НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ім. Ігоря СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ

Звіт з виконання кваліфікаційного дослідження

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №2

З КУРСУ

МЕТОДИ КРИПТОАНАЛІЗУ 1

Виконала студентка
групи ФІ-32мн
Міснік Аліна Олексіївна

Викладач:
Ядуха Д.В.

Київ — 2024

ВСТУП

Метою роботи є засвоєння статистичних методів розрізнення змістовного тексту від випадкової послідовності, порівняння їх, визначення похибок першого та другого роду.

Задача полягає у реалізації різних методи спотворення – перетворення змістовного тексту у випадкову послідовність, а також у визначенні помилки 1-го та 2-го роду для кожного із критеріїв заданих варіантом.

Наведемо повний перелік завдань, які необхідно виконати в ході комп'ютерного практикуму:

1) На великому тексті українською мовою (>1MB), де:

а) літера «ґ» замінена на літеру «г»;

б) видалений символ апострофу та усі інші спецсимволи в тексті, включно з пробілами;

в) текст містить лише маленькі літери алфавіту.

необхідно розрахувати частоти літер і біграм, а також ентропію та індекс відповідності.

2) Отримати N текстів X українською мовою для довжин $L = 10, 100, 1000$ та 10000, для кожного з яких згенерувати спотворені тексти Y. 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;

в) y_i — рівномірно розподілена послідовність символів з $(Z_m)^l$

г) y_i обчислюється відповідно до такого співвідношення: $y_i = s_{i-1} + s_{i-2}$

3) Реалізувати критерії (1.0-1.3, 3.0, 5.1 + структурний) і перевірити їх роботу на згенерованих N текстах для кожної довжини L. Розрахувати ймовірності похибок першого і другого роду. Усі вищезгадані критерії (та

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

4) 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.

Будемо використовувати дані для варіанта №10.

1 ХІД РОБОТИ

Опишемо як відбувався аналіз тексту.

1.1 Опис множин заборонених/частих символів

Для знаходження множини одиничних літер, які рідко зустрічаються, було обрано значення частоти 0,007. Літерами, частота яких була менша, виявилися такі: ’є’, ’щ’, ’ф’.

Для формування множини заборонених біграм були вибрані ті, які жодного разу не зустрілися у тексті, а саме: ’фж’, ’щщ’, ’еь’, ’гч’, ’сш’, ’фє’, ’шь’, ’лч’, ’фь’, ’щф’, ’жь’, ’щє’, ’фч’, ’фщ’, ’цє’, ’йь’, ’кь’, ’єь’, ’щц’, ’щх’, ’цж’, ’щї’, ’щш’, ’оь’, ’щк’, ’шю’, ’їь’, ’цг’, ’лї’, ’щл’, ’сї’, ’щг’, ’гї’, ’юь’, ’щм’, ’чє’, ’цй’, ’пю’, ’цб’, ’щд’, ’щь’, ’щт’, ’єи’, ’щз’, ’гж’, ’фї’, ’цщ’, ’щв’, ’цх’, ’шж’, ’цш’, ’щю’, ’цч’, ’пж’, ’щч’, ’чь’, ’єе’, ’щб’, ’щр’, ’юи’, ’шї’, ’щж’, ’бь’.

У випадку знаходження частих символів використовувалася частота більша за 0,06 і були обрані такі літери: ’н’, ’а’, ’о’, ’и’. Біграмами, які зустрілися найчастіше, стали: ’на’, ’ов’, ’ро’, ’ст’, ’го’, ’ко’, ’ог’, ’по’, ’ві’, ...

