} + s_{i-2}) \mod m^l,$$

де $s_0, s_1 \in_R (Z_m)^l$.

4. Реалізувати критерії (відповідно до варіанту + структурний) і перевірити їх роботу на згенерованих N текстах для кожної довжини L. Розрахувати ймовірності похибок першого і другого роду. Номер варіанту Критерії

Номер варіанту	Критерії
Парний	1.0-1.3, 3.0, 5.1
Непарний	2.0-2.3, 4.0, 5.0

Усі вищезгадані критерії (та інші формули), які використовували значення l, мають приймати значення $l = 1$ та $l = 2$, тобто реалізувати символьний та біграмний критерії.

5. Згенерувати випадковий текст довжини $L = 10000$, який точно не є зв'язним текстом українською мовою (наприклад, текст, який складається з величезної кількості літер а: "ааааааааа..."). Застосувати один з варіантів спотворення (на вибір) до цього тексту, після чого застосувати один з реалізованих критеріїв (на вибір). Порівняти результати застосування критерію до різних текстів.
6. Оформити звіт до комп'ютерного практикуму.

3 Хід роботи

1. Ознайомлення з методичними вказівками та вимогами щодо виконання комп'ютерного практикуму
2. Ознайомлення, вивчення та систематизація необхідного теоретичного матеріалу для виконання комп'ютерного практикуму
3. Створення та налаштування репозиторію
4. Безпосередня реалізація поставлених задач комп'ютерного практикуму за допомогою програмування
5. Аналіз та систематизація отриманих результатів
6. Підготовка звіту з виконання комп'ютерного практикуму з детальним описом отриманих результатів

4 Варіант №3

4.1 Опис множин заборонених/частих символів, які було отримано при виконанні завдання

Опишемо загальну характеристику, вибірку з п'яти найчасчтіших символів та символів, які зустрічаються дуже рідко:

- **Найчастіші символи: о, а, н, в, и**
- **Найрідкіші символи: щ, ц, є, ї, ф**

4.2 Таблиці

- $\alpha = P(H_1|H_0)$ - ймовірність помилки 1-го роду (англ.*false positive*), ймовірність назвати відкритий текст випадковою послідовністю;
- $\beta = P(H_0|H_1)$ - ймовірність помилки 2-го роду (англ.*false negative*), ймовірність прийняти випадкову послідовністю за відкритий текст.

Шифрування за Віженером (r = 1)					
		Монограма		Біграма	
L	Номер критерію	$\alpha(l=1)$	$\beta(l=1)$	$\alpha(l=2)$	$\beta(l=2)$
10	2.0 60000/9000	43,6%	5,51%	85,6%	0,63%
	2.1 40000/9000,$k_{(f_1)}$=11, $k_{(f_2)}$=400	87,68%	47,22%	96,19%	31,5%
	2.2 40000/9000,$k_{(x_1)}$=2, $k_{(x_2)}$=1	65,21%	33%	87,67%	1,89%
	2.3 40000/9000,$K_{(f_1)}$=11, $K_{(f_2)}$=7	38,76%	17,26%	56,7%	3,24%
	4.0 0,02/0,04,$k_{(I_1)}$=5,2, $k_{(I_2)}$=0,9	46,71%	53,32%	100%	0%
	5.0 4000/50,$k_{(empt_1)}$=5, $k_{(empt_2)}$=40	2,03%	73,2%	0%	56,8%
100	2.0 15000/7000	-%	%	-%	%
	2.1 15000/7000,$k_{(f_1)}$=3, $k_{(f_2)}$=520	-%	-%	-%	-%
	2.2 15000/7000,$k_{(x_1)}$=6, $k_{(x_2)}$=2	-%	-%	-%	-%
	2.3 15000/7000,$K_{(f_1)}$=320, $K_{(f_2)}$=40	-%	-%	-%	-%
	4.0 0,02/0,004,$k_{(I_1)}$=4,87, $k_{(I_2)}$=0,92	-%	-%	-%	-%
	5.0 4000/50,$k_{(empt_1)}$=3, $k_{(empt_2)}$=41,25	-%	-%	-%	-%
1000	2.0 20000/7000	-%	%	-%	-%
	2.1 2000/6000,$k_{(f_1)}$=7, $k_{(f_2)}$=250	-%	-%	-%	-%
	2.2 15000/7000,$k_{(x_1)}$=20, $k_{(x_2)}$=25	-%	-%	-%	-%
	2.3 15000/7000,$K_{(f_1)}$=857, $K_{(f_2)}$=35	-%	-%	-%	-%
	4.0 0,00002/0,0003,$k_{(I_1)}$=0,005, $k_{(I_2)}$=0,015	-%	-%	-%	-%
	5.0 7000/50,$k_{(empt_1)}$=3, $k_{(empt_2)}$=0,54	-%	-%	-%	-%
10000	2.0 20000/5000	-%	-%	-%	-%
	2.1 2000/5000,$k_{(f_1)}$=7, $k_{(f_2)}$=250	-%	-%	-%	-%
	2.2 15000/5000,$k_{(x_1)}$=20, $k_{(x_2)}$=25	-%	-%	-%	-%
	2.3 15000/5000,$K_{(f_1)}$=800, $K_{(f_2)}$=35	-%	-%	-%	-%
	4.0 0,005/0,002,$k_{(I_1)}$=0,015, $k_{(I_2)}$=0,015	-%	-%	-%	-%
	5.0 30000/60,$k_{(empt_1)}$=1, $k_{(empt_2)}$=200	-%	-%	-%	-%