1.2 Таблиці

Спотворення за допомогою шифру Віженера ($r = 1$)					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
10	1.0	0.12	0.36	0	0.46
	1.1	0.004	0.69	0	0.89
	1.2	0.12	0.17	0	0.57
	1.3	0.17	0	0	0.0003
	3.0	0.44	0.55	0.96	0.03
	5.1	0.19	0.24	0	0.78

Спотворення за допомогою шифру Віженера ($r = 5$)					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
10	1.0	0.12	0.43	0	0.52
	1.1	0.004	0.86	0	0.88
	1.2	0.12	0.42	0	0.58
	1.3	0.17	0.002	0	0
	3.0	0.44	0.51	0.96	0.06
	5.1	0.19	0.25	0	0.8

Спотворення за допомогою шифру Віженера ($r = 10$)					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
10	1.0	0.12	0.43	0	0.58
	1.1	0.004	0.87	0	0.89
	1.2	0.12	0.43	0	0.58
	1.3	0.17	0.001	0	0
	3.0	0.44	0.54	0.96	0.07
	5.1	0.19	0.25	0	0.8

Спотворення за допомогою Афінної підстановки					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
10	1.0	0.12	0.36	0	0.58
	1.1	0.004	0.74	0	0.84
	1.2	0.12	0.39	0	0.56
	1.3	0.17	0,02	0	0
	3.0	0.44	0.53	0.96	0.06
	5.1	0.19	0.31	0	0.8

Рівномірно розподілена послідовність					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
10	1.0	0.12	0.27	0	0.61
	1.1	0.004	0.71	0	0.90
	1.2	0.12	0.27	0	0.61
	1.3	0.17	0	0	0
	3.0	0.44	0.44	0.96	0.009
	5.1	0.19	0.12	0	0.8

Співвідношення					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
10	1.0	0.12	0.27	0	0.61
	1.1	0.004	0.7	0	0.92
	1.2	0.12	0.22	0	0.6
	1.3	0.17	0	0	0
	3.0	0.44	0.43	0.96	0.01
	5.1	0.19	0.12	0	0.8

Спотворення за допомогою шифру Віженера ($r = 1$)					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
100	1.0	0.6	0.001	0	0.006
	1.1	0.26	0.08	0	0.006
	1.2	0.6	0.001	0	0.006
	1.3	0.62	0	0	0
	3.0	0.4	0.59	0.52	0.47
	5.1	0.12	0.22	0.004	0.09

Спотворення за допомогою шифру Віженера ($r = 5$)					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
100	1.0	0.6	0.0	0	0.007
	1.1	0.26	0.004	0	0.007
	1.2	0.6	0	0	0.007
	1.3	0	0	0	0
	3.0	0.4	0.001	0.52	0.02
	5.1	0.12	0.004	0.004	0.12

Спотворення за допомогою шифру Віженера ($r = 10$)					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
100	1.0	0.6	0	0	0.002
	1.1	0.26	0.003	0	0.01
	1.2	0.6	0.001	0	0.002
	1.3	0.62	0	0	0
	3.0	0.4	0	0.52	0.027
	5.1	0.12	0.002	0	0.12

Спотворення за допомогою Афінної підстановки					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
100	1.0	0.6	0.04	0	0.004
	1.1	0.26	0.004	0	0.007
	1.2	0.6	0	0	0.007
	1.3	0	0	0	0
	3.0	0.4	0.59	0.52	0.47
	5.1	0.12	0.14	0.004	0.18

Рівномірно розподілена послідовність					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
100	1.0	0.6	0	0	0.003
	1.1	0.26	0.001	0	0.02
	1.2	0.6	0	0	0.03
	1.3	0	0	0	0
	3.0	0.4	0	0.52	0.001
	5.1	0.12	0	0.004	0.12

Співвідношення					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
100	1.0	0.6	0	0	0
	1.1	0.26	1	0	0
	1.2	0.6	0	0	0
	1.3	0	0	0	0
	3.0	0.4	0.62	0.52	0.08
	5.1	0.12	0.25	0.004	0.13

Спотворення за допомогою шифру Віженера ($r = 1$)					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
1000	1.0	1	0	0	0
	1.1	0.99	0	0	0
	1.2	0.84	0	0	0
	1.3	0.41	0	0	0
	3.0	0.61	0.38	0.31	0.68
	5.1	0.63	0.28	0	0.21