Шифрування за Віженером (r = 5)					
		Монограма		Біграма	
L	Номер критерію	**FP**$(l=1)$	**FN**$(l=1)$	**FP**$(l=2)$	**FN**$(l=2)$
10	2.0 60000/9000	43,6%	5,51%	85,6%	0,63%
	2.1 40000/9000,$k_{(f_1)}$=11, $k_{(f_2)}$=400	87,68%	-%	-%	-%
	2.2 40000/9000,$k_{(x_1)}$=2, $k_{(x_2)}$=1	65,21%	-%	-%	-%
	2.3 40000/9000,$K_{(f_1)}$=11, $K_{(f_2)}$=7	38,76%	26,26%	-%	-%
	4.0 0,02/0,04,$k_{(I_1)}$=5,2, $k_{(I_2)}$=0,9	46,71%	-%	-%	-%
	5.0 4000/50,$k_{(empt_1)}$=5, $k_{(empt_2)}$=40	2,03%	-%	-%	-%
100	2.0 15000/7000	-%	-%	-%	-%
	2.1 15000/7000,$k_{(f_1)}$=3, $k_{(f_2)}$=520	-%	-%	-%	-%
	2.2 15000/7000,$k_{(x_1)}$=6, $k_{(x_2)}$=2	-%	-%	-%	-%
	2.3 15000/7000,$K_{(f_1)}$=320, $K_{(f_2)}$=40	-%	-%	-%	-%
	4.0 0,02/0,004,$k_{(I_1)}$=4,87, $k_{(I_2)}$=0,92	-%	-%	-%	-%
	5.0 4000/50,$k_{(empt_1)}$=3, $k_{(empt_2)}$=41,25	-%	-%	-%	-%
1000	2.0 20000/7000	-%	-%	-%	-%
	2.1 2000/6000,$k_{(f_1)}$=7, $k_{(f_2)}$=250	-%	-%	-%	-%
	2.2 15000/7000,$k_{(x_1)}$=20, $k_{(x_2)}$=25	-%	-%	-%	-%
	2.3 15000/7000,$K_{(f_1)}$=8570, $K_{(f_2)}$=35	-%	-%	-%	-%
	4.0 0,00002/0,0003,$k_{(I_1)}$=0,005, $k_{(I_2)}$=0,015	-%	-%	-%	-%
	5.0 7000/50,$k_{(empt_1)}$=3, $k_{(empt_2)}$=0,54	-%	-%	-%	-%
10000	2.0 20000/5000	-%	-%	-%	-%
	2.1 2000/5000,$k_{(f_1)}$=7, $k_{(f_2)}$=250	-%	-%	-%	-%
	2.2 15000/5000,$k_{(x_1)}$=20, $k_{(x_2)}$=25	-%	-%	-%	-%
	2.3 15000/5000,$K_{(f_1)}$=800, $K_{(f_2)}$=35	-%	-%	-%	-%
	4.0 0,005/0,002,$k_{(I_1)}$=0,015, $k_{(I_2)}$=0,015	-%	-%	-%	-%
	5.0 30000/60,$k_{(empt_1)}$=1, $k_{(empt_2)}$=200	-%	-%	-%	-%

Шифрування за Віженером (r = 10)					
		Монограма		Біграма	
L	Номер критерію	**FP**$(l=1)$	**FN**$(l=1)$	**FP**$(l=2)$	**FN**$(l=2)$
10	2.0 60000/9000	43,6%	5,51%	-%	0,63%
	2.1 40000/9000,$k_{(f_1)}$=11, $k_{(f_2)}$=400	87,68%	47,22%	76,19%	-%
	2.2 40000/9000,$k_{(x_1)}$=2, $k_{(x_2)}$=1	65,21%	-%	-%	-%
	2.3 40000/9000,$K_{(f_1)}$=11, $K_{(f_2)}$=7	38,76%	26,26%	-%	-%
	4.0 0,02/0,04,$k_{(I_1)}$=5,2, $k_{(I_2)}$=0,9	46,71%	-%	-%	-%
	5.0 4000/50,$k_{(empt_1)}$=5, $k_{(empt_2)}$=40	2,03%	-%	-%	-%
100	2.0 15000/7000	-%	-%	-%	-%
	2.1 15000/7000,$k_{(f_1)}$=3, $k_{(f_2)}$=520	-%	-%	-%	-%
	2.2 15000/7000,$k_{(x_1)}$=6, $k_{(x_2)}$=2	-%	-%	-%	-%
	2.3 15000/7000,$K_{(f_1)}$=320, $K_{(f_2)}$=40	-%	-%	-%	-%
	4.0 0,02/0,004,$k_{(I_1)}$=4,87, $k_{(I_2)}$=0,92	-%	-%	-%	-%
	5.0 4000/50,$k_{(empt_1)}$=3, $k_{(empt_2)}$=41,25	-%	-%	-%	-%
1000	2.0 20000/7000	-%	-%	-%	-%
	2.1 2000/6000,$k_{(f_1)}$=7, $k_{(f_2)}$=250	-%	47,22%	-%	-%
	2.2 15000/7000,$k_{(x_1)}$=20, $k_{(x_2)}$=25	-%	-%	-%	-%
	2.3 15000/7000,$K_{(f_1)}$=857, $K_{(f_2)}$=35	-%	-%	-%	-%
	4.0 0,00002/0,0003,$k_{(I_1)}$=0,005, $k_{(I_2)}$=0,015	-%	-%	-%	-%
	5.0 7000/50,$k_{(empt_1)}$=3, $k_{(empt_2)}$=0,54	-%	-%	-%	-%
10000	2.0 20000/5000	-%	-%	-%	-%
	2.1 2000/5000,$k_{(f_1)}$=7, $k_{(f_2)}$=250	-%	-%	-%	-%
	2.2 15000/5000,$k_{(x_1)}$=20, $k_{(x_2)}$=25	-%	-%	-%	-%
	2.3 15000/5000,$K_{(f_1)}$=800, $K_{(f_2)}$=35	-%	-%	-%	-%
	4.0 0,005/0,002,$k_{(I_1)}$=0,015, $k_{(I_2)}$=0,015	-%	-%	-%	-%
	5.0 30000/60,$k_{(empt_1)}$=1, $k_{(empt_2)}$=200	-%	-%	-%	-%