Спотворення за допомогою шифру Віженера ($r = 5$)					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
1000	1.0	1	0	0	0
	1.1	0.9	0	0	0
	1.2	0.84	0	0	0
	1.3	0.41	0	0	0
	3.0	0.61	0	0.31	0
	5.1	0.63	0	0	0.03

Спотворення за допомогою шифру Віженера ($r = 10$)					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
1000	1.0	1	0	0	0
	1.1	0.9	0	0	0
	1.2	0.84	0	0	0
	1.3	0.41	0	0	0
	3.0	0.61	0	0.31	0
	5.1	0.63	0	0	0.06

Спотворення за допомогою Афінної підстановки					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
1000	1.0	1	0	0	0
	1.1	0.9	0	0	0
	1.2	0.84	0	0	0
	1.3	0.41	0	0	0
	3.0	0.61	0.38	0.31	0.68
	5.1	0.63	0.001	0	0.72

Рівномірно розподілена послідовність					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
1000	1.0	1	0	0	0
	1.1	0.9	0	0	0
	1.2	0.84	0	0	0
	1.3	0.41	0	0	0
	3.0	0.61	0	0.31	0
	5.1	0.63	0	0	0.01

Співвідношення					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
1000	1.0	1	0	0	0
	1.1	0.9	0	0	0
	1.2	0.84	0	0	0
	1.3	0.41	0	0	0
	3.0	0.61	1	0.31	0.15
	5.1	0.63	0.9	0	0.08

Спотворення за допомогою шифру Віженера ($r = 1$)					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
10000	1.0	1	0	0	0
	1.1	1	0	0	0
	1.2	0.85	0.15	0	0
	1.3	0.03	0.9	0	0
	3.0	0.46	0.53	0.25	0.74
	5.1	0.9	0	0	0

Спотворення за допомогою шифру Віженера ($r = 5$)					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
10000	1.0	1	0	0	0
	1.1	1	0	0	0
	1.2	0.85	0	0	0
	1.3	0.03	0	0	0
	3.0	0.46	0	0.25	0
	5.1	0.9	0	0	0

Спотворення за допомогою шифру Віженера ($r = 10$)					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
10000	1.0	1	0	0	0
	1.1	1	0	0	0
	1.2	0.85	0	0	0
	1.3	0.03	0	0	0
	3.0	0.46	0	0.25	0
	5.1	0.9	0	0	0

Спотворення за допомогою Афінної підстановки					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
10000	1.0	1	0	0	0
	1.1	1	0	0	0
	1.2	0.85	0	0	0
	1.3	0.03	0	0	0
	3.0	0.46	0.53	0.25	0.74
	5.1	0.9	0	0	0

Рівномірно розподілена послідовність					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
10000	1.0	1	0	0	0
	1.1	1	0	0	0
	1.2	0.85	0	0	0
	1.3	0.03	0	0	0
	3.0	0.46	0	0.25	0
	5.1	0.9	0	0	0

Співвідношення					
L	**Номер критерію**	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
10000	1.0	1	0	0	0
	1.1	1	1	0	0
	1.2	0.85	0	0	0
	1.3	0.03	0	0	0
	3.0	0.46	1	0.25	0.15
	5.1	0.9	0.9	0	0.06

2 СТРУКТУРНИЙ КРИТЕРІЙ

2.1 Алгоритм стиснення zlib

Для розробки структурного критерію був обраний алгоритм zlib. Zlib — це бібліотека для стиснення даних, яка використовує алгоритм Deflate. Алгоритм стиснення полягає в використанні двох етапів: перший — скорочення повторюваних фрагментів даних, і другий — використання стандартного алгоритму стиснення, такого як Huffman coding.

Основні етапи алгоритму стиснення Zlib:

1) **Підготовка блоків:**

– Дані подаються на вхід алгоритму блоками фіксованого розміру або в режимі потоку.

– Кожен блок обробляється незалежно.

2) **Створення Huffman-дерев:**

– Для кожного блоку визначається множина унікальних символів.

– Розраховується частота входження кожного символу у блок.

– Створюється Huffman-дерево на основі цих частот.