Шифрування за допомогою афінної підстановки з ключами a=5, b=7					
		Монограма		Біграма	
L	Номер критерію	**FP**$(l=1)$	**FN**$(l=1)$	**FP**$(l=2)$	**FN**$(l=2)$
10	2.0 60000/9000	-%	-%	-%	-%
	2.1 40000/9000,$k_{(f_1)}$=11, $k_{(f_2)}$=400	-%	-%	-%	-%
	2.2 40000/9000,$k_{(x_1)}$=2, $k_{(x_2)}$=1	-%	-%	-%	-%
	2.3 40000/9000,$K_{(f_1)}$=11, $K_{(f_2)}$=7	-%	-%	-%	-%
	4.0 0,02/0,04,$k_{(I_1)}$=5,2, $k_{(I_2)}$=0,9	-%	-%	-%	-%
	5.0 4000/50,$k_{(empt_1)}$=5, $k_{(empt_2)}$=40	-%	-%	-%	-%
100	2.0 15000/7000	-%	-%	-%	-%
	2.1 15000/7000,$k_{(f_1)}$=3, $k_{(f_2)}$=520	-%	-%	-%	-%
	2.2 15000/7000,$k_{(x_1)}$=6, $k_{(x_2)}$=2	-%	-%	-%	-%
	2.3 15000/7000,$K_{(f_1)}$=320, $K_{(f_2)}$=40	-%	-%	-%	-%
	4.0 0,02/0,004,$k_{(I_1)}$=4,87, $k_{(I_2)}$=0,92	-%	-%	-%	-%
	5.0 4000/50,$k_{(empt_1)}$=3, $k_{(empt_2)}$=41,25	-%	-%	-%	-%
1000	2.0 20000/7000	-%	-%	-%	-%
	2.1 2000/6000,$k_{(f_1)}$=7, $k_{(f_2)}$=250	-%	-%	-%	-%
	2.2 15000/7000,$k_{(x_1)}$=20, $k_{(x_2)}$=25	-%	-%	-%	-%
	2.3 15000/7000,$K_{(f_1)}$=857, $K_{(f_2)}$=35	-%	-%	-%	-%
	4.0 0,00002/0,0003,$k_{(I_1)}$=0,005, $k_{(I_2)}$=0,015	-%	-%	-%	-%
	5.0 7000/50,$k_{(empt_1)}$=3, $k_{(empt_2)}$=0,54	-%	-%	-%	-%
10000	2.0 20000/5000	-%	-%	-%	-%
	2.1 2000/5000,$k_{(f_1)}$=7, $k_{(f_2)}$=250	-%	-%	-%	-%
	2.2 15000/5000,$k_{(x_1)}$=20, $k_{(x_2)}$=25	-%	-%	-%	-%
	2.3 15000/5000,$K_{(f_1)}$=800, $K_{(f_2)}$=35	-%	-%	-%	-%
	4.0 0,005/0,002,$k_{(I_1)}$=0,015, $k_{(I_2)}$=0,015	-%	-%	-%	-%
	5.0 30000/60,$k_{(empt_1)}$=1, $k_{(empt_2)}$=200	-%	-%	-%	-%

Шифрування за рівномірно розподіленою послідовністю					
		Монограма		Біграма	
L	Номер критерію	**FP**$(l=1)$	**FN**$(l=1)$	**FP**$(l=2)$	**FN**$(l=2)$
10	2.0 60000/9000	-%	-%	-%	-%
	2.1 40000/9000,$k_{(f_1)}$=11, $k_{(f_2)}$=400	-%	-%	-%	-%
	2.2 40000/9000,$k_{(x_1)}$=2, $k_{(x_2)}$=1	-%	-%	-%	-%
	2.3 40000/9000,$K_{(f_1)}$=11, $K_{(f_2)}$=7	-%	26,26%	-%	-%
	4.0 0,02/0,04,$k_{(I_1)}$=5,2, $k_{(I_2)}$=0,9	-%	-%	-%	-%
	5.0 4000/50,$k_{(empt_1)}$=5, $k_{(empt_2)}$=40	-%	-%	-%	-%
100	2.0 15000/7000	-%	-%	-%	-%
	2.1 15000/7000,$k_{(f_1)}$=3, $k_{(f_2)}$=520	-%	-%	-%	-%
	2.2 15000/7000,$k_{(x_1)}$=6, $k_{(x_2)}$=2	-%	-%	-%	-%
	2.3 15000/7000,$K_{(f_1)}$=320, $K_{(f_2)}$=40	-%	-%	-%	-%
	4.0 0,02/0,004,$k_{(I_1)}$=4,87, $k_{(I_2)}$=0,92	-%	-%	-%	-%
	5.0 4000/50,$k_{(empt_1)}$=3, $k_{(empt_2)}$=41,25	-%	-%	-%	-%
1000	2.0 20000/7000	-%	-%	-%	-%
	2.1 2000/6000,$k_{(f_1)}$=7, $k_{(f_2)}$=250	-%	-%	-%	-%
	2.2 15000/7000,$k_{(x_1)}$=20, $k_{(x_2)}$=25	-%	-%	-%	-%
	2.3 15000/7000,$K_{(f_1)}$=857, $K_{(f_2)}$=35	-%	-%	-%	-%
	4.0 0,00002/0,0003,$k_{(I_1)}$=0,005, $k_{(I_2)}$=0,015	-%	-%	-%	-%
	5.0 7000/50,$k_{(empt_1)}$=3, $k_{(empt_2)}$=0,54	-%	-%	-%	-%
10000	2.0 20000/5000	-%	-%	-%	-%
	2.1 2000/5000,$k_{(f_1)}$=7, $k_{(f_2)}$=250	-%	-%	-%	-%
	2.2 15000/5000,$k_{(x_1)}$=20, $k_{(x_2)}$=25	-%	-%	-%	-%
	2.3 15000/5000,$K_{(f_1)}$=800, $K_{(f_2)}$=35	-%	-%	-%	-%
	4.0 0,005/0,002,$k_{(I_1)}$=0,015, $k_{(I_2)}$=0,015	-%	-%	-%	-%
	5.0 30000/60,$k_{(empt_1)}$=1, $k_{(empt_2)}$=200	-%	-%	-%	-%

Шифрування за псевдовипадковою послідовністю					
		Монограма		Біграма	
L	Номер критерію	**FP**$(l=1)$	**FN**$(l=1)$	**FP**$(l=2)$	**FN**$(l=2)$
10	2.0 60000/9000	-%	-%	-%	-%
	2.1 40000/9000,$k_{(f_1)}$=11, $k_{(f_2)}$=400	-%	%	-%	-%
	2.2 40000/9000,$k_{(x_1)}$=2, $k_{(x_2)}$=1	-%	-%	-%	-%
	2.3 40000/9000,$K_{(f_1)}$=11, $K_{(f_2)}$=7	-%	-%	-%	-%
	4.0 0,02/0,04,$k_{(I_1)}$=5,2, $k_{(I_2)}$=0,9	-%	-%	-%	-%
	5.0 4000/50,$k_{(empt_1)}$=5, $k_{(empt_2)}$=40	-%	-%	-%	-%
100	2.0 15000/7000	-%	-%	-%	-%
	2.1 15000/7000,$k_{(f_1)}$=3, $k_{(f_2)}$=520	-%	-%	-%	-%
	2.2 15000/7000,$k_{(x_1)}$=6, $k_{(x_2)}$=2	-%	-%	-%	-%
	2.3 15000/7000,$K_{(f_1)}$=320, $K_{(f_2)}$=40	-%	-%	-%	-%
	4.0 0,02/0,004,$k_{(I_1)}$=4,87, $k_{(I_2)}$=0,92	-%	-%	-%	-%
	5.0 4000/50,$k_{(empt_1)}$=3, $k_{(empt_2)}$=41,25	-%	-%	-%	-%
1000	2.0 20000/7000	-%	-%	-%	-%
	2.1 2000/6000,$k_{(f_1)}$=7, $k_{(f_2)}$=250	-%	-%	-%	-%
	2.2 15000/7000,$k_{(x_1)}$=20, $k_{(x_2)}$=25	-%	-%	-%	-%
	2.3 15000/7000,$K_{(f_1)}$=857, $K_{(f_2)}$=35	-%	-%	-%	-%
	4.0 0,00002/0,0003,$k_{(I_1)}$=0,005, $k_{(I_2)}$=0,015	-%	-%	-%	-%
	5.0 7000/50,$k_{(empt_1)}$=3, $k_{(empt_2)}$=0,54	-%	-%	-%	-%
10000	2.0 20000/5000	-%	-%	-%	-%
	2.1 2000/5000,$k_{(f_1)}$=7, $k_{(f_2)}$=250	-%	-%	-%	-%
	2.2 15000/5000,$k_{(x_1)}$=20, $k_{(x_2)}$=25	-%	-%	-%	-%
	2.3 15000/5000,$K_{(f_1)}$=800, $K_{(f_2)}$=35	-%	-%	-%	-%
	4.0 0,005/0,002,$k_{(I_1)}$=0,015, $k_{(I_2)}$=0,015	-%	-%	-%	-%
	5.0 30000/60,$k_{(empt_1)}$=1, $k_{(empt_2)}$=200	-%	-%	-%	-%

4.3 Опис алгоритму стиснення, що був обраний для розробки структурного критерію

Стиснення LZMA — це тип алгоритму стиснення даних, який був розроблений у рамках проекту 7z. Назва «LZMA» розшифровується як «алгоритм ланцюга Лемпеля-Зіва Маркова».