3) **Кодування символів:**

– Використовуючи Huffman-дерево, кожен символ замінюється відповідним бітовим кодом.

– Створюється таблиця кодів для швидкого знаходження коду для кожного символу.

4) **Стиснення за допомогою Deflate:**

– Використовуючи стандартний алгоритм стиснення, наприклад, LZ77 (знаходження повторень в даних), створюються блоки стиснених даних.

– Ці блоки об'єднуються в один або кілька фінальних блоків, які інкапсулюють стиснені дані.

5) **Додавання заголовку і футера:**

– Додається заголовок, який містить необхідні метадані (наприклад,

режим стиснення, розмір блоків тощо).

– Додається футер, який містить контрольну суму (Adler-32) для перевірки цілісності даних.

2.2 Формулювання структурного критерія

Запропонований структурний критерій полягає у наступному:

1) Генерується випадкова послідовність, яка складається з такої ж кількості фрагментів тексту L і такої ж їх довжини, як і змістовний текст.

2) За допомогою алгоритма zlib стискається згенерована послідовність і вихідний текст.

3) Обчислюється довжина кожного фрагмента тексту L у обох стиснених послідовностях.

4) Якщо довжина стисненого згенерованого тексту більша довжини стисненого вихідного тексту, то приймається гіпотеза H_0, інакше — H_1.

Основна ідея цього методу полягає у тому, що змістовний текст легше стискати, ніж випадкову послідовність літер, а тому він буде стиснений у більшій мірі і, відповідно, в середньому буде мати меншу довжину.

2.3 Опис труднощів

З метою зменшення громіздкості коду було вирішено реалізовувати всі необхідні функції таким чином, щоб вони могли працювати і для монограм, і для біграм одночасно, а не писати окрему функцію для кожного з випадків. Проте через це виникло багато труднощів з обробкою текстів. Було складно визначитися яким саме чином зберігати і передавати дані, і постійно виникали помилки, функція могла працювати для монограм/біграм, а для біграм/монограм ламалася. Знадобилося багато спроб для знаходження універсального метода для всіх варіантів даних. Вирішення цієї проблеми полягало в дебазі кожної функції і знаходження

рядка, де виникала проблема, і його виправленні. Великий час виконання коду також не сприяв швидкому написанню практикуму.

Окрім вищенаведеного, підбір параметрів у деяких критеріях, де це було необхідно, теж не був легким. Їх треба було підбирати окремо для біграм і монограм, що значно сповільнювало і ускладнювало роботу, адже не завжди цей підбір був інтуітивно зрозумілим.

ВИСНОВКИ

У роботі було програмно реалізовано алгоритми спотворення текстів, а саме шифр Віженера з випадковим ключем довжини $r = 1, 5, 10$, шифр афінної та афінної біграмної підстановки з випадковими ключами, рівномірно розподілена послідовність символів з $(Z_m)^l$, а також шифр заданий співвідношенням $y_i = s_{i-1} + s_{i-2}$. Було також згенеровано випадкові послідовності і імплементовано критерії перевірки гіпотез чи є вхідна послідовність символів або біграм змістовним текстом, чи випадковою послідовністю. Для кожного критерію була розрахувана ймовірность похибок першого і другого роду. Отримані значення були представлені у вигляді таблиць для кожного критерію з застосуванням різних шифрів.

Аналізуючи результати можна побачити, що найкраще працює критерій 1.3 для всіх шифрів і будь-якої довжини текста. При цьому, починаючи з $L = 1000$ всі критерії мають однаково гарні результати крім, можливо, шифра, заданого співвідношенням $y_i = s_{i-1} + s_{i-2}$, який інколи показує помилку другого роду рівну одиниці. На маленькій довжині тексту гіршим виявився критерій 3.0, який не розпізнає змістовний текст і вважає його випадковим.

Для структурного критерію був запропонован алгоритм, який полягає у порівнянні довжини стисненого згенерованого і реального тексту, оскільки змістовний текст стискається краще і ефективніше.