LZMA стискає файли за допомогою методів статистичного моделювання та словників. Статистичне моделювання дозволяє аналізувати цілі блоки тексту, тоді як словникові методи стискають невеликі фрагменти одночасно.

Як і більшість подібних алгоритмів, він працює шляхом заміни рядків, які часто зустрічаються в нестиснутих даних, покажчиками на попередні входження.

У стисненні LZMA оцінки частоти можна обчислити за набором символів або підрядків. Зробивши це, ви використовуєте отримані оцінки, щоб знайти збіги та замінити їх покажчиками.

LZMA розроблено для надзвичайно швидкої роботи на різноманітному апаратному забезпеченні, від традиційних жорстких дисків до сучасних твердотільних накопичувачів і вбудованих пристроїв з обмеженою потужністю ЦП.

4.4 Опис запропонованого структурного критерію, що базується на основі результатів стиснення

Основна ідея для побудови структурного критерію будується на різниці довжин вхідного тексту та отриманого тексту після застосування алгоритму стистення. Оскільки у природній мові певні слова, словосполучення та літери зустрічаються частіше, очевидно що стиснувши текст написаний природною мовою матиме меншу довжину, ніж стиснутий текст випадкового набору символів.

Запустивши даний алгоритм стиснення для природних текстів різних довжин, ми отримали дві множини: множина довжин текстів до стиснення (множина X) та множина текстів після стиснення (множина Y). Зобразивши дані точки на координатній площині ми отримали певну кількість точок. Імпіричним методом було визначено параметри та рівняння функції, яка б містила дані точки.

В результаті маємо функцію, яка приймає на вхід параментр довжини тексту, а на виході повертає орієнтовний розмір даного тексту після стиснення.

Структурний критерій працює наступним чином. Ми запускаємо алгортим стиснення для вхідного тексту, а потім обраховуємо його довжину. Після цього ми обраховуємо функцію, описану вище для отримання значення довжини вхідного тексту після стиснення, якщо б він був природним. Порівнявши дані значення можнемо зробити висновки стосовно того, чи є вхідний текст природним. Якщо його довжина є більшою, ніж значення, отримане в результаті обрахунку функції, даний текст не є текстом, написаним приодною мовою, і навпаки.

4.5 Опис труднощів, що виникали при виконанні комп'ютерного практикуму, та шляхи їх розв'язання

Завдання, виконані в межах даного комп'ютерного практикуму є досить об'ємними, хоча і не є дуже складними. Реалізація статистичних критеріїв зайняла у нашої команди багато часу, але ще більше часу знадобилось щоб отримати, систематизувати та проаналізувати отримані дані.

Також ми стикнулись з певними складнощами при розподілі завдань та комунації між учасниками бригади при виконанні ріних етапів даного комп'ютерного практикуму.

4.6 Висновки (аналіз ефективності реалізованих критеріїв, порівняння їх між собою, порівняння результатів для різних значень L, r, алгоритмів спотворення тощо)

У цій роботі було програмно реалізовано алгоритми для спотворення текстів і генерації випадкових послідовностей. Крім того, було створено критерії перевірки гіпотез, які визначають, чи є вхідна послідовність символів або біграм осмисленим текстом, чи випадковою послідовністю. Також розроблено власний структурний критерій, який використовує алгоритм стиснення даних LZMA.

Результати роботи представлені у вигляді таблиць із оцінками помилок першого та другого роду. Аналізуючи результати для критеріїв 2.0 та 2.1, встановлено, що вони найбільш ефективні для невеликих довжин послідовностей. Це зумовлено тим, що для великих довжин усі часті біграми швидше за все зустрінуться в тексті, тоді як для монограм шанс їхньої зустрічі стає ще вищим.

Критерій 4.0 показує хороші результати для довгих послідовностей, але вимагає надзвичайно точної оцінки, яку важко отримати. Критерій 2.2 демонструє покращення зі збільшенням довжини послідовності, однак для коротких текстів, наприклад, довжиною 10 або

100, він виявився неефективним. Значна ефективність цього критерію спостерігається лише при довжинах, що перевищують 1000.

Стосовно власного статистичного критерію, запропонованого нами у межах даної роботи, найкраще всього він працює на великих довжинах (від 1000 символів і ще краще від 10000 символів). Адже змістовний та незмістовний текст довжиною 10 символів матимуть однаку довжину в бітах після стиснення.